GEOSTATISTICAL MODELLING AND MAPPING OF AIR POLLUTION

BAYARMAA ENKHTUR

February, 2013

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

GEOSTATISTICAL MODELLING AND MAPPING OF AIR POLLUTION

BAYARMAA ENKHTUR
Enschede, The Netherlands, February, 2013

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:
Prof. Dr. Ir. George Vosselman (Chair)
Dr. Gerard Heuvelink (External Examiner, Wageningen University)

ABSTRACT

In this rapidly developing industrialized world, several factors negatively influence the environment with problematic consequences. One of the major problems is in air quality. Understanding pollutants' spatial distribution and monitoring the air quality by applying geostatistical approaches is challenging and topical in data quality research field. Prediction of air pollutant distribution and air pollution mapping could be improved by extending geostatistical approaches in spatio-temporal domain with the help of atmospheric models, in-situ field measurements and remote sensing data and techniques.

The research aims to model and to map air pollution data by applying geostatistical space-time approach integrating secondary information from different sources.

Data pre-processing and combining all available dataset, primary in-situ measurements of PM10, grid data of CTM outcome, raster dataset of DEM and land cover map have been done prior to the implementation of spatio-temporal modelling.

Overall, 580 monitoring stations in the north-western European five countries were used in the spatio-temporal modelling. Observed PM10 concentrations vary in space and time. And it's interesting to see that these variations are different in different month spatially and temporally. During the months of the warm seasons (May, June, July and August), concentrations are relatively lower, and in colder seasonal times, extreme high concentration were observed. This could be explained by factors that influence the air pollution become active in cold seasonal days. On the other hand, PM10 concentrations in April are observed to be high concentrations and even their CTM outcomes were high. Reason of this was explained by huge clouds of ash in the atmosphere due to the volcanic eruption in northern Iceland in between end of March and April.

Multiple regression analysis was carried out in order to choose significant explanatory variables as covariates. Due to insignificant result from DEM, it is neglected from the further analysis. And CORINE Land cover LABEL1 category and CTM outcome were used as explanatory variables and factor.

Separable spatio-temporal model was implemented. It was assumed for the modelling that spatial structure of the daily PM10 concentrations is constant over a year, but unknown (second-order stationarity). Spatio-temporal variogram function is defined as a function of spatial and temporal distance h_s and h_t respectively. Kriging has been performed based on this variogram estimation using the complete dataset.

Prediction map was created by spatio-temporal universal kriging based on the fitted separable spatio-temporal model at unsampled locations of CTM grid dataset using stations that have continuous observation during whole month (no missing value). As for the remaining stations, there were used for the validation.

Probability map of exceedance has been achieved by spatio-temporal indicator kriging based on given threshold value of daily PM10 concentrations. And exceedance map has been created by defining probability threshold as 0.697 in January.

Key words: PM10 concentrations, geostatistical mapping, separable model, spatio-temporal model, indicator kriging.

ACKNOWLEDGEMENTS

Words are inadequate to express my gratitude to my both supervisors, Dr. Nicholas Hamm and Prof. Dr. Ir. Alfred Stein. It is an honour for me doing this MSc thesis under their supervisions within this research period. Their supervisions enable me to think widely and wisely, and to see the results from different sides.

I take this opportunity towards to Dr. Nicholas Hamm for his excessive support, valuable remarks and critical suggestion. And his great support on data collection, handling R code and technical problem really helped and encouraged me to accomplish this thesis work successfully.

I would like to express my sincere gratitude to Prof. Dr. Ir. Alfred Stein. His critical thoughts and constructive remarks motivated and led me through the right direction.

I would like to thank State Training Fund of Mongolian government for offering me the government grant, which made my studying possible at ITC.

I am greatly indebted to my parents, Sarantuya Dorjkhand and Enkhtur Dulamjav, who always support me in any way at any time no matter what. I am always thankful to my family for being there for me.

And last but not least, special thanks to my fellow cluster-mates and friends who supported me without knowing during this whole period of MSc-life for making this time more interesting and memorable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

PM	Particulate Matter
CTM	Chemical Transport Model
AOT	Aerosol Optical Thickness
MODIS	MODerate resolution Imaging Spectroradiometer
NetCDF4	Network Common Data Form 4
STFDF	Space-Time Full Data Frame
EEA	European Environmental Agency
ETOPO5	Earth TOPOgraphy 5-minute
DEM	Digital Elevation Model
ETRS 1989 LAEA	The European Terrestrial Reference System 1989 Lambert Azimuthal Equal Area
CORINE	Coordination of Information on the Environment

1. INTRODUCTION

1.1. Motivation and problem statement

In this rapidly developing industrialized world, several factors negatively influence the environment with problematic consequences. One of the major problems is in air quality. Polluted atmosphere has a harmful impact on human health and quality of our life. Dickey (2000) explained role of primary pollutant groups, such as ozone (O_3), sulfur dioxide (SO_2), nitrogen oxides (NO_x), carbon monoxide (CO), particulate matter (PM) and other air pollutants, and mentioned that they typically present together pollution but vary by location, source activity, season, weather and year. Understanding pollutants' spatial distribution and monitoring the air quality by applying geostatistical approaches is challenging and topical in data quality research field.

Geostatistics is a branch of science that applies statistical methods to spatial interpolation. It has broad application in different disciplines and its methods could be integrated with geoinformation technologies to improve the quality of mapping. Nature of spatial variables varies in space and in time, and understanding their spatial distribution at any given location and at specific time gives an opportunity to predict the events over a particular area. Prediction of air pollutant distribution and air pollution mapping could be improved by extending geostatistical approaches in spatio-temporal domain with the help of atmospheric models, in-situ field measurements and remote sensing data and techniques.

Research has suggested that accuracy of mapping could be improved by integrating those different information sources, specifically using geostatistical modelling. van de Kassteele, Stein*, et al.* (2006) focused on statistical techniques for detailed mapping of major air pollutants: ozone, NO_2 and PM and its uncertainties, and concluded that additional information such as chemical transport model (CTM) and aerosol optical thickness (AOT) from Moderate Resolution Imaging Spectroradiometer (MODIS) led to more accurate and precise spatial interpolation result. Other types of geographic information such as elevation, land cover, traffic and so on could also be used to improve modelling and mapping (Mwenda, 2011; Desta, 2012). Although, additional source data is integrated in the modelling and prediction, temporal aspect of the variables were neglected in the most cases.

Nevertheless, spatio-temporal context is not a new field of research. It has been actively studied over decades. In order to integrate the multiple data sources and to model the local and/or global variation, model-based analysis with spatio-temporal context could be applied (Lark & Cullis, 2004; Denby *et al.*, 2008; Mwenda, 2011; Desta, 2012; Gräler *et al.*, 2012). In integrating the air pollution data with the other additional data, the data quality of the different sources draws attention and it would effect on estimating the model parameters, further the spatial and temporal prediction of the air pollution data.

Data analysis and predictions which based on joint spatial and temporal dependencies between observations are provided by geostatistical spatio-temporal models in probabilistic framework, and the joint space-time dependencies is not fully modelled in estimation of the unknown value at unmonitored location (after Kyriakidis & Journel, 1999). Moreover the most recent paper reviewed that spatial aspect of the variables has been extensively studied by applying geostatistical approaches with the help of remote sensing and image analysis, but the time aspect has been disregarded in most cases (van der Meer, 2012). Since the PM concentrations vary in short temporal and spatial scale (Koelemeijer *et al.*, 2006; Kloog *et al.*,

2011) and given datasets are rich in temporal context, geostatistical joint space-time models, which are supported by the secondary sources, are needed to be taken into consideration further.

As modelling the air pollutants' concentrations, health impacts of air pollution could be quantified with the help of identifying certain thresholds. In 2005, World Health Organization introduced *Air quality guidelines* (AQGs) which are intended for offering global guidance on reducing the health impacts of air pollution (WHO, 2011). As defined in the AQGs (World Health Organization, 2006), no threshold for PM has been identified, the lowest PM concentration need to be achieved as much as possible. Thus, exceeding of the PM consideration over the specified thresholds needs to be mapped and its implication for human health should be emphasized.

Therefore, by extending the previous research, geostatistical space-time models should be implemented based on the exploring the spatial patterns of the PM concentrations integrating with additional sources such as land cover, and elevation data at different spatial resolution and temporal aggregation. This would give improved results in geostatistical mapping of air pollution since temporal aspect of the distribution of PM concentrations is involved in the modelling.

1.2. Research objectives

1.2.1. Overall research objective

The research aims to model and to map air pollution data by applying geostatistical space-time approach integrating secondary information from different sources. In order to achieve the overall objective, the following research specific objectives and research questions are stated.

1.2.2. Specific research objectives

1. To investigate spatial correlations between in-situ measurements of PM concentrations, CTM and additional sources such as elevation and land cover over the regions in different temporal scales.
2. To examine temporal correlation between in-situ measurements of PM concentrations, CTM and additional sources.
3. To integrate data with different spatial resolution and different qualities.
4. To create map of PM concentrations, including specific threshold exceedance.

1.2.3. Research questions

Q1. What is the spatial distribution of the PM concentrations over study area?
Q2. What is the spatial relationship between in-situ measurements, CTM and additional sources such as elevation and land cover over the regions in different temporal scales?
Q3. What is the temporal correlation between in-situ measurements, CTM and additional sources such as elevation and land cover over the regions in different temporal scales?
Q4. How to integrate data with different spatial resolution and different qualities?
Q5. How can we develop space-time models of PM concentrations?
Q6. How can exceedance map of PM concentration be made based on the defined threshold value?
Q7. What area of Europe has a high probability of exceeding PM concentrations?

1.3. Innovation

Innovation of the research aims at improved geostatistical space-time modelling and mapping of air pollution data by integrating secondary information from different sources considering temporal aspects.

1.4. Thesis structure

The thesis consists of seven chapters. Research motivation, objectives and questions are described in Chapter 1. Chapter 2 provides literature review on previous studies related to the air pollutants and space-time models of air pollution. Chapter 3 defines description of study area and given datasets. Methodology part and results are provided in Chapter 4 and Chapter 5 respectively. Discussion and limitation of the research are provided in Chapter 6. Chapter 7 delivers conclusion and recommendation of the research.

2. RELATED WORK

2.1. Data related work

As exposure to air pollutants has harmful impact on quality of air, human health and ecosystem (WHO, 2011; EEA, 2012), limit values of air pollutants are defined. European Commission has set daily[1] and yearly limit values for PM10 which are $50\mu g/m^3$ and $40\mu g/m^3$ (EEA, 2012) over Europe. To control the exceedance of PM10 over a region, geostatistical modelling and mapping of air pollutants have been actively studied (Koelemeijer *et al.*, 2006; van de Kassteele, Koelemeijer, *et al.*, 2006; Denby *et al.*, 2008; Dadvand *et al.*, 2011; Kloog *et al.*, 2011; Lee *et al.*, 2011).

Different countries use different measurement and calibration methods for obtaining air pollution data which complicates the regional modelling and mapping (van de Kassteele, Koelemeijer, *et al.*, 2006; Denby *et al.*, 2008). Statistical method for standardizing PM measurements from different countries was proposed by van de Kassteele, Koelemeijer, *et al.* (2006). They used three factors called internal explanatory variables and included them in the linear model in order to achieve standardized PM10.

In order to improve understanding of air pollutants, secondary information from other sources e.g., CTM, meteorological condition, elevation and AOT, has been used in the statistical modelling of air pollution (van de Kassteele, Koelemeijer, *et al.*, 2006; Konovalov *et al.*, 2009; Kloog *et al.*, 2011). CTM is a 3D computer simulation which is designed for prediction of air pollutants' concentrations. AOT products are derived from MODIS satellite imagery to support climate modelling. These products have been found to be advantageous for the modelling in many studies. van de Kassteele, Koelemeijer, *et al.* (2006) studied the interpolation of PM10 concentrations using these secondary information. They achieved improved results from adding either CTM outcome or AOT to the statistical modelling. Koelemeijer *et al.* (2006) compared spatial and temporal variations of yearly and monthly averaged AOT and limited PM concentration respectively, over Europe which concludes AOT products could be useful to improve monitoring of PM distribution as well.

Lee *et al.* (2011) proposed the calibration approach of MODIS AOT to investigate spatial patterns of the PM2.5 concentrations, extended their work by considering temporal aspect of the PM2.5 concentrations on land use regression model (Kloog *et al.*, 2011). In their studies, daily PM2.5 concentrations and AOT values are used.

As studying the previous related work, the research considered CTM outcome and land cover and elevation as secondary information in order to support limited in-situ measurements of PM10 concentrations.

2.2. Space-time modelling

To capture the high level of PM concentrations, hourly and daily measurements are taken in fixed monitoring stations. This enables statistical analysis in both space and time domains. Previous studies have developed and proposed geostatistical models are mostly neglect temporal aspect of the data (van de

[1] Not to be exceeded on more than 35 days per year.

Kassteele, Koelemeijer, *et al.*, 2006; Lee *et al.*, 2011). Since both spatially and temporally available data are at hand, space-time geostatistical model are preferable. Moreover, recent studies are more considered on this subject and approaches has been developed in various application, for example, space-time distribution of soil water in Jost *et al.* (2005), for human health in Gething *et al.* (2007); de Fouquet *et al.* (2011), Romanowicz *et al.* (2006) etc.

