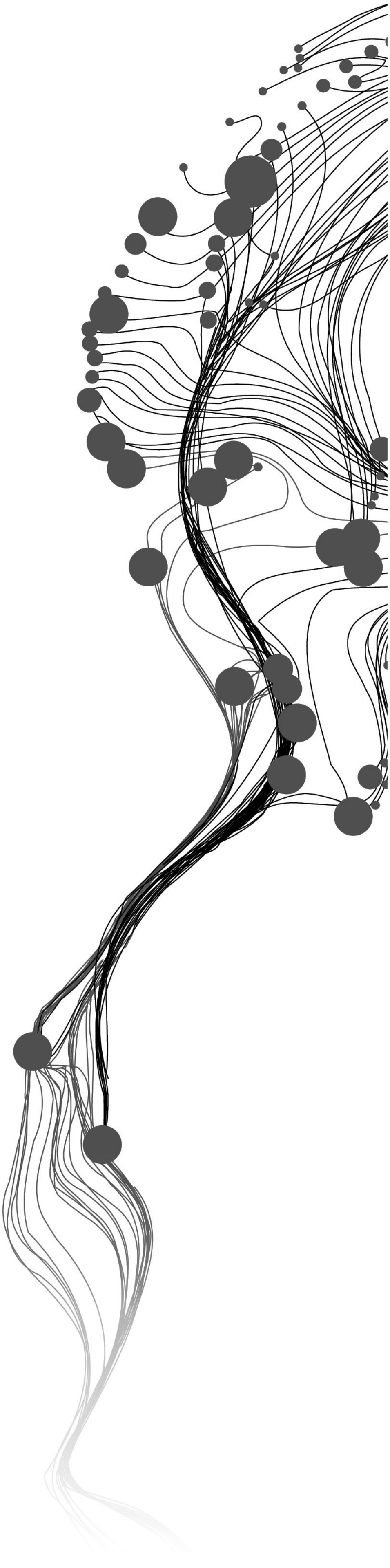
DECOMPOSITION OF PARTICULATE MATTER IN TO ITS COMPONENTS AND THEIR PREDICTION: BAYESIAN HIERARCHICAL MODELING

SHARAD GORAKH SHINGADE
February, 2012

SUPERVISORS:

Dr. Nicholas Hamm
Prof. Dr. Ir. Alfred Stein



DECOMPOSITION OF PARTICULATE MATTER IN TO ITS COMPONENTS AND THEIR PREDICTION: BAYESIAN HIERARCHICAL MODELING

SHARAD GORAKH SHINGADE
Enschede, The Netherlands, February, 2012

Thesis submitted to the Faculty of Geo-information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:

Dr. Nicholas Hamm
Prof. Dr. Ir. Alfred Stein

THESIS ASSESSMENT BOARD:

Dr. Ir. Alfred Stein (chair)
Dr. Ir. G. B. M. Heuvelink

Disclaimer

This document describes work undertaken as part of a programme of study at the Faculty of Geo-information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Particulate Matter (PM) receives global attention due to their association with human health and environment. The effects caused by PM depends on the chemical composition, origin and particle size. Detailed knowledge of PM and its components are required for understanding their effects, source appointment studies and policy making. PM and its components measured by in situ measurement techniques are limited at few locations which leads to uncertainty in prediction. To overcome the above mentioned problem, a study which could efficiently predict the decomposition of Particulate Matter into its components was required. Thus, in this study, PM components were modeled in Bayesian hierarchical paradigm with added strength from a densely gridded covariate like CTM (chemical transport model) and AOT (aerosol optical thickness). Bayesian hierarchical modeling have an advantage over classical geostatistical modeling as it takes into account the parameter uncertainty during prediction. In this research we develop models in Bayesian paradigm considering different approach. To understand the potential of adding covariable in to modeling, a model was developed with adding CTM given covariable and another model developed with CTM covariable along with AOT data. The PM component (PM_{10}) predicted with one model (RMSE = 0.5646) and the other (RMSE= 0.5632) shows similar value of RMSE. To incorporate PM components relationship; three models, namely, Model A, Model B and Model C were developed. Model A does not incorporate PM component relationship in to modeling and shows RMSE 0.6701. Model B incorporates the PM components relationship via adding prior knowledge about the parameter in modeling and as a result shows RMSE 0.6691. Model C incorporates PM relationship in to the mean of process as a covariable and gives RMSE 1.2676. Based on comparing the above mentioned models it was concluded that CTM and AOT both added strength in to modeling.Regarding the PM components relationship added in to modeling based on Model A, Model B and Model C we conclude that adding PM components into the mean of the process leads to a bias in prediction. Moreover, model B, which was developed with prior knowledge proved to be the most feasible approach with the least RMSE.

Keywords

Particulate Matter, Bayesian Hierarchical modeling, Multisource data, Geostatistical modeling

ACKNOWLEDGEMENTS

This thesis would not be a reality without the support and wishes of various people and organizations.

First, I would like to express my sincere thanks to my first supervisor Dr. Nicholas Hamm for his scholarly advice, continuous suggestion, comments, and encouragement.

A special thanks to my second supervisor Prof.Dr. Ir. Alfred Stein for his comments, advice and recommendation that help me a lot to overcome my academic challenges.

I would like to acknowledge Erasmus Mundus External Corporation Window, Lot 13 India, for providing fellowship to pursue M.Sc. degree in Netherlands, and ITC/UT for providing platform to develop profession in the field of Geo-Information Science and Earth observation.

Thanks to all my classmates for sharing a long period together and helping me directly or indirectly in various aspects of this academic program.

Finally, I would like to show my gratitude from the bottom of my heart to my family members for their love, support and blessing.

Thank you all !!!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Introduction

1.1 MOTIVATION AND PROBLEM STATEMENT

1.1.1 Motivation

Atmospheric particulate matter (PM) has received global attention because several studies indicate its association with human health, regional and global climate change (Schwartz, 1994; Pilewskie, 2007; Ramanathan et al., 2007). Moreover, epidemiological studies shows PM affects daily mortality, cardiovascular and respiratory diseases including asthmatic symptoms, pulmonary inflammation, cardiopulmonary mortality and lung cancer (Dockery et al., 1992; Atkinson et al., 2001; Pope et al., 2002). PM and their constituents change radiative forcing of the atmosphere resulting in cooling and heating of atmosphere and constantly receiving attention from the scientific community (Buseck and Pósfai, 1999; Huang et al., 2006; Pilewskie, 2007; Ramanathan et al., 2007; Ramanathan and Carmichael, 2008).

To understand the effect of PM on human health and environment, detailed knowledge of PM composition is required (Veefkind et al., 2011). PM is a complex mixture of solid particles and liquid droplets suspended in the air, with different size, chemical composition and origin (Cackvoic et al., 2008). Particle size includes fine and coarse particles known as $PM_{2.5}$ and PM_{10} respectively. Fine particles have a diameter less than 2.5 μm and coarse particles have diameter less than 10 μm . PM consists of several chemical species such as inorganic species (particulate sulphate, particulate nitrate and particulate ammonium etc.), carbon species (volatile organic compounds and elemental carbon etc.) and trace metallic elements (Cr, Cu, Ni, Cd etc.). These are known as "components of PM". However, PM composition depends on the type of emission source and emitted pollutants known as precursors reactions in atmosphere (e.g. emitted precursor pollutant sulphur dioxide reacted with other chemical species and form particulate sulphate). Chemical composition is a major factor that controls the atmospheric effects of PM; like particulate sulphate resulting in to a cooling effect and particulate organic carbon resulting in to a heating effect (Huang et al., 2006; Pilewskie, 2007; Ramanathan and Carmichael, 2008).

The PM, PM components and precursors are measured with the help of detailed in situ measurement techniques. In situ measurements are considered as accurate measurement and they are used to predict at unsampled locations (van de Kassteele, 2006). In situ measurement stations equipped with instruments (such as respirable dust sampler and gaseous sampler) measures PM and precursor directly. For PM component detection detail chemical analytical techniques are required. At the same time, the in situ procedure is time consuming and limited to few locations due to economical constraints. The density of in situ measurement stations affects the interpolation and leads to uncertainty in prediction and hence, it affects the policy makers decision.

To overcome the above mentioned problem and improve the prediction spatially other supplementary data such as chemical transport models (CTM) and satellite remotely sensed data are used. In chemical transport models (e.g. LOTOS-EUROS, AERMOD and CALPUFF) chemical and transport processes are described by physical laws and empirical relationship (van de Kassteele, 2006). The emission source of the precursor are an input for this model. However, uncertainties associated with models output are high due to various input sources, uncertain model parameters and model structure (Martin, 2008). Satellite remotely sensed data (e.g. MODIS and MERIS) have been used for retrieving aerosol in atmosphere in last decade (King et al., 1999).

Developing model for prediction using multi source data in air quality field is an active area of research. Recently Veefkind et al. (2011) showed relation of satellite retrieved component precursors to components. Data integration and PM prediction using multisource (in situ measurement, CTM and remote sensing data) has been successfully done based on the geostatistical method (van de Kassteele et al., 2006). Thus, it seems to be possible to integrate different data sources and to develop model of prediction for PM components with the help of geostatistics.

1.1.2 Problem statement

PM and their composition is an important step to understand their impact on health, environment and their source identification in a time. Existing prediction methods like CTM gives prediction of PM components but uncertainty associated with CTM output are high due to various input sources, uncertain model parameter, coarse resolution and model structure. PM and PM component data are available at only some locations. CTM provides grid model output of PM components. Satellite remote sensing techniques provide raster data of potential covariates.

Considering the above mentioned problem it is necessary to develop alternative prediction model for predicting PM components with the help of multisource observations (in situ measurements, CTM and satellite remote sensing data).

1.2 RESEARCH IDENTIFICATION

1.2.1 Research objectives

To develop and evaluate geostatistical prediction model in Bayesian paradigm for predicting PM components with the help of multisource observations (in situ measurements, Chemical Transport model and remote sensing).

Specific objectives

1. To build a geostatistical prediction model to predict PM components with the help of multisource observations.

2. To evaluate the uncertainty of geostatistical prediction model.

1.2.2 Research questions

1. What is the spatial structure of PM components and their associated covariable observed from multisource observations ?

2. Which covariable (CTM or remote sensing) gives more accurate prediction?

3. How can PM components relationship be incorporated in to the Bayesian hierarchical model?

4. Does PM components relationship improve the model prediction ? Why?

5. What is the accuracy of model developed for prediction ?

1.2.3 Innovation aimed at

In this research, innovation is aimed at developing Bayesian hierarchical model for decomposing PM into its components and their prediction considering multisource observations.

1.2.4 Thesis structure

The thesis divided in to 7 chapters. Chapter 1 give information about the research topic, motivation, problem statement, research objectives etc. Chapter 2 incorporates literature review. Chapter 3 gives information about study area and data used in thesis. Chapter 4 give information about research methodology. Chapter 5 incorporates result obtained during the completion of research. Chapter 6 includes the discussion and chapter 7 includes conclusion and recommendations.

Chapter 2
Literature review

2.1 PARTICULATE MATTER(PM) AND ITS EFFECT ON HEALTH AND CLIMATE CHANGE

Particulate Matter(PM) present in the atmosphere continuously receiving science communities attention due to their association with health and climate change. The anthropogenic emission of primary pollutants like sulphur dioxide, Nitrogen dioxide, organic compounds coming from different pollutant source defines the chemical composition and structure of PM (Pöschl, 2005). Dockery et al. (1992) showed the relation of PM and other associated air pollutants like particulate sulphate, ozone etc to daily mortality rate and concluded with the effect of PM mass concentration on mortality. Atkinson et al. (2001) reported the PM and respiratory disease related admissions in hospital of European cities. Babak and Deutsch (2009b) showed the dominated sulfate particle (component of PM) present in the atmosphere of remote oceanographic area and and their associated cooling effect. Ramanathan et al. (2007); Ramanathan and Carmichael (2008) observed the contribution of black carbon and organic particulates in heating the atmosphere by absorbing the solar radiation over Asia region. Huang et al. (2006) showed the effect of anthropogenic sulfate particle on surface temperature and precipitation with increasing downward long wave surface forcing. Pöschl (2005) illustrates the effect of PM on atmospheric, oceanographic and bio-geochemical cycle through the radiative forcing, changing flux of solar radiation etc. Their illustration of direct and indirect effect and feedback loop of PM on climate system given at figure 2.1(adopted from same article).

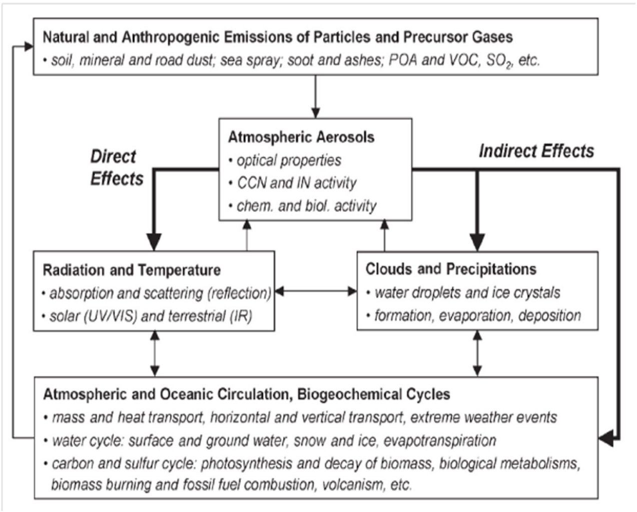

Natural and Anthropogenic Emissions of Particles and Precursor Gases
• soil, mineral and road dust; sea spray; soot and ashes; POA and VOC, SO_2, etc.
Atmospheric Aerosols
• optical properties
• CCN and IN activity
• chem. and biol. activity
Direct Effects
Indirect Effects
Radiation and Temperature
• absorption and scattering (reflection)
• solar (UV/VIS) and terrestrial (IR)
Clouds and Precipitations
• water droplets and ice crystals
• formation, evaporation, deposition
Atmospheric and Oceanic Circulation, Biogeochemical Cycles
• mass and heat transport, horizontal and vertical transport, extreme weather events
• water cycle: surface and ground water, snow and ice, evapotranspiration
• carbon and sulfur cycle: photosynthesis and decay of biomass, biological metabolisms, biomass burning and fossil fuel combustion, volcanism, etc.

Figure 2.1: Direct and indirect effect of Particulate Matter(PM) and feedback loop on climate system source (Pöschl, 2005).

2.2 PARTICULATE MATTER

The Section 2.1 identifies the importance of PM and need of its decomposition. PM consist of fine ($PM_{2.5}$) and coarse (PM_{10}) particles. Fine ($PM_{2.5}$) particles are result of fuel combustion, residential fire places, wood stoves, power generation and industrial facilities, where as coarse (PM_{10}) particles resulted from traffic, motor vehicles, dust from paved and unpaved roads, construction and demolition, bare ground, material handling ,crushing and grinding operation , industrial complexes, wind blown dust (Morawska et al., 2001; Chow and Watson, 2002; Fang et al., 2002). These particle size and their composition coming from different emission source play important role in their interaction with environment.

Weijers et al. (2011) showed that fine ($PM_{2.5}$) particles mass concentration over Netherlands is dominated by anthropogenic emission as compared to the coarse (PM_{10}) particle. Van Dingenen et al. (2004) showed PM physical characteristics over Europe and concluded with no universal ratio between mass concentration of $PM_{2.5}$ and PM_{10} except constant ratio existed at individual sampling site.

2.3 RELATED WORK

Several authors show the relationship of satellite retrieved Aerosol Optical Thickness (AOT) to PM measurement using empirical linear model (Gupta and Christopher, 2009; Péré et al., 2009; Emili et al., 2010; Li et al., 2011). Veefkind et al. (2011) shows spatio-temporal correlation between AOT and precursor gas (Nitrogen dioxide, sulphur dioxide and formaldehyde) to infer the composition of particulate matter.

Regression kriging, cokriging and Bayesian hierarchical modelling are usefull geostatistical approach for improving predection of sparsly sampled primary variable from a densly sampled secondary variables. van de Kassteele et al. (2006) showed improved prediction of primary variable PM10 using secondary variables information from dispersion modeling and satellite observations using external drift kriging method. Hengl et al. (2007)discussed the strength and limitations of regression kriging. This paper shows limitation of the method resulting in to bias prediction to the data coming from different sources, sparse samples and uneven relation of response variable to explanatory variables.

Singh et al. (2011) showed cokriging approach to improve the primary variables ozone and PM10 prediction using secondary information of chemical transport model. Cokriging is a multivariate geostatistical method that uses the spatial dependencies within the variables as well as cross spatial dependencies between variables. Huang et al. (2009) generalized the cross covariance function to quantify the spatial cross dependencies for multivariate intrinsic random functions and this helps for implementing cokriging when the process is intrinsic random function.

Babak and Deutsch (2009a,b) shows a novel approach of merging multiple secondary data in to super secondary variable and then implementing collocated cokriging with the single variable. The geostatistical modeling is improved when the estimation is constrained to all available secondary data. cokriging handle the multisource observations while implementation of collocated cokriging are limited to single most correlated or most relevant secondary variable. In this study author assume the structures of spatial correlation in variables are proportional to each other.

Liu et al. (2008)developed Bayesian hierarchical model for urban air quality prediction. In this study three pollutants variable and four external driving factors variable were used. The structure of air quality model and prior distributions of model parameters defined with the help of correlation analysis, classification and regression trees, hierarchical cluster analysis and discriminant analysis. For finding the relationship between pollutant concentration and driving variables multiple linear regressions was proposed. This paper shows Bayesian hierarchical model is useful for predicting urban air quality from related contributing factors.

These studies use different approaches such as the relation between PM and their precursor to infer PM composition, geostatistical prediction of PM with the help of multisource data. Veefkind et al. (2011) suggested importance of model development using satellite data to infer PM composition. To fill out this research gap, this study proposes to use Bayesian hierarchical model for particulate matter decomposition.

