**Master's Thesis of Geography**

# Spatial-temporal PM2.5 Prediction Using MODIS AOD Products

**February 2018**

**Graduate School of Social Science**

**Seoul National University**

**Geography Major**

**Wang Yifan**

# Spatial-temporal PM2.5 Prediction Using MODIS AOD Products

지도 교수 박기호

이 논문을 지리학석사 학위논문으로 제출함

2018 년 1 월

서울대학교 대학원

지라학전공

WANG YIFAN

왕이판의 석사 학위논문을 인준함

2018 년 1 월

위 원 장 Edo Han Siu Andriesse (인)

부위원장 박기호 (인)

위 원 김대현 (인)

## Abstract

# Spatial-temporal PM2.5 Prediction Using MODIS AOD Products

**WANG YIFAN**

**Geography**

**College of Social Science**

**Seoul National University**

In recently decade haze in China has severely hurt its economy and threatened the health of its population. There is often strong demand from the Ministry for the Environment for assessing, predicting, and trying to reduce the levels of PM2.5 around the country. In practice, PM2.5 data is difficult to measure. Monitor sites are not distributed uniformly, most of them built in urban area. Traditional air pollution epidemiology studies being conducted in large cities can be limited by the availability of monitoring. Satellite Aerosol Optical Depth (AOD) measurements offer the possibility of exposure estimates for the entire population. In this situation, the 10 km MODIS Aerosol Optical Depth (AOD) product can be used as predictor since recent studies has proved the statistical relationship between AOD and PM2.5. The traditional statistical study on AOD and PM2.5 are primarily Geographic

Weighted Regression. Based on Gaussian process regression, this study developed a new regression approach to predict PM2.5 distribution in a Bayesian hierarchical setting from October 2016 to October 2017. The spatial non-stationarity was modeled by a Gaussian process with exponential covariance function. Parameters to explain factors like AOD, spatial random effects and non-spatial factors were estimated via a Bayesian hierarchical framework. The result illustrated that our model showed a good daily prediction on unknow sites by giving a 0.76 $R^2$ under 10 cross validation and a precise annual prediction with $R^2$ equal to 0.90. For daily model, we compared our result with GWR and a machine learning method support vector machine (0.68 and 0.75 respectively), which showed modeling spatial random effects via Gaussian process was able to improve the accuracy PM2.5 predicting using MODIS AOD data.

**Keyword: AOD, PM2.5, Gaussian Process, Bayesian hierarchical modeling, GIS**

**Student Number: 2016-22068**

# Contents

Chapter 1 Introduction ..... 1

1.1 Research Motivation ..... 1

1.2 Problem Description ..... 2

1.3 Research Objective and Research Question ..... 4

1.4 Methodology ..... 4

1.5 Contribution ..... 7

Chapter 2 Literature Review ..... 9

2.1 Introduction to PM2.5 ..... 9

2.2 Aerosol Optical Depth ..... 12

2.3 Satellite Data and Algorithms for AOD retrieval ..... 14

2.3.1 The MODIS AOD product ..... 15

2.3.2 Validation on MODIS AOD in China ..... 16

2.4 PM2.5 Estimation based on AOD ..... 20

2.4.1 Theoretical basis ..... 20

2.4.2 Estimation Models ..... 23

2.5 Machine Learning Methods ..... 27

Chapter 3 Study Area and Data ..... 33

3.1 Study Area ..... 33

3.2 Data Acquisition ..... 34

3.2.1 MODIS 10km Products ..... 34

3.3.2 PM2.5 ground monitoring data ..... 35

3.3.3 Supplementary Data ..... 37

Chapter 4 Model ..... 40
4.1 Overview of Workflow ..... 40
4.2 Data Pre-processing ..... 41
4.3 Model Construction ..... 43
4.3.1 Gaussian Process Regression Model ..... 43
4.3.2 Geographically Weighted Regression Model ..... 48
4.3.3 Support Vector Regression Model ..... 49

Chapter 5 Results and Analysis ..... 51
5.1 Descriptive Statistics on dataset ..... 51
5.2 Model validation ..... 52

Chapter 6 Conclusions and Limitations ..... 61
5.1 Conclusion ..... 61
5.2 Limitation of this study ..... 63

Bibliography ..... 64

# List of Tables


Table 1 Band designation for MODIS ..... 19
Table 2 Meteorological data acquired from ECMWF ..... 39
Table 3 descriptive statistics of dataset ..... 51
Table 4 Person correlation among all input variables in this study ..... 52
Table 5 descriptive statistics of seasonal dataset ..... 52
Table 6 Statistical results of all daily models ..... 56

# List of Figures


Figure 1 Global satellite-derived PM2.5 averaged from 2001 to 2006. ..... 3

Figure 2 Ambient particles' size distribution, patterned after Chow (1995) and Watson (2002) ..... 10

Figure 3 AERONET sites (Source, AERONET, 2016a) ..... 17

Figure 4 Global mean PM2.5 concentrations from 2001 to 2010 ..... 24

Figure 5 HazehoveredovereasternChinaon October 20, 2012..... 33

Figure 6 Acomparison of the MODIS True Colour Image ..... 35

Figure 7 Aqua predicted pass time ..... 36

Figure 8 Ground monitoring stations' locations and the averaged PM2.5 of every city' all stations during the study period..... 37

Figure 9 Work flow of Methodology ..... 38

Figure 10 Daily Global Boundary Layer map..... 40

Figure 11 Box-Plot of the whole year'sPM2.5 monitoring site data..... 41

Figure 12 10 folds cross validation ..... 42

Figure 13 parameters traceplot after 500 iterations..... 54

Figure 14 parameterst raceplot after 5000 iterations..... 55

Figure 15 Scatter plot of predict values and real PM2.5 ..... 57

Figure 16 10 fold validation Scatter plot..... 58

Figure 17 Seasonal distributions of PM2.5 concentrations ..... 59

# Chapter1. Introduction

## 1.1 Research Motivation

Since 2013, as a part of the air quality improving program, a ground-level monitoring network to record ground-measured PM2.5 information was established by Chinese governments and public organizations. New plans for implementing air pollution enacted in next few years extended the monitoring network from 900 ground sites to over 1500.

Ground-based monitoring data is generally considered as an accurate record of real value. However, the data are quite sparse, merely representing a small part of whole territory of China (Tian et al., 2010). A tough problem is that distribution of this network is spatially unbalanced, which makes interpolating difficult (Hu et al., 2013). Relying on instrument operation period and functionality, the data integrity of time series of ground-level PM monitoring also highly varies (Benas et al., 2013). Although making more monitor site is definite a solution but will be both time and labor consuming.

PM estimation using remote sensing techniques is an efficient solution for issues above (Benas et al., 2013). Firstly, satellite taking image on its orbit can provide a complete, worldwide spatial resolution. (Hadjimitsis, 2009). Secondly, it provides information of 6 global air quality, which can be used to track the origin of urban air pollutant and global transportation (Wang et al., 2013). Without maintaining the whole monitoring network, this method is affordable for more

regions by saving labor and facility cost. Previous studies showed a strong correlation between retrieved AOD from satellite and ground-level PM2.5 concentration by various models (Chu et al., 2003; Wang, 2003). Regarding to shortcomings, the major discussed issue is that AOD is in whole atmospheric aerosol level, while ground-level PM2.5 data were observed on the Earth's surface (Benas et al., 2013). Furthermore, cloud, snow and ice cover can reduce the AOD availability and accuracy, thus unable to estimate PM2.5 concentrations (Lee et al., 2012).

The AOD product from remote sensing techniques provides a chance to predict PM2.5 concentrations in a high spatial resolution. Popular statistical models include the simple linear regression model, the multiple linear regression (MLR) model, the geographically weighted regression (GWR) model, artificial neural network (ANN) algorithms, generalized additive models (GAMs) or two-stage hierarchical models that include combinations of different statistical models.

In conclusion, finding the way using RS techniques like MODIS data in the estimation of PM2.5 over China will not only benefit local citizens' health and their quality of life, but also facilitate local government to take corresponding actions in regulating pollutants emission and protecting its local environment.

## 1.2 Problem Description

It is showed in previous studies that the relationship between PM2.5 and AOD

values varies in space (Hu, 2009). Van Donkelaar et al (2010) generated a map about global satellite-derived PM2.5 using the averaged AOD from MISR and MODIS during 2001 and 2006. Their results are shown in Figure. 1.2.

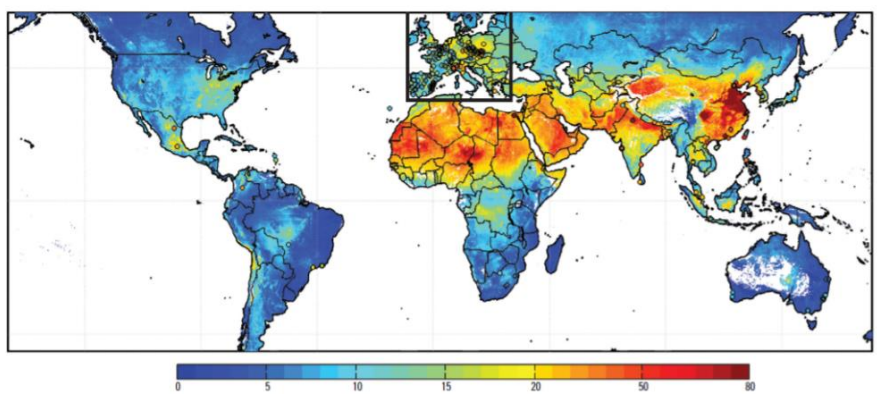


Figure 1.1 Global satellite-derived PM2.5 (μg/m3) averaged from 2001 to 2006.(Source: van Donkelaar et al., 2010)

Underlying spatially continuous phenomenon need to be modeled thus simple global regression methods performing poorly on this kind of problem. Spatial statistics is used to describe a wide range of statistical models and methods intended for the analysis of spatially referenced data. To addressing the spatial variability and non-stationarity of regression parameters, many studies have employed spatial statistics model to address the spatial heterogeneity of the PM2.5-AOD relationship.

In observation-based models, besides AOD, locations, meteorological

parameters and socio-economic factors have been widely utilized as inputs to perfect the performance. Explanation of spatial effect and these factors is a complex task. The hierarchical nature can help explain various sources of variations in PM2.5 while hyperparameters in such modeling, which usually set subjectively and empirically, require tedious trying for optimization in certain dataset.

## 1.3 Research Objective and Research Question

Overall, this study aims to explore implying statistical models on remote sensing datasets for estimating PM2.5 concentrations in China from Oct 1st, 2016 to Sep 31st, 2017. In order to achieve this goal, the following three main research questions need be addressed:

(1) How do we explain multiple sources of variation of PM2.5?

(2) How to treat spatial relationship between PM2.5 and AOD?

(3) How to optimize hyperparameters in hierarchical setting?

# 1.4 Methodology

To answer the three research questions outlined, the specific methodology of this thesis is as following:

Firstly, we analyzed relationship between AOD data and PM2.5, and selected planetary boundary layer and relative humidity in meteorology factors as other predictor in our model by reviewing definition of AOD. Analysis of feasibility of MODIS AOD data in china is conducted through literature review. MODIS aerosol product has been widely used in PM2.5 estimation in previous study. Based on validation works on MODIS AOD data by AERONET AOD conducted in those study, feasibility and error were also discussed.

Secondly, we reviewed that Gaussian processes are one of the most intuitive methods to model spatial surfaces as realization of stochastic processes and it has impressive performance on modeling spatial effect. We set a hierarchical model to help explain various sources of variations in PM2.5 with a linear group of intercept and coefficients of AOD, planetary boundary layer and relative humidity, and a spatial random effect to capture the geographic variation and a non-spatial random effect. Unlike traditional geostatistical methods, which rely on particular functions (such as wavelets and splines) to represent spatial relationships, Gaussian processes are one of the most intuitive methods to model spatial surfaces as realization of stochastic processes. Specifically, Gaussian processes consider the spatial effect as random variables by specifying their means and covariance functions, which is the major feature that distinguishes them from other traditional methods.

In particular, our model can be described in the following three stages: for the first stage, PM2.5 concentrations are conditional on the distribution of AOD values, spatial and non-spatial random effects, which is the basic foundation of our model;

the second stage mainly focuses on the distribution of spatial random effects, which are modeled by Gaussian processes with specific mean surface and covariance functions; the last stage concentrates on the conditional distribution of the covariance functions of Gaussian processes given by the hyperparameters we chose. This hierarchical approach is helpful when dealing with ambiguous variations. Comparatively, for GWR models, the coefficients of each independent variable (in our case, there is a single explanatory variable, AOD) and intercept are different at different locations, and the coefficients are intrinsically modeled as fixed numbers.

Bayesian methods is gaining popularity in recent environmental science, epidemiology and health policy management studies along with advancement of computing resources. It sounds reasoning of treating parameters as random quantities rather than fixed values. Parameters are updated by calculating the posterior distribution (prob(parameters|data)) by the incorporated external knowledge with respect to the distribution of parameters and the likelihood function (prob(data|parameters)). The Bayesian methodology is flexible because it allows non-informative priors, as well as informative priors acquired by relevant research or spatial variogram analysis. In our study we employed Bayesian approach by using Pymc3 to optimize all parameters with an empirical prior setting. Daily PM2.5-AOD models for China from Oct 1$^{st}$ 2016 to Sep 31$^{st}$ 2017 were constructed. Spatial distribution and seasonal variation were examined. We used cross validation to analyze over-fitting in our model.