There are studies that reviewed how geostatistics is used and developed. van der Meer (2012) reviewed how geostatistics is played role in remote sensing studies which concluded space-time analysis is neglected most of the time. Kyriakidis and Journel (1999) and Sahu and Mardia (2005) provided clear reviews on research in geostatistical space-time domain. They extensively studied previous and recent geostatistical work of spatio-temporal domain in different field of disciplines. Kyriakidis and Journel (1999) reviewed various stochastic space-time models, and their pros and cons were highlighted. They discussed joint spatiotemporal models and gave formulation by adding additional time dimension (T) to the two-dimensional space (R^2)[2]. Random function of variables Y at locations s and instants t in time are expressed as,

$$\{Y(s,t),(s,t) \in R^2 \times T\}$$

Sahu and Mardia (2005) reviewed and discussed about textbooks, articles and history of development of geostatistical space-time models in number of disciplines which gives clear direction of recent trend of space-time models. After publication of these reviews, spatio-temporal models have been studied actively in mapping of air pollutants' concentrations. Such researches could be found in Cesare *et al.* (2001), De Iaco *et al.* (2002), de Fouquet *et al.* (2011), De Iaco *et al.* (2011), De Iaco and Posa (2012), Gräler *et al.* (2012), Gerharz *et al.* (2013), among others.

Cesare *et al.* (2001) introduced the product-sum covariance model which requires additional three parameters in the calculation of covariance function. These parameters could be achieved by computing covariance functions at zero spatial and time, and should be positive definitive. On basis of this model, De Iaco *et al.* (2002) estimated total air pollution in the Milan district, Italy by proposing new generalized space-time functional model.

Temporal variability of ozone concentrations has been studied in de Fouquet *et al.* (2011). In this paper, they gave clear description of geostatistical model and applied in hourly ozone observation, and compared the result with chemistry-transport model outcome over France. From their studies, temporal variability of the air pollutants is important for the modelling.

Sampson *et al.* (2011) presented spatio-temporal modelling of air quality data to provide long-term prediction of the PM2.5 concentrations using data from Air Quality System (AQS) fixed site monitors. Recent study, Gräler *et al.* (2012), analysed and interpolated daily and annual mean PM10 concentrations in European scale. They investigated and compared types of variogram models and concluded that interpolation using daily mean PM10 concentration gave improved result than using yearly mean concentrations.

Other than spatio-temporal covariance/variogram models, hierarchical Bayesian space-time models are studied as well (Banerjee *et al.*, 2003; Riccio *et al.*, 2006; Cocchi *et al.*, 2007). Textbook by Banerjee *et al.* (2003) explained hierarchical models in spatiotemporal context and gave formulation of its models and prediction.

[2] Some notation of the formulae differs

While spatio-temporal models are being developed, proper software or packages that can handle the new models and spatio-temporal data structure are needed. Integrated packages and programs have been developed by the software developers. Some are *gstat* with *spacetime* package in R (Pebesma, 2004, 2012), *R-INLA* package (Blangiardo *et al.*), and *GSLib* routine in FORTRAN (De Iaco *et al.*, 2010; De Iaco & Posa, 2012). *GSLib* is the customized routine for kriging based on generalized product-sum model proposed in De Iaco *et al.* (2001). *R-INLA* package is designed for Integrated Nested Laplace Approximation approach. *spacetime* package enables data structure in space and time and allows gstat functions to run spatio-temporal analysis (Pebesma, 2012) in R.

3. STUDY AREA AND DATA DESCRIPTION

3.1. Study area

EEA (2012) reported air quality of previous years over Europe. From their compared statistics for each country, it's shown that exceedance of PM10 concentrations is relatively reduced. However, spatio-temporal interpolation is required for investigating spatial and temporal variability of air pollutants in order to limit the concentrations as much as possible. Hence north-western European five countries were selected as study area in this research. Namely, they are Belgium, Czech Republic, Germany, The Netherlands and Poland (Figure 3-1).

Countries have different types of methods of correction factors to account for underestimation of the PM mass (van de Kassteele, Koelemeijer, *et al.*, 2006). Basically, to avoid the heterogeneity in space, study area has been narrowed down from European scale to regional even though data is available in most of the European territory. Monitoring stations are spatially limited in space but continuous in time; however, measurements are averaged on a daily basis and there are some missing values as well. Therefore, in-situ measurements of PM10 concentrations are fixed in space and discrete in time.

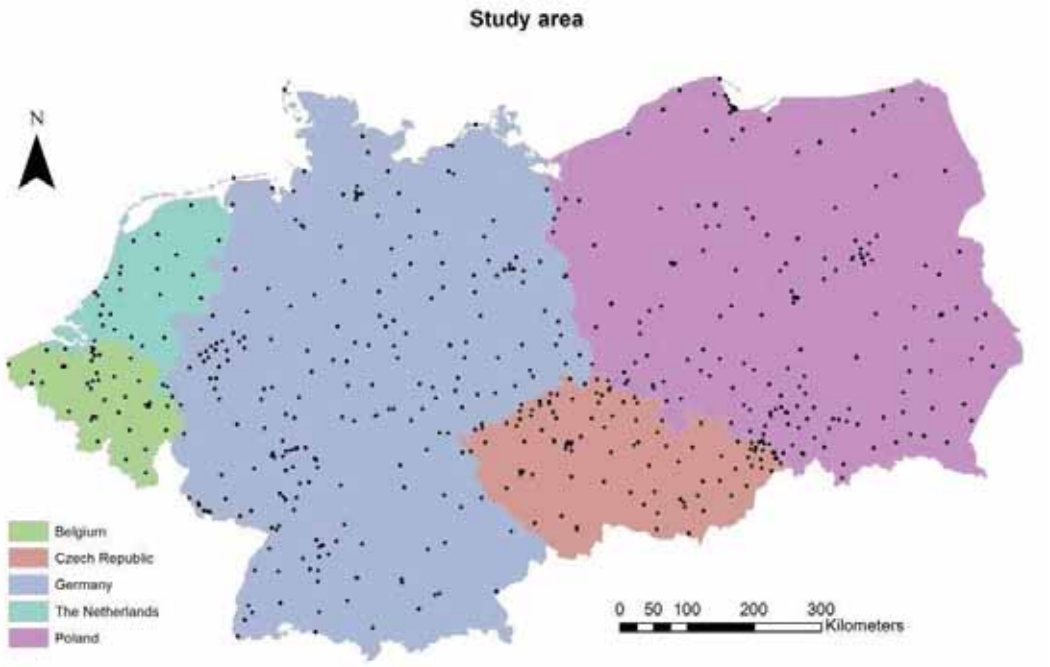


Figure 3-1 Study area and location of monitoring stations

Northern part of the area is shown to be lower elevated than southern part (Figure 3-3). According to the CORINE[3] digital elevation model (DEM), the highest elevated monitoring station is "PL0304A" with 983m and the lowest is "PL0168A" at -39m both in Poland.

[3] Coordination of information on the environment

3.2. Data description

The research handles number of spatially and temporally available data including in-situ measurements and CTM output of PM10 concentrations, and remote sensing products such as land cover map and DEM. Data descriptions of these are explained in the following sub-sections.

3.2.1. In-situ measurement of PM10 concentration

The whole European data is archived in NetCDF4 format by air quality data base (AirBase). With the help of ncdf4 library and given R codes, archived data has been extracted and reconstructed into R environment. Later, measurements in study region are extracted from the whole dataset. In-situ measurements of PM10 concentrations from 580 monitoring stations over the study area are used for the geostatistical modelling and mapping. The provided original data contained whole three years (2008, 2009 and 2010) daily-mean of PM10 concentrations. According to research interest, measurements from 2010 have been selected out of the whole dataset. Since the data is available in both space (location and its attribute) and time (daily measurements over a year), data has been reconstructed into STFDF format which is a class for spatio-temporal data with full space-time grid using *spacetime* package in R environment (Pebesma, 2012). Further explanation of the data is in the section 4.3 Data Exploratory analysis. Daily in-situ measurements are not available for all stations continuously. Missing values exist. AirBase monitoring stations have its type of area regards to the location of the stations, which are rural, suburban and urban.

3.2.2. CTM outcome

As CTM is designed for the assessment of particulate air pollutants (LOTOS-EUROS, 2011a), this model outcome is used as a covariate to improve the accuracy of modelling and as a grid for prediction as well. Along with the in-situ measurements of PM10 concentrations, CTM outcome is provided at European scale including concentrations over ocean (Figure 3-2). Study area has been subset from this dataset. Grid dataset has 100" × 140" cell size and PM10 concentration is provided in unit of kg/m^{-3}.

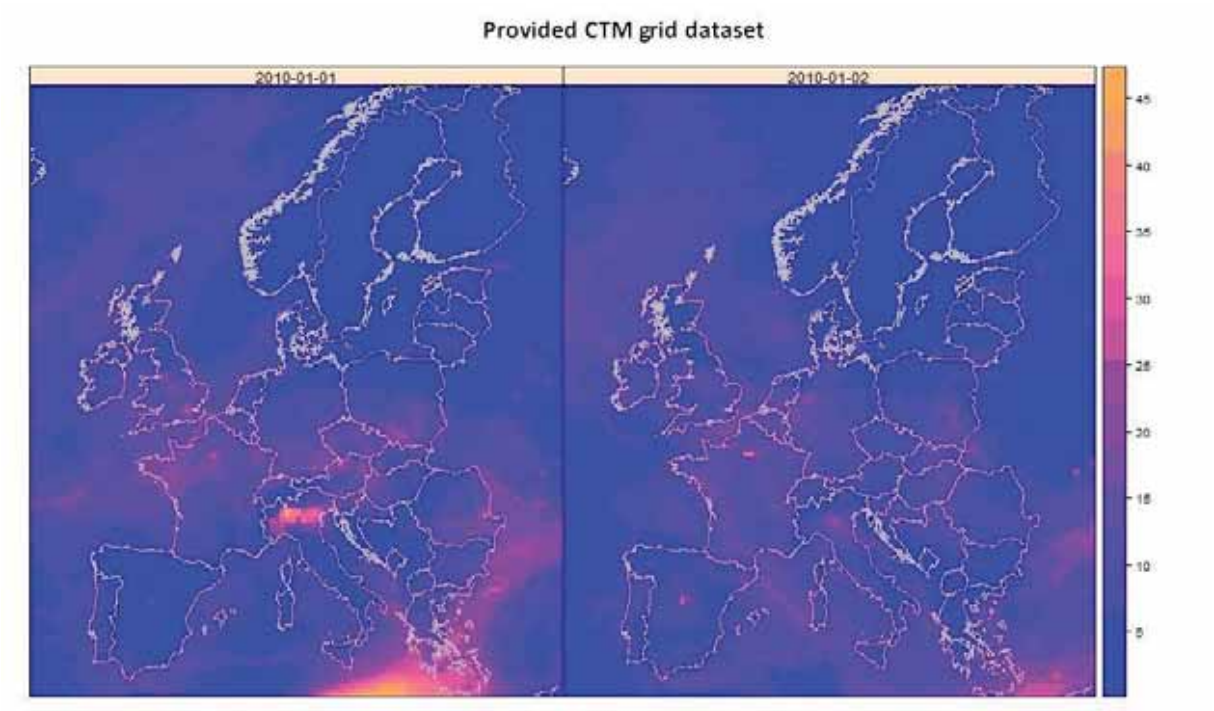


Figure 3-2 Provided grid dataset of PM10 concentration from CTM at European scale. Reference system is in geographical WGS84. Concentrations (kg/m^{-3}) are increasing from blue to orange.

3.2.3. Elevation data

European Environmental Agency (EEA) provides world digital elevation model (ETOPO5) and the data has been downloaded from EEA's website[4]. From this European DEM, study area's part has been subset (Figure 3-3). The resolution of the DEM varies from 5-minute for the ocean floors, the USA., Europe, Japan, and Australia to 1 degree in data-deficient parts of Asia, South America, northern Canada, and Africa (European Environment Agency). So approximately, the pixel size is 10 km × 10 km.