Chapter 3

Study area and data description

3.1 INTRODUCTION

This chapter describes the study area and dataset.

3.2 STUDY AREA

The study area selected for present study covers the countries namely Belgium, Netherlands, Luxemburg, France and Germany (Figure 3.1). The study area lies between longitude of -5 degree West to 15 degree East and latitude of 40 degree to 56 degree North. The study area is appropriate for addressing research problem of decomposing PM in to its components due to the study areas status of industrialization (source of PM emission), importance of associated effects on environment & health and multisource data availability from various sources in study area.

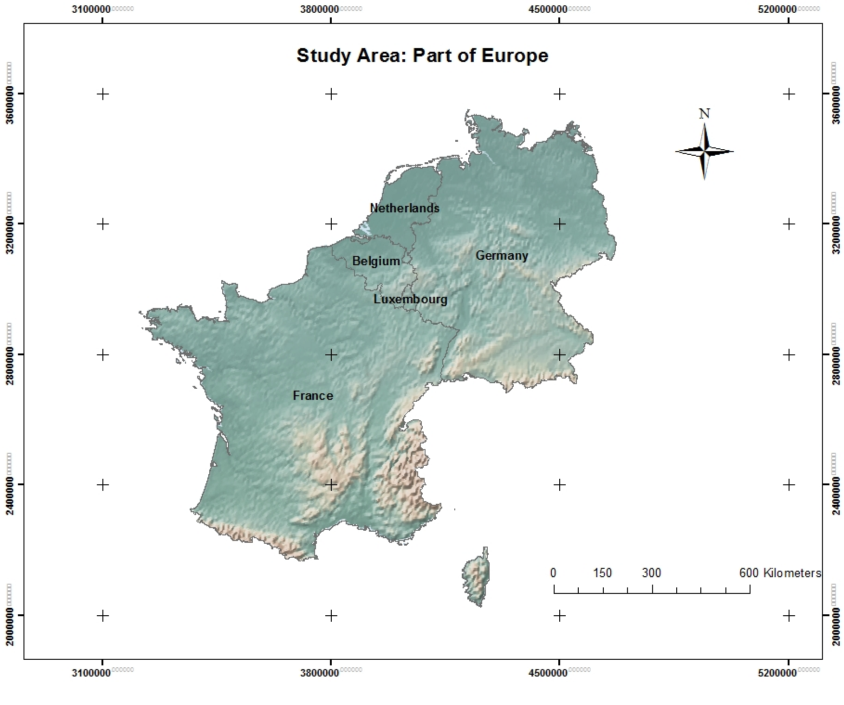


Figure 3.1: Study area : part of Europe (countries: Belgium, Netherlands, Luxemberg, France and Germany.

3.3 DATA DESCRIPTION

Data from different sources(multisource) like in situ measurements, chemical transport model(CTM) and satellite remote sensing data were selected for present study. Multisource data availability have an advantage in modeling spatial process because multisource data reduces the noise coming from single source data.

3.3.1 In situ data

Daily in situ measurements of PM_{10} and $PM_{2.5}$ for the year 2009 over study area provided by TNO (Netherlands Organization for Applied Scientific Research) extracted from Airbase database. These air pollutants have been measured by responsible organization of the respective country and submitted to Airbase database (public air quality database system of the European Union countries) as per the guidelines provided by European Union. PM_{10} and $PM_{2.5}$ measured with the help of in situ instrument located in measurement stations at the interval of 8 hours. These in situ measurements considered as accurate measurement (van de Kassteele, 2006) and used for modeling purpose. However, in situ measurements are sparsely measured over study area due to the economical constraint associated with it. PM_{10} and $PM_{2.5}$ measured at 555 and 171 measurement stations respectively (figure 3.2).

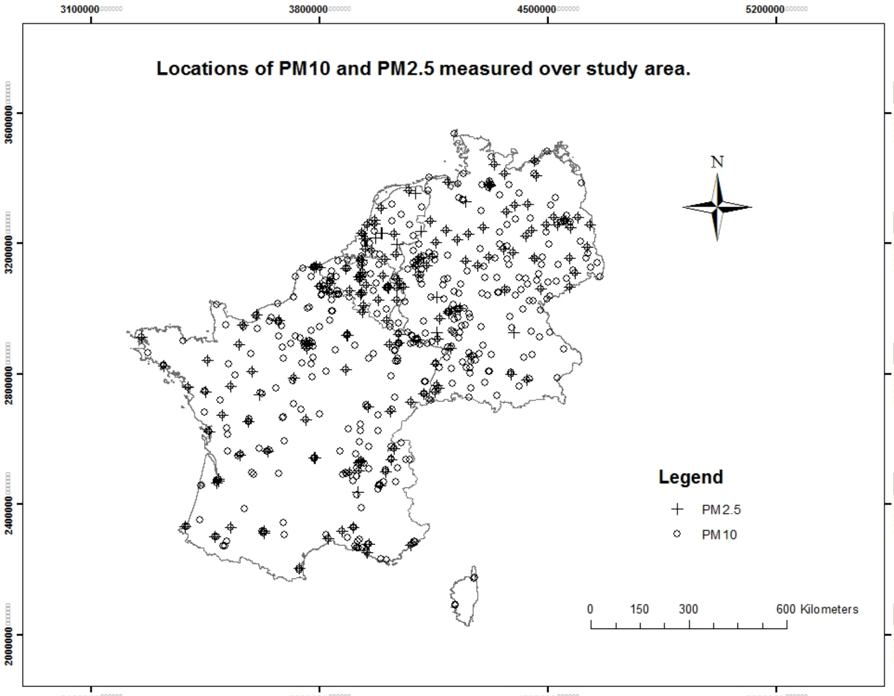


Figure 3.2: Locations of PM10 and PM2.5 measured over study area

3.3.2 CTM data (chemical transport model)

CTM model LOTUS-EUROS gridded data of PM_{10} and $PM_{2.5}$ for year 2009 provided by TNO for present study. The LOTUS-EURO is an operational 3D chemical transport model measures the composition of air quality in lower troposphere considering physical, chemical and empirical relationship between pollutants. The LOTOS-EUROS model surrounded over Europe at longitude of 10 degree West to 60 degree East and latitude of 35 degree to 70 degree North. The grid resolution of LOTUS-EURO is 0.50 degree longitude to 0.25 degree latitude, approximately 30 km by 30 km. PM_{10} and $PM_{2.5}$ is defined in to LOTUS-EUROS model by summing the respective individual components like PM_{10} is an sum of coarse primary emitted particles, sea salt and secondary inorganic components and $PM_{2.5}$ is an sum of fine primary emitted particles and secondary inorganic components (Schaap et al., 2009a). The anthropogenic emission data of pollutants (primary emitted particles,sea salt and secondary inorganic components) act as input data source in LOTUS-EURO model. The chemical reaction of the input pollutants is defined as per chemical mechanism of CBM-IV and TNO CBM-IV scheme and vertical and horizontal transport defined by adding meteorological data (Schaap et al., 2009a). Finally the model calculates the PM_{10} and $PM_{2.5}$ concentration considering chemical reactions, dry and wet deposition and transport and dispersion mechanism.

3.3.3 AOT (aerosol optical thickness) data

AOT is an degree of aerosol or PM particles which prevent the transmission of light in atmosphere due to the scattering and absorption processes. Several studies shows the relationship between AOT and PM concentration (Gupta and Christopher, 2009; Péré et al., 2009; Emili et al., 2010). Wang et al. (2010); van de Kassteele et al. (2006) shows potential of AOT data to infer PM concentration.

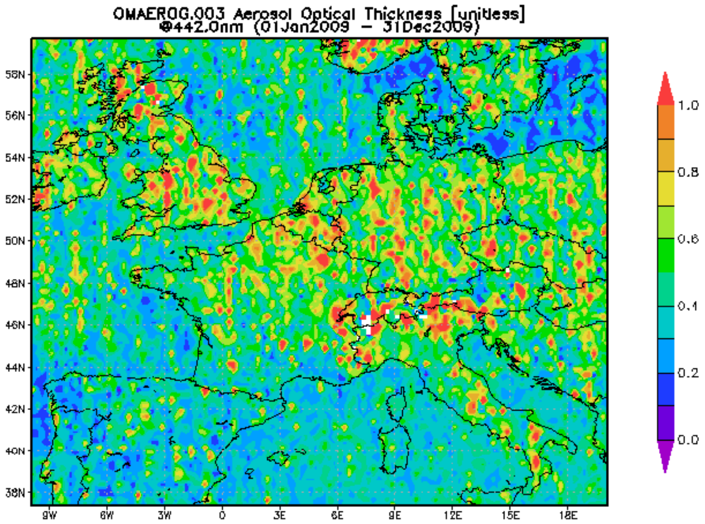

OMAEROG.003 Aerosol Optical Thickness [unitless]
@442.0nm (01Jan2009 - 31Dec2009)
58N
56N
54N
52N
50N
48N
46N
44N
42N
40N
38N
9W
6W
3W
0
3E
6E
9E
12E
15E
18E
1.0
0.8
0.6
0.4
0.2
0.0

Figure 3.3: Annual product of AOT for the year 2009 .

AOT data of OMI (Ozone Monitoring Instrument) sensor located on EOS AURA satellite downloaded from GES DISC (Goddard Earth Sciences Data Information Services Center) of NASA. The grid resolution of AOT is 0.25 degree longitude by 0.25 degree latitude. The annual product of AOT Level-2G dataset (wavelength 442nm)for the year 2009 (Figure 3.3) created from daily data through Giovanni tool (data exploration interface).

Chapter 4
Methodology

4.1 INTRODUCTION

Methodology for decomposition of particulate matter in to its components is given in this chapter.

4.2 MODELING APPROACH FOR DECOMPOSITION

Modeling decomposition of Particulate Matter(PM) in to its components requires understanding of the nature of PM formation. PM is a complex mixture of solid particles and liquid droplets suspended in the air, with different size, chemical composition and origin. PM composition depends on the type of direct emission from emission source (often known as primary pollutants) and through the chemical reactions between atmospheric pollutants (often known as secondary pollutants). According to the air pollution context and Environmental Protection Agency(EPA) terminology (EPA, 2012) PM size categorizes in to Total Suspended Particulate Matter (TSP), PM_{10}, $PM_{2.5}$ and Particles less than 0.1 μm. TSP ranging in size from 0.1 μm to about 30 μm and includes PM_{10}, $PM_{2.5}$ and Particles less than 0.1 μm (Figure 4.1).

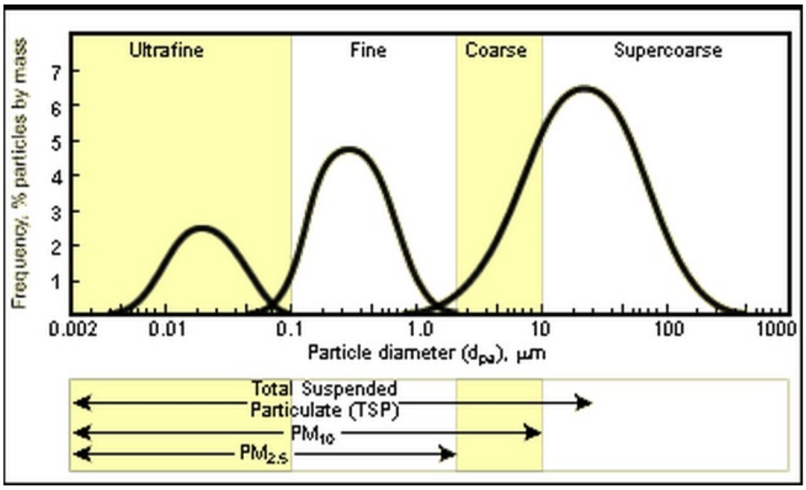


Figure 4.1: PM size distribution (EPA, 2012)

For modeling convenience consider TSP as upper level (topmost) or level 0 PM, now target is decomposing the PM (level 0) in to finer level and assume such levels as level I, level II and level III decomposition(figure 4.2). Decomposition of level I is a PM_{10} defined as the sum of primary emitted particles (PPM) and secondary inorganic components (SO_4, NO_3, carbonaceous particles and sea salt etc.); $PM_{10} = PPM_{2.5} + PPM_{2.5-10} + SO_4 + NO_3 +$ carbonaceous particle + other (sea salt etc.) (Schaap et al., 2009a). Decomposition of level I further decomposes in to level II as $PM_{2.5}$, SO_4 (PM Sulphate), NO_3 (PM Nitrate) and carbonaceous particle (PM carbon) because

PM_{10} is made up of these components. Level II $PM_{2.5}$ is further decomposed in to level III as SO_4 , NO_3 and carbonaceous particle because $PM_{2.5}$ is made up by this components (Schaap et al., 2009a). Finally decomposition levels each component modeled in Bayesian hierarchical paradigm separately.

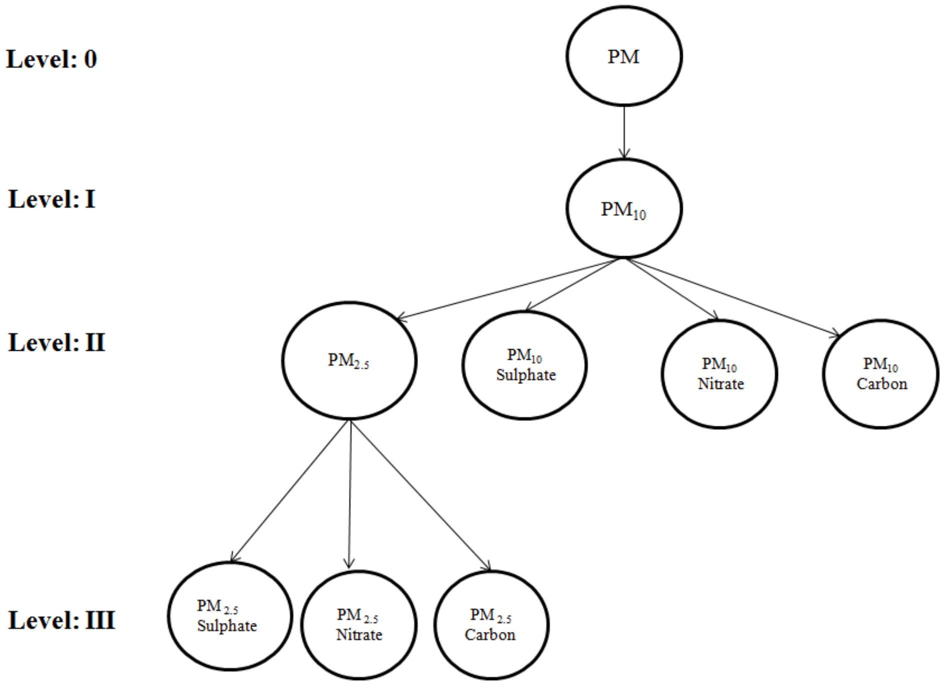


Figure 4.2: Modeling approach of PM decomposition

4.3 PREPROCESSING AND EXPLORATORY ANALYSIS OF DATASET

4.3.1 Available dataset for decomposition

As per the modeling approach adopted in section 4.2 for decomposition of PM in to its components, dataset for decomposition level I (PM_{10}) and decomposition level II ($PM_{2.5}$) available. Detailed description of the study area and dataset was given in chapter 3. In situ observations of PM_{10} and $PM_{2.5}$ considered as accurate measurement (van de Kassteele, 2006) and it acts as response variable in modeling of subsequent decomposition level but these observations are sparsely sampled over geographic area. Densely sampled CTM and AOT data added in modeling as covariable (explanatory variable) for adding strength to response variable.

Table 4.1 Available dataset for decomposition

Decomposition level	Response variable	Explanatory variable	
	(in situ observations)	(CTM data)	(Remote sensing data)
Level I	PM_{10}	CTM PM_{10}	AOT
Level II	$PM_{2.5}$	CTM $PM_{2.5}$	AOT

4.3.2 Software and tools

- ArcGIS Desktop 10
- Statistical software R version 2.13.2 : R packages spBayes, GeoR, gstat, rgdal, MBA, CODA.

4.3.3 Preprocessing of dataset

Preprocessing of dataset is an important step to obtain qualitative result from experiment. Daily In situ measurements of PM_{10} and $PM_{2.5}$ were retrieved from netcdf file and their annual average for the year 2009 calculated in R software. Daily CTM data at in situ measured locations provided by data provider and their annual average calculated in similar way of in situ data. Daily Gridded data of CTM was provided in netcdf file format , annual average calculated and final product was converted from kg/m^3 to microgram/m^3 for the convenience of data attribute storage in programming interface of R and ArcGIS. Gridded annual AOT product of OMI data created using Giovanni interface and downloaded in ASCII format. In situ, CTM and AOT data in Geographic coordinate system is not suitable for modeling purpose due to distance difference in longitude and latitude. For overcoming above mentioned problem all data projected on Lambert Azimuth Equal Area 1989 (ETRS LAEA 1989) projection. After preprocessing more than 25% of data subseted using random sampling for validation purpose and kept seperated from the process of exploratory analysis and modeling. For PM_{10} (decomposition level I) and PM_{10} (decomposition level II) 125 and 40 measurement points subseted for validation purpose and 430 and 131 measurement points used for constructing model respectively.

4.3.4 Exploratory analysis of dataset

Exploratory analysis of dataset is an integral part of geostatistical modeling for understanding the data structure as well as it is an important step to take decision of data transformation. Non-spatial aspects of data like summary statistics, histogram, box plots, and normal Q-Q plot calculated using R software. For understanding the spatial aspects of data empirical variogram plotted using equation 4.1. Based on visual inspection variogram model fitted to empirical variogram.

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{Y(s_i) - Y(s_i + h)\}^2 \tag{4.1}$$

Where $Y(s_i)$ and $Y(s_i + h)$ represent the values of observation Y separated by lag distance h. $N(h)$ is the number of pairs of data points at particular lag distance h. lag distance need to be change for obtaining set of semi variances which constructs the empirical variogram.