We anticipated the GPR model in Bayesian hierarchical setting increased the

accuracy so that a comparison with other methods were conducted. We matched our daily prediction result with Geographic weighted regression and Support vector regression because the GWR model has been widely used in AOD-PM2.5 relationship with spatial effect represented as linear coefficient on each predicting factors, while the SVR and our method, GPR, has the same powerful way to deal with non-linear data which is called Kernel trick.

## 1.5 Contribution

The results proved our research has increased accuracy on modeling PM2.5-AOD relationship compared with traditional method GWR and machine learning method with Kernel trick SVR. It works remarkably accurate on training data and showed little over-fitting but also acceptable performance on test data.

We also discussed the relationship of PM2.5 between planetary boundary layer height and relative humidity via reviewing definition of AOD, and used them to perfect our model. The person correlation result showed that correlation among variables are low and no collinearity exists thus able to predict PM2.5.

Our model treated the spatial relationships as random variables and used gaussian process to depict. We gave a hierarchical explanation of multiple sources of PM2.5 variation. Although fitting the hierarchical models is always considered time-consuming owing to the large sample size and high cost of matrix decomposition, our research showed that MCMC algorithm performed computed

effectively on a national scaled data with over 300 inputs in daily model.

# Chapter 2. Literature Review

## 2.1 Introduction to PM2.5

It has been globally recognized that air pollution poses a threat to public health and the steady worldwide increase since 1990 of the burden of disease is attributed to ambient air pollution (Forouzanfar et al., 2015). WHO has reported that 3.7 million people died in 2012 caused by ambient air pollution, and the Southeast Asian and Western Pacific regions bear most of the burden (WHO, 2012).

The major pollutants consist of carbon monoxide, sulfur oxides, particulates, nitrogen oxides, and ground level ozone. The particulate matter is formed by liquid and solid airborne particles with different diameters and complicated components (Gupta et al., 2006). PM includes coarse particles (diameter greater than 2.5 μm), fine particles (PM2.5, particles less than diameter < 2.5 μm) and ultrafine particles (particles less than diameter < 0.1 μm) (Wilson et al., 1997). Measurement of Coarse, fine and ultrafine particles is based on its size, source, formation mechanism, lifetime and spatial- distribution (Wilson et al., 1997). Atmospheric life time of PM2.5 lasts days to weeks compared with minutes to hours life time of coarse-mode particle. The travel range of PM2.5 is wider than coarse as well, 100 to 1000 kilometers comparing 10 to 100 kilometers. (Wilson et al., 1997). The small size, long life time and wide travel range make it more dangerous and larger uncertainty in its distribution.

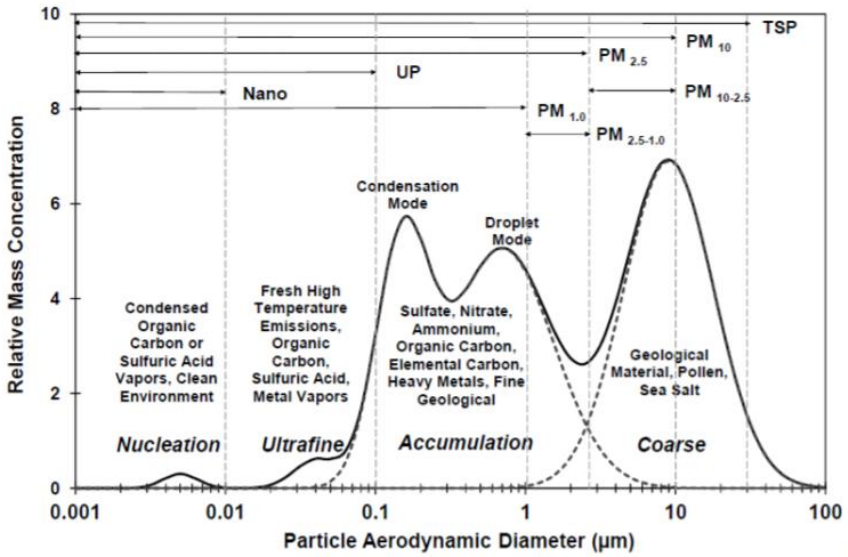


Figure 2.1 Ambient particles' size distribution, patterned after Chow (1995) and Watson (2002). (Source: Cao et al., 2013)

Composition of PM2.5 differ from its source: natural and anthropogenic. The natural source includes sea salt, dust, volcanic eruptions, forest and grassland fires (Emili et al., 2010; Beh et al., 2013), and the anthropogenic source contains industrial processes, transportation, fossil fuel combustion (coal, gasoline and diesel), and uncertain sources (Emili et al., 2010; Wang et al., 2016). Figure 1.1 shows the size range and some of the major components of PM2.5 and PM10. Generally, PM2.5 contains nanoparticles (condensed organic carbon and sulfuric acid vapors), ultrafine particles (fresh high temperature emissions, organic carbon and metal vapors), while PM10 contains the components of PM2.5, and other components such as geological material, pollen and sea salt. (Watson et al., 2002;

Cao et al., 2013). In addition, atmospheric chemical reactions also occur among primary particles and result in secondary particles (Franklin et al., 2008).

The increasing PM2.5 effects negatively on population health and hinders economic development. What's more, climate change has been impacted by PM2.5's effects directly and indirectly. Through directly interacting the solar radiation and terrestrial surface radiation like absorbing and scattering, PM2.5 makes the radiation budget balance and temperature abnormal (Sokolik et al., 1996). Indirectly, PM2.5 influence climate through effecting on the chemical composition and density of the atmosphere. (Schwartz et al., 1995). It is also proved that formation of acid rain can partly attributed PM2.5 (EPA .n.d) thus reduces agricultural productivity (Chameides et al., 1999). PM2.5 also reduce air visibility because of its hygroscopic properties of constituent Sulphur (Deng et al., 2011).

Due to its size, PM2.5 can be breathed deeply into the lungs and would never come out (Pope III et al., 2000). Long term and short exposure to PM2.5 has been associated with hospital admissions for 3 pneumonia, emergency department visits, asthma, bronchitis, cardiovascular problems, respiratory infections, lung cancer, heart disease and premature deaths. (Wellenius et al., 2005; Baccarelli, 2009; Jones et al., 2015; Kioumourtzoglou et al., 2016; Zanobetti et al., 2015). According to a survey in OECD Environmental Outlook To 2050, it is estimated that in 2010, 1.4 million people died due to PMs and this number is expected to increase to 2.3 in 2030 and 3.6 in 2050. Most of the premature deaths are elderly with weaker

immune systems (EPA. n.d). Children are also at high health risks because their immune and respiratory systems are premature: 40% of asthma cases are children, while the population of children only occupies 25% of the whole world's population (EPA. n.d).Recent research also shows the health risks attributed to PM2.5 differ for men and women: the increase of PM2.5 is associated with a higher increase of heart rhythm disturbance admission to hospital for women than for men (Bell et al., 2015). In addition, PM2.5 can even damage DNA in human cell (Sørensen et al., 2003; Corsini et al., 2013).

In addition to the influence on climate change and human health, PM2.5 also brings economic loss. According to Ontario Ministry of the Environment (MOE) (2005), Ontario was burdened with approximately $9.6 billion CAD economic loss due to the high concentration of ozone and PMs in 2003. $5.28 billion CAD loss was due to U.S. emissions, while the rest, $4.32 billion CAD, is attributed to provincial air pollution. It was also estimated that in the Yangtze River Delta, China, the total economic loss caused by the high concentration of PM2.5 was ¥22.10 billion CNY in 2010 (Wang et al., 2015a). Gao et al (2015) assessed that Beijing's economic loss resulted from the haze in January 2013 was more than $250 million USD.

## 2.2 Aerosol Optical Depth

Aerosol represents particles suspending in atmosphere with diameter range

from $10^{-2}\sim10^{-3}\mu m$, which is an important part of atmospheric system. Aerosols plays a significant role in climate change and environmental affections in three ways, (Liu et al, 2008) scattering and absorbing radiation, influencing cloud formation as condensation nucleus, change greenhouse gas by involving chemical process.

Aerosol optical depth (AOD) is a parameter of aerosol, representing the extinction of electromagnetic radiation in a certain wavelength (Chudnovsky et al., 2014). Basically, values of AOD is in a range of 0 to 2. Values smaller than 0.1 illustrate an extreme clean air with quite good visibility, and those larger than 1 means thick hazy air condition (NASA). Atmospheric particles in any form like dust, faze and PM2.5 are able to block sunlight by absorbing or scattering (NOAA). The degree of attenuation can be described by AOD (NOAA).

Techniques of AOD monitoring has been developed fast in last two decades. There are two main approach to gaining AOD data, ground station and remote sensing. Main facility capturing ground AOD is sun spectrophotometer. However limited number of facilities lead to a limited geographic information scale. Representative of ground level station network are Aerosol Bo botic Network (AEORNET) and Sky Radiometer Network (SKYNET). On the contrary, AOD data acquired via remote sensing has larger coverage. The most popular AOD product in academic utilization is MODIS.

## 2.3 Satellite Data and Algorithms for AOD retrieval

Remote sensing techniques utilized on aerosol started since 1970s. Now a complete satellite monitoring system for AOD has been built, users can access to a full-available spatial and temporal AOD dataset (Guo et al., 2009). Here we reviewed basic retrieving AOD algorithm in satellite sensor.

The radiation characteristics of solar radiation varies while going through atmosphere and being receiving remote sensor because of scattering and reflecting. Information received by sensor includes two parts, atmosphere and earth's surface. Given surface reflectance and certain absorption and scattering, AOD data can be retrieved using spectral characteristic (Liu et al, 2001).

Aerosol detected by remote sensor is based on atmosphere surface reflectance $\rho^*$, (Kaufman et al, 1997)

$$\rho^* = \pi L / \mu_s F_s$$

Where L is top atmosphere's spectral radiance, $\mu_s$ is cos of solar zenith angle, $F_s$ is flux density of the direct solar radiation. The atmosphere surface reflectance $\rho^*$ has following relationship with surface bi-reflectance,

$$\rho^*(\theta_v, \theta_s, \varphi) = \rho_a(\theta_v, \theta_s, \varphi) + \frac{\rho(\theta_v, \theta_s, \varphi) F_d(\theta_s) T(\theta_v)}{1 - S\rho'}$$

Where $\theta_v$ is remote sensor's zenith angle, $\theta_s$ is solar zenith angle, $\varphi$ is the relative angle between the former two, $\rho_a(\theta_v, \theta_s, \varphi)$ is Atmospheric path radiation, which triggered by molecule and aerosol in atmosphere, $F_d(\theta_s)$ is down direct radiant flux, $T(\theta_v)$ is total transmittances, S is back-scattering ratio depending on single scattering albedo $\omega_0$, aerosol optical depth and aerosol Scattering Phase Function $P_a(\theta_v, \theta_s, \varphi)$.

Depend on this formula, $\rho_a(\theta_v, \theta_s, \varphi)$ in the right part is the atmosphere contribution in remote sensor's observation, and the second part is surface reflectance contribution. When surface reflectance contribution is low, $\rho^*$ is mostly depend on atmosphere contribution so precision is high. Therefore, retrieving AOD perform well in low surface reflectance regions. (Liu et al 2001).

### 2.3.1 The MODIS AOD product

Since the development of remote sensing techniques from 1980s, satellite images have been explored for AOD retrieval. Moderate Resolution Imaging Spectroradiometer (MODIS) is carried on both Terra and Aqua launched in 1999 and 2002, respectively. The band designation for MODIS can be found in Table 2.1: there are seven well-calibrated channels for spectral information ranging from visible to SWIR wavelength (470, 550, 670, 870, 1240, 1640 and 2100 nm) (Chu et al., 2003).

MODIS derives an AOD product (Terra: MOD04_L2; Aqua: MYD04_L2) at

10 km resolution using “Deep Blue” (DB) and “Dark Target” (DT) algorithms. DT is adopted over ocean and dark land, such as vegetated area, while DB is applied over the entire land areas including both dark and bright surfaces in MODIS Collection 6 (C6) product. “Collection” means a MODIS dataset and previous collections include 001, 003, 004, 005 and 051. Data user can choose the parameter when downloading data online, such as “AOD 550 Dark Target Deep Blue Combined and “Deep Blue Aerosol Optical Depth 550 Land”. For more detailed product information, please refer to MODIS Website (http://modis.gsfc.nasa.gov/). In 2014, DT algorithm team released 3 km MODIS AOD product in a separate file (Terra: MOD04_3K; Aqua: MYD04_3K) as a part of MODIS C6 production. Xie et al (2015) estimated PM2.5 within urban region in Beijing, China using 3 km MODIS AOD product. In the same year, Retails et al. (2015) identified the correlations between 3 km MODIS AOD product and ground-based PM10 measurements in the area of Athens, Greece. AOD retrieved from MODIS by using visible spectrum and infrared spectrum can reduce errors caused by a single band calibration (Xie et al., 2011). Meanwhile, the high temporal resolution (twice a day provided by Terra and Aqua) is another advantage of MODIS AOD product over others. However, cloud, snow and ice still affect the accuracy of AOD retrieval from MODIS (Gupta et al. 2006).