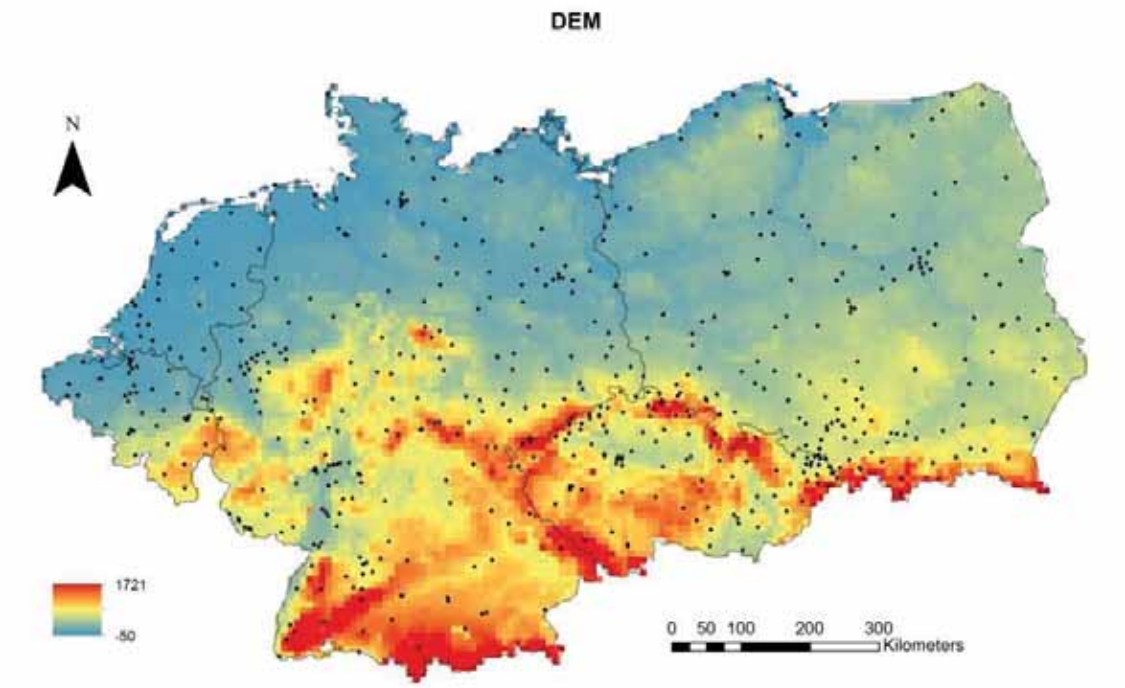


Figure 3-3 CORINE DEM over study region.

3.2.4. Land cover/use map

CORINE Land Cover 2006 raster data (Figure 3-4) has been downloaded from EEA's website[5]. This land cover product is derived from SPOT-4/5 and IRS P6 LISS III satellite images having geometric accuracy better than 100m (Büttner *et al.*, 2012). The data is provided in GeoTIFF format with 100m×100m resolution and it has three hierarchal categories. LABEL1 category (Table 3-1) is land cover classification and other two are land use information. All categories nomenclature attached in the Appendix-1. First category is chosen for the modelling because of its significant relationship with the observed PM10 concentrations than other two levels.

Table 3-1 CORINE Land cover LABEL1 category with corresponding area in square km over study region

No.	LABEL1	Area (km^2)
1	Artificial surfaces	58911.8
2	Agricultural areas	495745.8
3	Forest and semi natural areas	245256.4
4	Wetlands	3548.6
5	Water bodies	13188.0

[4] http://www.eea.europa.eu/data-and-maps/data/world-digital-elevation-model-etopo5
[5] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-2

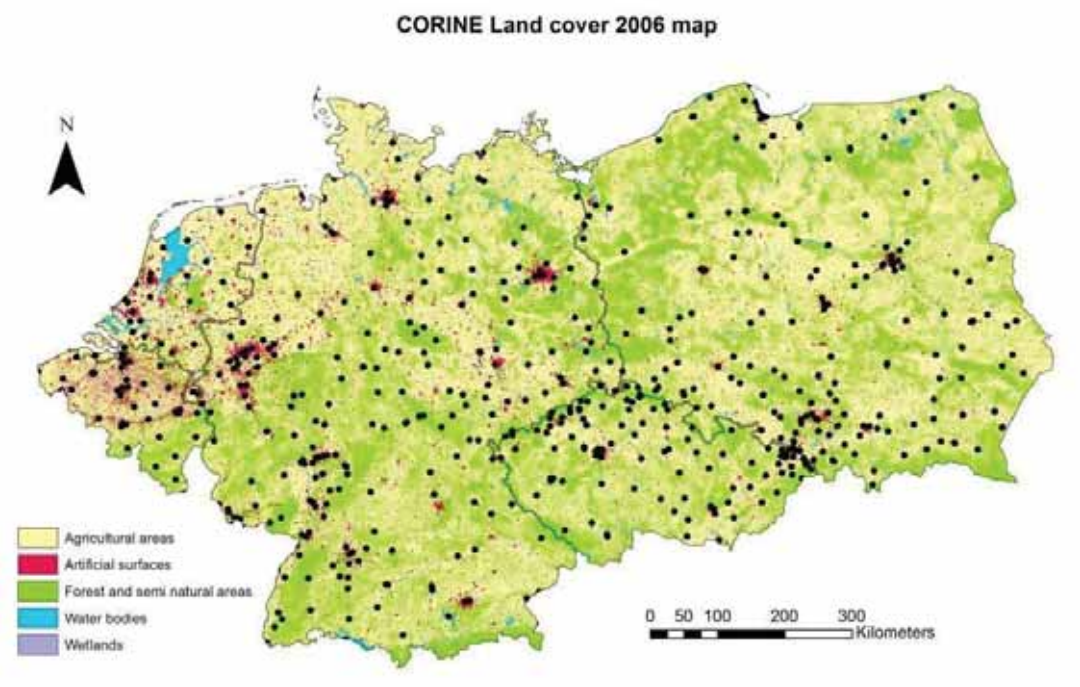


Figure 3-4 CORINE Land cover 2006 map

4. METHODOLOGY

4.1. General methodology

The overall methodology to answer the research questions is illustrated in Figure 4-1. Basically, it has three main stages including Data pre-processing and integration, Spatio-temporal modelling and Air pollution mapping.

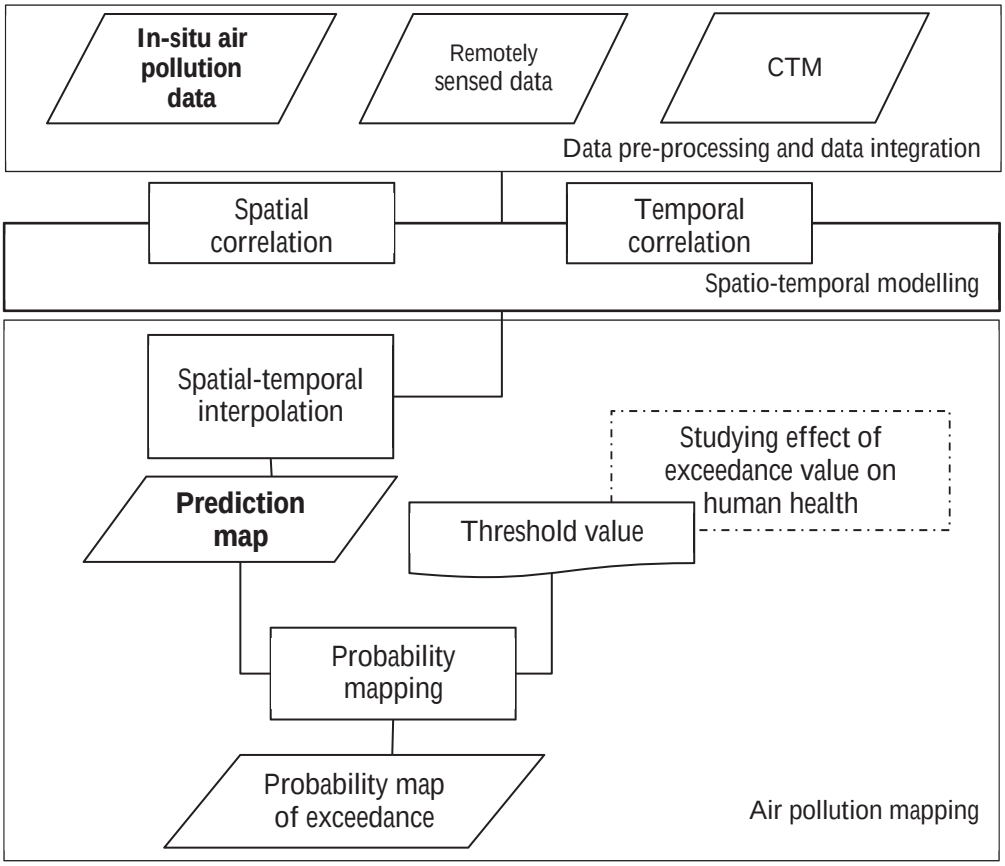


Figure 4-1 Overall methodology leads to geostatistical mapping of air pollution

4.2. Data pre-processing

4.2.1. Primary data

Prior to the exploratory data analysis (EDA), all available data are needed to be prepared in a convenient and consistent way of structure for the further analysis. First of all, required primary data, in-situ measurements, is extracted to the R environment and its geographic coordinates are transformed into ETRS 1989 LAEA which is proved to be suitable for statistical mapping (Annoni *et al.*, 2003).

After getting required data and products, EDA is applied. Since the primary data are available in space-time context, it is needed to be properly stored as one. *spacetime* package from R project team makes this possible. This is newly developed package which handles spatio-temporal data in fully gridded space-time class (Pebesma, 2012), and geostatistical methods can be applied using *gstat* package.

STFDF is a class for spatio-temporal data with full space time grid and is constructed as follows (Pebesma, 2012),

STFDF(Spatial object, Time, Data, endTime)

Where, there are *n* number of spatial locations and *m* times. The size of the dataset should be equal to product of *n*×*m*. *endTime* is there for the setting the time is true instance or interval. So the in-situ measurement is reconstructed as STFDF class having 580 spatial locations, 365 times and 211700 observations. More details of structure of this class can be found in Pebesma (2012) and its vignette. Illustration of the structure of how STFDF class handles spatio-temporal data is attached in the Appendix-2.

Missing values in the data are problematic and are needed to be taken care of carefully. Proper handling procedure is required; otherwise, it leads to wrong computation. If the monitoring station has no observation over a year, it was removed from the dataset.

4.2.2. Secondary dataset

Three types of secondary products have been used in the research. These products came from different sources and having inconsistent formats. Pre-processing is needed in order to have consistent structure, projection, and spatial resolution.

CTM outcome is extracted where in-situ measurements are available. And for the prediction, grid data of model outcome is prepared. As same as in-situ measurements, model outcome is given as NetCDF4 format, so the extraction procedure has been done and is stored as STFDF class file. This procedure is quite complex since provided CTM grid file does not have any attribute information but model outcomes and spatial locations. To assign country names attribute to the spatial locations, European boundary shapefile is used. All the steps (Figure 4-2) are done in combination of R and ArcGIS environments.

Land cover information and DEM for each monitoring stations can be derived from EEA datasets using Spatial Analyst extension in ArcGIS and imported into R. These secondary information are attached into both in-situ observations and grid data. This would allow universal kriging at unsampled locations over study area.

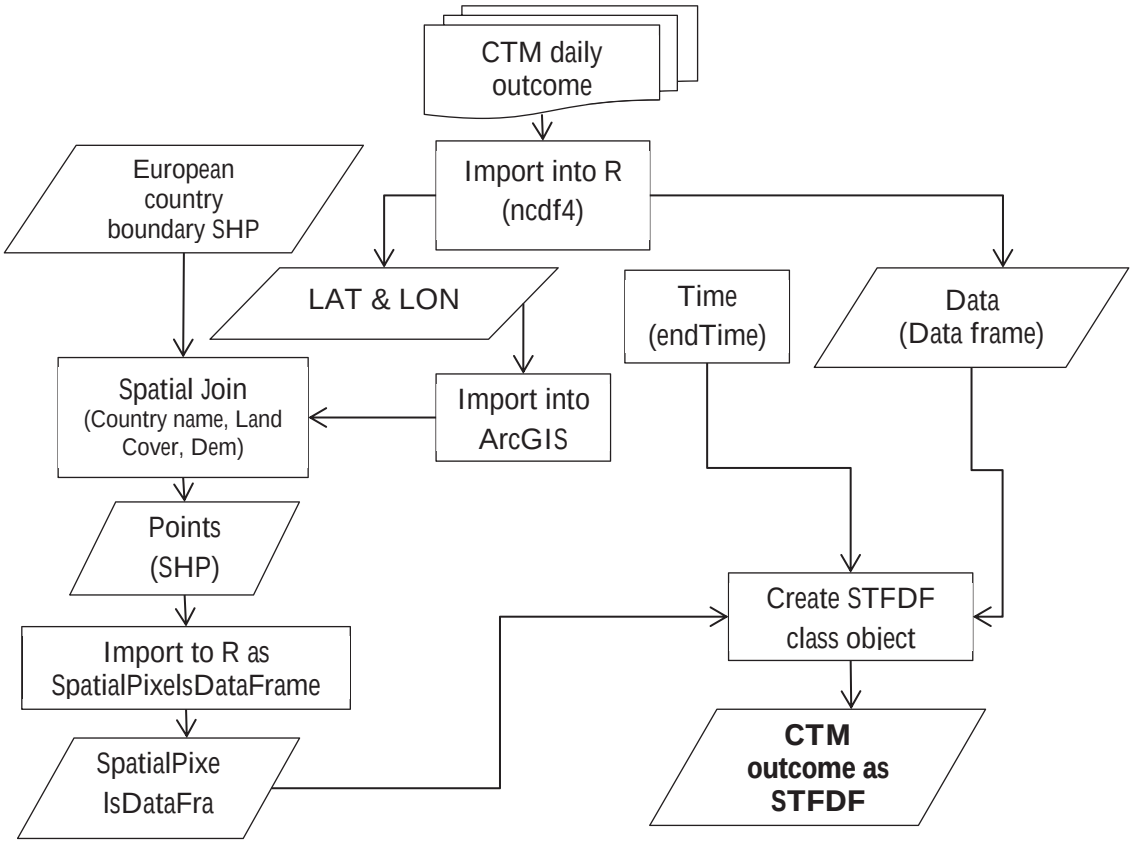


Figure 4-2 Flowchart of preparation of CTM grid file for prediction over study region.