4.4 GEOSTATISTICAL MODELING

4.4.1 Basic model

In the geostatistical modeling often interest is to understand the spatial process S at unsampled location. The knowledge of any spatial process obtained through the measuring random variable Y because spatial process is not directly observable and it is also known as realization of spatial process. However,measurement of random variable $Y = (Y_1, ..., Y_n)$ is noisy version (measurement error) of spatial process. The basic spatial linear regression model for point-referenced data given in equation 4.2.

$$Y(s) = \mu(s) + W(s) + \epsilon(s) \tag{4.2}$$

Where $Y(s)$ is an $n \times 1$ vector of observed response variable at generic location s. The mean structure is $\mu(s) = X^T(s)\beta$. The residual has two components one is spatial error $W(s)$ and another is non-spatial error $\epsilon(s)$. The spatial error $W(s) = f(\sigma^2, \phi)$ are considered as realizations from a zero-centered stationary Gaussian spatial process and it captures residual spatial association (Banerjee et al., 2004) and it introduces the parameters partial sill σ^2 and range ϕ. The non-spatial error $\epsilon(s) = f(\tau^2)$ is uncorrelated pure error term and it introduces the nugget effect τ^2. As per basic model given in equation 4.2 our interest is to estimate the parameters (σ^2, ϕ and τ^2) which defines the covariance structure of the process.

Considering classical or conventional geostatistical approach for interpolation i.e. kriging, the covariance structure of the data is estimated first, then the parameters estimated from covariance model plug in to theoretical prediction equation as they were the true values. However, classical geostatistical approach ignores the uncertainty of estimated parameters leads to uncertainty in subsequent prediction. As opposite to the classical geostatistical approach the Bayesian approach for interpolation of spatial processes will provide a general methodology for taking in to account the uncertainty about parameters on subsequent predictions. This approach leads to same answers as the standard kriging predictor when the model parameters are known, but it also extends to the case where these parameters are unknown. This is one major reason for viewing the problem in Bayesian terms as well as it have ability to build more flexible model than other .

4.4.2 Bayesian Hierarchical model

Bayesian inference treats unknown parameters as random variables and during prediction it considers parameter uncertainty. This approach leads to more realistic estimates of the prediction variance. Let us consider now approach of decomposition of particulate matter in to level I and level II decomposition as described in modeling approach section. Now each decomposition level modeled in separately in hierarchical manner. Note in this section all equations are adopted according to Banerjee et al. (2004) and Diggle and Ribeiro (2007). The basic model given in equation 4.2 can be rewritten as equation 4.3 assuming Gaussian spatial process Y (recall measurements Y is an realization of spatial process) conditional on some parameter θ:

$$Y|\theta \sim N(X\beta, \sigma^2 R(\phi) + \tau^2 I) \tag{4.3}$$

Where Y is an $n \times 1$ vector of response variable. X is an $n \times p$ matrix of explanatory variables associated with response variable. β is an $p \times 1$ vector of trend parameters or associated regression parameter. $\sigma^2 R(\phi) + \tau^2 I$ is an covariance structure of the process defined by parameters partial sill σ^2 , range ϕ and nugget τ^2 . I is an $n \times n$ identity matrix. $R(\phi)$ is an $n \times n$ correlation matrix given by equation 4.4 .

$$R_{ij}\phi = \rho(||s_i - s_j||; \phi) \tag{4.4}$$

Where R_{ij} are given by an authorized correlation function ρ in geographic space, which depends on the distance between $||s_i - s_j||$ the location of $_i$ and $_j$ indexed by parameter ϕ.

Now consider $\theta = (\beta, \sigma^2, \phi, \tau^2)$ be the set of model parameters. Using Bayes theorem for obtaining the posterior probability density of parameters denoted by $p(\theta|Y)$ given by equation 4.5.

$$p(\theta|Y) = \frac{f(Y|\theta)\pi(\theta)}{\int f(Y|\theta)\pi(\theta)d\theta} \tag{4.5}$$

where $f(Y|\theta)$ is the likelihood associated with equation 4.3. $\pi(\theta)$ is an prior(π) distribution of parameters θ as well as Bayesian solution requires an appropriate prior distribution. The Bayes theorem equation 4.5 can be rewritten as equation 4.6, where denominator of equation 4.5 drops out because calculations (numerical and algebraic) are required only up to a proportionally constant as well as it does not add any extra information to obtain posterior probability density of parameters $p(\theta|Y)$.

$$p(\theta|Y) \propto f(Y|\theta)\pi(\theta) \tag{4.6}$$

However, the computation of the likelihood $f(Y|\theta)$ require $(\sigma^2 R(\phi) + \tau^2 I)^{-1}$, which creates the problem of matrix inversion if n is a large number. Therefore, it is convenient to work with a hierarchical model (Banerjee et al., 2004).

The hierarchical model defined at three stages and their specification is as follows;
First stage:

$$Y|\beta, \tau^2, W \sim MVN(X\beta + W, \tau^2 I) \tag{4.7}$$

Second stage:

$$W|\sigma^2, \phi \sim MVN(0, \sigma^2 R(\phi)) \tag{4.8}$$

Third stage:

$$\theta = (\beta, \sigma^2, \phi, \tau^2) \tag{4.9}$$

The expression of Gaussian spatial process given by equation 4.3 rewritten as a hierarchical model by writing the first stage specification 4.7 as Y conditional not only on the parameters β and τ^2 but also on the vector of spatial random effects $W = (W(s_1), ..., W(s_n))$ (Banerjee et al., 2004). The second stage specification 4.8 of model is for spatial random effects W conditional on parameters σ^2 and ϕ which defines spatial dependence, where $R(\phi)$ is as per equation 4.4 .The hierarchical model completes at the third stage specification 4.9 by adding priors for β and τ^2 as well as for σ^2 and ϕ. The spatial dependence parameters σ^2 and ϕ added at third stage may be viewed as hyperparameters.

According to Bayes theorem we can write down the hierarchical model as per equation 4.10

$$p(\beta, \tau^2, \sigma^2, \phi|Y) \propto \int f(Y|\beta, \tau^2, W)\pi(\beta)\pi(\tau^2)f(W|\sigma^2, \phi)\pi(\sigma^2)\pi(\phi)dW \tag{4.10}$$

where hierarchical nature of model defined with the help of adding first stage prior and hyperprior. First stage prior $f(W|\sigma^2, \phi)$ defines the spatial random effects conditional on vector of hyperparameters σ^2 and ϕ. Hyperprior (prior for prior) controls the variation of spatial random effects but in practice it is unknown, so hyperprior distribution $\pi(\sigma^2)$ and $\pi(\phi)$ required in hierarchical model formulation. Posterior distribution of parameter $p(\beta, \tau^2, \sigma^2, \phi|Y)$ obtained from hierarchical model 4.10 is same as the non-hierarchical model 4.6 posterior distribution $p(\theta|Y)$ (recall $\theta = \beta, \sigma^2, \phi, \tau^2$). Posterior realization of spatial random effect W obtained via sampling using posterior distribution of σ^2 and ϕ during the process of fitting the model. Success of Bayesian paradigm models is highly depends on the prior specification and it incorporates the prior opinion of modeler regarding parameters distribution and this makes Bayesian inference subjective. To avoid the misleading Bayesian inference of model safest strategy is to choose informative prior for σ^2, ϕ and τ^2 based on primary knowledge of parameter distribution. However,as a general rule flat prior adopted for β since even it give the proper posterior (Banerjee et al., 2004). Bayesian model solved by MCMC (Monte Carlo Markov Chain) simulation.

Through the hierarchical model written in Bayesian paradigm as per equation 4.10, the posterior $p(\theta|Y)$ ($\theta = (\beta, \sigma^2, \phi, \tau^2)$) be the set of model parameters) estimate of parameter θ is some measure of centrality. According to the Banerjee et al. (2004) and their given equation familiar choices are the posterior mean 4.11 and posterior median 4.12.

$$\widehat{\theta} = E(\theta|Y) \tag{4.11}$$

$$\widehat{\theta} : \int_{-\infty}^{\widehat{\theta}} p(\theta|Y)d\theta = 0.5 \tag{4.12}$$

However, posterior mean often highly influenced by the outliers so posterior median be the best and safest point to estimate as well as posterior allows to make direct probability statement about parameters.

4.4.3 Bayesian predictive process

After building model in Bayesian paradigm described in section 4.4.2 next procedure is to predict response variable $Y(s_0)$ at new location s_0 taking consideration of associated covariate vector $X(s_0)$. Now assume $Y_0 \equiv Y(s_0)$, $X_0 \equiv X(s_0)$ and θ is an set of model parameters as described in section 4.4.2 for convenience. The prediction model for response variable at unsampled location written as per equation 4.13 in Bayesian framework.

$$\begin{aligned} p(Y_0|Y,X,X_0) &= \int f(Y_0,\theta|Y,X,X_0)d\theta \\ &= \int f(Y_0|Y,\theta,X_0)p(\theta|Y,X)d\theta \end{aligned} \tag{4.13}$$

Where $p(Y_0|Y,X,X_0)$ has an conditional normal distribution arising from joint distribution of Y_0 and Y taking full advantage of densely covariate over geographic space or predictive space (predictive target locations).

4.4.4 Bayesian hierarchical model for decomposition level I (PM10)

As per the Bayesian hierarchical model framework described in section 4.4.2 and modeling approach adopted in section 4.2 decomposition level I of PM is PM_{10}.Assuming PM_{10} (decomposition level I) concentration over study area is an Gaussian process Y_{PM10} conditional on some parameter θ_1 written as equation 4.14 is similar with basic model described at equation 4.3. Note subscript $_1$ used for defining the each parameter in writing the Bayesian Hierarchical model for PM_{10} denotes parameter belonging to model PM_{10} and it only used to avoid confusion with Bayesian paradigm described in section 4.4.2 and 4.4.3.

$$Y_{PM10}|\theta_1 \sim N(X_1\beta_1, \sigma_1^2R(\phi_1) + \tau_1^2I) \tag{4.14}$$

Where Y_{PM10} is an $n \times 1$ vector of response variable PM_{10}. X_1 is an $n \times p$ matrix of explanatory variable of CTM PM_{10} and AOT associated with response variable. All other explanation is same as equation 4.3. The first stage hierarchical model for PM_{10} (decomposition level I) written as equation 4.15 according to equation 4.7.The second and third stage of hierarchical model of PM_{10} is written as equation 4.16 and 4.17 is same as hierarchical stages explained by equation 4.8 and 4.9 respectively.

$$Y_{PM10}|\beta_1, \tau_1^2, W_1 \sim MVN(X_1\beta_1 + W_1, \tau_1^2I) \tag{4.15}$$

$$W_1|\sigma_1^2, \phi_1 \sim MVN(0, \sigma_1^2 R(\phi_1)) \tag{4.16}$$

$$\theta_1 = (\beta_1, \sigma_1^2, \phi_1, \tau_1^2) \tag{4.17}$$

Complete Bayesian hierarchical model of PM10 (decomposition level I) write down as per equation 4.18 according to equation 4.10

$$p(\beta_1, \tau_1^2, \sigma_1^2, \phi_1|Y_{PM10}) \propto \int f(Y_{PM10}|\beta_1, \tau_1^2, W_1)\pi(\beta_1)\pi(\tau_1^2)f(W_1|\sigma_1^2, \phi_1)\pi(\sigma_1^2)\pi(\phi_1)dW_1 \tag{4.18}$$

Where $p(\beta_1, \tau_1^2, \sigma_1^2, \phi_1|Y_{PM10})$ is an posterior distribution of parameter updated on PM_{10} (decomposition level I). After building hierarchical model for PM10 (decomposition level I) the prediction of $Y_{PM10}(s_0)$ at new location s_0 with the help of associated covariate vector $X_1(s_0)$ of CTM PM_{10} and AOT at new location. The prediction model according to equation 4.13 for PM_{10} decomposition level I written as per following equation.

$$\begin{aligned} p(Y_{PM10}(s_0)|Y_{PM10}, X_1, X_1(s_0)) &= \int f(Y_{PM10}(s_0), \theta_1|Y_{PM10}, X_1, X_1(s_0))d\theta_1 \\ &= \int f(Y_{PM10}(s_0)|Y_{PM10}, \theta_1, X_1(s_0))p(\theta_1|Y_{PM10}, X_1)d\theta_1 \end{aligned}$$

Where $p(Y_{PM10}(s_0)|Y_{PM10}, X_1, X_1(s_0))$ has an conditional normal distribution arising from the joint distribution of $Y_{PM10}(s_0)$ and original data Y_{PM10} taking full advantage of densely covariate $X_1(s_0)$ CTM PM_{10} and AOT over geographic space. For decomposition level I three different models namely Model 1, Model 2 and Model 3 are constructed and their workflow given in Figure 4.3.

4.4.5 Bayesian hierarchical model for decomposition level II (PM2.5)

During the decomposition of PM in to level II ($\text{PM}_{2.5}$) need to be consider the relationship between decomposition level I (PM10) and II ($\text{PM}_{2.5}$) because $\text{PM}_{2.5}$ is an component (part) of PM_{10}. Assume $\text{PM}_{2.5}$ (decomposition level II) concentration over study area is an Gaussian process $Y_{PM2.5}$ conditional on some parameter θ_2 and it is written as per equation 4.19 according to basic model described in equation 4.3. Note subscript $_2$ used for defining the each parameter in writing the Bayesian Hierarchical model for $\text{PM}_{2.5}$ denotes parameter belonging to model $\text{PM}_{2.5}$ and it only used to avoid confusion with Bayesian paradigm described in section 4.4.2, 4.4.3 and 4.4.4.

$$Y_{PM2.5}|\theta_2 \sim N(X_2\beta_2, \sigma_2^2 R(\phi_2) + \tau_2^2 I) \tag{4.19}$$

Where $Y_{PM2.5}$ is an $n \times 1$ vector of response variable PM2.5. X_2 is an $n \times p$ matrix of explanatory variable of CTM $\text{PM}_{2.5}$ and AOT associated with response variable. Now consider the relationship between $\text{PM}_{2.5}$ and PM10 in air pollution context and modeling approach adopted for decomposition in section 4.2, assume process $Y_{PM2.5}$ conditional not only on parameter θ_2 (equation 4.19) but also on upper level process Y_{PM10} (decomposition level I). Now defining the first stage of hierarchical model for $Y_{PM2.5}$ according to equation 4.7 the upper level process Y_{PM10} need to be modeled on separate next level like random effect modeled in section 4.4.2 (see the hierarchical nature of model defined at equation 4.7,4.8 and 4.9). However, upper level process Y_{PM10} is already modeled at decomposition level I and predicted over study area as described in section 4.4.4. Instead of modeling the upper level process Y_{PM10} separately , the approach of modeling the same process in the mean of the process $Y_{PM2.5}$ is suitable because it act as additional $n \times 1$ vector in matrix of explanatory variable. However, during the prediction process of PM$_{2.5}$ it take full advantage of predicted surface PM10 which is modeled at upper level. Now in first stage specification of hierarchical model of the process $Y_{PM2.5}$ written as equation 4.20 according to equation 4.7 and additional vector of process Y_{PM10} act as covariable (remember this process (decomposition level I) modeled in the mean of the process of PM$_{2.5}$(decomposition level II).

$$Y_{PM2.5}|\beta_2, \tau_2^2, W_2 \sim MVN(X_2\beta_2 + W_2, \tau_2^2 I) \tag{4.20}$$

The second and third stage of hierarchical model of PM$_{2.5}$ is written as equation 4.21 and 4.22 is same as hierarchical stages explained by equation 4.8 and 4.9 respectively.

$$W_2|\sigma_2^2, \phi_2 \sim MVN(0, \sigma_2^2 R(\phi_2)) \tag{4.21}$$

$$\theta_2 = (\beta_2, \sigma_2^2, \phi_2, \tau_2^2) \tag{4.22}$$

Complete framework of Bayesian hierarchical model of PM$_{2.5}$ (decomposition level II) write down as per equation 4.23 according to equation 4.10

$$p(\beta_2, \tau_2^2, \sigma_2^2, \phi_2|Y_{PM2.5}) \propto \int f(Y_{PM2.5}|\beta_2, \tau_2^2, W_2)\pi(\beta_2)\pi(\tau_2^2)f(W_2|\sigma_2^2, \phi_2)\pi(\sigma_2^2)\pi(\phi_2)dW_2 \tag{4.23}$$

Where $p(\beta_2, \tau_2^2, \sigma_2^2, \phi_2|Y_{PM2.5})$ is an posterior distribution of parameter updated on PM2.5 (decomposition level II). After building hierarchical model for PM2.5 (decomposition level II) the prediction of $Y_{PM2.5}(s_0)$ at new location s_0 with the help of associated covariate vector $X_2(s_0)$ of CTM PM2.5, AOT and level I process Y_{PM10} at new location. The prediction model according to equation 4.13 for PM2.5 decomposition level II written as per following equation.

$$\begin{aligned} p(Y_{PM2.5}(s_0)|Y_{PM2.5}, X_2, X_2(s_0)) &= \int f(Y_{PM2.5}(s_0), \theta_2|Y_{PM2.5}, X_2, X_2(s_0))d\theta_2 \\ &= \int f(Y_{PM2.5}(s_0)|Y_{PM2.5}, \theta_2, X_2(s_0))p(\theta_2|Y_{PM2.5}, X_2)d\theta_2 \end{aligned}$$

Where $p(Y_{PM2.5}(s_0)|Y_{PM2.5}, X_2, X_2(s_0))$ has an conditional normal distribution arising from the joint distribution of $Y_{PM2.5}(s_0)$ and original data $Y_{PM2.5}$ taking full advantage of densely covariate $X_2(s_0)$ CTM PM$_{2.5}$, AOT and modeled level I process Y_{PM10} over geographic space.For decomposition level II three different models namely Model A, Model B and Model C are constructed and their workflow given in Figure 4.4.