### 2.3.2 Validation on MODIS AOD in China

In ground level, long-term international AOD observation network have been

built. The AEORNET program, mentioned in Chapter 1, started up by NASA and PHOTONS, then extended by national institutes, agencies, universities, etc. National observation networks joined AEORNET successively, including Chinese Sun Hazemeter Network (CASHNET), Finnish Meteorological Institute (FMI), German Weather Service (DWD), Japan Meteorological Agency, U.S. (ARM and SURFRAD), Australia Bureau of Meteorology (BOM) (Levin et al., 2008). A global scale, long-term continuous AOD observation with 15 min temporal resolution and 0.01-0.02 low uncertainties is offered by AERONET network. (Sayer et al., 2013).

The Chinese part AEORNET network, China meteorological administration Aerosol Remote Sensing NETwork (CARSNET), was established in 2002. (Che et al., 2009). he instrument deployed by CARSNET is automatic Cimel sun and sky scanning radiometer (Cimel Electronique Cimel-318), the same instrument used by AERONET.

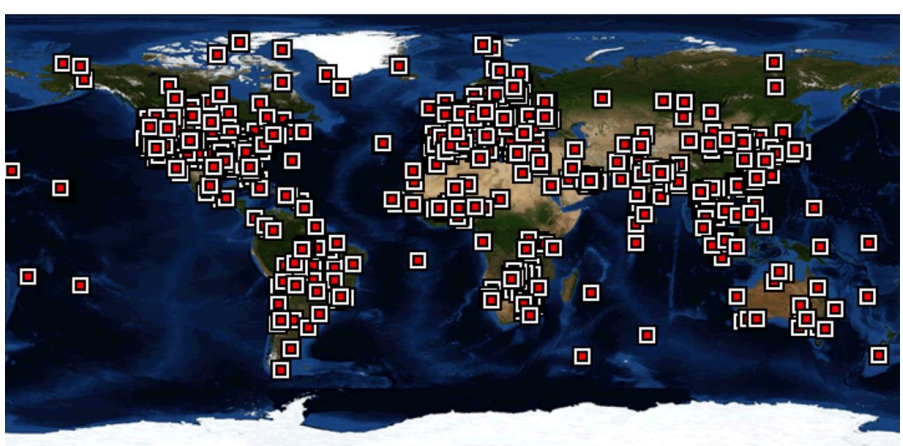

Figure2-2, AERONET sites (Source, AERONET, 2016a)

The long-term reliable and consistent measurements of CARSNET provide an unprecedented opportunity to study aerosol properties and validate MODIS retrieved AODs over various terrestrial regions in China.

Previous validation work showed a high corelated linear relationship between MODIS AOD and ground observation AOD. Xie et al, 2011 matched MODIS retrieval AOD data with interpolated CARSNET monitor AOD in national scale and proved both DT and DB AOD Fall within the expected error envelope. He et al, 2010 indicated that MODIS AODs are in good agreement with observation sites in Yangtze River Delta region with a correlation coefficient of 0.85 and RMS of 0.15, showing MODIS AOD product are generally suitably reasonable for aerosol retrieval in YRD.

| Primary Use | Band | Bandwidth (Band1-19:nm; Band20-36:um) | Spectral Radiance (W/m2 -μm-sr) | Spatial Resolution(m) |
|---|---|---|---|---|
| Land/Cloud/Aerosols Boundaries | 1 | 620 - 670 | 21.80 | 250 |
| | 2 | 841 - 876 | 24.70 | |
| Land/Cloud/Aerosols Boundaries | 3 | 459 - 479 | 35.30 | 500 |
| | 4 | 545 - 565 | 29.00 | |
| | 5 | 1230 - 1250 | 5.40 | |
| | 6 | 1628 - 1652 | 7.30 | |
| | 7 | 2105 - 2155 | 1.00 | |
| Ocean Color/Phytoplankton/Biogeochemistry | 8 | 405 - 420 | 44.90 | 1000 |
| | 9 | 438 - 448 | 41.90 | |
| | 10 | 483 - 493 | 32.10 | |
| | 11 | 526 - 536 | 27.90 | |
| | 12 | 546 - 556 | 21.00 | |
| | 13 | 662 - 672 | 9.50 | |
| | 14 | 673 - 683 | 8.70 | |
| | 15 | 743 - 753 | 10.20 | |
| | 16 | 862 - 877 | 6.20 | |
| Atmospheric Water Vapor | 17 | 890 - 920 | 10.00 | |
| | 18 | 931 - 941 | 3.60 | |
| | 19 | 915 - 965 | 15.00 | |
| Surface/Cloud Temperature | 20 | 3.70 - 3.84 | 0.45(300K) | |
| | 21 | 3.93 - 3.99 | 2.38(335K) | |
| | 22 | 3.93 - 3.99 | 0.67(300K) | |
| | 23 | 4.02 - 4.08 | 0.79(300K) | |
| Atmospheric Temperature | 24 | 4.43 - 4.50 | 0.17(250K) | |
| | 25 | 4.48 - 4.55 | 0.59(275K) | |
| Cirrus Clouds Water Vapor | 26 | 1.36 - 1.39 | 6.00 | |
| | 27 | 6.54 - 6.90 | 1.16(240K) | |
| | 28 | 7.18 - 7.48 | 2.18(250K) | |
| Cloud Properties | 29 | 8.40 - 8.70 | 9.58(300K) | |
| Ozone | 30 | 9.58 - 9.88 | 3.69(250K) | |
| Surface/Cloud Temperature | 31 | 10.78 - 11.28 | 9.55(300K) | |
| | 32 | 11.77 - 12.27 | 8.94(300K) | |
| Cloud Top Altitude | 33 | 13.19 - 13.49 | 4.52(260K) | |
| | 34 | 13.49 - 13.79 | 3.76(250K) | |
| | 35 | 13.79 - 14.09 | 3.11(240K) | |
| | 36 | 14.09 - 14.39 | 2.08(220K) | |

Table 2.1 Band designation for MODIS (Source: NASA, 2016)

Wang evaluated MODIS AOD performance over different ecosystem in China, showing that most agreement between the MODIS data and that of the CSHNET was in farmland sites in central-southern China with high R2>0.82, and moderate

agreement, with R0.64–0.80 in temperate forest, coastal regions, and northeast while poorest agreement existed in northern arid and semiarid regions, in remote northeast farmlands, in the Tibetan and Loess Plateau, and in southern forests, with 13–54% of retrieval data falling within the expected errors. Over different ecological and geographic regions in China, Wang et al, 2007 indicated that performance MODIS AOD is poor in Tibetan Plateau, northern desert area, and northeast corner of China, while it is moderate in forest area and performed greatly in agricultural, vegetated areas and eastern seashore area. Overall, validation studies indicated that MODIS AOD in China has varies bias in different region though, it performed satisfactorily in an overall national perspective. Therefore, it is feasible to model on MODIS AOD data in China.

## 2.4 PM2.5 Estimation based on AOD

### 2.4.1 Theoretical basis

Numerous studies have focused on constructing statistical relationships between satellite AOD retrievals and ground-level PM2.5 measurements that can then be used to estimate PM2.5 concentrations in places where AOD data are available.

AOD is defined as the integration of aerosol extinction coefficient in vertical direction. It is a physical dimensionless quantity about counts of aerosol particles. Commonly used PM2.5 value means concentration of PM2.5 particles per cubic

meter. Their physical significance though, previous studies showed there is correlation between them and it is feasible to estimate PM2.5 using AOD (Li et al, 2003). The theocratic basic of modeling is that AOD retrieved by visible and near-infrared light correspond to 0.1-0.2μm particles, which is pretty close to PM2.5 (Kahn et al, 1998). This provides the theoretical basis of PM2.5-AOD modeling.

The AOD recorded by MODIS is the integration of aerosol extinction coefficient in vertical direction, while PM2.5 represent the concentration of dried ground particles. Based on this, two factors can be found working in relationship between AOD and PM2.5, vertical distribution of aerosol and relative humidity. The formula representing this relationship is as following (Jia et al, 2014),

$$\mathrm{AOD} = \mathrm{PM}_x \cdot \mathrm{H} \cdot \mathrm{f(RH)} \cdot \frac{3 < Q_{exi} >}{4\alpha \cdot \rho \cdot \mathrm{r}_{eff}}$$

Where $\mathrm{PM}_x$ is concentration of particles with diameter less than x, H is the aerosol scale height, f(RH) is a function on relative humidity, $< Q_{exi} >$ represents normalized extinction efficiency particle, $\mathrm{r}_{eff}$ is effective radius, ρ is density, $\alpha$ is ratio of aerosol depth in scale height in total aerosol depth. This formula indicates aerosol scale height and f(RH) plays a significant role in PM-AOD relationship.

The aerosol scale height is defined as the height when concentration of atmosphere aerosol decreases to 1/e of the ground concentration, namely the height

where aerosol concentration become constant with height increasing. This figure is difficult to gain. Research showed that most particles distribute in Planetary boundary layer, therefore the height of Planetary boundary layer is good alternative of aerosol scale height working as a predicting factor in models (Liu et al, 2005).

The other factor is relative humidity. Effect on extinction coefficient by humidity can be described as hygroscopic growth factor f(RH). Defination of this function is ratio of extinction coefficient in nature and extinction coefficient in environment humidity less than 40% (Kotchenruther et al, 1999).

$$\mathrm{f(RH)} = \frac{k(\lambda)}{k(RH \leq 40\%)}$$

Where k represents extinction coefficient. f(RH) varies a lot in regions with different environmental humidity. What's more, relative humidity helps the formation of ammonium nitrate (Tai et al., 2010).

Besides height of Planetary boundary layer and relative humidity, other meteorological variables and other natural factors like Visibility, precipitation, temperature, wind speed, elevation, pressure affect the formation and dispersion of PM2.5 as well (Tai et al., 2010). Human activities also generate considerable amount of PM2.5. Socio economic factors like GDP, population and land use data are used to reflect the impact of human activities.

### 2.4.2 Estimation Models

Basically, models estimating PM2.5-AOD are classified into two types: simulation-based and observation-based methods (Lin et al., 2015). Simulation-based models are usually on chemical transport theoretical basis (Liu et al., 2004; Martin, & Park, 2006; Van Donkelaar et al., 2010).

3D chemical transport is most used in Simulation-based models (Martin, & Park, 2006). These models are composed of meteorological driver and chemical transport module. Goddard Earth Observing System Atmospheric Chemistry Transport(GEOS-Chem) is driven by meteorological variables from the GEOS of NASA. Van Donkelaar et al. (2010) applied GEOS-Chem model and calculated the PM2.5 concentration with MODIS and MISR data for 2001 to 2005 at a global level. In 2015, van Donkelaar built a global model for PM2.5 concentrations from 2001 to 2010 by utsing GEOS-Chem model and MODIS data, which is shown in Figure 2.4. In his research the PM2.5 concentration in China is more than 80 μg/m3.

The Weather Research and Forecasting model coupled with Chemistry (WRF-Chem) is a numerical weather prediction system for atmospheric research needs (Tie et al., 2007). WRF-Chem models results present a good association with the correct emission database. Eta-CMAQ and MM5-CMAQ model have also been applied in estimating PM2.5 (Yu et al., 2004).

The major advantage of simulating models is simulating the process of factors

forming PM2.5 (such as chemical composition and particulate size) and explaining the correlation between AOD and PM2.5. Whereas the major shortcoming is that their principle is complex thus made modeling difficult.

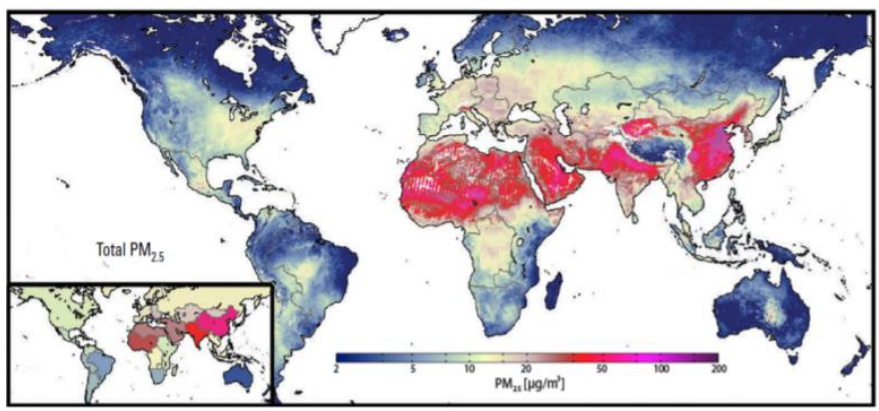


Figure 2.3 Global mean PM2.5 concentrations from 2001 to 2010 (Source: van Donkelaar et al., 2015)

Observation-based models are mainly based on statistical regression methods (Lin et al., 2015). Simple linear regression method was firstly employed to estimate PM10 concentration by inputting daily averaged AERONET AOD observation in Italy. Later in research in Alabama, U.S., PM2.5 concentrations were estimated via simple linear regression with dataset composed of 7 observe stations and MODIS AOD product (Wang & Christopher, 2003). The research presented AOD as a useful tool for modeling PM estimation with a result of 0.49 R2.