4.3. Data Exploratory analysis

4.3.1. Data Exploration

Understanding the nature of the data is important before processing and analysing it. EDA and quantification of the data are handled by summary statistics such as histogram, boxplots, scatter plots and so forth. If the distributions of the in-situ measurements and the CTM outputs of PM10 concentration do not follow the normal distribution, logarithm transformation method will be applied in order to describe distribution with mean and variance.

In order to visualize the pattern and to give overview of the PM10 concentrations from in-situ measurements and model outcome, spatial plots are plotted in monthly-wise. In this case prepared space-time daily data need to be aggregated into monthly basis.

4.3.2. Regression analysis

Regression and multiple regression analysis are carried out in order to examine the linear relation between variables. Linear relation between response variable and covariates is formulated as follows,

$$Y = X\beta + \varepsilon$$

where, Y is response variable, X is matrix form of explanatory independent variable, β is the parameter the model.. The following regression analysis are identified,

a) In-situ measurement regressed on model outcome and DEM
b) In-situ measurement regressed on model outcome and land cover LABEL1 category

c) In-situ measurement regressed on model outcome, DEM and land cover LABEL1 category

Based on which of above gives significant result would be used in further analysis. *gstat* library is required. To simplify the procedure, monthly aggregated dataset is used for this regression analysis, and regression is carried out through whole months. Regression analysis only meant for choosing covariates.

4.4. Spatio-temporal modelling

4.4.1. Spatial correlation

To examine the spatial structure of the space-time data, spatial variogram is computed for averaging variograms over every time lag. In this case temporal component is disregarded. Assuming spatial structure of daily measurement over year is constant, mean of all time lagged semi-variance within each lag is calculated. This computation is known as pooled variogram (Gräler *et al.*, 2012) and it's defined by:

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

Where, $\gamma(h)$ is semi-variance of lag h, $Y(s_i)$ is ith $(i = 1,2 \dots N)$ observation at location s. Pooled variogram of each month is computed.

4.4.2. Temporal autocorrelation and cross correlation

To understand the temporal correlation at individual stations, autocorrelation and cross correlation are computed. Let univariate time series $\{y_1 \dots y_N\}$ is given at known locations, then the mean μ and the autocorrelation function R_t are estimated (Kitagawa, 2010) as,

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} y_N$$

$$\hat{R}_t = \frac{\hat{C}_t}{\hat{C}_0} = \frac{\frac{1}{N}\sum_{n=t+1}^{N}(y_n - \hat{\mu})(y_{n-t} - \hat{\mu})}{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{\mu})^2}$$

Where, n is number of time series y_n, y_{n-t} is the time series with time lag t, and $\hat{C}_t$ is the autocovariance function at given lag t. It's clear that at zero lag t, autocorrelation equals 1,

$$\hat{R}_0 = \frac{\hat{C}_0}{\hat{C}_0} = \frac{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{\mu})(y_n - \hat{\mu})}{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{\mu})^2} = 1$$

Among autocorrelation function, crosscorrelations are examined. Let $\{y_1(s) \dots y_N(s)\}$ be a multivariate time series at locations s, so estimates of the mean $\mu(s)$ and cross-correlation function $R_t(i,j)$ are formulated by

$$\hat{\mu}(s) = \frac{1}{N} \sum_{n=1}^{N} y_N(s)$$

$$\hat{R}_t(i,j) = \frac{\hat{C}_t(i,j)}{\sqrt{\hat{C}_0(i,i)\hat{C}_0(j,j)}} = \frac{\frac{1}{N}\sum_{n=t+1}^{N}\big(y_n(i) - \hat{\mu}(i)\big)\big(y_{n-t}(j) - \hat{\mu}(j)\big)}{\frac{1}{N}\sum_{n=1}^{N}\big(y_n(i) - \hat{\mu}(i)\big)\big(y_n(j) - \hat{\mu}(j)\big)}$$

Where, $i = 1, \dots s$ and , $j = 1, \dots s$. Detailed explanation of correlation and covariance functions of time series data can be found in Kitagawa (2010).

4.4.3. Spatio-temporal variogram and fitting model

Geostatistical approach to model the spatio-temporal air quality data for estimation is derived from models that decompose observation into spatio-temporal trend and spatio-temporal residuals (Sampson *et al.*, 2011). As the joint space-time framework defined in Kyriakidis and Journel (1999), a finite space domain R^2, and a finite time domain T will be considered. So the spatial and temporal correlations are studied in terms of factors that affect the variables in space and daily, seasonal and yearly variance of the variables respectfully.

Let Y be a spatio-temporal random process with $R^2 \times T$ domain. It could be expressed by

$$Y = \{Y(s,t): s \in R^2, t \in T\}$$

$Y(s,t)$ is the observation of PM10 concentration at time t and at location s (Cressie & Wikle, 2011). Location of each time lagged random process can be assumed as constant since each monitoring stations are located at fixed geographical location. And it is evolving through time t.
Modelling is applied at each month separately since the PM10 concentration varies on a monthly basis rather than seasonal.

Types of spatio-temporal covariance models have proposed and discussed in the literature (Kyriakidis & Journel, 1999; Cesare *et al.*, 2001; De Iaco *et al.*, 2002; Gräler *et al.*, 2012). In this research, separable model is used for analysing spatio-temporal structure of the PM10 concentration. This is one of the simple covariance models to separate independencies by adding spatial and temporal covariances (Cesare *et al.*, 2001). Separable semi-variance function is given by,

$$\gamma_{s,t}(h_s, h_t) = \gamma_s(h_s) + \gamma_t(h_t)$$

Where,

γ_s – Spatial variogram
γ_t – Temporal variogram
$\gamma_{s,t}$ – Spatio-temporal variogram
h_s – Spatial lag
h_t – Temporal lag

Above separable semi-variance equation would be rewritten as,

$$\begin{aligned}\gamma_{s,t}(h_s, h_t) &= 0.5Var[Y(s+h_s, t+h_t) - Y(s,t)] \\ &= 0.5(Var[Y(s+h_s) - Y(s)] + Var[Y(t+h_t) - Y(t)])\end{aligned}$$

4.5. Prediction map

Once spatio-temporal variogram is achieved, universal kriging applied. Due to the computational restriction, prediction is done separately for each country. However, all observations are used in separate prediction to avoid border effect. Current version of *gstat* package requires full set of STFDF data for the prediction. Hence stations that have full time values were used. Remaining stations (have at least one missing value) are used to validate the prediction. Validation is assessed using root mean square error (RMSE) and mean error (ME) which are determined by,

$$RMSE = \left[\frac{1}{N}\sum_{n=1}^{N}\left(Y - \hat{Y}\right)^2\right]^{0.5}$$

and

$$ME = \frac{1}{N}\sum_{n=1}^{N}\left(Y - \hat{Y}\right)$$

where N is number of predicted values $\hat{Y}$.

For the prediction, new grid data is prepared in SpatialPoint format; therefore prediction map would be represented as spatial points. Yet, those points represent their spatial location and attributes.

4.6. Probability map of exceedance

Probability map of exceedance is generated by using indicator geostatistics. As defined by European Commission, the limit value of daily PM10 concentrations is used in the exceedance mapping for defining the indicator. Since prediction is based on log observed PM10 concentrations, limit value was also log transformed. Goovaerts *et al.* (1997) gave a clear description about indicator kriging approach to account for probability of contamination in soil. Probability of exceedance of PM10 concentration is computed based on the method described in this paper.

Let PM10 concentration y exceed a given threshold y_c at an unsampled point s_0. Then unknown PM10 concentration $y(s_0)$ is regarded as a realization of $Y(s_0)$. Hence conditional probability can be written as, given data $y(s)$, that $Y(s_0)$ exceeds y_c (Goovaerts *et al.*, 1997):

$$Prob\{Y(s_0) > y_c \mid y(s)\} = 1 - Prob\{Y(s_0) \leq y_c \mid y(s)\}$$

For given threshold y_c, new indicator $i(s_0; y_c)$ is created. Indicator takes value 1 if concentration is less than y_c, and otherwise 0:

$$i(s_0; y_c) = \begin{cases} 1 \;\; if \; y(s) \leq y_c \\ 0 \;\; otherwise \end{cases}$$

And based on the created indicator, indicator kriging has been computed. Probability threshold p is defined by counting how many observed concentrations exceeded given threshold value.

4.7. Data post processing

To ensure the normal distribution of the data, dataset are transformed by logarithm, and kriging predictor is applied on the residuals form those transformed data. After kriging, back transformation is used. Back transformation has been done using formulation explained in Denby *et al.* (2008). Expected back transformed concentration is computed by

$$\hat{Y}(s) = exp\left(Y(s) + \frac{\sigma(s)^2}{2}\right)$$

Where $Y(s)$denotes predicted concentration from log-transformed data at location s, $\sigma^2(s)$ is the kriging variance at each predicted location s. The back transformed kriging variance can be expressed by

$$var\left(\hat{Y}(s)\right) = (exp(\sigma(s)^2) - 1) \cdot exp(2Y(s) + \sigma(s)^2)$$

4.8. Used software and package

The following software and packages are used in the processing, analysing and visualization purposes Table 4-1. Used and implemented code are attached in the Appendix-4.

Table 4-1 Description of used software packages

Software	Package/Extension	Description/usage
R	colorspace	Color space manupilations
	GISTools	GIS capabilities in R
	gplots	For plotting data
	gstat	Spatial and spatio-temporal modelling and prediction
	maps	Support visualizations, graphics
	maptools	Handling spatial point, line and point objects
	ncdf4	Read NetCDF4 file and import in to R environment
	PBSMapping	Reading and visualizing spatial point and line objects
	RColorBrewer	For colour ramp
	rgdal	For handling SHP files and reference system transformation
	shapefiles	Import export SHP files
	sp	Methods for handling spatial objects
	spacetime	Create STFDF object class
	xts/zoo	Create and construct time class
	lattice	Lattice graphics
ArcGIS	Spatial Analyst	For Extract Raster cell value to the overlaid point data
		Subset study area region
MS Excel		Handling text files

5. RESULTS AND ANALYSIS

5.1. Data pre-processing

Using the *spacetime* package, given dataset of in-situ measurements of PM10 were reformatted into the STFDF class in R. Given data contains values are lower or equal to zero. PM10 concentration cannot be observed negatively. Within whole dataset of 2010, 63 values detected, so those values are set as missing values. From these full dataset of Europe, study areas' observations and CTM outcome were extracted. Monitoring stations that have no records over whole year are neglected from the dataset. Null values exist in the dataset and are problematic in regard to the modelling and prediction procedure.

CORINE Land cover map over study region is prepared in ArcGIS. Available attribute information is examined and joined to the map using Spatial Join tool of ArcGIS for the analysis. CORINE DEM data is prepared in a same way of land cover map. Study area region was subset.

All available dataset were projected into same reference system, ETRS 1989 LAEA.

5.2. Data exploratory analysis

5.2.1. Data exploration

5.2.1.1. Primary data

Over the study area, total 580 stations (Figure 5-1) have recorded PM10 concentration within the year of 2010. These in-situ measurements are daily mean of PM10 concentration. In total, 211700 measurements including null observations of year 2010 are stored. Not all monitoring stations have continuous measurements.

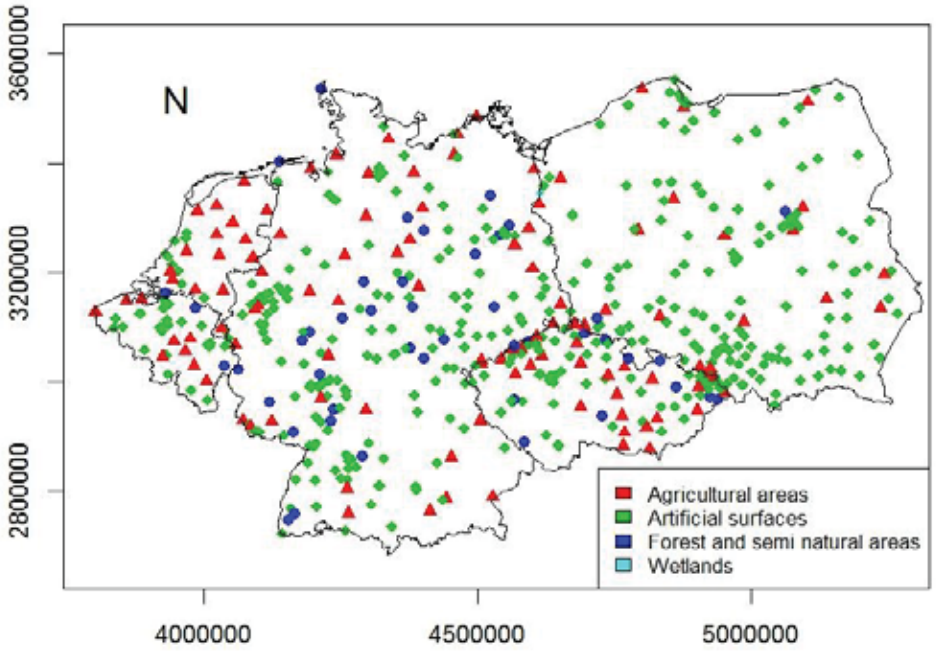


Figure 5-1 Monitoring stations over the study area. Stations are distinguished by their corresponding CORINE Land cover LABEL1 category

Number of measurements and their summary statistics by CORINE land cover LABEL1 category is shown in Table 5-1. About 72% of all stations over study region fall into "Artificial surfaces". About 20% and 8% are in Agricultural Areas and Forest and semi natural areas, respectively. Only two stations are in "Water bodies" category which is the places where water bodies surrounded by urban areas. However, with regards to temporal resolution, these two stations have daily measurements with 38 days of null values (Figure 5-13). No stations are in "Wetlands" category. And within 580 stations, 144, 122, and 314 stations fall into rural, suburban and urban respectively.