4.5 VALIDATION OF MODEL

Validation of model with independent dataset is an important aspect of modeling to check feasibility of model. As per described in section 4.3.3 validation data subseted using random sampling and kept separate from modeling.. In validation subseted measurement points of PM$_{10}$ and PM$_{2.5}$ compared with respective prediction and Mean Error (ME),Sum of Square Error (SEE) and Root Mean Square Error (RMSE) calculated using equations 4.24, 4.25 and 4.26 respectively.

$$ME = \frac{1}{N}\sum_{i=1}^{N} Y^*(s_i) - Y(si) \tag{4.24}$$

$$SSE = \sum_{i=1}^{N} (Y^*(s_i) - Y(s_i))^2 \tag{4.25}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y^*(s_i) - Y(s_i))^2} \tag{4.26}$$

Where $Y^*(s_i)$ is an estimated value at location s_i and $Y(s_i)$ is the observed value at location s_i.

For PM_{10} (decomposition level I) total 555 measurements points available out of 125 measurement points subseted for validation purpose and 430 measurement points used for constructing model . Decomposition level II or for $PM_{2.5}$ total 171 measurements points available out of 40 points used for validation and 131 measurement points used for building model.

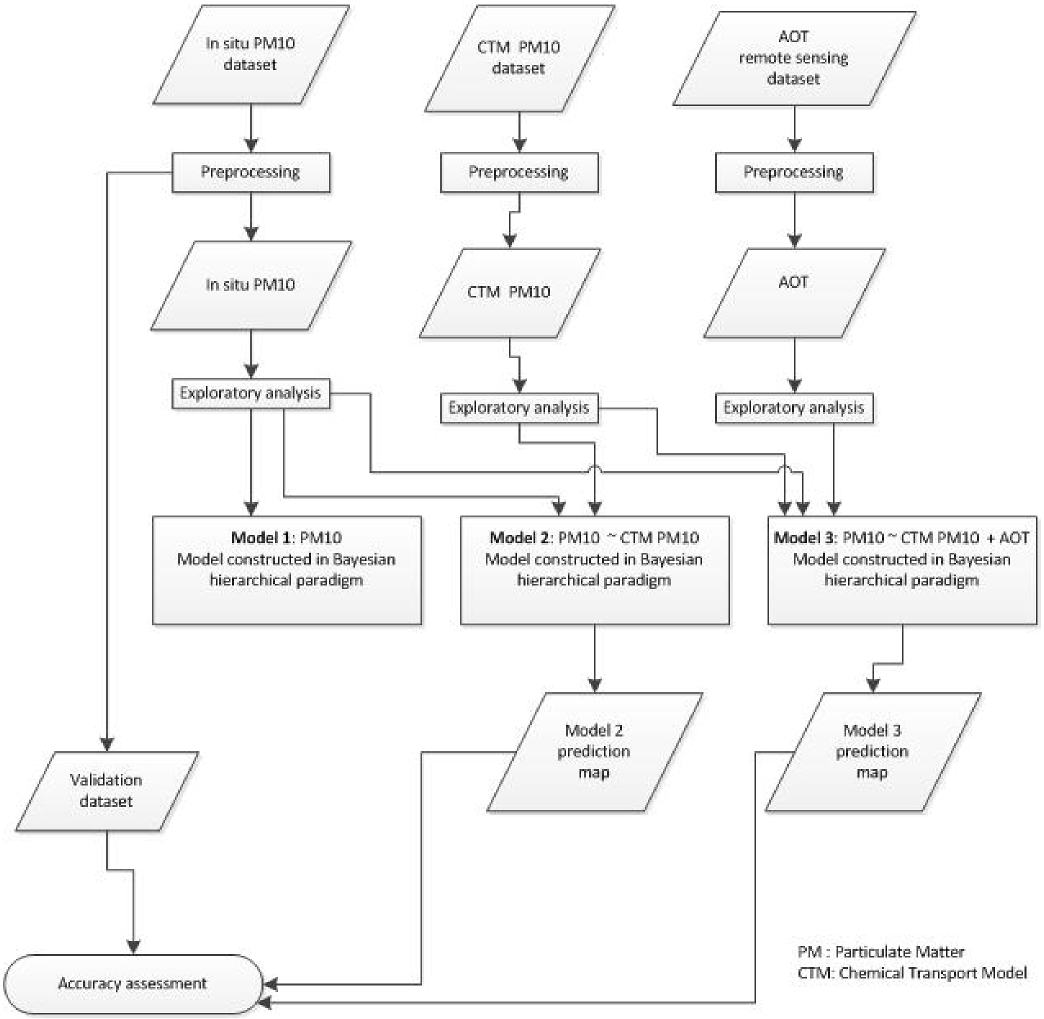

In situ PM10 dataset
CTM PM10 dataset
AOT remote sensing dataset
Preprocessing
Preprocessing
Preprocessing
In situ PM10
CTM PM10
AOT
Exploratory analysis
Exploratory analysis
Exploratory analysis
Model 1: PM10 Model constructed in Bayesian hierarchical paradigm
Model 2: PM10 ~ CTM PM10 Model constructed in Bayesian hierarchical paradigm
Model 3: PM10 ~ CTM PM10 + AOT Model constructed in Bayesian hierarchical paradigm
Model 2 prediction map
Model 3 prediction map
Validation dataset
Accuracy assessment
PM : Particulate Matter
CTM: Chemical Transport Model

Figure 4.3: Workflow of decomposition level I

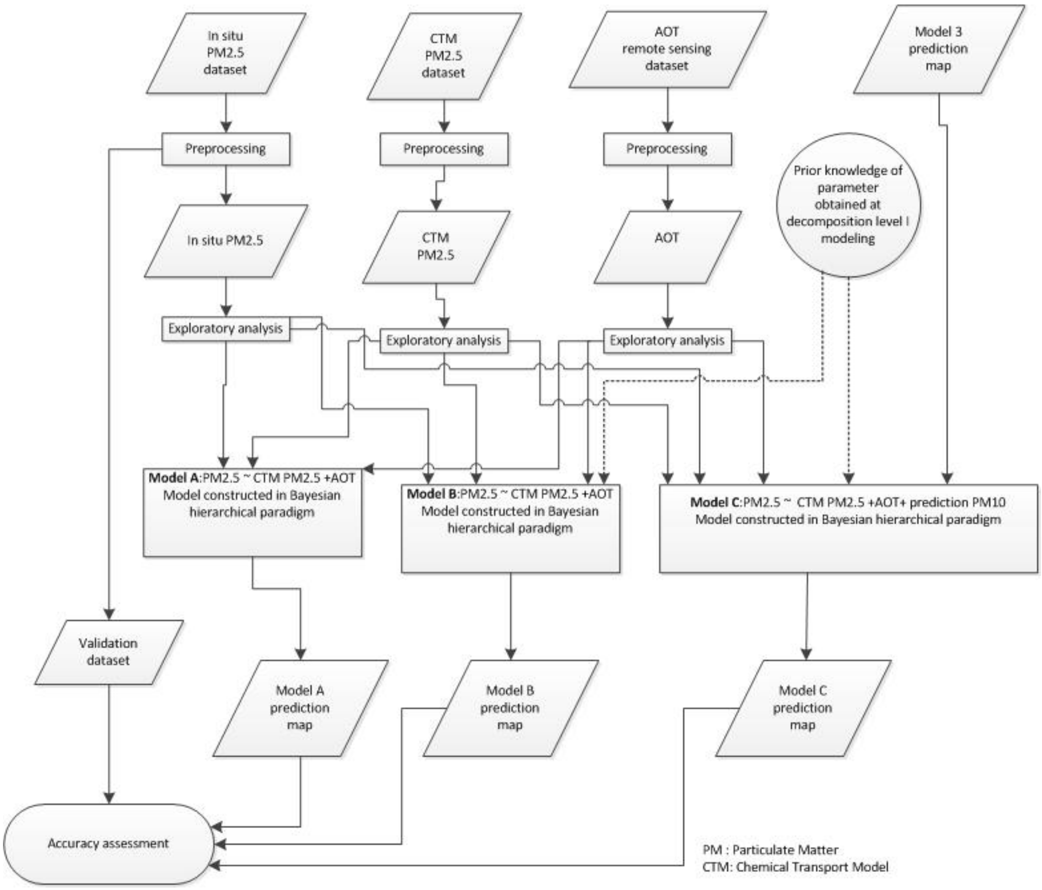

In situ PM2.5 dataset
CTM PM2.5 dataset
AOT remote sensing dataset
Model 3 prediction map
Preprocessing
Preprocessing
Preprocessing
Prior knowledge of parameter obtained at decomposition level I modeling
In situ PM2.5
CTM PM2.5
AOT
Exploratory analysis
Exploratory analysis
Exploratory analysis
Model A:PM2.5 ~ CTM PM2.5 +AOT Model constructed in Bayesian hierarchical paradigm
Model B:PM2.5 ~ CTM PM2.5 +AOT Model constructed in Bayesian hierarchical paradigm
Model C:PM2.5 ~ CTM PM2.5 +AOT+ prediction PM10 Model constructed in Bayesian hierarchical paradigm
Validation dataset
Model A prediction map
Model B prediction map
Model C prediction map
Accuracy assessment
PM : Particulate Matter
CTM: Chemical Transport Model

Figure 4.4: Workflow of decomposition level II

Chapter 5
Results

5.1 INTRODUCTION

Results obtained in the process of PM decomposition are described in this chapter.

5.2 DECOMPOSITION LEVEL I (PM10)

PM is decomposed in to its level I decomposition as per modeling approach adopted in section 4.2 and obtained results during the process are as Follows.

5.2.1 Exploratory analysis of decomposition level I

Exploratory analysis of dataset before modeling is an important aspect of geostatistical analysis. It give the primary idea of dataset and help to make primary assumption about the dataset. Descriptive statistics of the decomposition level I dataset calculated and their results are given in table 5.1 (In bracket value after log transformation given). Histogram and normal Q-Q plot of PM_{10} (Figure 5.1) indicate that data are asymmetrically (Non normal) distributed as well as summary statistics (Table 5.1) shows mean (23.60) is greater than median (22.62) which confirms positive skewness of PM_{10}. Standard deviation (7.26) of PM_{10} is high compared to the standard deviation(SD) of CTM PM_{10} (3.06) and AOT (0.23) (Table 5.1).

Table 5.1 Summary statistics of variable PM_{10}, CTM PM_{10} and AOT(In bracket value after log transformation given).

Parameter	PM_{10}	CTM PM_{10}	AOT
Mean	23.60 (3.12)	13.22 (2.56)	0.59 (-0.60)
Standard deviation	7.26 (0.25)	3.06 (0.20)	0.23 (0.40)
Median	22.62 (3.12)	12.32 (2.51)	0.56 (-0.59)
1st quartile	19.76 (2.98)	11.41 (2.41)	0.14 (-1.99)
3rd quartile	25.95 (3.26)	14.65 (2.68)	0.74 (-0.30)
Minimum value	9.94 (2.30)	8.12 (2.09)	0.14 (-1.96)
Maximum value	87.05 (4.46)	27.43 (3.31)	1.43 (0.36)

Histogram of CTM PM_{10} (Figure 5.1) shows positive skewness as well as Normal Q-Q plot (Figure 5.1) indicate shifting of data points over line which is indicator of non normal distribution. Mean (13.22) and median (12.32) confirms the skewness of CTM PM_{10} (Table 5.1). Histogram and Normal Q-Q plot (Figure 5.1) of AOT indicates positive skewness of data as well as mean(0.59) and median(0.56) also confirms the skewness. After the log transformation of PM_{10}, CTM PM_{10} and AOT data Histogram, Normal Q-Q plot and summary statistics of log transformed data (Figure 5.1, 5.1 and Table 5.1)shows approximated normal distribution of data.

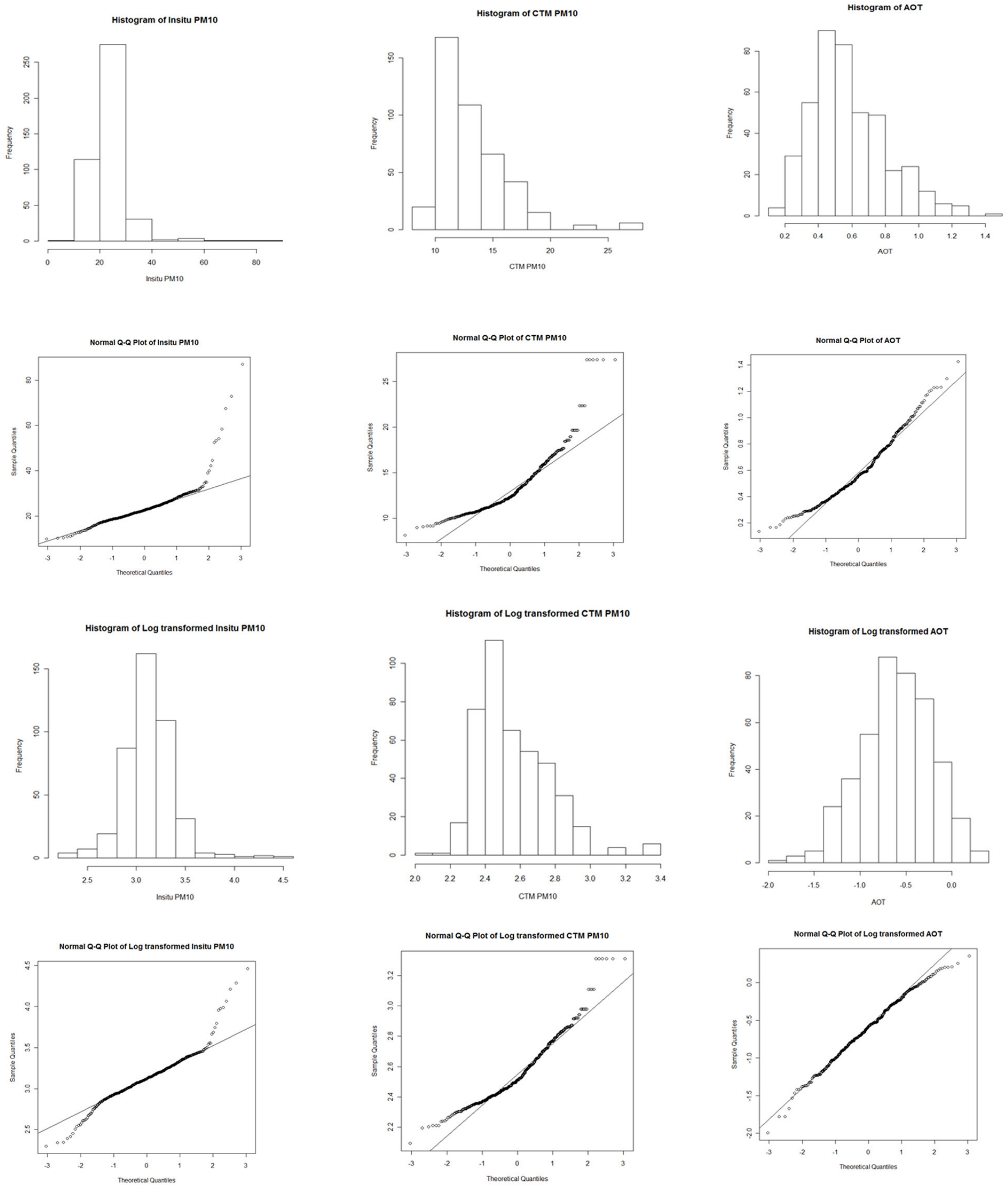


Figure 5.1: Histogram and Normal Q-Q plots of Variable PM10,CTM PM10 and AOT.

5.2.2 Correlation between variables

Log to log Pearson correlation between variable PM_{10}, CTM PM_{10} and AOT calculated and given in table 5.2, and scatter plot shown in figure 5.2. The highest positive correlation (0.39) is observed between PM_{10} and CTM PM_{10}. however, lowest positive correlation (0.04) observed in between PM_{10} and AOT. CTM PM_{10} shows 0.09 correlation with AOT.

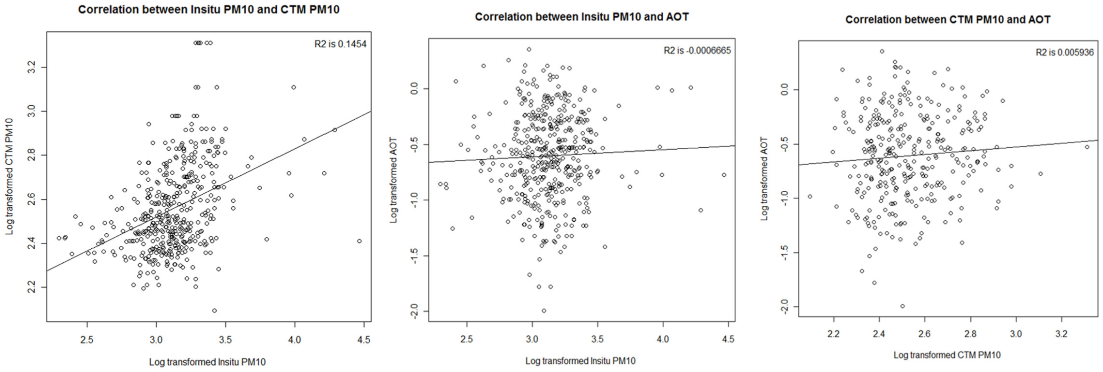


Figure 5.2: Scatter plots of log transformed PM_{10}, CTM PM_{10} and AOT

Table 5.2 log to log correlation between variable PM_{10}, CTM PM_{10} and AOT

	PM_{10}	CTM PM_{10}	AOT
PM_{10}	1		
CTM PM_{10}	0.39	1	
AOT	0.04	0.09	1

5.2.3 Variogram modeling

To understand the spatial structure of decomposition level I variables variogram was modeled. A fitted variogram to empirical variogram is shown in figure 5.3 and estimated parameter given in table 5.3. Estimated Range value varies among the variable PM_{10}, CTM PM_{10} and AOT. Highest spatial dependence observed for variable PM_{10} (360000 Meter) and lowest for variable AOT (283920 Meters). CTM PM_{10} shows range up to 320000 Meters. Non-spatial variability or nugget effect observed for variable PM_{10} and CTM PM_{10} is 0.035 and 0.010 respectively. However, no nugget effect (0.000) observed for variable AOT.