Nevertheless, most researchers agree that AOD can poorly retrieve surface

distribution of PM2.5 in satisfied accuracies because the inconsideration of other factors may affect AOD-PM2.5 relationship (Paciorek et al., 2008b).

The multi linear regression was introduced in order to incorporate more predictors (Gupta & Christopher, 2009a). Variables that directly relate to PM2.5 include temperature, relative humidity, height of the planetary boundary layer and wind speed. Gupta et al. (2009a) presented that correlation coefficients increased up to threefold from simple linear regression to MLR model in their model of MLR equations with AOD and meteorological factors over the southern U.S. MODIS AOD and NCEP (National Centers for Environmental Prediction) data were applied on PM2.5 modeling in study of Li et al, 2011. They also compared the R square in simple linear regression and MLR. Though the remarkable lifting of MLR result showing a better capacity due to the consideration of more relevant variables, most MLR models in previous study using meteorological factors were biult on a global, national or regional level, the spatial variability are rarely considered in modeling structures.

Regression method were proved as a strong tool to estimate PM2.5 by assuming a linear relationship between PM2.5 and predictors, though, Liu et al., 2005 indicated non-linear model works as well. Li et al. (2011) run a non-liear model and showed a better performance (R2 =0.49) than simple and multiple linear regression (R2 = 0.24; 0.44). In small scale study, for example, Yu et al., 2006's study in semi-arid area in northern China, the performance has been improved when using non-linear regression model. Besides nonlinear regression model, a

more complex model, generalized additive model (GAM) also allows non-linear function of variables. This model is developed for each scaling method at each site (Liu et al., 2011). By allowing some of all variables to be non-linear related to dependent variable, GAM improves the capacity of traditional linear regression (Liu et al., 2011). Though non-linear model is able to improve the accuracy in these studies, it only works for certain areas and or seasons (Li et al., 2011). Moreover, similar to linear regression, this model does not consider local variables: this is because the correlations between AOD and PM2.5 are non-stationary, so the dependent and independent variables are not spatially constant (Engel-Cox et al., 2004; Hu et al., 2009).

To solve this problem, spatial regression model, such as the GWR model is also applied to build a local relationship between AOD and PM2.5 (Hu et al., 2013). Instead of assuming global geographic uniformity, GWR estimates PM2.5 in consideration of local variability. Hu et al. (2013) adopted both Ordinary Least Squares and the GWR model to estimate PM2.5 in U.S., and R squre was slightly improved when using the GWR model. The GWR model has also been applied in China at a national level in 28 previous studies (Ma et al., 2014). You et al. (2016) used the GWR model with the 3 km AOD product to estimate PM2.5 concentrations at a national-scale. Their model could explained 81% of the dairy PM2.5 variations.

## 2.5 Machine Learning Methods

Linear regression models did performe well in short-term forecasting based on daily or weekly time resolution, but not for long period forecasting at a seasonal or annual time resolution, neither can they handle nonlinearity exhibited relationship in variables well. Since the computer's performance is developing vary fast in recent decade, various machine learning models have been developed and harnessed to model spatial issues in geography studies.

Machine learning methods and, in particular, supervised learning methods, refer broadly to statistical techniques for developing predictive models using training data. Unlike physics-based models, machine learning methods are data-driven and rely almost exclusively on information embedded in training datasets. Artificial neural network (ANN) is one of the earliest machine learning methods adopted in PM2.5 AOD modelling. However, despite its popularity and remarkable accuracy fitting result on training data, crucial issues of Artificial neural network are its tendency to overfit training data and instability with short training data records (Sun et al, 2014).

In machine learning studies, recent decades have witnessed a soaring interest in the development and application of kernel-based methods. In particular, the support vector machine (SVM) algorithm was proposed to deal with two issues alluded above, in other words, how to establish a relationship between the size of training data and generalization performance of a trained model and how to

incorporate such knowledge in the training process to prevent overfitting. Via kernel trick, SVM projects the model's inputs into a higher dimensional or even infinite-dimensional space, such that the projected training data exhibit linearity and linear regression methods can be applied (Bishop, 2006). An elegant feature of SVM is that the actual form of nonlinear mapping does not need to be known, and only their inner products (i.e., the so-called kernel function) are required to train an SVM model. This is known as the "kernel trick" in machine learning, which has served as a building block in all kernel-based methods. The Support vector regression is a variation of support vector machine particularly for regression problem which has been already introduced in spatial PM10 forecasting and wind predicting.

Although the SVR can found a satisfactory regression line linear or non-linear dataset, a main limitation of the SVR method is that it can yield unreliable results when there is a test data point deviating far from the relevance vectors, in which case the predictive distribution will be a Gaussian with mean close to zero and variance also close to zero (Rasmussen and Williams, 2006). The Gaussian Process Regression (GPR) was developed to address this issue.

The GPR is a full Bayesian learning algorithm that has received significant attention in the machine learning community for applications such as model approximation, multivariate regression, and experiment design (Rasmussen and Williams, 2006).

Gaussian processes (GP) assume that the joint probability distribution of

model outputs is Gaussian. The notion of GP is not new in the geospatial analysis literature. In fact, GP is underlying the kriging algorithm in classical geostatistics, the autoregressive moving average models (ARMA), Kalman filters, geostatistical inversion methods, and radial basis function networks (Bishop, 2006). The ensemble Kalman filter and Gaussian particle filter may also be regarded as sequential versions of GP-based learning algorithms. Gaussian stochastic processes are widely used in practice as models for geostatistical data

The GPR was originally formulated by Rasmussen and his coworkers, provides a "principled, practical, and probabilistic approach to learning in kernel machines" (Rasmussen, 1996; Rasmussen and Williams, 2006). The advantage of GPR over many other machine learning methods lies in its seamless integration of several machine learning tasks, including hyperparameter estimation, model training, and uncertainty estimation which strengthen the model result and explanation of variables in practical studies; thereby, the regression process is streamlined significantly and the results are less affected by subjectivity and more interpretable. Along with surging popularity of GPR, a suite of GPR tools packages are now available in the public domain for various applications. In comparison, similar machine learning methods mentioned above usually only address certain aspects of the regression/prediction problem.

GPR can be viewed in weight space, thus be considered as multivariate regression techniques. In this sense, it is closely related to generalized least squares, which has been used extensively in the so-called regional regression analysis in

hydrology (Sun et al, 2014). The difference between GPR and general MLR method is that most existing studies parameterize the predicted values as a linear combination of predictors and then estimate the linear coefficients while GPR expresses the unknown as a linear combination of nonlinear basis functions. The Bayesian joint probability method proposed recently by Yu and his coworkers (Yu et al, 2017) used Bayesian inference to predict PM2.5 in China. The authors mainly focused on learning parameters of an enhanced Box-Cox transform using Monte Carlo Markov chain sampling.

Previous studies have found out that the relationship between PM2.5 and AOD values varies in space. The varying spatial surfaces is the critical issue to be addressed for lifting PM2.5 predicting into higher level.

In Gaussian process, training points in dataset that near test points should be more informative than far points on giving prediction. This is closely related to geostatistic principle, near all attribute values on a geographic surface are related to each other, but closer values are more strongly related than are more distant ones thus can be used to model the non-stationary in spatial data. From the perspective of Gaussian process, it is the covariance function that define the nearness or similarity in data (Rasmussen et al, 2006). Gaussian process is one of the most intuitive methods to model spatial surfaces as realization of stochastic processes. Specifically, Gaussian processes consider the spatial effects as random variables by specifying their means and covariance functions, which is the major feature that distinguishes them from other traditional methods.

Gaussian processes are one of the most intuitive methods to model spatial surfaces as realization of stochastic processes. Specifically, Gaussian processes consider the spatial effects as random variables by specifying their means and covariance functions, which is the major feature that distinguishes them from other traditional methods. The hierarchical setting in a GPR model can explain diverse sources of variations in PM2.5. The hierarchical approach is helpful when dealing with ambiguous variations (Finley, A. O., 2007).

However, historically, few studies have developed Gaussian process models for PM2.5-AOD modeling. Along with the advancement of Geographical Information Systems (GIS), large spatiotemporal datasets were adopted in studies in areas like environmental science, epidemiology and health policy management, which is a challenge for modelling. In the existing spatial statistical methods, Bayesian methods have gained in popularity because of its sound reasoning of treating parameters as random quantities rather than fixed values. Parameters are updated by calculating the posterior distribution by the incorporated external knowledge with respect to the distribution of parameters and the likelihood function. The Bayesian methodology is flexible because it allows non-informative priors, as well as informative priors acquired by relevant research or spatial variogram analysis.

In recent years, several studies have employed Bayesian methods to improve satellite PM2.5 modeling. For example, Chang et al. applied a unified Bayesian hierarchical framework to improve PM2.5-AOD modeling that allows the model to

calculate the prediction uncertainties, which are invaluable in further health impact analyses, Yu et al. utilized GPR in a Bayesian hierarchical setting to improve satellite based PM2.5-AOD estimates.

# Chapter 3. Study Area and Data

## 3.1 Study Area

The People's Republic of China has an area of about 9,600,000 km2. The eastern plains and southern coasts of the country consists of fertile lowlands and foothills and is the location of most of China's agricultural output and human population. The southern areas of the country (South of the Yangtze River) consists of hilly and mountainous terrain. The west and north of the country is dominated by sunken basins (such as the Gobi and the Taklamakan), rolling plateaus, and towering massifs. It contains part of the highest tableland on earth, the Tibetan Plateau, and has much lower agricultural potential and population.

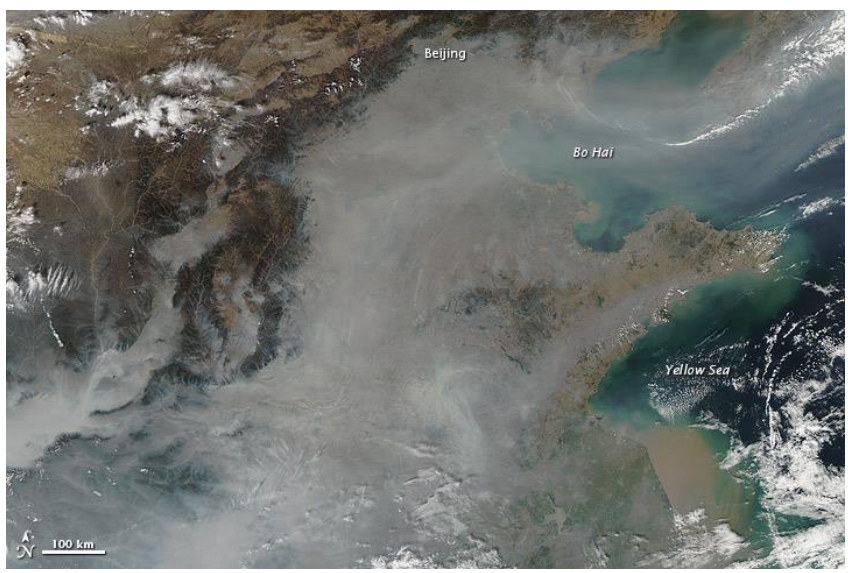


Figure 3.1 Haze hovered over eastern China on October 20, 2012 (Source: Image took by NASA's Aqua satellite)

The dense population, high-speed economic development and urbanization, industrial process, congested local traffic and coal consumption for winter heating all make the China the most concentrated region of PM2.5 over the world these days.

## 3.2 Data Acquisition

### 3.2.1 MODIS 10km products

Although the MODIS AOD started releasing its 3KM spatial resolution product since 2014 which has been recently utilized in fine scale PM2.5 prediction, displaying a richer variation than 10KM AOD, the miss value portion is so big that even difficult to interpolate in regions like Tibet, Xin Jiang, Inner Mongolia province, etc., thus not appropriate in our whole national scale model. On the other hand, the traditional 10KM works well in climate related application (Leigh et al., 2014) and providing information for more area in China.

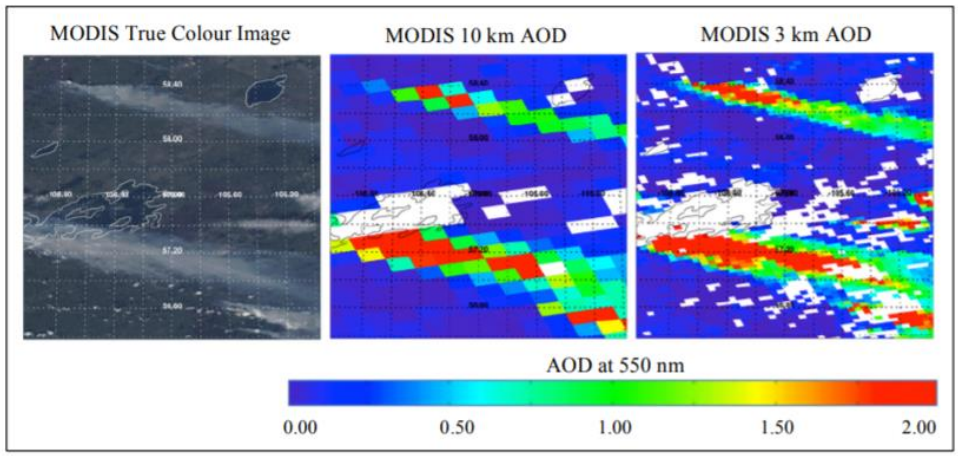


Figure 3.2 A comparison of the MODIS True Colour Image, MODIS10 km AOD and 3 km AOD Products (Source: Leigh et al., 2014)

Here we use MYD04_L2 - MODIS/Aqua Aerosol 5-Min L2 Swath 10km product that produced daily level 2 data at spatial resolution of a 10 × 10 KM pixel array and 5 minutes temporal resolution. Considering two satellites shows same performance on image quality here we only use the Aqua AOD product to control the size of whole dataset, whose time of passing eastern China is 1:30 PM and 1:30 AM every day (Bouarar et al., 2017). MODIS AOD product files are stored in Hierarchical Data Format (HDF-EOS).