Table 5-1 Number of measurements and their summary statistics for each country

	Count	Min	Median	Mean	Max	NA's
Agricultural areas	114	0.30	18.00	22.60	299.50	3929
Artificial surfaces	419	0.20	21.29	27.51	455.00	15026
Forest and semi natural areas	45	0.32	13.22	16.30	256.00	1546
Water bodies	2	3.70	23.14	25.43	155.50	38
Wetlands	0	-	-	-	-	-
Total	**580**	**0.20**	**20.00**	**25.66**	**455.00**	**20539**

In order to understand the nature of the data, histograms and normal QQ plot were plotted. These plots show that data is positively skewed (Figure 5-2). To achieve the normality, logarithmic data transformation has been applied (Figure 5-3).

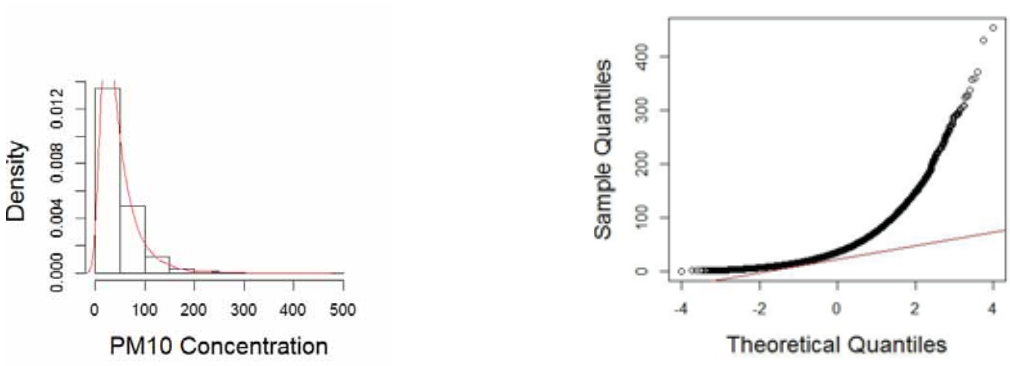


Figure 5-2 Histogram and normal QQ plot of the in-situ measurements in January, 2010

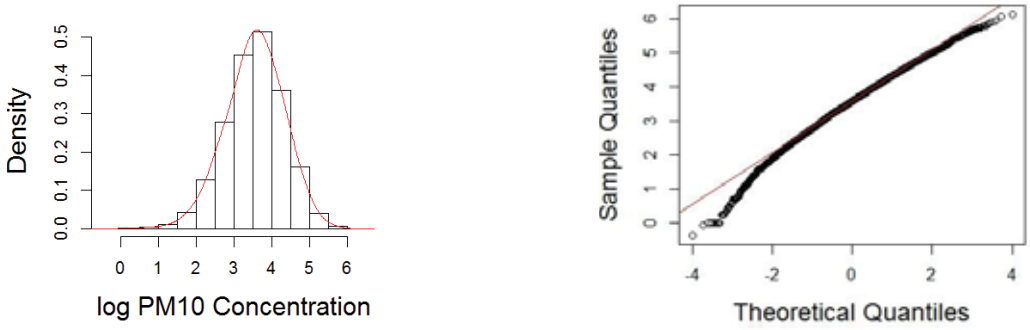


Figure 5-3 Histogram and normal QQ plot of the log transformed in-situ measurements in January, 2010

After logarithm transformation has been applied, the distribution of the data followed a log normal distribution (Figure 5-3). Histograms are displayed as probability densities.

Spatial plot of PM10 concentrations are shown in Figure 5-4 and concentrations varies gradually. It shows that mean of daily observations are not constant over a month. PM10 concentrations are high on 1st, 13th-15th and lower in end of the month.

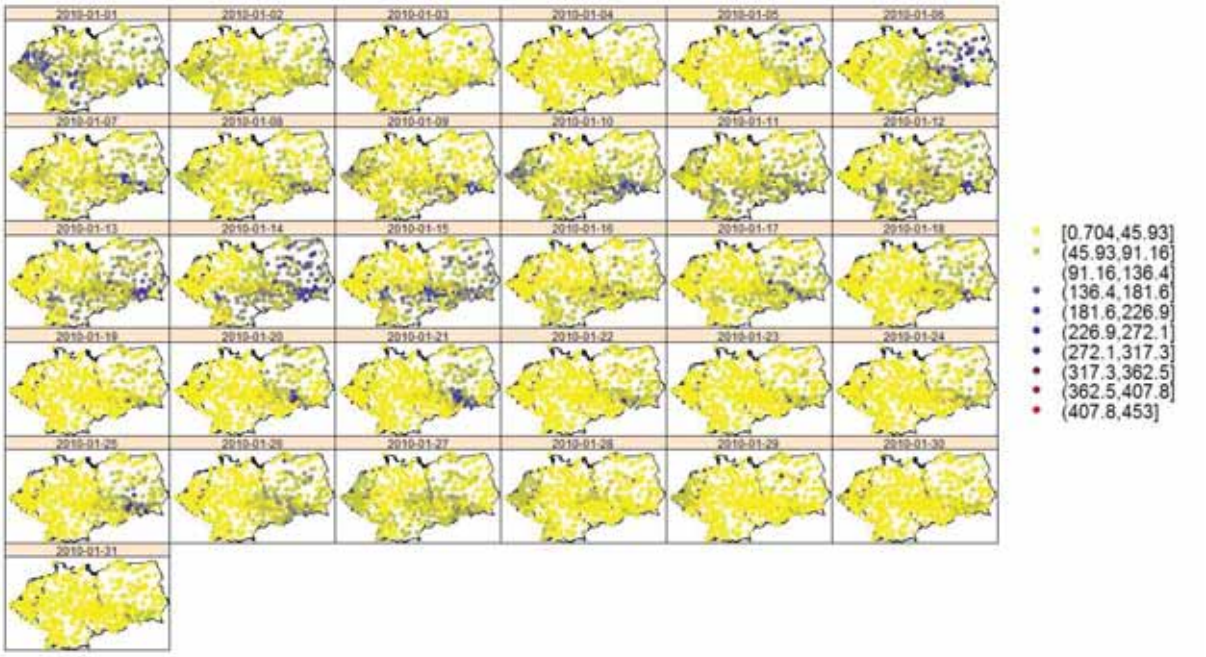


Figure 5-4 Daily observed PM10 concentrations in January, 2010. Lower to higher observations are symbolized by colour composition between yellow, blue and red.

Daily measurements are averaged to monthly base and are plotted to see if there is any pattern over a year. Monthly aggregation is applied only because overall visualization purposes and multiple regression for choosing covariates. Highest value (455$\mu g/m^3$) is recorded in February in Poland. From May to August, PM10 concentrations are relatively lower, and increase from September till April. Interestingly, concentration in April suddenly gets higher than March and from May they became lower again (Figure 5-5).

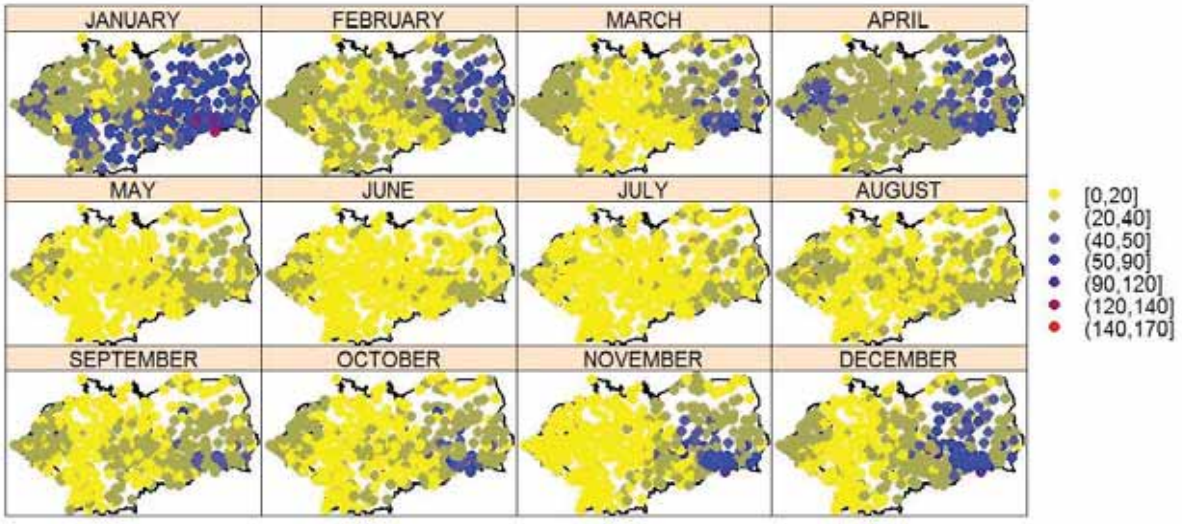


Figure 5-5 Monthly averaged observed PM10 concentrations, 2010

In order to visualize the PM10 concentration's variation over the year, daily measurements at each station were plotted (Figure 5-6). As well as shown in Figure 5-5, in the plots PM10 concentrations became

relatively higher in April in all stations (Figure 5-6). In-situ measurements of PM10 concentrations showed that PM10 concentrations are higher during the cold season and become lower during the warm season.

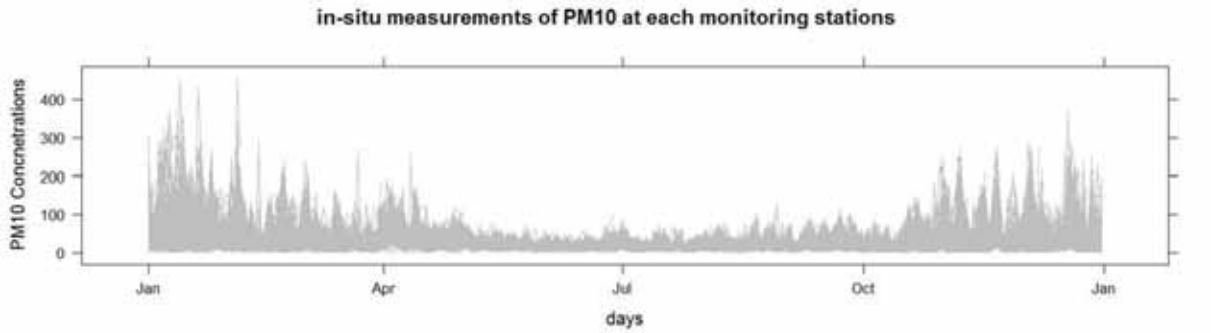


Figure 5-6 Daily measurements of observed PM10 concentrations at each monitoring stations over a year, 2010

5.2.1.2. Secondary information

CTM outcome at monitoring stations

CTM outcome covers whole study area at each station where in-situ measurement is available. Histogram and normal QQ plot of CTM outcome has been plotted in Figure 5-7 and Figure 5-8. Its summary statistics are shown in Table 5-2. This table only represents CTM outcome at each monitoring stations where in-situ observation are available.

Table 5-2 Summary statistics of CTM outcome by CORINE Land cover LABEL1 category at monitoring stations

	Count	Min	Median	Mean	Max	NA's
Agricultural areas	114	0.92	11.06	12.44	75.92	0
Artificial surfaces	419	0.68	10.93	12.40	75.92	0
Forest and semi natural areas	45	1.45	10.29	11.66	67.91	0
Water bodies	2	1.99	12.41	13.39	55.91	0
Wetlands	0	-	-	-	-	-
Total	**580**	**0.68**	**10.90**	**12.36**	**75.92**	**0**

Model outcome at each location over a year ranges between 0.68$\mu g/m^3$ and 75.92$\mu g/m^3$. The highest values occur in April in area of Belgium and The Netherlands. From the table, there is no clear difference between observations.

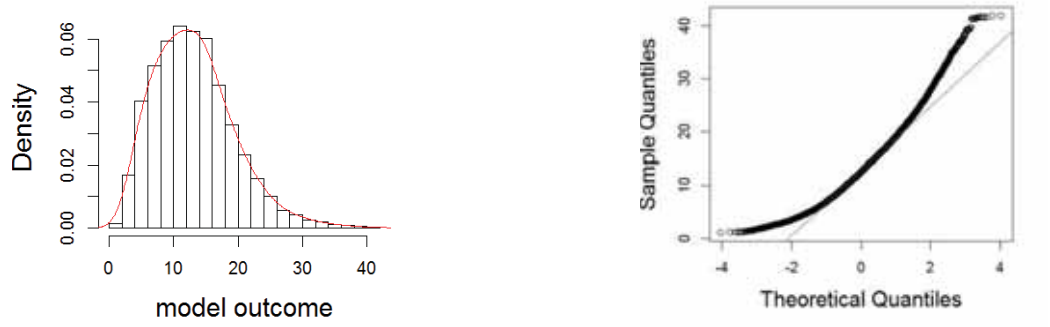


Figure 5-7 Histogram and normal QQ plot of CTM outcome at monitoring stations

Figure 5-8 Histogram and normal QQ plot of log transformed CTM outcome at monitoring stations

To see the variation over a year, monthly average outcomes plotted over the study area (Figure 5-9).