Table 5.3 Estimated variogram parameter for variable PM_{10}, CTM PM_{10} and AOT

Variable	Model	Nugget	Partial sill	Range
PM_{10}	Exponential	0.035	0.062	360000
CTM PM_{10}	Exponential	0.010	0.048	320000
AOT	Exponential	0.000	0.132	283920

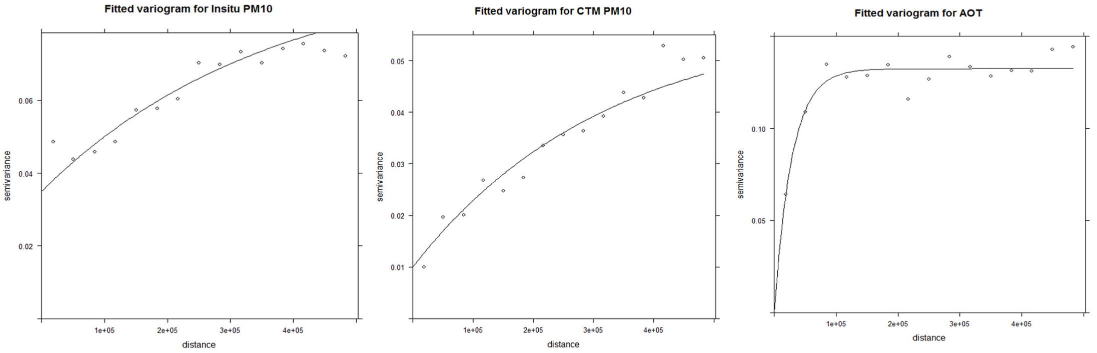


Figure 5.3: Variogram of PM_{10}, CTM PM_{10} and AOT

5.2.4 Bayesian hierarchical modeling of decomposition level I

As per modeling approach adopted for decomposing PM in to its components. For decomposition level I (PM_{10}) three different models constructed considering covariable to understand the effect of adding covariable in modeling. Three different models constructed in Bayesian paradigm namely **Model 1**: considering only response variable PM_{10}, **Model 2**: considering response variable PM_{10} and predictor covariable CTM PM_{10} and **Model 3**: considering response variable PM_{10} and predictor covariable CTM PM_{10} and AOT. Each model runs for 50,000 MCMC iterations to convergence of MCMC chain reached at homogeneous stationary distribution in parameter space first 40,000 iterations burn in. Last 10,000 iterations or samples used for calculation of posterior parameters summary statistics and prediction. Trace and density plot of each models parameter given in Figure 5.4 and 5.5. Summary statistics and percentiles of posterior parameters of each model given in table 5.4 and 5.5 respectively.

Table 5.4 Posterior summary statistics of each parameter (Decomposition level I : PM_{10})

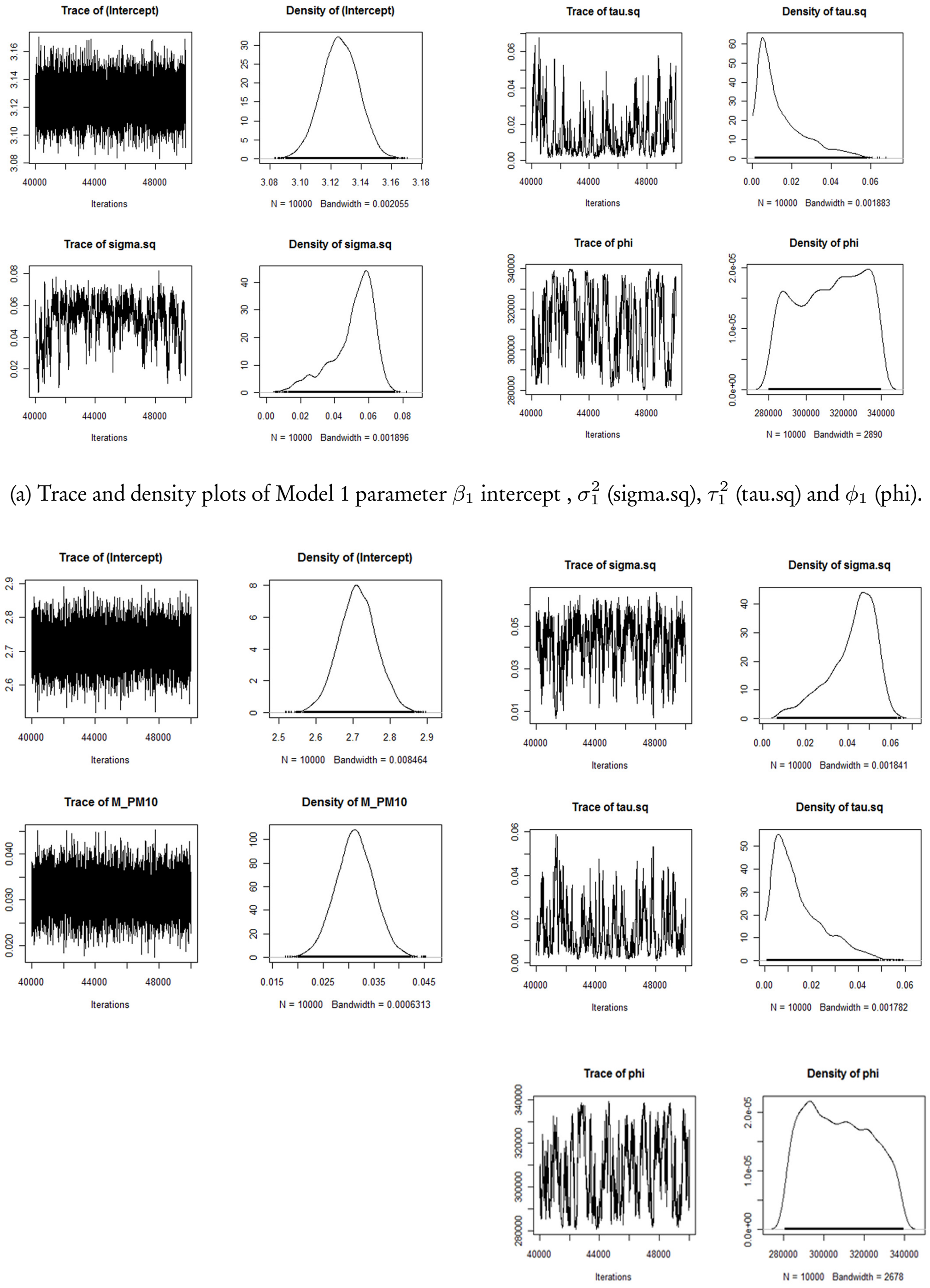
Parameters	Mean	SD	Naive SE
Model 1: PM_{10} intercept only			
β_1 intercept	3.126	0.012	1.223×10^{-04}
σ_1^2	0.050	0.013	1.267×10^{-04}
τ_1^2	0.015	0.012	1.224×10^{-04}
ϕ_1	312300	17200	172
Model 2:PM_{10} with predictor covariable CTM PM_{10}			
β_1 intercept	2.713	0.051	5.118×10^{-04}
β_1 CTM PM_{10}	0.031	0.004	3.785×10^{-05}
σ_1^2	0.042	0.011	1.176×10^{-04}
τ_1^2	0.015	0.010	1.123×10^{-04}
ϕ_1	307500	15940	159
Model 3:PM_{10} with predictor covariable CTM PM_{10} and AOT			
β_1 intercept	2.702	0.059	5.847×10^{-04}
β_1 CTM PM_{10}	0.031	0.004	3.783×10^{-05}
β_1 AOT	0.023	0.051	5.103×10^{-04}
σ_1^2	0.043	0.011	1.081×10^{-04}
τ_1^2	0.013	0.010	1.042×10^{-04}
ϕ_1	309500	13960	140

The mean of posterior parameter β_1 intercept (considering only response variable PM_{10}) decreasing from Model 1 to Model 3 (Table 5.4) as a result of adding of covariable in to modeling. However, standard deviation of β intercept increasing from Model 1 to Model 3 (Table 5.4). The highest mean (0.050) of the posterior parameter σ_1^2 (partial sill) observed for Model 1. The similar mean of posterior parameter τ_1^2 (nugget or non spatial variability) observed for Model 1 and Model 2 and lowest for Model 3 (Table 5.4). The highest mean of posterior range parameter ϕ_1 (range) observed for Model 1 and followed by Model 3 and Model 2. However, standard deviation of parameter ϕ_1 decreases from Model 1 to Model 3 (Table 5.4). Percentiles of the posterior parameter of each model given in table 5.5. The highest 95% credibile (2.5% percentile to 97.5% percentile) interval of parameter β_1 intercept observed for Model 1 and followd by Model 3 and Model 2 (Table 5.5) . The lowest 95% credibile interval for covariance parameter σ_1^2, τ_1^2 and ϕ_1 observed for Model 3 followed by Model 2 and Model 1 (Table 5.5). Naive SE is consistant for each parameter in all models (Model 1, Model 2 and Model 3).

Table 5.5 Percentiles of the posterior distribution of each parameter (Decomposition level I : PM_{10})

Parameters	2.5%	25%	50%	75%	97.5%
Model 1: PM_{10} intercept only					
β_1 intercept	3.102	3.118	3.126	3.134	3.150
σ_1^2	0.019	0.045	0.054	0.060	0.068
τ^2	0.002	0.005	0.010	0.020	0.047
ϕ_1	282861	297435	313576	327296	339042
Model 2:PM_{10} with predictor covariable CTM PM_{10}					
β_1 intercept	2.612	2.679	2.712	2.746	2.814
β_1 CTM PM_{10}	0.024	0.029	0.031	0.034	0.039
σ_1^2	0.014	0.036	0.045	0.050	0.058
τ_1^2	0.002	0.006	0.011	0.020	0.042
ϕ_1	282530	293451	306574	320870	336343
Model 3:PM_{10} with predictor covariable CTM PM_{10} and AOT					
β_1 intercept	2.587	2.662	2.701	2.742	2.817
β_1 CTM PM_{10}	0.024	0.029	0.031	0.034	0.039
β_1 AOT	-0.077	-0.012	0.022	0.057	0.124
σ_1^2	0.019	0.036	0.046	0.052	0.059
τ_1^2	0.002	0.005	0.009	0.020	0.038
ϕ_1	283741	298760	309416	320582	334158

To make sure that MCMC chain of each parameter converged to the stationary distribution in parameter space trace and density plot of each parameter plotted considering last 10,000 MCMC iterations after burning the first 40,000 MCMC iterations. After visually inspecting the trace and density plot of parameter σ_1^2 and τ_1^2 for Model 1, Model 2 and Model 3 shows unimodal distribution (Figure 5.4 and 5.5). However, trace and density plot of parameter ϕ_1 shows multimodal distribution for Model 1 and Model 2 as compared to Model 3 (Figure 5.4 and 5.5).



(a) Trace and density plots of Model 1 parameter β_1 intercept , σ_1^2 (sigma.sq), τ_1^2 (tau.sq) and ϕ_1 (phi).

(b) Trace and density plots of Model 2 parameter β_1 intercept , β_1 CTM PM_{10} (M PM10), σ_1^2 (sigma.sq), τ_1^2 (tau.sq) and ϕ_1 (phi).

Figure 5.4: Trace and density plot of Model 1 (a) and Model 2 (b) parameter.

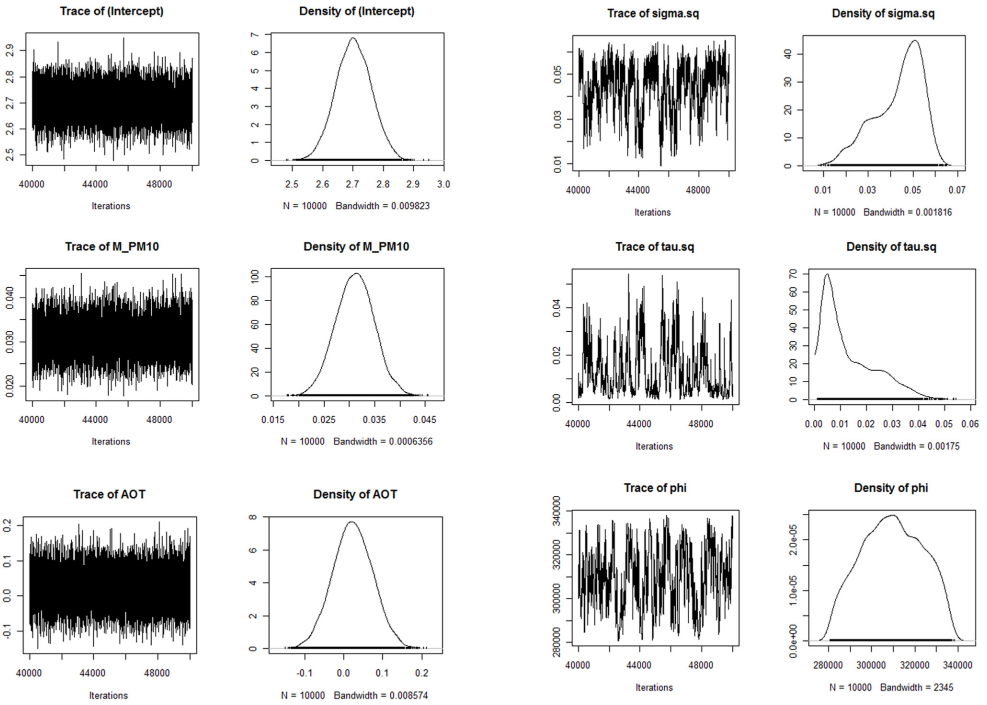


Figure 5.5: Trace and density plots of Model 3 parameter β intercept , β_1 CTM PM$_{10}$ (M PM10), β_1 AOT, σ_1^2 (sigma.sq), τ_1^2 (tau.sq) and ϕ_1 (phi).

5.2.5 Model selection

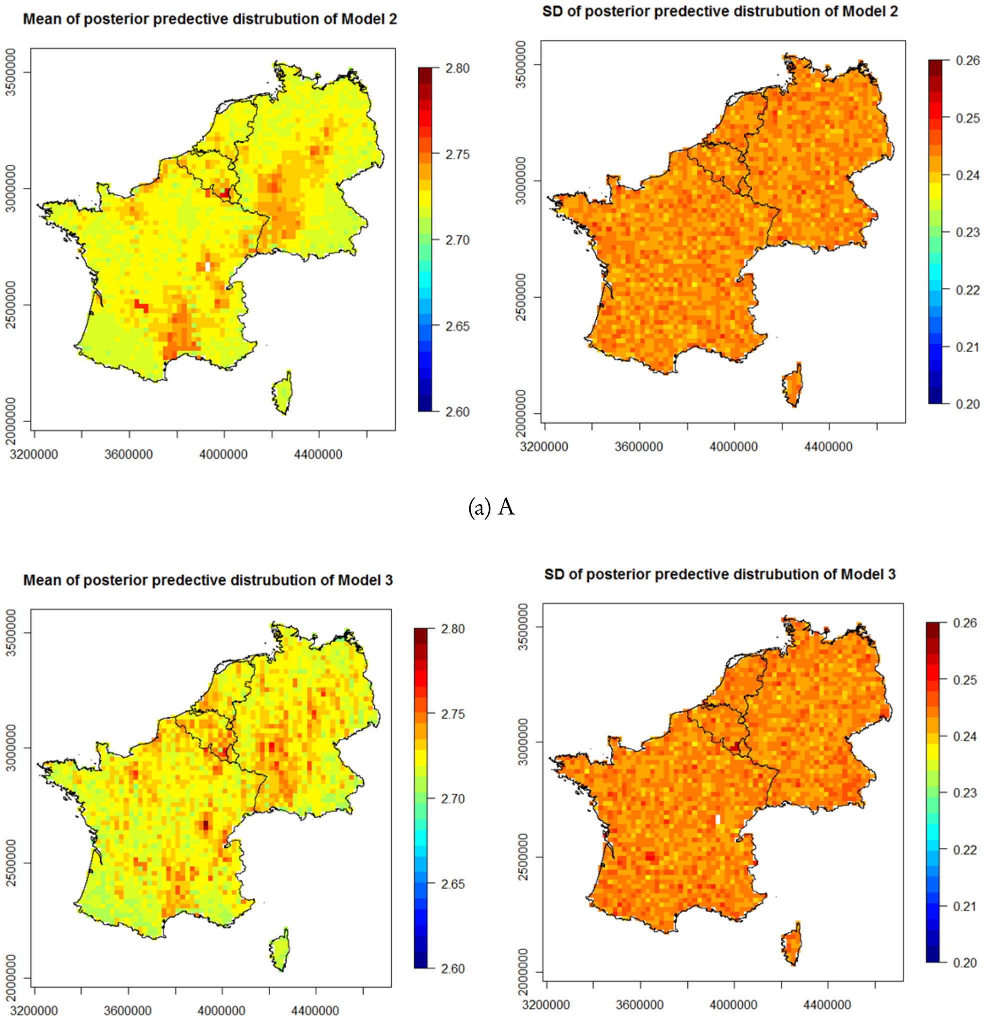
To compare the model for best fit and checking model adequacy deviance information criteria (DIC) and posterior predictive loss (D) criteria calculated for each model and their comparison in table 5.6. Smaller value of DIC and posterior predictive loss criteria indicate better fit of model. Lowest value of DIC and D observed for Model 2 followed by Model 3 (Table 5.6). Based on DIC and D criteria it seems Model 2 is an best model among other models. According to Banerjee et al. (2004) DIC is useful for when the objective is explaining the model and predictive loss criteria is useful when the objective of model is prediction. However, in present study our objective is prediction of PM component. Considering predictive loss criteria value of Model 2 (11.31) and Model 3 (13.68) it seems both model have close value and it makes selection of model difficult. Both Model 2 and Model 3 selected for prediction of response variable PM$_{10}$ (decomposition level I) at unsampled location.