### 3.2.2 PM2.5 ground monitoring data

All ground level monitoring PM2.5 data is crawled from website, https://www.aqistudy.cn. Chinese government has established the network for

PM2.5 monitoring since 2013, the number of monitoring sites increase from 946 to 1497 in 2016. Our study crawled an hourly recorded dataset from Oct 1st, 2016 to Oct 1st, 2017 in all available sites. Because our AOD data was recorded by various times a day, for simplify our data process, we use the mean daily PM2.5 monitor data to match AOD data on with same date and location.

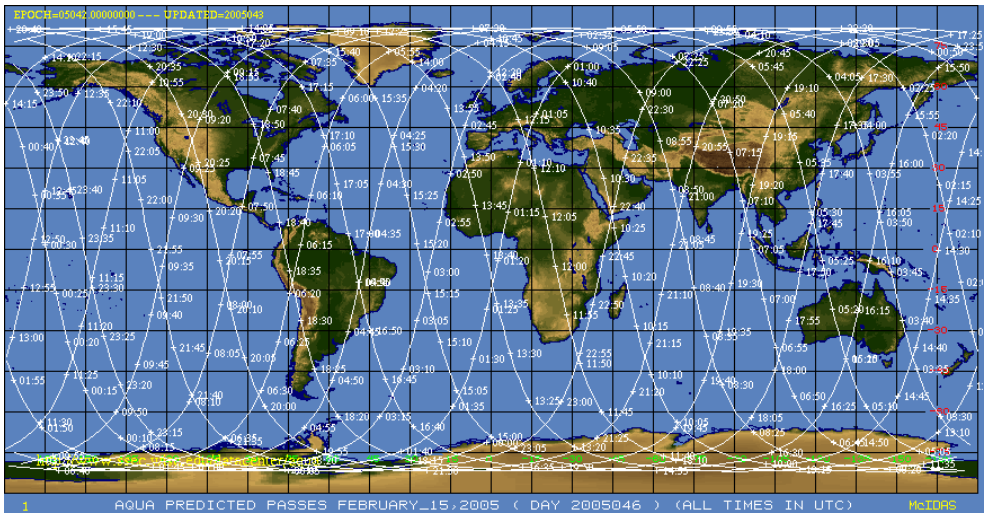


Figure3.3, Aqua predicted pass time (Source: NASA distributed active archive center)

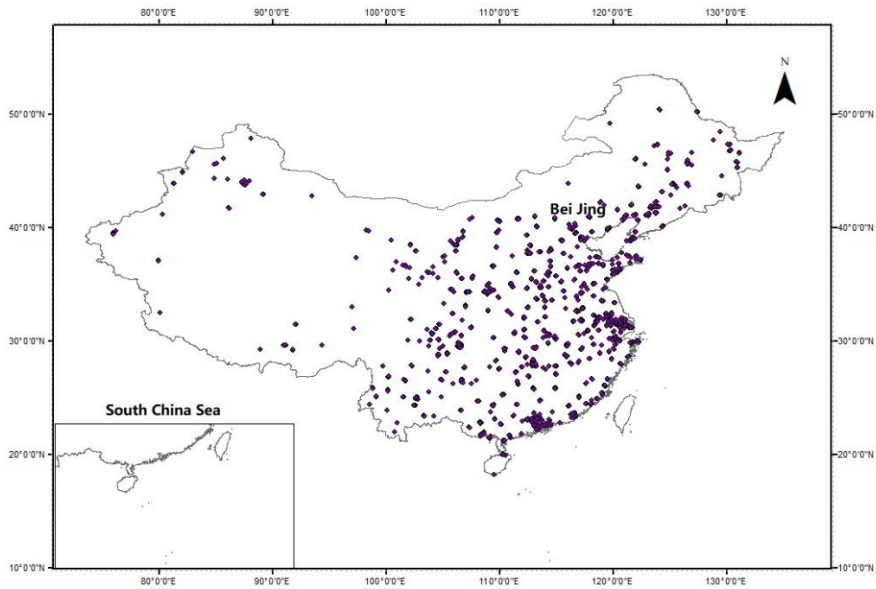


Figure 3.4 Ground monitoring stations' locations and the averaged PM2.5 of every city' all stations during the study period

The monitoring sites distribute as showed in figure. Ground-level PM2.5 concentrations were mainly measured by the TEOM and BAM instruments as introduced in Chapter 2. On the basis of the Environmental Protection Standard of China (HJ 618-2011), all the measurements had been processed with calibration and quality control (MEPCN, 2011).

### 3.2.3 Supplementary Data

Elevation data was obtained from the digital elevation model (DEM) of the Shuttle Radar Topography Mission (SRTM) with a resolution of 90m. The China

elevation map is presented in Figure 3.1.

For meteorological data, we downloaded daily ERA Interim dataset from the European Centre for Medium-Range Weatherm Forecasts (ECMWF). ECMWF uses its data assimilation systems and forecast models to re-analyze observation datasets. As one of the ECMWF's reanalysis datasets, ERA Interim is a global atmosphere reanalysis from 1979. Meteorological data calibrating in our model in the period of Oct 1st, 2016 to Sep 31$^{st}$ , 2017 includes only relative humidity (RH) and boundary layer height (BLH), which has been illustrated as main predict factor in Chapter 2.

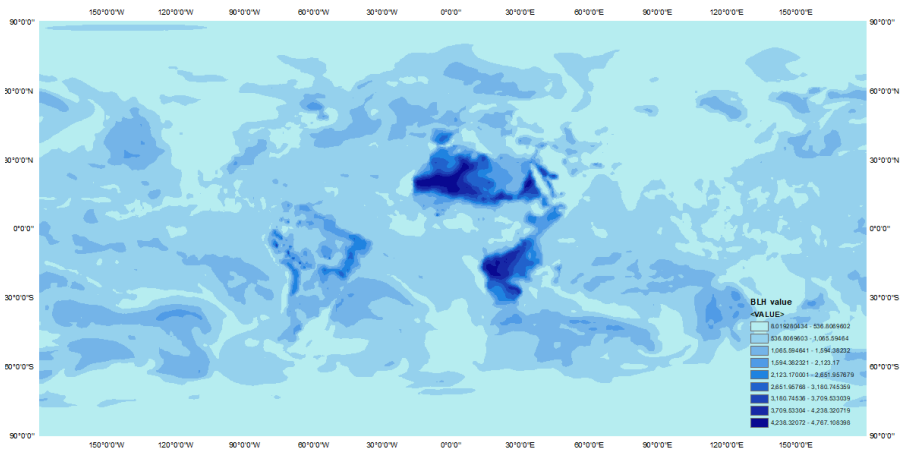


Figure 3.5, daily Global Boundary Layer map, interpolated via IDW method, point data sourced from ECMWF

| Meteorological Data | Beijing Time | Spatial Resolution |
|---|---|---|
| Relative humidity | 12:00 | 0.7 |
| Boundary Layer Height | | |

Table3.1, Meteorological data acquired from ECMWF

Data offered by ERA Interim model do not contain RH data. With Dewpoint and temperature values, the relative humidity data value was calculated using the August-Roche-Magnus approximation (Alduchov et al. 1996).

# Chapter 4. Model

## 4.1 Overview of Workflow

Methodology includes two phases: data pre-processing and model construction.

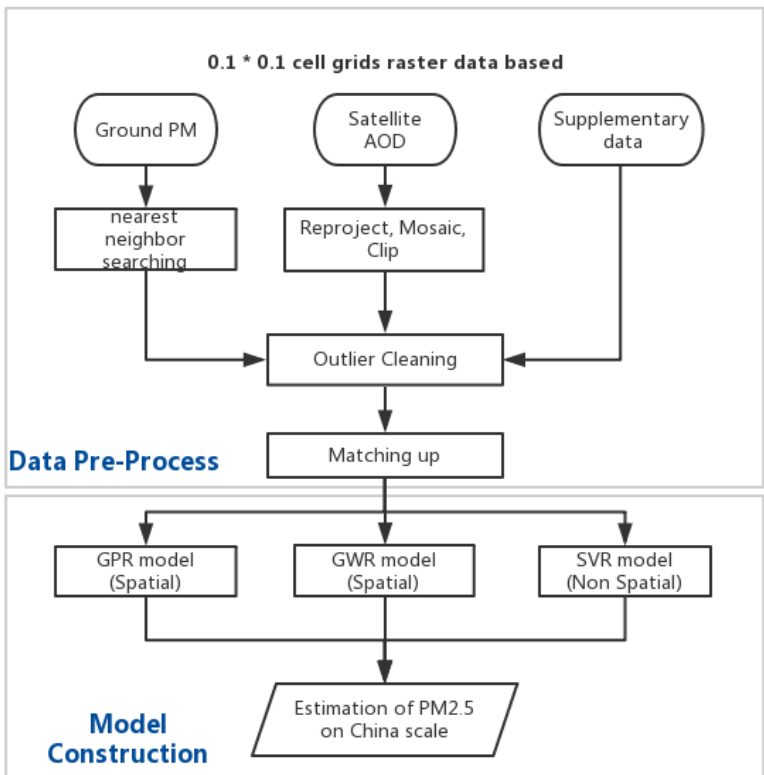


Figure4.1, Work flow of Methodology

## 4.2 Data Pre-processing

Data pre-processing is primarily integrating data from all kinds of sources into one large dataset. A 0.1° × 0.1° grid with 100,699 grid cells was created that covers all of China. All data were resampled into grids by longitude and latitude and conducted outlier identification and removal steps. PM 2.5 Data recorded over 3000 are treated as invalid data and mean of neighbor value were given while None values were given 0 while extracting from monitoring sits files.

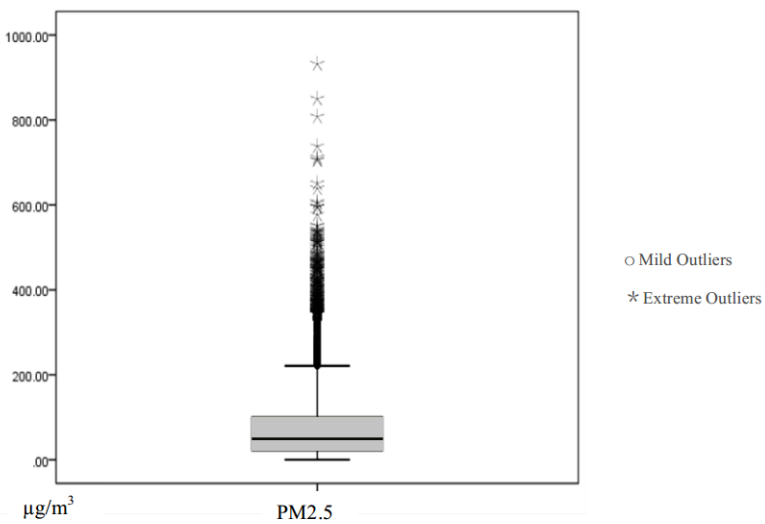


Figure4.2, Box-Plot of the whole year's PM2.5 monitoring site data

The width of AOD satellite image is 2330 km. There are over 10 images per day contains China territory took by Aqua and we reproject and mosaic the daily meta data into one image in ENVI 5.3 IDL. Resample on AOD data is also implied

because the meta data were in 10 × 10 KM resolution and cells near edge of each images tends to be distorted.

Daily data in Locations with small daily sample size of PM2.5-AOD matchups values were used as training and testing dataset. Considering the spares AOD data leads to a poor match-up result, we implied a grid based neighbor searching algorithm with bandwidth of 0.5° latitude-longitude grid, which ameliorate our matching result from a 96 average daily pairs to over 104 pairs.

A 10 folds cross validation was conducted in them. Those locations without monitoring PM2.5 values were regarded as predict matrix to generate our estimation of daily model. We resample data by extracting all other data on locations where monitor site was built by GDAL package in Python 3.6.


Figure4.3, 10 folds cross validation

The model construction module consists of the GPR model construction, Bayesian Hierarchical nature setting and hyper parameters' prior arrangement. Also, GWR model and support vector regression model were fit for result comparison.