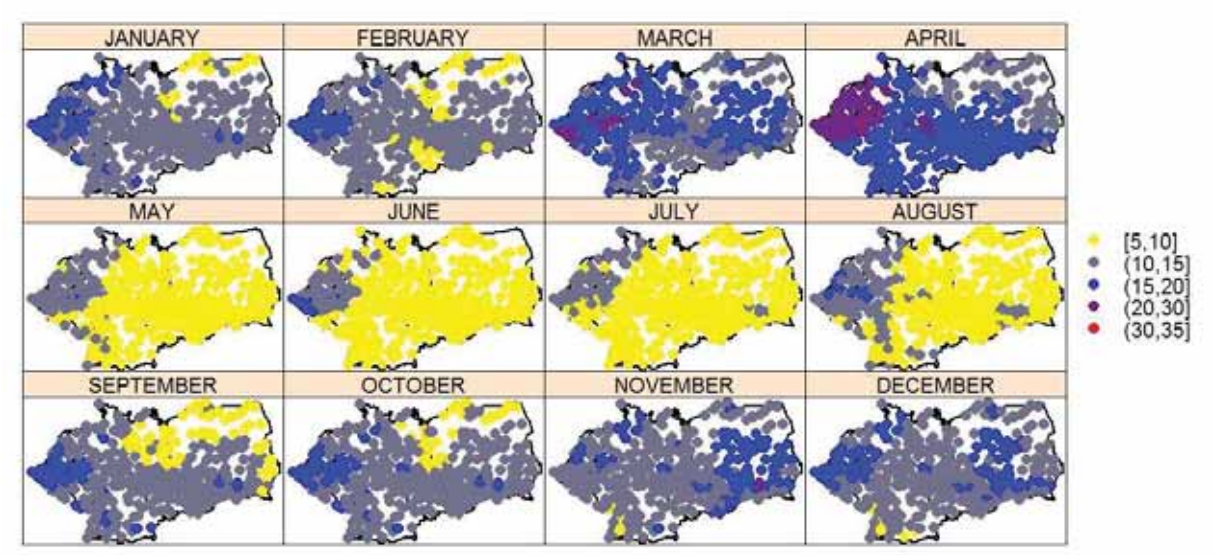


Figure 5-9 Monthly average CTM outcome at monitoring stations, 2010

Time versus measurement plot (Figure 5-10) shows how outcome results over a year vary. At each location, values in April are higher than other. To relate with the PM10 in-situ measurements, in April, either PM10 concentration becomes higher. This emergent phenomena draw attention. The reason could be due to either the meteorological condition, huge smoke in the atmosphere or instrument error or there could be bias in the measurements. The cause of this phenomenon has been investigated. Meteorological condition has been examined over the study area, however nothing attention-grabbing found out. But between 20th March and 20th April of 2010, the Eyjafjallajökull volcano which is located in the northern part of European Iceland country, Iceland, has been erupted ("The 2010 Eruptions of Eyjafjallajökull," 2011). According to the news (The New York Times, 2010), the volcano had two active phases. First phase started in the late evening of March 20th and ended on April 12th in 2010. Within this time, olivine-basaltic andesite lavas were flowing into the air. After two days gap, second phase started intensively and it lasted from April 14th to 20th in 2010. Clouds of ash were exploded up to several kilometres in the atmosphere and causes air traffic in northern Europe (The New York Times, 2010). This explains the higher PM10 concentrations in April. The two phases of the eruption are being indicated by the two peaks in the both Figure 5-6 and Figure 5-10.

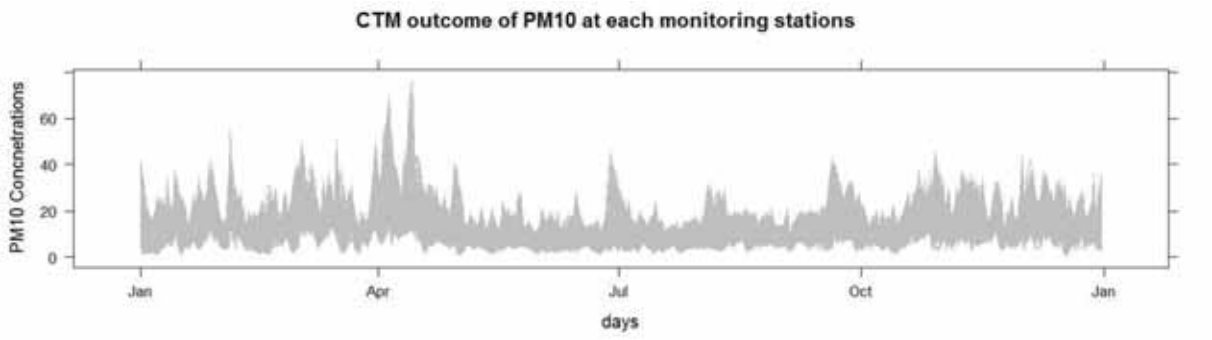


Figure 5-10 Daily CTM outcome of PM10 concentrations at each monitoring stations over a year, 2010

Grid data of CTM outcome

Grid data of CTM outcome are available in study region and its summary statistics are presented in Table 5-3. It has outcome values range between minimum 0.59μg/m^3 and maximum 58.98μg/m^3. Its spatial plot in January is illustrated in Figure 5-11. Higher PM concentrations were occurred on 6th, 7th and 19th in January.

Table 5-3 Summary statistics of CTM grid data by country

	Counts	Min	Median	Mean	Max	NA's
Belgium	33	2.66	13.20	14.55	53.89	-
The Netherlands	38	2.48	13.04	14.22	47.64	-
Germany	369	0.93	10.49	11.50	58.98	-
Poland	322	0.59	8.92	9.92	54.09	-
Czech Republic	81	1.52	9.61	10.56	39.65	-
Total	**843**	**0.58**	**10.02**	**11.05**	**58.98**	**-**

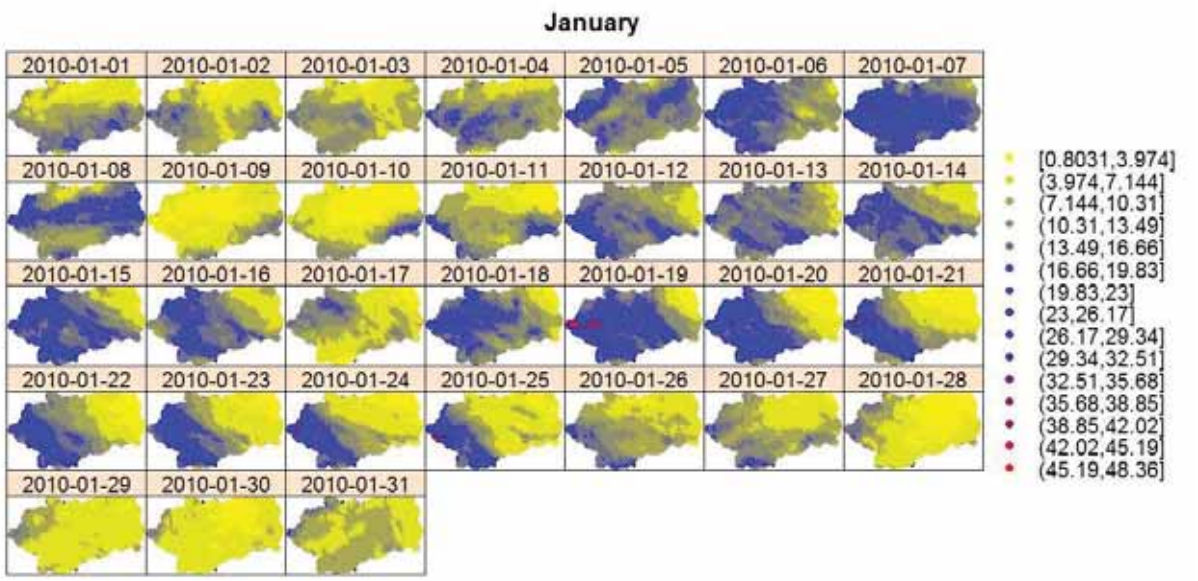


Figure 5-11 Spatial plot of CTM grid product of PM10 concentrations in January, 2010

CORINE DEM

Whether there is linear relationship between PM10 concentration and elevation, CORINE DEM is analysed (Table 5-4). The distribution of the elevation values could not be assumed normal because of

irregular variation of the surface. Due to lower elevations than mean sea level, data transformation produces null values. Therefore, in further analysis raw DEM values is used.

Table 5-4 Summary statistics of CORINE DEM

	Min	Median	Mean	Max
Elevation	-39.00	209.00	230.10	983.00

Figure 5-12 illustrates the histogram and the normal QQ plot of DEM.

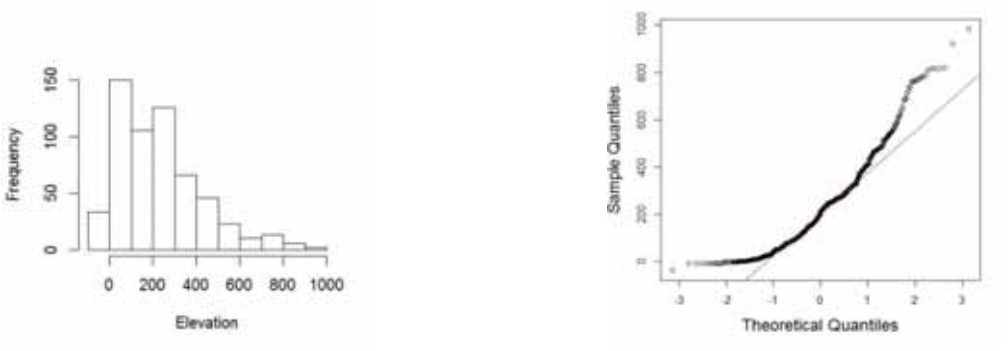


Figure 5-12 Histogram and normal QQ plot of DEM of monitoring stations

5.2.2. Regression analysis

Regression analysis has been done as following ways, and although, only result from regression in January is shown, all months' regression results are considered for the final analysis.

a) In-situ measurement regressed on CTM outcome and DEM
b) In-situ measurement regressed on CTM outcome and land cover LABEL1 category
c) In-situ measurement regressed on CTM outcome, DEM, and land cover LABEL1 category

Table 5-5 Partial results of regression analysis, Significance codes in R: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; and 0.1 ' ' 1

	a)				b)			c)			
	signf.code (DEM)	Adj.R	F-stat	Res. SE	Adj.R	F-stat	Res. SE	signf.code (DEM)	Adj.R	F-stat	Res. SE
January		0.49	0.93	0.00	0.43	46.50	0.24		0.43	37.70	0.24
February	***	0.46	26.37	0.08	0.42	41.39	0.22		0.41	40.43	0.26
March	***	0.40	32.87	0.10	0.37	45.64	0.24	**	0.36	38.34	0.25
April	***	0.29	21.65	0.07	0.28	20.57	0.12	***	0.28	19.70	0.14
May	***	0.26	10.06	0.03	0.25	19.89	0.12	*	0.25	17.02	0.12
June	**	0.25	64.17	0.18	0.24	54.64	0.28		0.24	44.38	0.28
July	*	0.26	57.50	0.17	0.25	42.73	0.23		0.25	34.36	0.23
August	**	0.25	28.51	0.09	0.24	27.27	0.16		0.24	22.13	0.16
September		0.32	17.95	0.06	0.30	26.64	0.15		0.30	21.73	0.15
October	***	0.36	16.96	0.05	0.33	40.25	0.22	*	0.33	33.41	0.22
November	**	0.42	147.20	0.34	0.39	108.30	0.43	***	0.39	93.59	0.45
December	***	0.48	81.74	0.22	0.45	65.35	0.32	***	0.44	60.62	0.35

Regression results from (a), PM10 in-situ measurements regressed on CORINE DEM, gave significant modelling result despite of insignificance of DEM from January and September. And for the (c), significance lever for DEM became even lower, only 6 months showed significant result out of 12. Results

from (b) showed, Land Cover categories are significant for the modelling, however, "Water Bodies" category provided insignificance in each month. This is could be due to only 2 monitoring stations fall in this category. Even in the analysis of (c), this category gives insignificant result. Values from this category are relatively higher (Figure 5-13) and they are located in water area surrounded by urban/settlement. So it is decided that replace these two stations' category into "Artificial surfaces" and DEM is not used as a covariate. LABEL1 information from CORINE Land cover product and CTM outcome are chosen as covariates.