Table 5.6 Model comparison using DIC and posterior predictive loss(D)criteria (Decomposition level I : PM$_{10}$)

Model	DIC	D
Model 1:PM$_{10}$ intercept only	-977	22.84
Model 2:PM$_{10}$ with predictor covariable CTM PM$_{10}$	-1347	11.31
Model 3:PM$_{10}$ with predictor covariable CTM PM$_{10}$ and AOT	-1281	13.68

5.2.6 Prediction of decomposition Level I

After building the model in Bayesian paradigm next procedure is to predict response variable PM_{10} with the help of densely sampled covariate CTM PM_{10} and AOT over geographic space. However, covariate CTM PM_{10} and AOT is not available at equal grid (perfect rectangular shape) due to longitude and latitude difference. For prediction purpose 20 km × 20 km grid created and gridded data of covariate attached to prediction grid. The prediction map at unsampled location for Model 2 and Model 3 shown in figure 5.6.


Mean of posterior predective distrubution of Model 2
SD of posterior predective distrubution of Model 2
Mean of posterior predective distrubution of Model 3
SD of posterior predective distrubution of Model 3
2000000
2500000
3000000
3500000
3200000
3600000
4000000
4400000
2.60
2.65
2.70
2.75
2.80
0.20
0.21
0.22
0.23
0.24
0.25
0.26

(a) A

(b) B

Figure 5.6: Mean and Standard deviation (SD) of posterior predictive distribution; A: Model 2 and B: Model 3

5.2.7 Validation of decomposition level I models

Accuracy assessment of prediction Model 2 and Model 3 (section 5.2.6) done using independent validation dataset and results given in table 5.7. Mean Error (ME) , SSE and RMSE value of both model shows quite similar (No large difference in values of Model 2 and Model 3)

Table 5.7 Accuracy assessment of decomposition level I models

Model	ME	SSE	RMSE
Model 2 :PM10 with covariable CTM PM10	0.4200	39.85	0.5646
Model 3: PM10 with covariable CTM PM10 and AOT	0.4176	39.65	0.5632

5.3 DECOMPOSITION LEVEL II (PM2.5)

PM is decomposed in to its level II decomposition as a $PM_{2.5}$ according to modeling approach adopted in section 4.2 and obtained results during the process given here.

5.3.1 Exploratory analysis of decomposition level II

Descriptive statistics of the decomposition level II dataset calculated and their results are given in table 5.8 (In bracket value after log transformation given). Histogram and normal Q-Q plot of $PM_{2.5}$ shows data are negatively skewed as well as median(16.77) is greater than mean(2.78) confirms negative skewness of data (Figure 5.7 and Table 5.8). Standard deviation of variable $PM_{2.5}$, CTM $PM_{2.5}$ and AOT is 4.39, 2.33 and 0.23 respectively defines the spread of data around mean(Table 5.8). Histogram and normal Q-Q plot of variable CTM $PM_{2.5}$ and AOT indicate data are positively skewed and summary statistics shows mean is greater than median which confirms positive skewness of both variable.

Table 5.8 Summary statistics of variable $PM_{2.5}$, CTM $PM_{2.5}$ and AOT(In bracket value after log transformation given).

Parameter	$PM_{2.5}$	CTM $PM_{2.5}$	AOT
Mean	16.72 (2.78)	9.79 (2.26)	0.61 (-0.56)
Standard deviation	4.39 (0.25)	2.33 (0.21)	0.23 (0.39)
Median	16.77 (2.82)	9.20 (2.22)	0.59 (-0.53)
1st quartile	14.23 (2.65)	8.30 (2.12)	0.45 (-0.80)
3rd quartile	18.50 (2.62)	10.90 (2.39)	0.75 (-0.28)
Minimum value	6.94 (1.90)	6.42 (1.86)	0.21 (-1.57)
Maximum value	44.75 (3.80)	19.64 (2.98)	1.23 (0.20)

Data of all variable ($PM_{2.5}$, CTM $PM_{2.5}$ and AOT) shows approximated normal distribution after log transformation and their Histogram and Normal Q-Q plots given in figure 5.7.

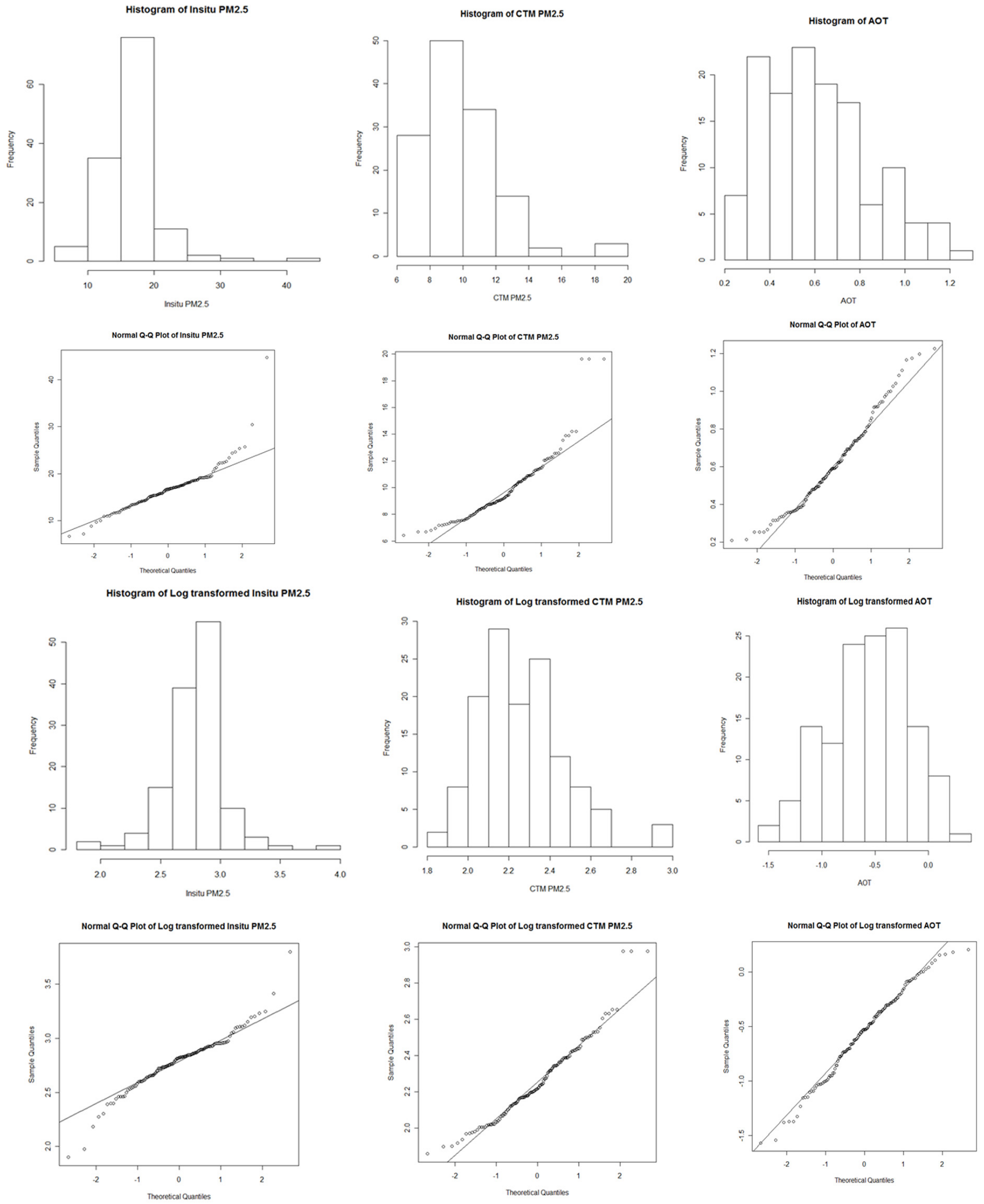


Figure 5.7: Histogram and Normal Q-Q plots of Variable $PM_{2.5}$,CTM $PM_{2.5}$ and AOT.

5.3.2 Correlation between variables (decomposition level II)

To understand the relationship between variables log to log Pearson correlation calculated and given in table 5.9. The highest positive correlation(0.33) observed between variable $PM_{2.5}$ and CTM $PM_{2.5}$ and lowest positive correlation(0.18) shown between CTM $PM_{2.5}$ and AOT. However, negative correlation(-0.44) observed between variable $PM_{2.5}$ and AOT.

Table 5.9 log to log correlation between variable $PM_{2.5}$, CTM $PM_{2.5}$ and AOT

	$PM_{2.5}$	CTM $PM_{2.5}$	AOT
$PM_{2.5}$	1		
CTM $PM_{2.5}$	0.33	1	
AOT	-0.14	0.18	1

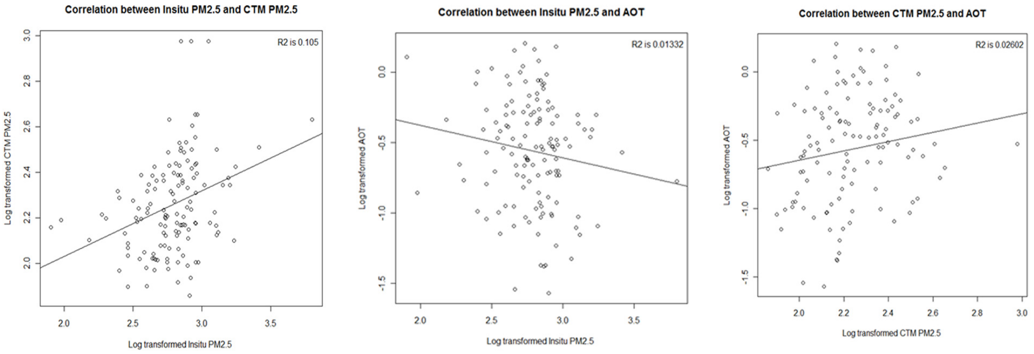


Figure 5.8: Scatter plots of log transformed $PM_{2.5}$, CTM $PM_{2.5}$ and AOT

5.3.3 Variogram modeling (decomposition level II)

To understand the spatial structure of variable $PM_{2.5}$, CTM $PM_{2.5}$ and AOT empirical variogram plotted and authorized variogram model fitted. The estimated parameter of variogram given in table 5.10 and fitted variogram of each variable shown in figure 5.10. Highest range observed for variable CTM $PM_{2.5}$ (384000 Meters) followed by $PM_{2.5}$ (290000 Meters) and AOT (266260 Meters). No nugget effect (non spatial variability) observed for AOT, However highest non spatial variability observed for $PM_{2.5}$ (0.028) followed by CTM $PM_{2.5}$ (0.002).

Table 5.10 Estimated variogram parameter for variable $PM_{2.5}$, CTM $PM_{2.5}$ and AOT

Variable	Model	Nugget	Partial sill	Range
$PM_{2.5}$	Exponential	0.028	0.065	290000
CTM $PM_{2.5}$	Exponential	0.002	0.049	384000
AOT	Exponential	0.000	0.128	266260

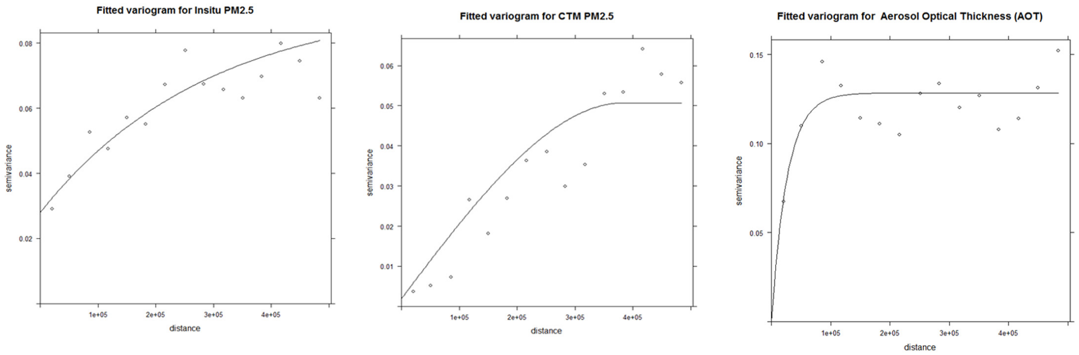


Figure 5.9: Variogram of $PM_{2.5}$, CTM $PM_{2.5}$ and AOT

5.3.4 Bayesian Hierarchical modeling of decomposition Level II

The decomposition level II ($PM_{2.5}$) was modeled separately in Bayesian paradigm considering their relationship with decomposition level I (PM_{10}) as per modeling approach adopted in section 4.2. For the decomposition level II ($PM_{2.5}$) three different models constructed to understand the feasibility of adopted modeling approach to model $PM_{2.5}$ by adding PM_{10} in the mean of process (predicted PM_{10} act as additional vector in covariable matrix of $PM_{2.5}$) as per II level decomposition methodology defined in section 4.4.5. Three different models constructed for $PM_{2.5}$ namely **Model A** : $PM_{2.5}$ as a function of predictor covariable CTM $PM_{2.5}$ and AOT without adding any information from PM_{10} (decomposition Level I), **Model B**: $PM_{2.5}$ as a function of predictor covariable CTM $PM_{2.5}$ and AOT with adding posterior parameter (σ_1^2, τ_1^2 and ϕ_1) of PM_{10} model (decomposition Level I)as the prior for parameter (σ_2^2, τ_2^2 and ϕ_2) in $PM_{2.5}$ modeling and **Model C** : $PM_{2.5}$ as a function of predictor covariable CTM $PM_{2.5}$, AOT and prediction of PM_{10} (additional covariable) as well as posterior parameter (σ_1^2, τ_1^2 and ϕ_1) of PM_{10} model (decomposition Level I) as the prior for parameter (σ_2^2, τ_2^2 and ϕ_2) in $PM_{2.5}$ model.

Table 5.11 Posterior summary statistics of each parameter (Decomposition level II : $PM_{2.5}$)

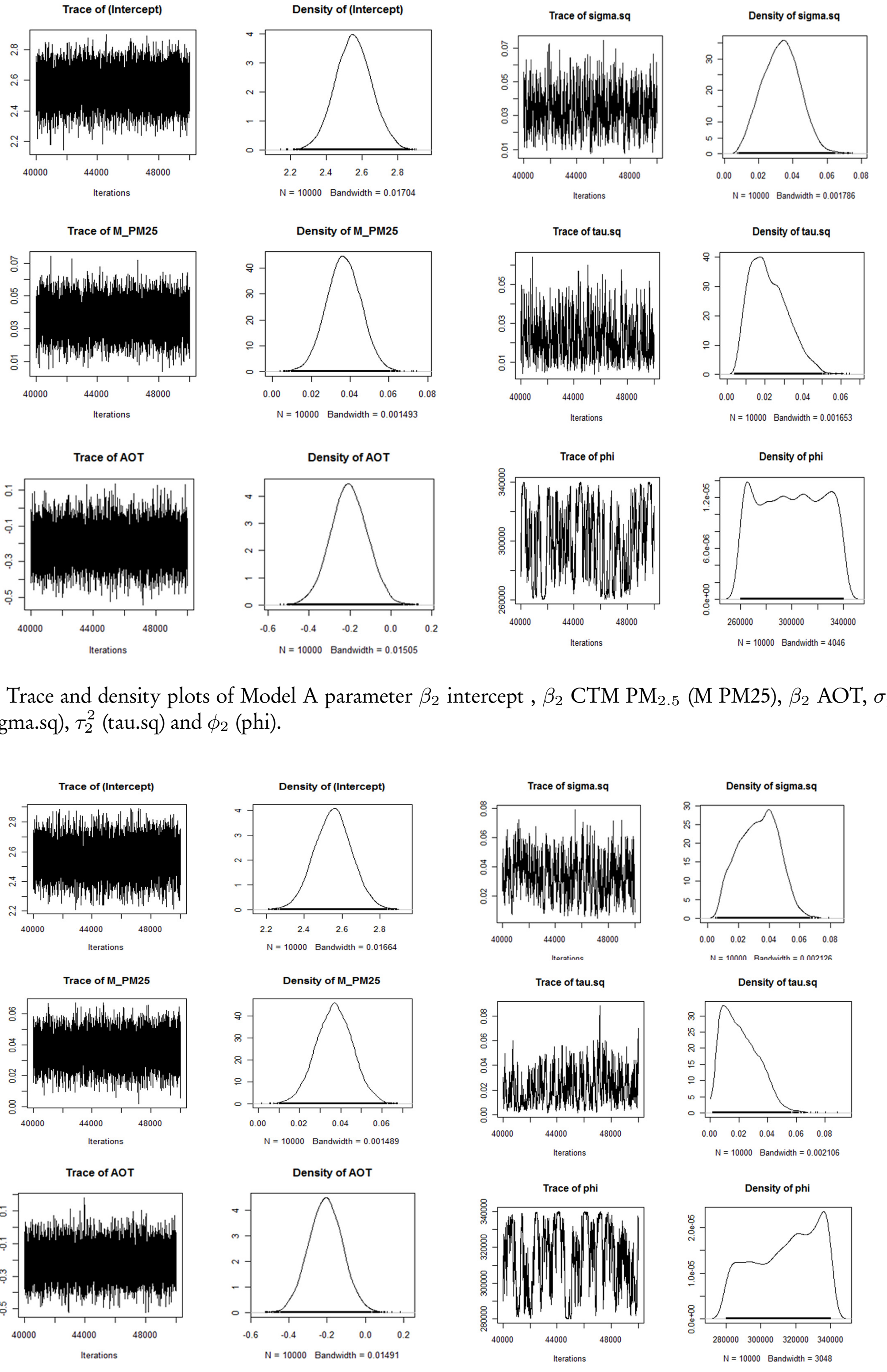
Parameters	Mean	SD	Naive SE
Model A			
β_2 intercept	2.551	0.101	1.014×10^{-03}
β_2 CTM PM 25	0.036	0.008	8.900×10^{-05}
β_2 AOT	-0.204	0.090	9.065×10^{-04}
σ_2^2	0.033	0.010	1.063×10^{-04}
τ_2^2	0.022	0.009	9.840×10^{-05}
ϕ_2	299500	24090	240
Model B			
β_2 intercept	2.553	0.099	9.943×10^{-04}
β_2 CTM PM 25	0.037	0.009	8.862×10^{-05}
β_2 AOT	-0.208	0.088	8.874×10^{-04}
σ_2^2	0.033	0.013	1.265×10^{-04}
τ_2^2	0.021	0.012	1.253×10^{-04}
ϕ_2	314400	18140	181
Model C			
β_2 intercept	1.951	0.260	2.602×10^{-03}
β_2 CTM PM 25	0.035	0.009	8.686×10^{-05}
β_2 AOT	-0.277	0.086	8.611×10^{-04}
β_2 predicted PM10	0.232	0.093	9.339×10^{-04}
σ_2^2	0.033	0.012	1.188×10^{-04}
τ_2^2	0.019	0.011	1.140×10^{-04}
ϕ_2	311000	16530	165

These three different models (Model A, Model B and Model C) runs for 50,000 MCMC iterations and first 40,000 MCMC iterations burn out to make sure convergence of MCMC chain in parameter space. Last 10,000 iterations used for calculating posterior summary of parameter and prediction. Summary statistics and percentiles of posterior parameters of each model given in table 5.11 and 5.12 respectively as well as trace and density plot of parameters given in Figure 5.10 and 5.11. The lowest mean with highest standard deviation of posterior parameter β_2 intercept observed for Model C as compared with Model A and Model B (5.12). However, Model A and Model B shows quite similar mean and standard deviation for parameter β_2 intercept (Table 5.12). The mean and standard deviation of the covariance parameters σ_2^2 and τ_2^2 shows variation of 0.001 to 0.003 between Models (Model A, Model B and Model c) (Table 5.12). The highest mean of parameter ϕ_2 observed for Model B followed by Model C and Model A. However, lowest standard deviation of parameter ϕ_2 shown by Model C followed by Model B and Model A.