## 4.3 Model Construction

### 4.3.1 Gaussian Process Regression Model

Gaussian stochastic processes are widely used in practice as models for geostatistical data (Gelfand, 2016). Physical justification rarely appeared in such model. Rather, they are used as convenient empirical models which can capture a wide range of spatial behavior according to the specification of their correlation structure (Diggle P J, 1998). According to previous studies, one very good reason for concentrating on Gaussian models was that they are uniquely tractable as models for dependent data. With the increasing use of computationally intensive methods, and in particular of simulation-based methods of inference, the analytic tractability of Gaussian models is becoming a less compelling reason to use them (Rasmussen and Williams, 2006).

Different from traditional geostatistical methods, which are based on certain functions, such as wavelets and splines, to depict spatial relationships, Gaussian processes are one of the most intuitive methods to model spatial surfaces as realization of stochastic processes. Specifically, Gaussian processes consider the

spatial effect as random variables by specifying their means and covariance functions, which is the major feature that distinguishes them from other traditional methods. What's more, the hierarchical nature can help explain various sources of variations in PM2.5.

In our model, the Gaussian process for PM2.5 predicting can be interpreted as following: firstly, PM2.5 concentrations in China follows a conditional distribution of AOD values, spatial and non-spatial random effects, which is the basic foundation in the hierarchical setting; the second stage is mainly aimed at specifying the distribution of spatial random effects in PM2.5 and AOD relationship. It is modeled by Gaussian processes with specific mean surface and covariance functions; Last stage focus on the conditional distribution of the covariance functions we set for the GPR given by the hyperparameters we chose empirically. This hierarchical approach is helpful when dealing with ambiguous variations.

Comparatively, in GWR models, coefficients of each predictor variable, AOD, RH, PBLH and intercept varies along with locations. In Gaussian processes settings, these coefficients and the intercept remain the same in each daily mode. While the geographical variation was simulated by the spatial random effect. Thus, compared to GWR, Gaussian processes separate out different sources of variation (the independent variable AOD, RH, PBLH, spatial random effects and non-spatial random effects) in explaining PM2.5.

Our Gaussian process model in Bayesian hierarchical setting is as following:

Giving training data $x_1, x_2 \ldots x_n$ and corresponding observe value $y_1, y_2 \ldots y_n$, in Gaussian process, function of $y$ was not assumed as specific formula like $f(x) = \text{mx} + \text{c}$ or $f(x) = ax^2 + bx + c$ but considered as an Infinite dimension point from Gaussian process.

For observation with noise:

$$Y = f(x) + N(0, \sigma^2)$$

a gaussian process prior is given to $f(x)$

$$f(x) \sim \text{GP}(0, \mathbf{K})$$

K is the covariance function. With noise, the k is

$$\mathbf{K} = \mathbf{k}(x_n, x_m) + \sigma^2 \boldsymbol{\delta}_{n,m}$$

Because of conjunction of Gaussian, the joint distribution of training data and test data is still Gaussian. With new input vector $\boldsymbol{x}^*$, the joint Gaussian of $y$ and $y^*$ is

$$\begin{bmatrix} \boldsymbol{y} \\ y^* \end{bmatrix} \sim N(0, \begin{matrix} \boldsymbol{K} & \boldsymbol{K}_*^{\boldsymbol{T}} \\ \boldsymbol{K}_* & \boldsymbol{K}_{**} \end{matrix})$$

The distribution of $y^*$ is conditional distribution of $P(y, y^*)$, a Gaussian as well

$$y^* | \boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{K}_* \boldsymbol{K}^{-1} \boldsymbol{y}, \boldsymbol{K}_{**} - \boldsymbol{K}_* \boldsymbol{K}^{-1} \boldsymbol{K}_*^{\boldsymbol{T}})$$

Basically, the mean of distribution above is used as prediction

$$\overline{y^*} = \boldsymbol{K}_* \boldsymbol{K}^{-1} \boldsymbol{y}$$

In this study, the PM2.5 daily estimation Y on site i is supposed to be followed a Gaussian model as following:

$$Y_i = \boldsymbol{X}_i{}^T \boldsymbol{\beta} + \omega$$

Where $\boldsymbol{X}_i$ is the input vector at location i, including AOD data, relative humidity, boundary layer height, $\boldsymbol{\beta}$ is the weight vector including intercept and slopes corresponding to input $\boldsymbol{X}_i$. The spatial random effect $\omega$ follows a multivariate Gaussian process with covariance function $\boldsymbol{K}$, the function is specified as a exponential model as following, this function outputs a covariance matrix, D is distance between two sites i and j and the covariance. We use two parameters, $\eta^2$and $\rho^2$ to define a squared distance function, which is a common assumption. In the last piece, $\delta_{ij}\sigma$ can be treated as the nugget together, with $\boldsymbol{\delta}$ being a diagonal unit matrix and $\sigma$ being random error. If i not equal to j then this part well not matter because $\delta_{ij}$ is zero.

$$\omega \sim Multivariate\ Gaussian\ process(\mathbf{0}, \boldsymbol{K})$$

$$\boldsymbol{K} = \eta^2 exp(-\rho^2 \boldsymbol{D}^2) + \boldsymbol{\delta}\sigma$$

Where $K_{ij} = \eta^2 exp\left(-\rho^2 D_{ij}^2\right) + \delta_{ij}\sigma$

$$\sigma \sim Normal(0, \tau^2)$$

The prior of hyper parameters are initially set as:

$$\beta_0 \sim Normal(0,100)$$

$$\beta_1 \sim Normal(0,100)$$

$$\beta_2 \sim Normal(0,100)$$

$$\beta_3 \sim Normal(0,100)$$

$$\tau^2 \sim HalfCauchy(0, 1)$$

$$\eta^2 \sim HalfCauchy(0, 1)$$

$$\rho^2 \sim HalfCauchy(0, 1)$$

We also defined the prior distributions for each parameter. Specifically, the mean parameters βfollow normal distributions with assigned means and covariances. The variance parameters $\tau^2$ $\eta^2$, and $\rho^2$ all obey half Cauchy distributions with shape hyperparameters equal to 2 (thus, the variance is infinite, by definition). Reasons for selecting the corresponding prior distribution for each parameter are twofold. On the one hand, the type of distribution of each parameter was chosen by referencing previous studies. On the other hand, some of the values of hyperparameters were selected so that each parameter has a broad range of potential values (greater variance), which allows for daily variations. The selection of prior distributions is mainly heuristic and subject to change. The parameters were updated using the Metropolis-Hastings algorithm. A summary of the prior

distributions of all the parameters is presented in the Chapter 5.

We set the number of iterations for each parameter to 5,000. By monitoring the changes in these parameters, we found that they changed dramatically from the beginning (within 3,000 iterations) and gradually stabilized over time (See Supplementary Information, Text S2). Ten, we recovered the regression coefficients β and spatial random effects I from the parameters after a burn-in period of 3,000 iterations. Regarding the model fitting and cross-validation processes, we obtained the mean value of daily predictive PM2.5 according to the parameters of each iteration.

In our study, the model build, MCMC iterations and posterior prediction are all carried out in python 3.6 with Pymc3 package. Daily model was first trained with daily input and observed data, then test with 20% test points. For over-fit detection, 10-fold cross validation was conducted as well for daily model.

### 4.3.2 Geographically Weighted Regression Model

Geographically weighted regression (GWR) was adopted to explore the local spatial heterogeneity of the causal relationships between PM2.5 concentrations and geographic Factors (Luo, et al, 2017). The traditional GWR model on a daily basis can be expressed as

$$Y_i = \boldsymbol{\alpha_0}(\boldsymbol{u_i}, \boldsymbol{v_i}) + \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{\alpha_j}(\boldsymbol{u_i}, \boldsymbol{v_i})$$

Where $\boldsymbol{X}_i$ is the input vector at location i, including AOD data, relative humidity, boundary layer height, $\boldsymbol{\alpha}$ is the weight vector including intercept and slopes corresponding to input $\boldsymbol{X}_i$.

In GWR model, the regression coefficients show the local spatial variation, and the standard errors of the coefficients illustrate the reliability of the estimated coefficient (Gao et al, 2012). GWR v4.0 with the adaptive bandwidth and bi-square kernel was implemented to build the model. After the spatial autocorrelation analysis, a 10-fold Cross Validation (CV) was conducted to verify whether the GWR model was over-fitted or not.

### 4.3.3 Support Vector Regression Model

Support vector regression (SVR) has been proposed as a good alternative to fit relationship between PM25 and PM10 data and had a high generalization performance regardless of the big geographical characteristics differences of those stations (Song et al, 2014).

The SVR model solve following optimization problem:

$$\min_{w,b} \frac{1}{2} w^T w C \sum_{i=1}^{i} (\xi_i + \xi_i^*)$$

$$\text{s.t.}\begin{cases} \mathrm{y_i} - (w^T \emptyset(x) + b) \le \epsilon \xi_i \\ (w^T \emptyset(x) + b) - \mathrm{y_i} \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^*, i = 1 \dots l \end{cases}$$

where $\phi(x)$ is the kernel function, w is the margin and $x$ is input vector, including AOD data, relative humidity, boundary layer height, and coordinate of location as an aid spatial data, and $y$ is the observe data, $\epsilon$ is the lose function, $\xi$ and $\xi*$ are slack variables which quantify the estimation errors greater than $\epsilon$, penalty parameter C controls the norm of the weights w. As the result of equation above, where $\theta$ is Lagrange multiplier,

$$f(x) = \sum_{i=1}^{i} \theta \emptyset(x, x_i) + b$$

# Chapter 5. Results and Analysis

## 5.1 Descriptive Statistics on Dataset

Before modeling, to validate whether our dataset meets requirement of building a model, analysis of correlation among each variable is needed. The result of descriptive statistics is as following. In our dataset, the average AOD data and PM2.5 data is 0.497 and 56 respectively. Comparing with same time period annual mean and Std dv in the U.S., figures in China are higher, which showed that air quality is worse and AOD-PM2.5 model is more complex and more uncertain.

| Variables | Min | Max | Mean | Std dv |
|---|---|---|---|---|
| **AOD** | 0.001 | 3.095 | 0.497 | 0.322 |
| **BLH** | 0.118 | 2.587 | 1.120 | 0.487 |
| **RH** | 17.9 | 96.7 | 58.9 | 18.2 |
| **PM2.5** | 5 | 701 | 56 | 38.671 |

Table5.1, descriptive statistics of dataset

| Index | AOD | PBLH | RH |
|---|---|---|---|
| **AOD** | 1 | 0.067 | -0.116 |
| **BLH** | 0.067 | 1 | -0.192 |
| **RH** | -0.116 | -0.192 | 1 |

Table5.2, Person correlation among all input variables in this study

We had reviewed relationship of PM2.5 between BLH and RH in section 2 from perspective of AOD definition. The person correlation result showed that correlation of AOD between BLH and RH are low and no collinearity exists.

| | Mean PM2.5 | Mean AOD | SD PM2.5 | SD AOD |
|---|---|---|---|---|
| **Spring** | 54.30 | 0.48 | 42.65 | 0.321 |
| **Summer** | 41.87 | 0.47 | 40.19 | 0.287 |
| **Autumn** | 52.91 | 0.53 | 47.90 | 0.298 |
| **Winter** | 87.33 | 0.55 | 55.12 | 0.311 |

Table5.3, descriptive statistics of seasonal dataset

## 5.2 Model Validation

We fit our model on a daily scale, totally 365 models were trained for each day from Oct 1$^{st}$, 2016 to Sep 31$^{st}$, 2017. All parameters in our Bayesian hierarchical framework model were updated using the Metropolis-Hastings algorithm. For instance, in one daily model Oct 20$^{th}$, 2016, we set the number of iteration (n

samples) equal to 500, 5,000 to see how each parameter would react using the Metropolis-Hastings algorithm.

Changes were plotted for all the parameters over time to determine the appropriate number of iterations due to limitation on computing resource. An iteration number as small as possible meanwhile ensuring parameters converge in all 365-daily model is what we want. Notably, the trace of all the parameters does not go steady in 500 sample iterations and still varies drastically, which meant that 500 iterations were not sufficient for the parameters to converge. The 5000 samples trace showed a stable trend in all parameters. The mean of intercept and beta does not differ greatly in two sample results though, means of other three hyperparameters were quite different. Moreover, in second figure the trace reminded steady after 1000 iterations. Compared with trace of more iterations, we find that use of more iterations will not increase stability. In order not to hinder further application on large dataset of all daily models, we use mean of 1000 samples to estimate the parameter and predict unknown locations.