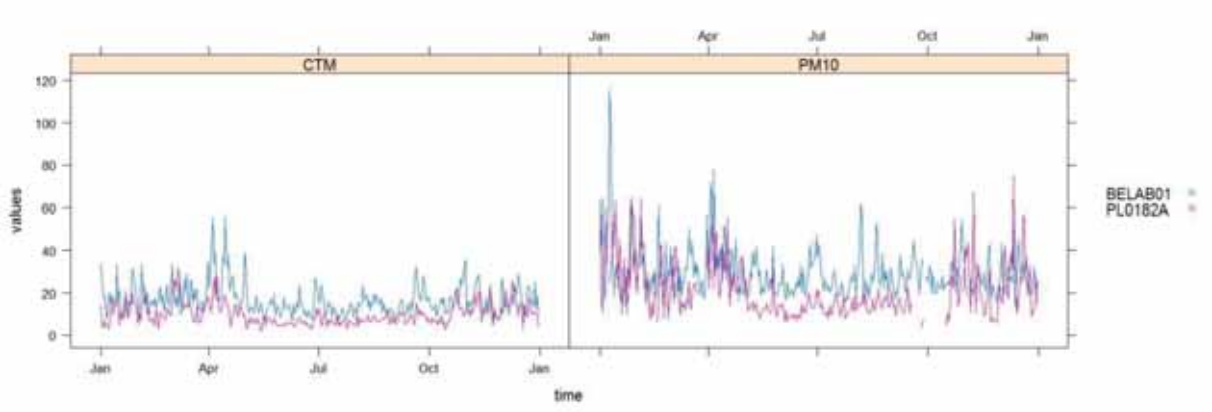


Figure 5-13 CTM outcome and in-situ measurements of two stations that are in "Water body" land cover category.

5.3. Spatio-temporal modelling

5.3.1. Spatial correlation

Sample variogram of sole log PM concentrations (a) and sample variogram of log PM10 concentrations regressed on log CTM model outcome and CORINE Land cover.(b) for each month are calculated at 580 monitoring stations. Exponential variogram model is fitted to each sample variogram with min and max sums of square error (SSE) of $11.6 \cdot 10^{-3}$ and $0.01 \cdot 10^{-3}$ in (a) and of $0.1 \cdot 10^{-3}$ and $4.9 \cdot 10^{-3}$ in (b) respectively. Sample variograms and their exponential variogram models are shown in Figure 5-14.

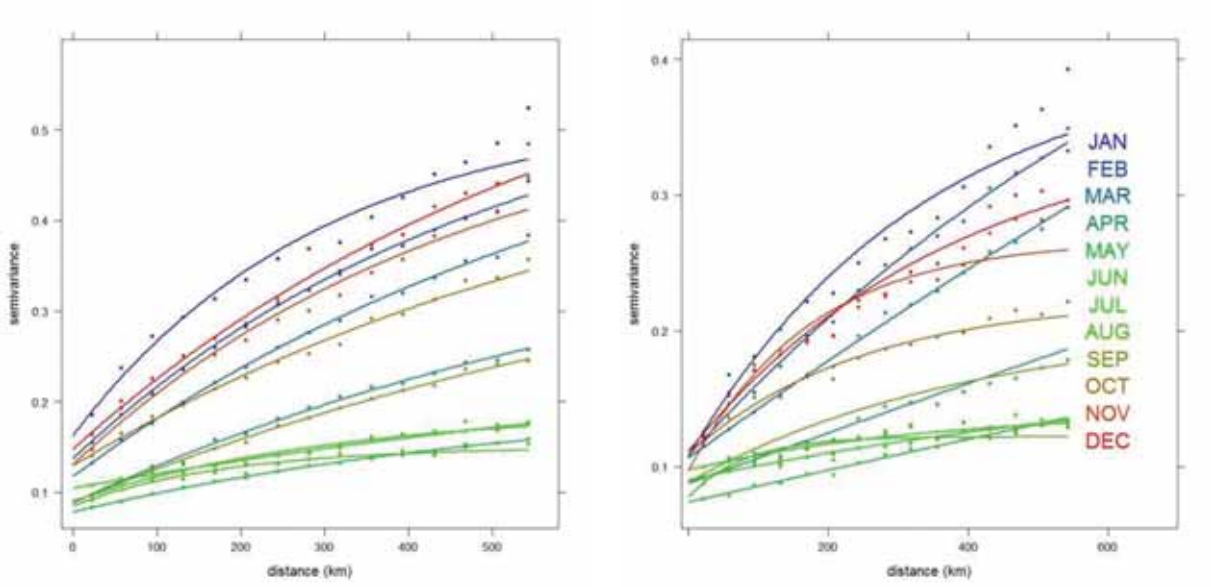


Figure 5-14 Spatial variogram, on the left log in-situ PM10 concentrations, and on the right covariates are used variogram and their fitted exponential model.

Estimated parameters of exponential model for each month are listed in Table 5-6. Depending on the variability of the observation of PM10 concentrations over months, variogram parameters vary.

From the result of sample variograms and their fitted exponential model, nugget values range between 0.08-0.16 in model (a) and 0.07-0.11 in model (b). Nuggets are measurement of non-spatial variability of the observations, and it is decreased in variogram (b) compared to (a) (Figure 5-15). It is proved that adding secondary information to the model is improved the model by decreasing the non-measurement error in the observed data.

Table 5-6 Fitted exponential model parameters

	nugget		Partial sill		Range (km)		SSE (10^{-3})	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
January	0.16	0.11	0.36	0.30	296	343	11.60	4.9
February	0.14	0.11	0.44	0.51	507	911	2.26	1.0
March	0.12	0.11	0.48	1.29	704	3505	0.52	0.2
April	0.09	0.09	0.35	1.73	798	9201	1.19	0.8
May	0.08	0.07	0.13	0.88	584	7395	0.01	0.1
June	0.09	0.09	0.10	0.09	321	780	1.49	0.3
July	0.10	0.10	0.18	0.04	1108	339	0.63	0.4
August	0.09	0.08	0.07	0.04	180	79	1.25	0.8
September	0.09	0.09	0.36	0.12	927	406	0.45	0.5
October	0.13	0.11	0.41	0.11	725	246	1.74	0.8
November	0.13	0.10	0.41	0.17	475	173	5.72	3.4
December	0.15	0.11	0.53	0.24	631	382	4.15	3.3

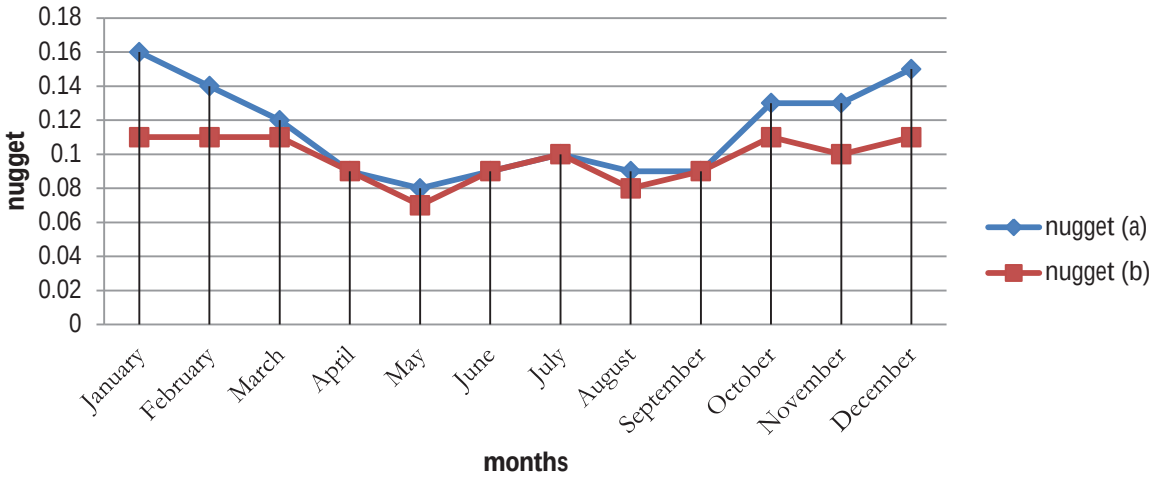


Figure 5-15 Variation of nuggets from variogram (a) and (b)

5.3.2. Temporal correlation

Temporal correlation has been examined through autocorrelation and cross-correlation function at each monitoring location. Results of randomly chosen four stations in Czech Republic are provided here. Daily PM10 concentrations over a year are plotted in Figure 5-16, and their spatial distances to each other provided in Table 5-7. CZ0BBNE and CZ0BBNY stations are located closely to each other. Others are separated by distance more than 120m.

Table 5-7 Spatial Distances (in meters) between 4 monitoring stations in Czech Republic

	CZ0BBNE	CZ0BBNY	CZ0CCBA	CZ0TCEL
CZ0BBNE	0	13072	160640	129850
CZ0BBNY	-	0	164193	128439
CZ0CCBA	-	-	0	290272
CZ0TCEL	-	-	-	0

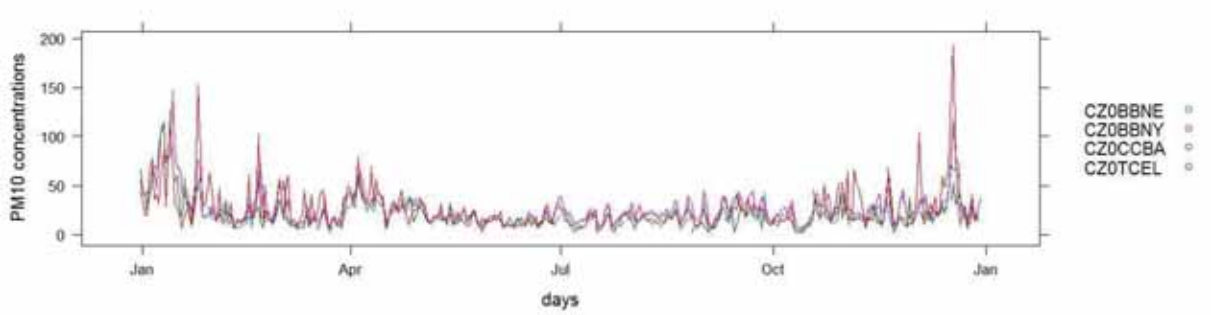


Figure 5-16 PM10 concentrations over a year at CZ0BBNE, CZ0BBNY, CZ0CCBA and CZ0TCEL monitoring stations

Autocorrelations at these stations are above confidence level at minimum four days-time lag. Although the resulting plot shows (Figure 5-17) that asymmetry of the cross-correlations is not really strong, it forms similar to autocorrelation function result. Cross-correlation of stations which are closer to each other is higher than that of stations that are far from each other. Hence the spatial distance played role in temporal correlation.

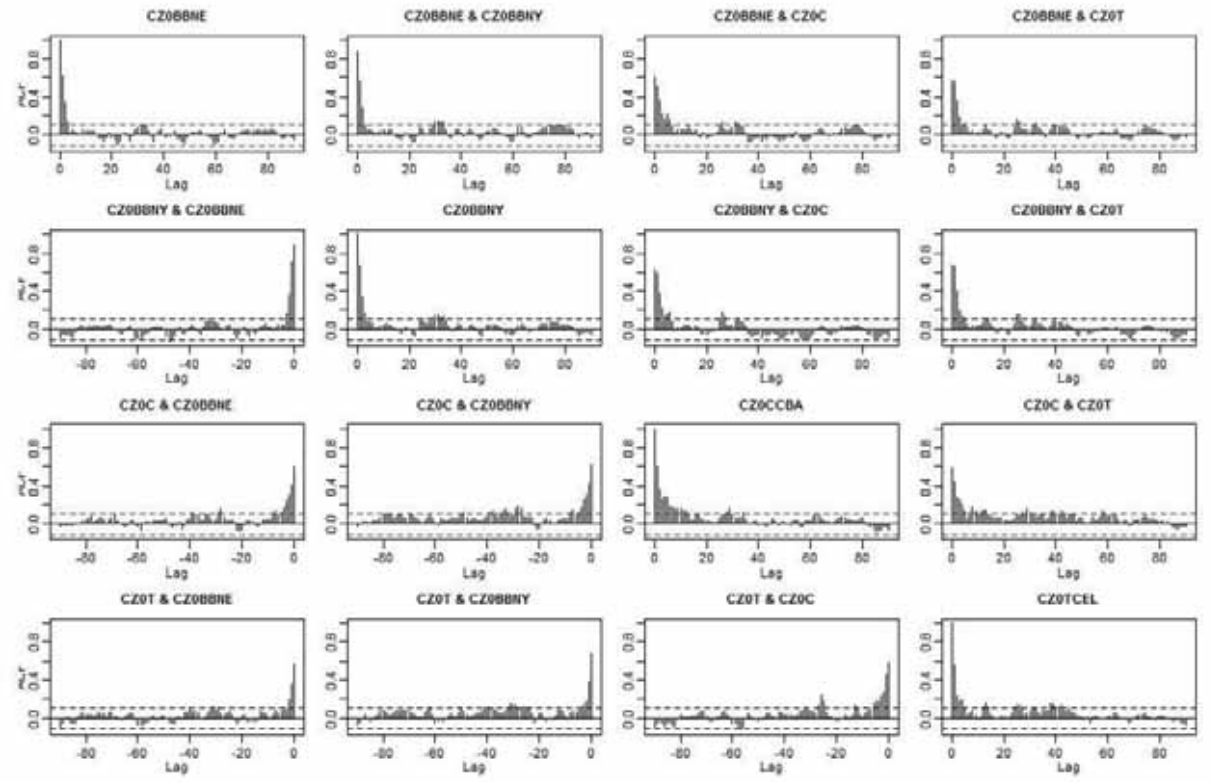


Figure 5-17 Temporal autocorrelation and cross correlation between CZ0BBNE, CZ0BBNY, CZ0CCBA and CZ0TCEL monitoring stations in Czech Republic. Time lag equals a day.