Table 5.12 Percentiles of the posterior distribution of each parameter (Decomposition level II : $PM_{2.5}$)

Parameters	2.5%	25%	50%	75%	97.5%
Model A					
β_2 intercept	2.354	2.482	2.551	2.619	2.755
β_2 CTM PM25	0.019	0.031	0.037	0.043	0.054
β_2 AOT	-0.383	-0.264	-0.205	-0.144	-0.027
σ_2^2	0.013	0.025	0.033	0.040	0.054
τ_2^2	0.007	0.014	0.021	0.029	0.044
ϕ_2	261511	278570	299771	320718	338801
Model B					
β_2 intercept	2.357	2.486	2.553	2.619	2.750
β_2 CTM PM25	0.020	0.031	0.037	0.043	0.054
β_2 AOT	-0.383	-0.268	-0.208	-0.148	-0.031
σ_2^2	0.010	0.023	0.034	0.043	0.056
τ_2^2	0.004	0.011	0.020	0.030	0.048
ϕ_1	280866	299145	317008	330630	339418
Model C					
β_2 intercept	1.442	1.779	1.953	2.127	2.453
β_2 CTM PM25	0.018	0.030	0.035	0.041	0.053
β_2 AOT	-0.397	-0.284	-0.224	-0.170	-0.058
β_2 predicted PM10	0.053	0.169	0.232	0.294	0.416
σ_2^2	0.011	0.024	0.034	0.042	0.055
τ_2^2	0.004	0.009	0.017	0.027	0.045
ϕ_2	283276	296617	311396	325106	338468

Percentiles of the posterior parameter of each model given in table 5.12 . The lowest 95 % credibile interval for covariance parameter σ_2^2 and τ_2^2 observed for Model A followed by Model C and Model B (Table 5.12). However, lowest 95 % credibile interval for parameter ϕ_2 shown by Model C followed by Model B and Model A (Table 5.12). The Model C shows highest (1.011) 95 % credibile interval for parameter β_2 intercept and lowest (0.4) for Model A and Model B.



(a) Trace and density plots of Model A parameter β_2 intercept , β_2 CTM PM$_{2.5}$ (M PM25), β_2 AOT, σ_2^2 (sigma.sq), τ_2^2 (tau.sq) and ϕ_2 (phi).

(b) Trace and density plots of Model B parameter β_2 intercept , β_2 CTM PM$_{2.5}$ (M PM10), β_2 AOT, σ_2^2 (sigma.sq), τ_2^2 (tau.sq) and ϕ_2 (phi).

Figure 5.10: Trace and density plot of Model A (a) and Model B (b) parameters.

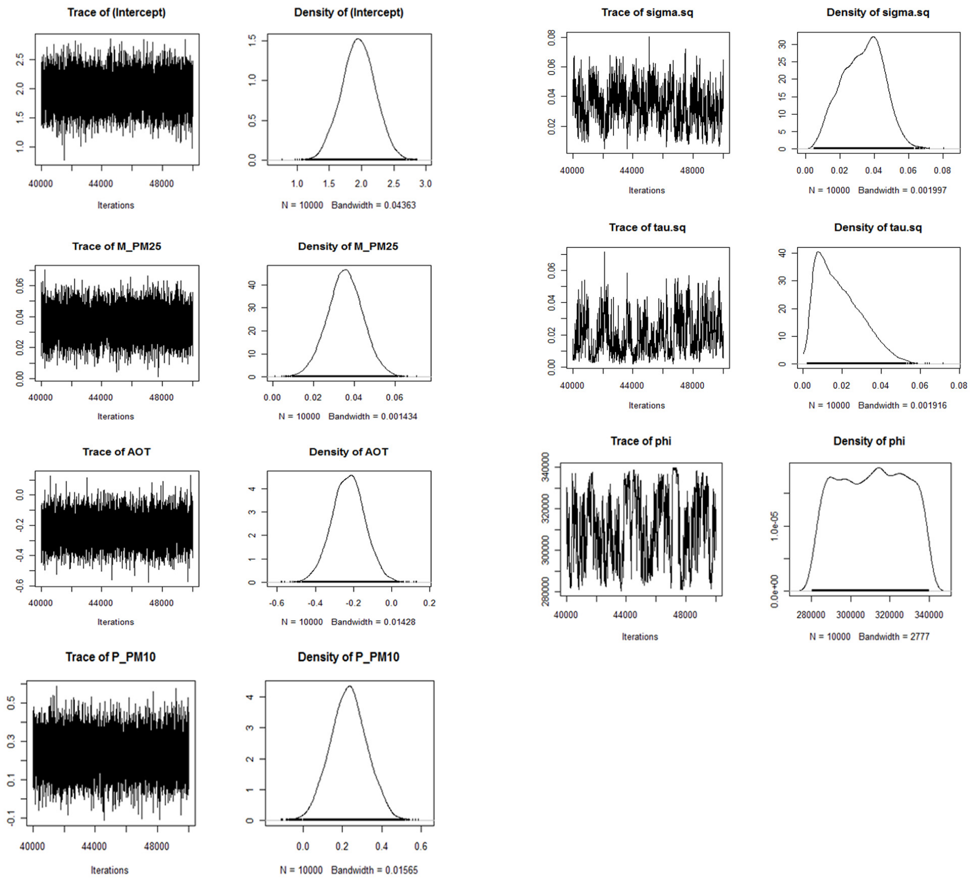


Figure 5.11: Trace and density plots of Model C parameter β_2 intercept , β_2 CTM PM$_{2.5}$ (M PM25), β_2 AOT, β_2 prediction PM$_{10}$, σ_2^2 (sigma.sq), τ_2^2 (tau.sq) and ϕ_2 (phi).

5.3.5 Model selection (decomposition level II)

To check the model for best fit and its adequacy DIC and predictive loss (D) criteria calculated for each model and their comparison given in table 5.13. Lowest value of DIC observed for Model A followed by Model B and Model C, though, Model C shows quite high value of DIC as compared with Model A and Model B (Table 5.13). However, posterior predictive loss criteria indicate lowest value for Model C followed by Model B and Model A.

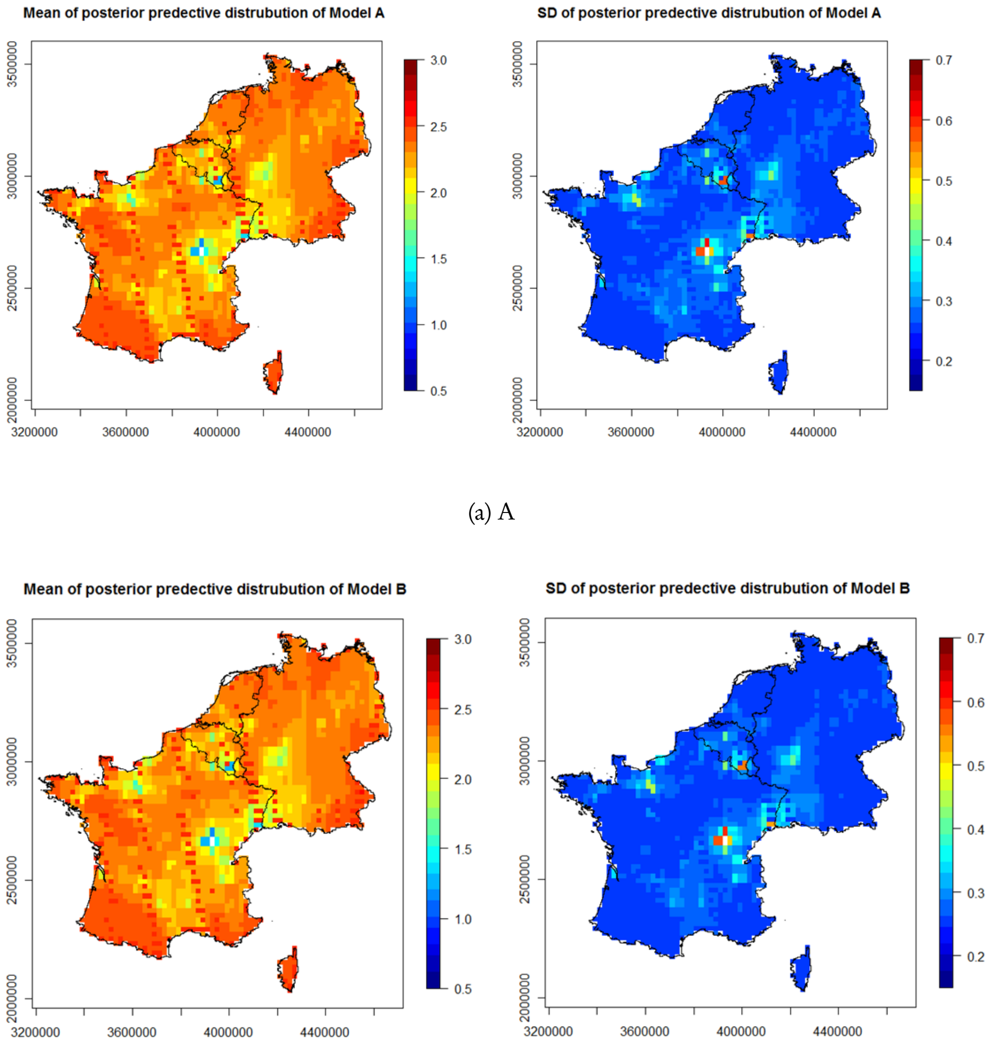
Table 5.13 Model comparison using DIC and posterior predictive loss(D)criteria (Decomposition level II : PM$_{2.5}$)

Model	DIC	D
Model A	-319	5.62
Model B	-346	5.46
Model C	-350	5.18

DIC and posterior predictive loss criteria usually used for selecting model for prediction. However, in level II decomposition (PM$_{2.5}$ modeling) objective is not only the prediction but also checking the adopted approach (Section 4.2 and 4.4.5) of level II decomposition. In view of this all three models (Model A, Model B and Model C) selected for predicting response variable $PM_{2.5}$ at unsampled location.

5.3.6 Prediction of decomposition level II

Response variable $PM_{2.5}$ predicted at unsampled location for Model A and Model B with the help of densely sampled covariate CTM $PM_{2.5}$ and AOT. However, for Model C prediction predicted surface of PM_{10} act as one additional covariate along with CTM $PM_{2.5}$ and AOT. Due to unavailability of covariate CTM $PM_{2.5}$ and AOT data on equal grid , 20×20 Km prediction grid created and covariate data attached with it for prediction purpose. The predicted surface of Model A, Model B and Model C shown in Figure 5.12 and 5.13.


Mean of posterior predective distrubution of Model A
SD of posterior predective distrubution of Model A
Mean of posterior predective distrubution of Model B
SD of posterior predective distrubution of Model B
2000000
2500000
3000000
3500000
3200000
3600000
4000000
4400000
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.3
0.4
0.5
0.6
0.7

(a) A

(b) B

Figure 5.12: Mean and Standard deviation (SD) of posterior predictive distribution; A: Model A and B: Model B

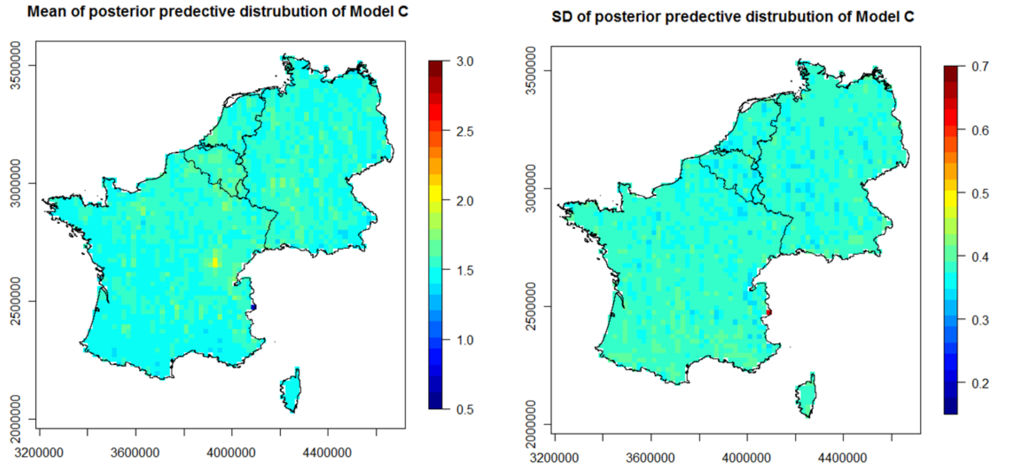

Mean of posterior predective distrubution of Model C
SD of posterior predective distrubution of Model C
3500000
3000000
2500000
2000000
3200000
3600000
4000000
4400000
3.0
2.5
2.0
1.5
1.0
0.5
0.7
0.6
0.5
0.4
0.3
0.2

Figure 5.13: Mean and Standard deviation (SD) of posterior predictive distribution of Model C.

5.3.7 Validation of decomposition level II models

Accuracy assessment of prediction Model A, Model B and Model C (Section 5.3.6) done using independent validation dataset and results given in Table 5.14.

Table 5.14 Accuracy assessment of decomposition level II models

Model	ME	SSE	RMSE
Model A	0.5168	17.96	0.6701
Model B	0.5159	17.90	0.6691
Model C	1.2334	64.27	1.2676

As per defined in section 5.3.4 Model A , Model B and Model C constructed to check the methodological approach of level II decomposition, in view of this validation result obtained in process is an important to make inferential statement on adopted methodology. ME, SSE and RMSE of Model A and Model B is quite similar (the difference observed in second digit). However, for Model C ME, SSE and RMSE is high as compared to the Model A and Model B as well as ME and RMSE of Model C is above 1.

Chapter 6
Discussion

This chapter discusses the formulated methodology for decomposition of PM and results obtained during decomposition process.

As per modeling approach defined for PM decomposition in section 4.2, PM decomposed in to finer level component considering decomposition level I and decomposition level II. In decomposition level I and II PM is modeled as PM_{10} and $PM_{2.5}$ component respectively. During the PM_{10} and $PM_{2.5}$ modeling densely gridded CTM and AOT data used as covariable for adding strength to response variable. The reason behind adding CTM and AOT data in PM (PM_{10} and $PM_{2.5}$) modeling due to the in situ observations of PM are available at few limited locations and it affect interpolation. van de Kassteele et al. (2006) showed the strength of adding CTM and AOT data in to PM_{10} mapping and concludes secondary source data (CTM and AOT) give more accurate and precise prediction.

6.1 CORRELATION AND SPATIAL STRUCTURE

In present study multisource (In situ, CTM and Remote sensing) data used for decomposing PM in to PM_{10} (decomposition level I) and $PM_{2.5}$ (decomposition level II). It is important to know the correlation and spatial structure of variable coming from different source. The highest positive correlation observed between variable PM_{10} and CTM PM_{10} (Section 5.2.2) for decomposition level I and between variable $PM_{2.5}$ and CTM $PM_{2.5}$ (Section 5.3.2) for decomposition level II. As evident, the high correlation present between the variable coming from in situ and CTM data may be due to the CTM data incorporates the anthropogenic emission data of air pollutants which contains the in situ observed PM component (Schaap et al., 2009a). The low correlation observed between in situ PM (PM_{10} and $PM_{2.5}$) and AOT (data downloaded from OMI sensor (section 3.3.2). As compare to the correlation between In situ PM and AOT , high correlation observed between CTM and AOT may be due to CTM data incorporates satellite data in assimilation (Schaap et al., 2009b).