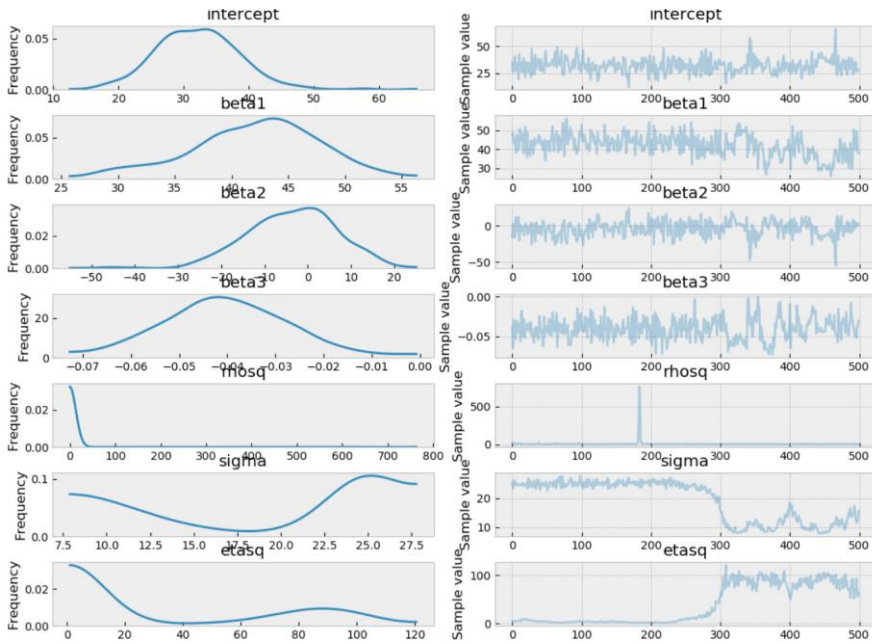


Figure5.1, parameters trace plot after 500 iterations

Scatter plot in Figure showed performance of our predictive model. The figure indicates that PM25-Predict match points in our model is centralized near the ideal line y=x. The R2 and RSME was 0.88 and 14.89, showing a pretty precise performance of fitting data.

Over fit problem is common in AOD-PM25 modeling study, which means model performs well in training data but does not work in test data. To validate whether our Gaussian process model has over-fit problem, a 10-fold cross validation was conducted in our training-test dataset. Because the long iterating time consumed by MCMC algorithm for parameter estimation, we only used 1/30 daily dataset.

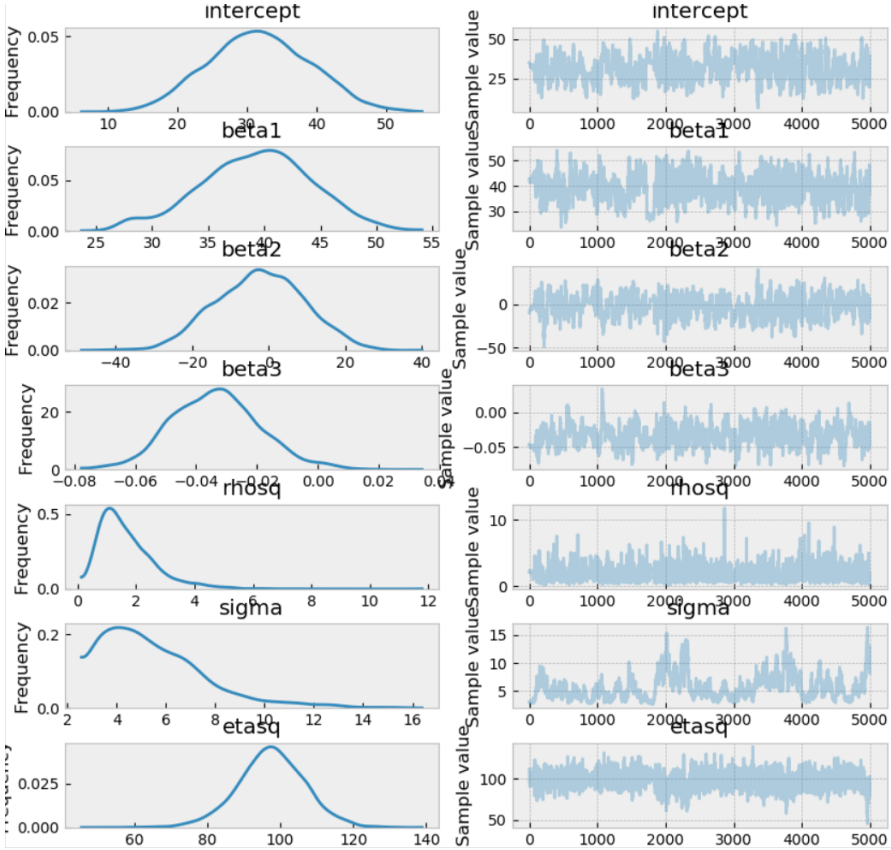


Figure5.2, parameters trace plot after 5000 iterations

The R2 and RMSE was 0.70 and 49.07 respectively, showing that Gaussian processes in the Bayesian hierarchical setting may provide an improved description of daily spatial variations and generate more precise model results, given appropriate data. The high R2 and small RMSE figures were found in recent machine learning AOD-PM2.5 estimation study.

In contrast with GWR and the SVR model, our GPR is more strongly over-

fitted. The value of R2 of GPR dropped from 0.88 to 0.70, whereas GWR showed a slightly small variation from 10-fold cross validation to one data set training. However, both the result of 10-fold cross validation and one data set training of our GPR model is better than the result of GWR one data set training. Nevertheless, previous studies of GWR on AOD-PM2.5 estimation showed a much more precise estimation result, in our study, the robust decays in large 400 sites national scale daily dataset.

| **Model** | **N** | **R2** | **RMSE (μg/m3 )** |
|---|---|---|---|
| **GPR** | 37960 | 0.88 | 14.89 |
| **GWR** | 37960 | 0.68 | 56.48 |
| **SVR** | 37960 | 0.57 | 98.41 |

Table 5.4 Statistical results of all daily models

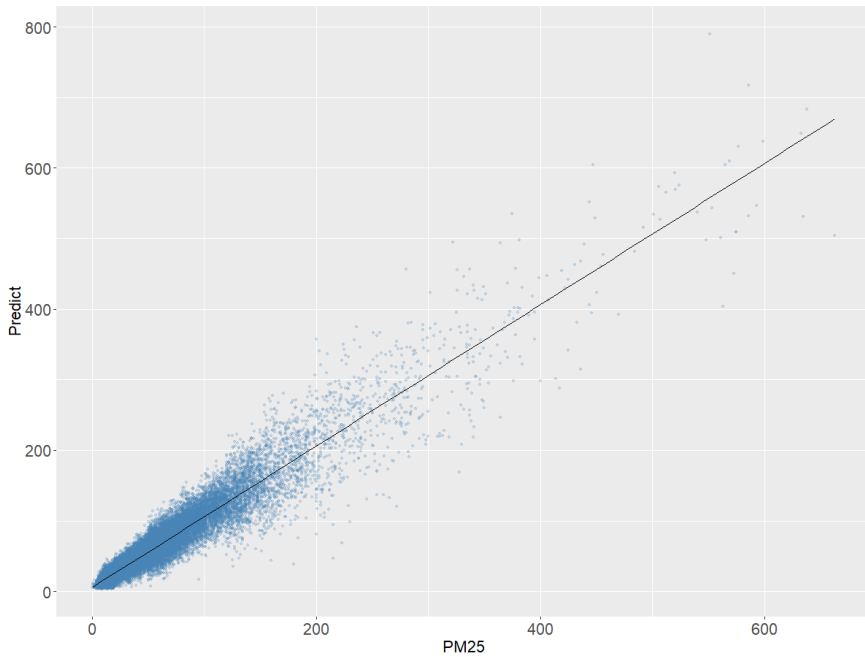


Figure5.3, Scatter plot of predict values and real PM2.5

To detect the seasonal variation of AOD-PM2.5, we applied Gaussian process model to seasonal mean dataset. The overall spatial patterns and local details of our model-estimated values are satisfactorily consistent with the ground-based recording data from monitoring sites in a seasonal scale. The contrast of spatial coverage showed that though satellite-retrieved AOD is not able to cover all china territory but possess better spatial coverage than ground-based PM2.5 monitoring sites can do. Covered area was lifted by a substantial extension in both spatial and temporal term using valid PM2.5 data on the national grid rather than air quality ground-based monitoring network only.

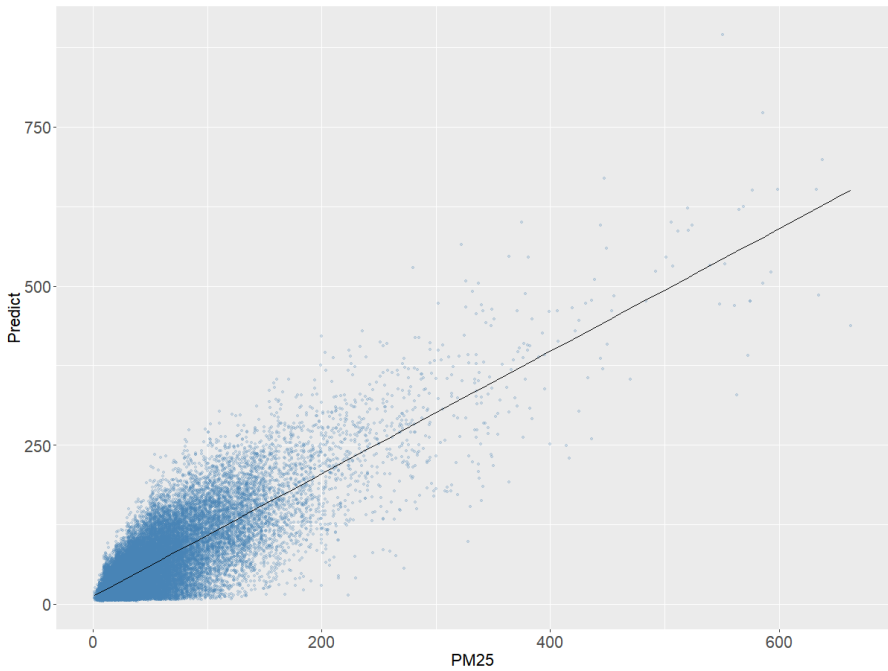


Figure5.4, 10-fold validation scatter plot of predict values and real PM2.5

Monitoring network are mainly concentrated in Beijing, Tinjing Hebei area and eastern coastal areas. Whereas our predicted value showed a more comprehensive coverage of the whole area of China. This indicates that a better spatial and temporal coverage is in our AOD-PM2.5 compared to the ground-based monitoring network.

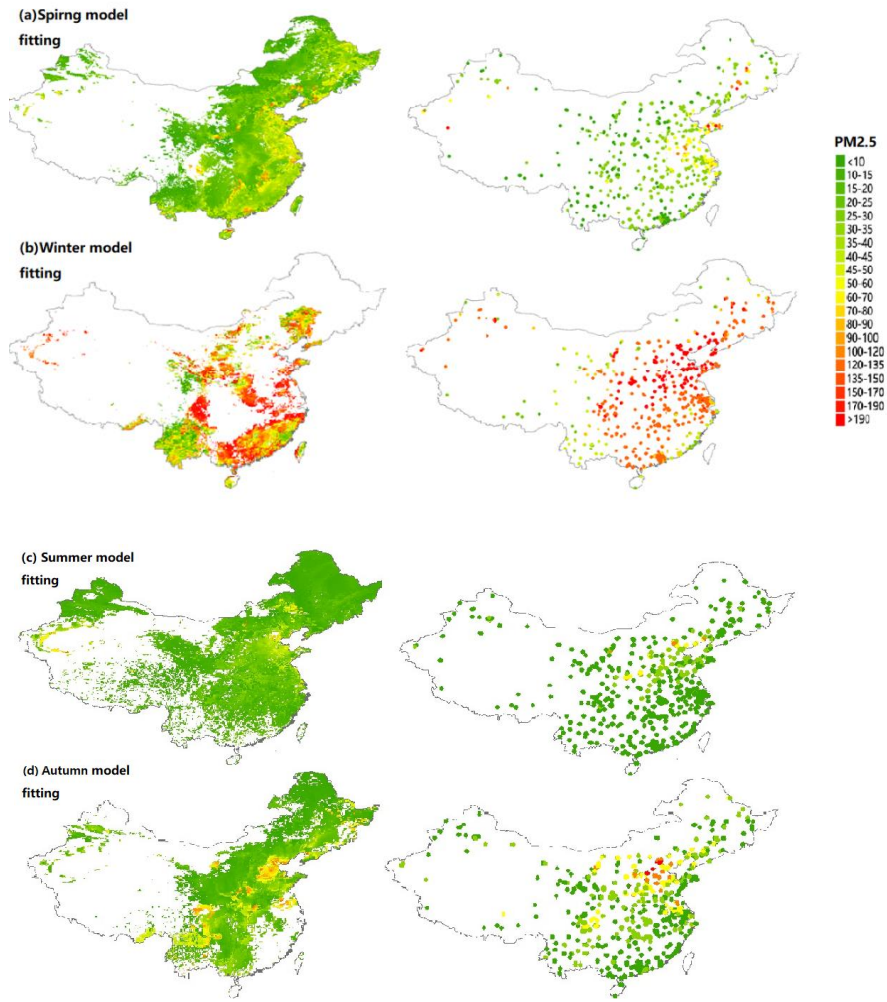


Figure5.5, Seasonal distributions of PM2.5 concentrations estimated using the Bayesian Gaussian process model

As for the difference in the seasonal trends, winter has the highest seasonal average PM2.5 value of all the seasons, with a value of 89.62 μg/m3. Spring and autumn has close values of 44.10 μg/m3 and 43.06 μg/m3, nevertheless summer is

the lowest among four seasons with average equal to 18.08 μg/m3. One notable reason can be inferred is the employment of heating system powered by coal fire. The spatial socio-economic development imbalance and diversity of Chinese landscape lead to an apparent geographic variation among the various parts of China. The central eastern coastal regions, central plain regions, as well as North China Plain and the Sichuan Basin, compared to other area, remarkably possess worse air quality.

Additionally, although fitting hierarchical models can be time-consuming owing to the large sample size and high cost of matrix decomposition, which is known as a "large-N" problem. Our model run on a PC with i7 6700k CPU and 8GB memory, 1000 iteration MCMC algorithm computation for each daily model takes 30mins, which is acceptable.

# Chapter 6. Conclusions and Limitations

## 6.1 Conclusion

This study aimed to improve the modeling performance on PM2.5-AOD relationship. We implied the Gaussian process regression Bayesian hierarchical AOD-PM2.5 model on a year scaled dataset and analyzed the performance of model, seasonal and spatial variation of PM2.5 values. The key findings responding to each specific objective are summarized below.

**Explanation of sources of PM2.5 variation**

We discussed relationship between AOD data and PM2.5. and selected Planetary boundary layer and relative humidity were chose as predictor factors among meteorology variables in our model by reviewing definition of AOD. We also discussed feasibility of MODIS AOD data in china is conducted through literature review. We used a hierarchical setting to help explain various sources of variations in PM2.5 with a linear group of intercept and coefficients of AOD, planetary boundary layer and relative humidity, and a spatial random effect to capture the geographic variation and a non-spatial random effect.

**MCMC algorithm runs efficiently for Bayesian hierarchical model on China national scale dataset**

Spatial relationships in our research were considered as random variables and

simulated by Gaussian process through giving a hierarchical explanation of multiple sources of PM2.5 variation. Although fitting the hierarchical models is always considered time-consuming owing to the large sample size and high cost of matrix decomposition, our research showed that MCMC algorithm performed computed effectively on a national scaled data with over 300 inputs in daily model.

**Model performance of Gaussian process regression**

Gaussian process model in this study exhibited remarkable performance. Compared to the commonly way of modeling such relationship, Geographic weighted regression, the Gaussian process model increased the model cross-validation R2. The Bayesian hierarchical setting helped we estimated a spatial random effect that captured the spatial variance in the non-stationary spatial data.

**Seasonal and spatial analysis of PM2.5 estimation result**

We trained our model with seasonal mean values and gave a specific analysis on seasonal and spatial variation of PM2.5 concentration. The seasonal estimation with a large spatial AOD coverage showed a high consistency with spatial patterns of PM2.5 ground monitoring data. The seasonal and spatial distribution and its contribution factors also been discussed.

## 6.2 Limitation of this study

Near the Earth's surface, ground-level PM2.5 are recorded while aerosol represents its whole distribution in atmosphere. However, we directly used AOD data regardless of the vertical structure and components of aerosol, which might reduce accuracy of estimation

If the Terra MODIS AOD was also explored and mosaic with Aqua data to generate a complete dataset of AOD, the non-retrieved day's AOD might be reduced and the PM2.5 might be estimated with a higher accuracy and a greater coverage.

Supplementary data like more meteorological parameters, population density, GDP, local industrial output and land use information were not wildly used in our Gaussian process model. More variables are expected to explain more details spatial variation and seasonal variation via model fitting. This is a limitation that needs to be further examined in our future research work.

Also, the computing consumption of Bayesian hierarchical solution need to be reduced by improving in an efficient programming perspective.

# Bibliography

Alduchov O A, Eskridge R E. Improved Magnus form approximation of saturation vapor pressure[J]. Journal of Applied Meteorology, 1996, 35(4): 601-609.

Baccarelli, A., 2009. Breathe Deeply into Your Genes! Genetic Variants and Air Pollution Effects. American Journal of Respiratory and Critical Care Medicine, 179(6), pp.431-432.

Bell, M.L., Son, J.Y., Peng, R.D., Wang, Y. and Dominici, F., 2015. Brief Report: Ambient PM2. 5 and Risk of Hospital Admissions: Do Risks Differ for Men and Women? Epidemiology, 26(4), pp.575-579

Bishop C. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn[J]. Springer, New York, 2007.

Bouarar I, Wang X, Brasseur G P. Air Pollution in Eastern Asia: An Integrated Perspective[J]. 2017.

Chen, J., Qiu, S., Shang, J., Wilfrid, O.M., Liu, X., Tian, H. and Boman, J., 2014. Impact of relative humidity and water soluble constituents of PM2. 5 on visibility impairment in Beijing, China. Aerosol Air Quality Research, 14, pp. 260-268.

Chu D A, Kaufman Y J, Ichoku C, et al. Validation of MODIS aerosol optical depth retrieval over land[J]. Geophysical research letters, 2002, 29(12).

Diggle P J, Tawn J A, Moyeed R A. Model-based geostatistics[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1998, 47(3): 299-350.

Engel-Cox, J.A., Holloman, C.H., Coutant, B.W. and Hoff, R.M., 2004. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. Atmospheric Environment, 38(16), pp. 2495-2509.

Finley, A. O., Banerjee, S. & Carlin, B. P. spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. Journal of Statistical Sofware 19, 1–24 (2007).

Gao, Y., Huang, J., Li, S. & Li, S. Spatial pattern of non-stationarity and scale-dependent relationships between NDVI and climatic factors—A case study in Qinghai–Tibet Plateau, China. Ecol. Indic. 20, 170–176 (2012).

Gelfand, Alan E., and Erin M. Schliep. "Spatial statistics and Gaussian processes: A beautiful marriage." Spatial Statistics 18 (2016): 86-104.

George, J.P., Harenduprakash, L. and Mohan, M., 2008, February. Multi year changes of Aerosol Optical Depth in the monsoon region of the Indian Ocean since 1986 as seen in the AVHRR and TOMS data. In Annales Geophysicae, 26 (1), pp. 7-11.

Guo,Yuling., Li-hua, X., Fang, W. and Jing, H., 2009, May. The progress and prospect of remote sensing for aerosol optical depth. Urban Remote Sensing Event, 2009, pp. 1-6.

Gupta, P. and Christopher, S.A., 2009a. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. Journal of Geophysical Research: Atmospheres, 114(D14).

Gupta, P., Christopher, S.A., Wang, J., Gehrig, R., Lee, Y.C. and Kumar, N., 2006. Satellite remote sensing of particulate matter and air quality assessment over global cities. Atmospheric Environment, 40(30), pp.5880-5892.

Hadjimitsis, D.G., 2009. Aerosol optical thickness (AOT) retrieval over land using satellite image-based algorithm. Air Quality, Atmosphere & Health, 2(2), pp. 89-97.

Holben, B.N., Eck, T.F., Slutsker, I., Tanre, D., Buis, J.P., Setzer, A., Vermote, E., Reagan, J.A., Kaufman, Y.J., Nakajima, T. and Lavenu, F., 1998. AERONET—A federated instrument network and data archive for aerosol characterization. Remote Sensing of Environment, 66(1), pp. 1-16.

Holben, B.N., Tanre, D., Smirnov, A., Eck, T.F., Slutsker, I., Abuhassan, N., Newcomb, W.W., Schafer, J.S., Chatenet, B., Lavenu, F. and Kaufman, Y.J., 2001. An emerging groundÑbased aerosol climatology: Aerosol optical depth from AERONET. Journal of Geophysical Research: Atmospheres, 106(D11), pp. 12067-12097.

He Q, Li C, Tang X, et al. Validation of MODIS derived aerosol optical depth over the Yangtze River Delta in China[J]. Remote Sensing of Environment, 2010, 114(8): 1649-1661.

Hu, Z. Spatial analysis of MODIS aerosol optical depth, PM2.5, and chronic coronary heart disease. International Journal of Health Geographics 8, 1–10, doi:10.1186/1476-072x-8-27 (2009).

Kaufman, Y.J. and Sendra, C., 1988. Algorithm for automatic atmospheric corrections to visible and nearIR satellite imagery. International Journal of Remote Sensing, 9(8), pp.1357-1381.

Kumar, N., Chu, A. and Foster, A., 2007. An empirical relationship between PM 2.5 and aerosol optical depth in Delhi Metropolitan. Atmospheric Environment, 41(21), pp.4492-4503.

Lang, J.L., Cheng, S.Y., Li, J.B., Chen, D.S., Zhou, Y., Wei, X., Han,

Lee, H.J., Coull, B.A., Bell, M.L. and Koutrakis, P., 2012. Use of satellite-based aerosol optical depth and spatial clustering to predict ambient PM 2.5 concentrations. Environmental Research, 118, pp. 8-15.

Li, C., Lau, A.K.H., Mao, J. and Chu, D.A., 2005. Retrieval, validation, and application of the 1-km aerosol optical depth from MODIS measurements over Hong Kong. IEEE Transactions on Geoscience and Remote Sensing, 43(11), pp. 2650-2658.

Li, Q. and Wang, J., 2015. Aerosol Retrieval Using Remote Sensing Data over Jing-Jin-Ji Area with SARA Algorithm. Scientific Journal of Earth Science, 5(3).

Lin, C., Li, Y., Yuan, Z., Lau, A.K., Li, C. and Fung, J.C., 2015. Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM 2.5. Remote Sensing of Environment, 156, pp.117-128.

Liu, Y., He, K., Li, S., Wang, Z., Christiani, D.C. and Koutrakis, P., 2012. A statistical model to evaluate the effectiveness of PM 2.5 emissions control during the Beijing 2008 Olympic Games. Environment International, 44, pp.100-105.

Luo J, Du P, Samat A, et al. Spatiotemporal Pattern of PM2. 5 Concentrations in

Mainland China and Analysis of Its Influencing Factors using Geographically Weighted Regression[J]. Scientific reports, 2017, 7: 40607.

NASA., n.d.b. MISR Introduction. https://www-misr.jpl.nasa.gov/index.cfm https://wwwmisr.jpl.nasa.gov/index.cfm. Accessed 20 March 2016.

NOAA., n.d. ESRL Global Monitoring Division - GRAD - Surface Radiation Budget Network (SURFRAD). http://www.esrl.noaa.gov/gmd/grad/surfrad/aod/. Accessed 20 March 2016.

Paciorek, C.J., Liu, Y., Moreno-Macias, H. and Kondragunta, S., 2008b. Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level PM2. 5. Environmental Science & Technology, 42(15), pp.5800-5806.

Remer, L.A., Mattoo, S., Levy, R.C. and Munchak, L., 2013. MODIS 3km aerosol product: algorithm and global perspective. Atmospheric Measurement Techniques Discussions (AMTD), 6; pp. 69-112

Roberts, G., Mauger, G., Hadley, O. and Ramanathan, V., 2006. North American and Asian aerosols over the eastern Pacific Ocean and their role in regulating cloud condensation nuclei. Journal of Geophysical Research: Atmospheres, 111(D13).

Song, Lei, et al. "Spatio-temporal PM 2.5 prediction by spatial data aided incremental support vector regression." Neural Networks (IJCNN), 2014 International Joint Conference on. IEEE, 2014.

Sun A Y, Wang D, Xu X. Monthly streamflow forecasting using Gaussian process regression[J]. Journal of Hydrology, 2014, 511: 72-81.

Tie, X., Madronich, S., Li, G., Ying, Z., Zhang, R., Garcia, A.R., Lee-Taylor, J. and Liu, Y., 2007. Characterizations of chemical oxidants in Mexico City: A regional chemical dynamical model (WRF-Chem) study. Atmospheric Environment, 41(9), pp. 1989-2008.

Van Donkelaar, A., Martin, R.V., Brauer, M., Kahn, R., Levy, R., Verduzco, C. and Villeneuve, P.J., 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environmental Health Perspectives, 118(6), pp.847.

Wang L, Xin J, Wang Y, et al. Evaluation of the MODIS aerosol optical depth retrieval over different ecosystems in China during EAST-AIRE[J]. Atmospheric Environment, 2007, 41(33): 7138-7149.

WHO., 2015. WHO Expert Consultation: Available evidence for the future update of the WHO Global Air Quality Guidelines (AQGs). http://www.euro.who.int/__data/assets/pdf_file/0013/301720/Evidence-future-update-AQGs-mtgreport-Bonn-sept-oct-15.pdf. Accessed 20 March 2016.

You, W., Zang, Z., Zhang, L., Li, Y., Pan, X. and Wang, W., 2016. National-Scale Estimates of GroundLevel PM2. 5 Concentration in China Using Geographically Weighted Regression Based on 3 km Resolution MODIS AOD. Remote Sensing, 8(3), pp. 184.

Yu, S., Mathur, R., Schere, K., Kang, D., Pleim, J., Young, J., Tong, D., Pouliot, G., McKeen, S.A. and Rao, S.T., 2008. Evaluation of realÑtime PM2. 5 foreca

Yu, S., Mathur, R., Schere, K., Kang, D., Pleim, J., Young, J., Tong, D., Pouliot, G., McKeen, S.A. and Rao, S.T., 2008. Evaluation of realÑtime PM2. 5 forecasts and process analysis for PM2.5 formation over the eastern United States using the EtaÑCMAQ forecast model during the 2004 ICARTT study. Journal of Geophysical Research: Atmospheres, 113(D6).