To understand the spatial structure of variable coming from different source variogram fitted to each decomposition level (I and II) (section 5.2.3 and 5.3.3). Through the variogram analysis highest nugget effect shown by in situ data for decomposition level I and level II may be due to the fact of in situ measurement observations techniques , calibration procedures are vary from country to country (even city from city of same country). As comparing the value of range range parameter between variable PM_{10} and $PM_{2.5}$, large range observed for PM_{10} (360 km) compared to $PM_{2.5}$ (290 km). This indicate that these two PM components (PM_{10} and $PM_{2.5}$) behave differently in atmosphere may be due to their emission source are different. Value of range parameter of AOT shows different in decomposition level I and II irrespective of AOT is an same dataset may be because range value of AOT sensitive to number of samples used in calculation as it is the only difference.

6.2 DECOMPOSITION LEVEL I (PM_{10})

For decomposing PM in to its component at level I decomposition (PM_{10}) three different models constructed in Bayesian paradigm to understand the effect of covariable in modeling and their detail description given in section 5.2.4. Covariable CTM PM_{10} is positively related to the PM_{10} in Model 2 and Model 3 as looking at the mean of the posterior parameter β_1 CTM PM_{10} (Table 5.4) as well as at the 95% credibile interval of parameter β_1 CTM PM_{10} excluded zero (Table 5.5) shows statistical significance (Jiang et al., 2009). van de Kassteele et al. (2006) shows strength of CTM data in mapping of PM_{10}. Covariable AOT also shows the positively relation to the PM_{10} (Table 5.4) but with no statistical significance as looking at 95% credibile interval of β_1 AOT includes zero (Table 5.5). The reason behind the AOT is not adding strength in Model 3 (no statistical significance) might be because the low correlation(0.04) observed between responce variable PM_{10} and AOT (Table 5.2). The low uncertainty as considering lowest 95% credibile interval for PM_{10} (parameter β_1 PM_{10}) shown by Model 2 as adding covariable CTM PM_{10}. The standard deviation of the range (spatial dependence) parameter ϕ_1 is decreasing from Model 1 to Model 3 as adding the covariable in to modeling. The lowest uncertainty (minimum 95% credibile interval) of the covariance parameter σ_1^2, τ_1^2 and ϕ_1 observed for Model 3 may be because of large portion of variation explained by the regressor CTM PM_{10} and AOT together.

After constructing the models for PM_{10} (decomposition level I) the adequacy of model defined by DIC and posterior predictive loss criteria indicate Model 3 is best fitted model among other models (Model 1 and Model 3) (Table 5.6). However, as comparing the posterior predictive loss criteria value between the Model 2 and Model 3 it seems both model have closer value. Rather looking at the DIC and posterior predictive loss criteria value for selecting model for prediction both model (Model 2 and Model 3) selected because densely gridded covariate incorporated in to the models. During the prediction of response variable PM_{10} at unsampled location with the help of covariate it takes full advantage of densely sampled covariate over predictive space , this approach have an advantage over sparsely sampled PM_{10}. Accuracy assessment of prediction shows similar (difference observed after second digit of value) ME and RMSE for Model 2 and Model 3 as comparing the ME and RMSE it seems most of inaccuracy coming from bias prediction (Table 5.7).

6.3 DECOMPOSITION LEVEL II ($PM_{2.5}$)

PM decomposed at level II as a component $PM_{2.5}$ considering their relationship with PM_{10} (decomposition level I). Three different models (Model A, Model B and Model C) constructed for $PM_{2.5}$ (section 5.3.4) to check the feasibility of adopted methodological approach. The standard deviation and 95% credibile interval of Parameter β_2 intercept shows highest for Model C as compared to the Model A and Model B suggest large portion of variation added by regressor predicted PM_{10} because its extra covariable added in Model C (section 5.3.4 and 4.4.5). Comparing the 95% interval of covariance parameter for parameter σ_2^2 and τ_2^2 shows lowest for Model A (indicate low uncertainty) but no large difference from Model B and Model c indicate different prior not affecting the posterior estimate of these parameters. The 95% credibile interval of parameter β_2 CTM $PM_{2.5}$, β_2 AOT and β_2 predicted PM_{10} excludes zero indicate these covariable adding information in to model with statistical significance.

After constructing models for decomposition level II for checking methodological approach three models selected for prediction as they builded with different approach(section 5.3.4). Validation results of Model A and Model B shows similar ME and RMSE value (the difference observed after second digit) indicate that prior given based on some understanding of parameter (Model A) and adopting prior knowledge (Model B) from related spatial process (spatial process PM_{10} and

$PM_{2.5}$ interrelated as both parts of PM) not affecting the prediction. However, Model C shows high value of ME and RMSE as compare to Model A and Model B indicate adding the decomposition level I prediction (PM_{10}) in to the mean of the process of decomposition level II ($PM_{2.5}$) leads bias prediction.

Chapter 7
Conclusion and Recommendations

The main objective of this study was to develop and evaluate geostatistical prediction model in Bayesian paradigm for predicting PM components with the help of multisource observations (in situ measurements, Chemical Transport model and remote sensing data). To achieve the main objective research questions formulated and their answers are given in this chapter.

7.1 CONCLUSION

What is the spatial structure of PM components and their associated covariable observed from multisource observations ?

In present study two PM components namely PM_{10} and $PM_{2.5}$ were modeled with added strength from multisource covariable. For PM_{10} multisource covariable namely CTM PM_{10} (coming from CTM source) and AOT (coming from satellite remote sensing) were used and for $PM_{2.5}$ covariable namely CTM $PM_{2.5}$ (CTM source) and AOT (satellite remote sensing source) were used. The spatial structure of PM_{10} and $PM_{2.5}$ with its covariable was evaluated based on variogram fitting, highest non-spatial variability (nugget effect) was observed for in situ PM components (PM_{10} and $PM_{2.5}$) followed by CTM given covariable (CTM PM_{10} and CTM $PM_{2.5}$) and no nugget effect observed for AOT. The highest value of partial sill (spatial variance) was observed for AOT covariable followed by in situ PM components (PM_{10} and $PM_{2.5}$) and covariable CTM (CTM PM_{10} and CTM $PM_{2.5}$). For component PM_{10} highest value for range parameter was observed for in situ PM_{10} followed by CTM PM_{10} and AOT. However, for component $PM_{2.5}$ highest value was observed for covariable CTM $PM_{2.5}$ followed by in situ PM_{10} and AOT.

As evident, the highest non-spatial variability observed for both in situ PM component (PM_{10} and $PM_{2.5}$) due to uncertainty associated with in situ measurement procedure. A difference was displayed by value of range parameter (which is a measure for the distance up to which spatial dependence present) between PM_{10} and $PM_{2.5}$. This indicates that spatial dependence of PM_{10} component is present at a large distance as compared to $PM_{2.5}$ component.

Which covariable (CTM or remote sensing) gives more accurate prediction?

To understand the effect of adding covariable in modeling; Model 1, Model 2 and Model 3 were constructed at decomposition level I (PM_{10}). The CTM given covariable (CTM PM_{10}) was added in to Model 2 and covariable AOT added in to Model 3 along with CTM PM_{10}. The validation results of model show that both Model 2 and Model 3 have similar RMSE 0.5646 and 0.5632 respectively. Looking at the RMSE values of models it is difficult to choose which covariable gives more accurate a prediction.

How can PM components relationship be incorporated in to the Bayesian hierarchical model?

Considering that the relationship between PM components (PM_{10} and $PM_{2.5}$) is a subset of a large spatial process, it is possible to add relationship in to the mean of the process of target PM component like PM_{10} was modeled in the mean of the $PM_{2.5}$ (Model C) by adding as a covariable. However, another approach towards adding PM components relationship via the prior knowledge about the parameter (mean and covariance parameter of the process) obtained during modeling individual PM is a starting point for modeling target PM component like Model B developed in this study.

Does PM components relationship improve the model prediction ? Why?

PM components relationship added in to the modeling considering two approaches as a Model C (adding relationship in to the mean of the process) and Model B (using the prior knowledge about the parameter as starting point for modeling) were developed. On comparison, Model C (RMSE = 1.2676) and Model B (RMSE = 0.6691) with independently developed Model A (RMSE = 0.6701) showed a difference between their RMSE values. There is no improvement observed in modeling when PM components relationship added in to the mean of process (Model C) while prior knowledge about the parameter added in to modeling (Model B) shows improvement. Model C approach shows no improvement in modeling because assumption of subset process of large spatial process and their joint distribution does not hold true due to different emission source, different atmospheric chemistry of these two components (PM_{10} and $PM_{2.5}$) and leads to in bias prediction. However, Model B approach shows improvement in modeling because relationship of PM component does not directly take part in the modeling but helps to understand the parameter (mean and covariance parameters) distribution of process.

What is the accuracy of model developed for prediction ?

Considering three different approach models developed for PM component prediction. The models (Model 2, Model 3 and Model A) developed without adding any information of PM components relationship shows RMSE between 0.5632 to 0.6701. Model developed with using PM relationships prior knowledge about the parameter as starting point for modeling (Model B) shows RMSE 0.6691 and model developed with adding PM component relationship in to mean of the process (Model C) shows RMSE 1.2676.

7.2 RECOMMENDATIONS

Considering the importance of the PM and its components and their association with health and environment and results obtained in present study, I recommend the following points for modeling the decomposition of PM in to its components.

1. Modeling PM components in Bayesian paradigm using precursor dataset considering precursor relationship with PM component like CTM models.
2. Modeling anisotropy considering air pollutants dispersion and transport depending on meteorological variable.
3. Developing model at finer level resolution.

Bibliography

Atkinson, R. W., Ross Anderson, H., Sunyer, J., Ayres, J., Baccini, M., Vonk, J. M., Boumghar, A., Forastiere, F., Forsberg, B., Touloumi, G., Schwartz, J., and Katsouyanni, K. (2001). Acute effects of particulate air pollution on respiratory admissions . results from aphea 2 project. *American Journal of Respiratory and Critical Care Medicine*, 164(10):1860–1866.

Babak, O. and Deutsch, C. (2009a). Collocated cokriging based on merged secondary attributes. *Mathematical Geosciences*, 41(8):921–926.

Babak, O. and Deutsch, C. V. (2009b). Improved spatial modeling by merging multiple secondary data for intrinsic collocated cokriging. *Journal of Petroleum Science and Engineering*, 69(1-2):93–99.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 1 edition.

Buseck, P. R. and Pósfai, M. (1999). Airborne minerals and related aerosol particles: Effects on climate and the environment. *Proceedings of the National Academy of Sciences*, 96(7):3372–3379.

Cackvoic, M., Sega, K., Vadic, V., and Beslic, I. (2008). Characterisation of major acidic anions in tsp and pm10 in zagreb air. *Bulletin of Environmental Contamination and Toxicology*, 80(2):112–114.

Chow, J. C. and Watson, J. G. (2002). Review of pm2.5 and pm10 apportionment for fossil fuel combustion and other sources by the chemical mass balance receptor model. *Energy and Fuels*, 16(2):222–260.

Diggle, P. and Ribeiro, P. (2007). *Model - based geostatistics*. Springer series in statistics. Springer, New York.

Dockery, D. W., Schwartz, J., and Spengler, J. D. (1992). Air pollution and daily mortality: Associations with particulates and acid aerosols. *Environmental Research*, 59(2):362–373.

Emili, E., Popp, C., Petitta, M., Riffler, M., Wunderle, S., and Zebisch, M. (2010). Pm10 remote sensing from geostationary seviri and polar-orbiting modis sensors over the complex terrain of the european alpine region. *Remote sensing of Environment*, 114(11):2485–2499.

EPA, U. (2012). Web site of the US EPA for Characteristics of Particles - Particle Size Categories. http://www.epa.gov/apti/bces/module3/category/category.htm, (11/02/2012).

Fang, G.-C., Chang, C.-N., Wu, Y.-S., Fu, P. P.-C., Yang, C.-J., Chen, C.-D., and Chang, S.-C. (2002). Ambient suspended particulate matters and related chemical species study in central taiwan, taichung during 1998-2001. *Atmospheric Environment*, 36(12):1921–1928.

Gupta, P. and Christopher, S. A. (2009). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *J. Geophys. Res.*, 114(D14):D14205.

Hengl, T., Heuvelink, G. B. M., and Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers and Geosciences*, 33(10):1301–1315.

Huang, C., Yao, Y., Cressie, N., and Hsing, T. (2009). Multivariate intrinsic random functions for cokriging. *Mathematical Geosciences*, 41(8):887–904.

Huang, Y., Dickinson, R. E., and Chameides, W. L. (2006). Impact of aerosol indirect effect on surface temperature over east asia. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4371–4376.

Jiang, P., He, Z., Kitchen, N., and Sudduth, K. (2009). Bayesian analysis of within-field variability of corn yield using a spatial hierarchical model. *Precision Agriculture*, 10(2):111–127.

King, M. D., Kaufman, Y. J., Tanré, D., and Nakajima, T. (1999). Remote sensing of tropospheric aerosols from space: Past, present, and future. *Bulletin of the American Meteorological Society*, 80(11):2229–2259.

Li, C., Hsu, N. C., and Tsay, S.-C. (2011). A study on the potential applications of satellite data in air quality monitoring and forecasting. *Atmospheric Environment*, 45(22):3663–3675.

Liu, Y., Guo, H., Mao, G., and Yang, P. (2008). A bayesian hierarchical model for urban air quality prediction under uncertainty. *Atmospheric Environment*, 42(36):8464–8469.

Martin, R. V. (2008). Satellite remote sensing of surface air quality. *Atmospheric Environment*, 42(34):7823–7843.

Morawska, L., He, C., Hitchins, J., Gilbert, D., and Parappukkaran, S. (2001). The relationship between indoor and outdoor airborne particles in the residential environment. *Atmospheric Environment*, 35(20):3463–3473.

Pilewskie, P. (2007). Climate change: Aerosols heat up. *Nature*, 448(7153):541–542. 10.1038/448541a.

Pope, C., Burnett, R., Thun, M., Calle, E., Krewski, D., Ito, K., and Thurston, G. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA: the journal of the American Medical Association*, 287(9):1132.

Péré, J. C., Pont, V., Mallet, M., and Bessagnet, B. (2009). Mapping of pm10 surface concentrations derived from satellite observations of aerosol optical thickness over south-eastern france. *Atmospheric Research*, 91(1):1–8.

Pöschl, U. (2005). Atmospheric aerosols: Composition, transformation, climate and health effects. *Angewandte Chemie International Edition*, 44(46):7520–7540.

Ramanathan, V. and Carmichael, G. (2008). Global and regional climate changes due to black carbon. *Nature Geosci*, 1(4):221–227. 10.1038/ngeo156.

Ramanathan, V., Ramana, M. V., Roberts, G., Kim, D., Corrigan, C., Chung, C., and Winker, D. (2007). Warming trends in asia amplified by brown cloud solar absorption. *Nature*, 448(7153):575–578. 10.1038/nature06019.

Schaap, M., Manders, A., Hendriks, E., Cnossen, J., Segers, A., Denier van der Gon, H., Jozwicka, M., Sauter, F., Velders, G., Matthijsen, J., and Builtjes, P. (2009a). *Regional modelling of particulate matter for the Netherlands*.

Schaap, M., Timmermans, R., Segers, A., and Eskes, H. (2009b). Lotos -euros,products, quality and background information. Technical report, Netherlands Organization for Applied Scientific Research.

Schwartz, J. (1994). Air pollution and daily mortality: A review and meta analysis. *Environmental Research*, 64(1):36–52.

Singh, V., Carnevale, C., Finzi, G., Pisoni, E., and Volta, M. (2011). A cokriging based approach to reconstruct air pollution maps, processing measurement station concentrations and deterministic model simulations. *Environmental Modelling and Software*, 26(6):778–786.

van de Kassteele, J. (2006). *Statistical air quality mapping*. Phd thesis.

van de Kassteele, J., Koelemeijer, R., Dekkers, A., Schaap, M., Homan, C., and Stein, A. (2006). Statistical mapping of pm10 concentrations over western europe using secondary information from dispersion modeling and modis satellite observations. *Stochastic Environmental Research and Risk Assessment*, 21(2):183–194.

Van Dingenen, R., Raes, F., Putaud, J. P., Baltensperger, U., Charron, A., Facchini, M. C., Decesari, S., Fuzzi, S., Gehrig, R., Hansson, H. C., Harrison, R. M., Huglin, C., Jones, A. M., Laj, P., Lorbeer, G., Maenhaut, W., Palmgren, F., Querol, X., Rodriguez, S., Schneider, J., ten Brink, H., Tunved, P., Torseth, K., Wehner, B., Weingartner, E., Wiedensohler, A., and Wahlin, P. (2004). A european aerosol phenomenology-1: physical characteristics of particulate matter at kerbside, urban, rural and background sites in europe. *Atmospheric Environment*, 38(16):2561–2577.

Veefkind, J., Boersma, K., Wang, J., Kurosu, T., Krotkov, N., Chance, K., and Levelt, P. (2011). Global satellite analysis of the relation between aerosols and short-lived trace gases. *Atmospheric Chemistry and Physics*, 11:1255–1267.

Wang, Z., Chen, L., Tao, J., Zhang, Y., and Su, L. (2010). Satellite-based estimation of regional particulate matter (pm) in beijing using vertical-and-rh correcting method. *Remote sensing of environment*, 114(1):50–63.

Weijers, E. P., Schaap, M., Nguyen, L., Matthijsen, J., van der Gon, H., ten Brink, H. M., and Hoogerbrugge, R. (2011). Anthropogenic and natural constituents in particulate matter in the netherlands. *Atmospheric Chemistry and Physics*, 11(5):2281–2294.