5.3.3. Spatio-temporal variogram and fitting model

Spatio-temporal variograms were calculated at each month separately. As defined in the methodology section, separable spatio-temporal variogram was calculated and exponential model was fitted at each month separately. Wireframe plot of spatio-temporal variogram, fitted model and sample variogram of January are illustrated in Figure 5-18 and Figure 5-20 respectively. Fitted model map and sample variogram maps are shown in Figure 5-19. For the spatial process, the model parameters: nugget, sill and range are estimated as 0.42, 0.58 and 300km respectively. For the temporal process, they are estimated as 0.29, 0.71 and 14days respectively. The result of the spatio-temporal variogram shows that as time lag increases, variogram increases and variograms reached the sill at the range, meaning that spatial and temporal correlation are decreased.

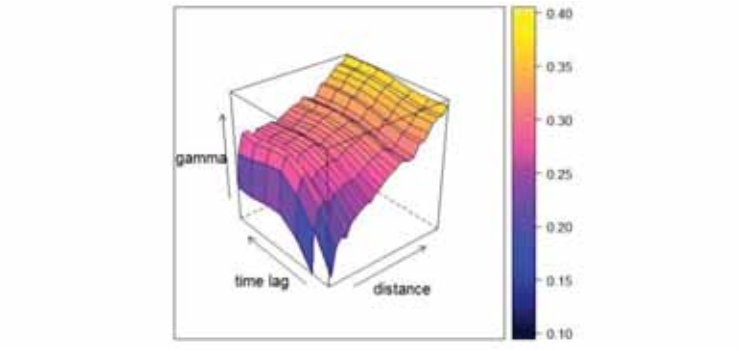


Figure 5-18 Wireframe plot of spatio-temporal variogram. Time lag in days

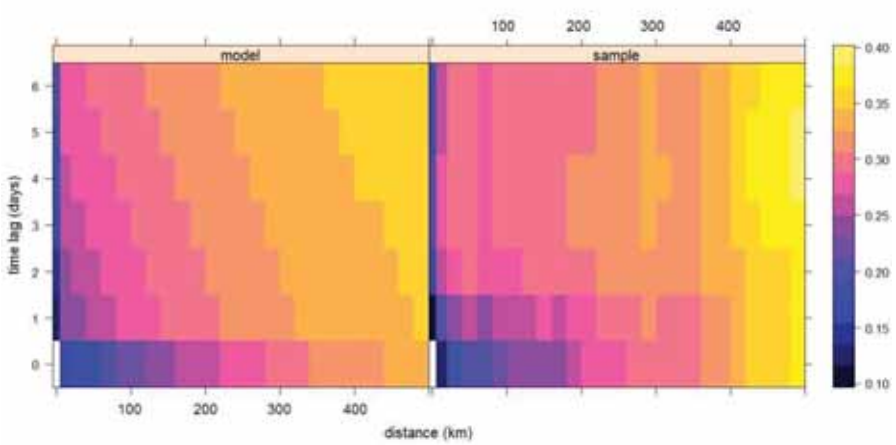


Figure 5-19 Fitted separable model (on the left) and sample spatio-temporal variogram map (on the right)

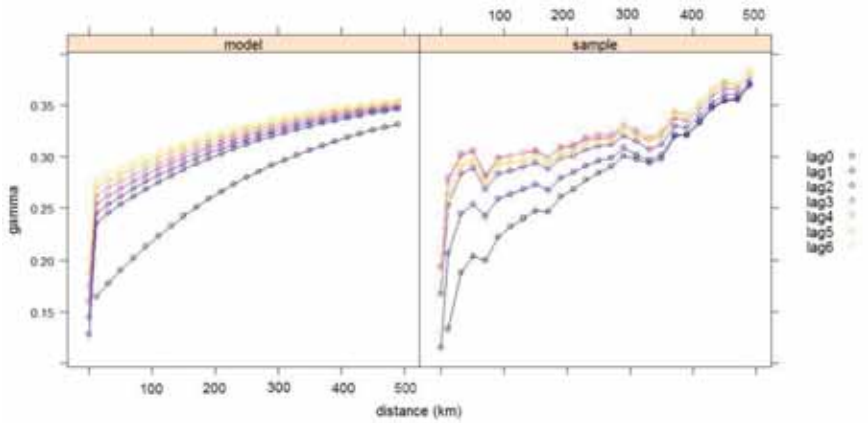


Figure 5-20 Fitted separable model (on the left) and sample spatio-temporal variogram (on the right)

5.4. Prediction map

Prediction map is computed based fully gridded space-time data at unsampled locations where CTM grid outcome is available. Due to the size of the prediction grid, computational progress is slow and requires much memory size. Hence spatio-temporal kriging has been done in each country separately and combined together afterwards. However, all observations are used for each separate kriging prediction. Prediction map of air pollution over study region in January is illustrated in the Figure 5-21. Figure 5-22 shows prediction variance. To relate the prediction variance with predicted value, variance is higher where PM10 concentration is higher.

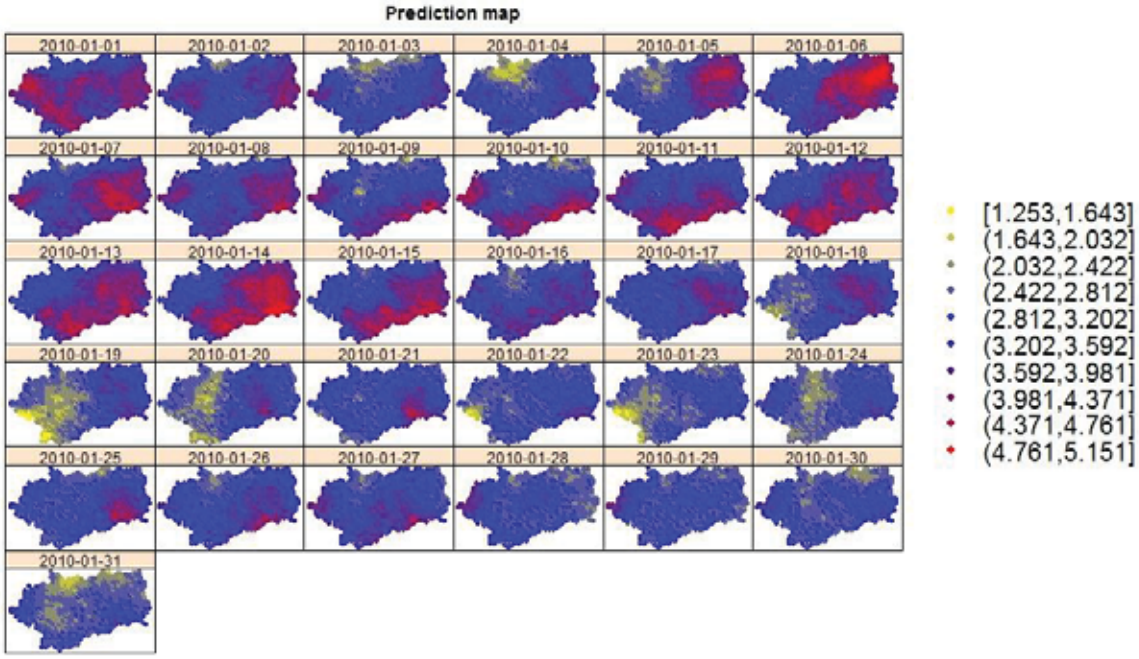


Figure 5-21 Prediction map of log transformed PM10 concentrations in January, 2010

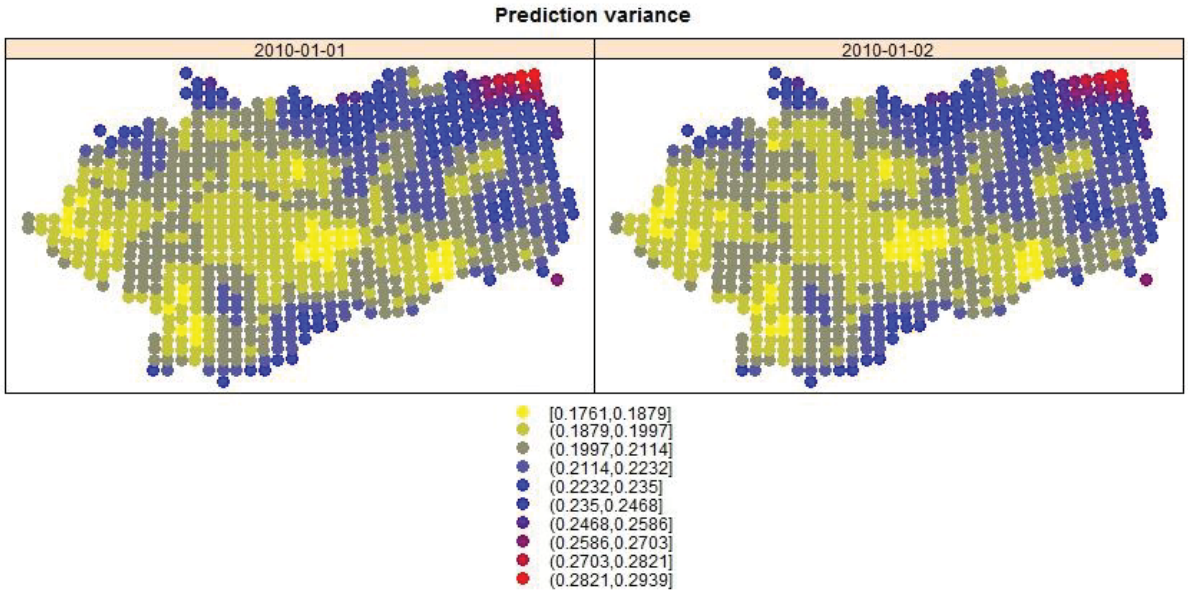


Figure 5-22 Prediction Variance of log transformed PM10 concentrations, January 1st and 2nd, 2010

As fully gridded data were used in the prediction, remaining data which contains missing values are used to assess the prediction. In January, 334 stations are used for prediction and 236 stations are used for validation. Spatio-temporal kriging is applied those 236 stations, and its predicted value is compared to observed ones at those stations. Numbers of stations used in the prediction and in the validation are shown in Appendix-3. Computed RMSE and ME are 0.48 and 0.01 respectively in January. And adjusted R^2 of linear model for predicted versus observed is computed as 0.63. Scatter plot of the result is shown in the Figure 5-23. Linear model for predicted versus observed showed reasonable fit of the relationship between them.

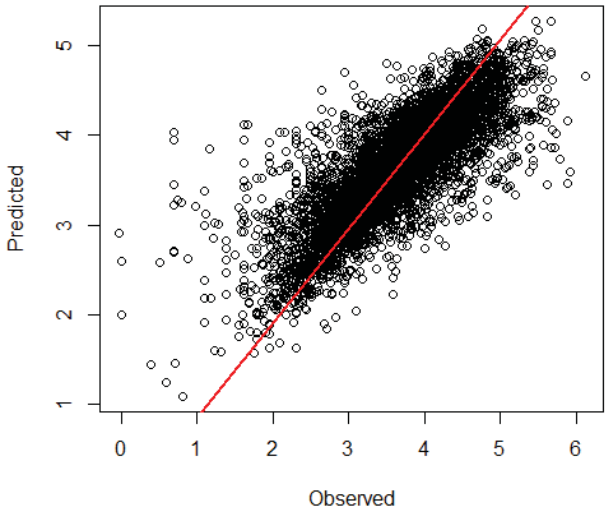


Figure 5-23 Scatter plot of cross validation, January, 2010

5.5. Probability map of exceedance

Indicator was defined by setting 1 where observed value is less or equal than given threshold, otherwise 0. In January, 69.7 percent of observation is below the threshold, hence probability threshold was defined as $p = 0.697$. Figure 5-25 shows indicator map of in-situ measurements of first four days of January. Number of exceedance varies day by day.

Semi-variances of indicator values are only can be 0 and 0.5 since it is half variance of binary value. These values have been averaged using indicator variogram.

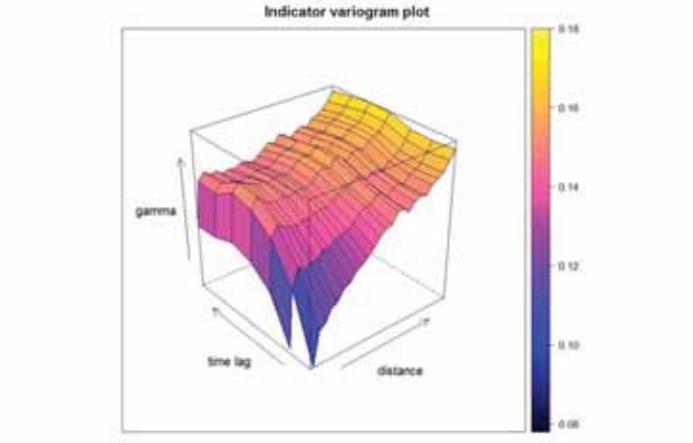


Figure 5-24 Indicator spatio-temporal variogram plot, in January, 2010

Beside this indicator value of in-situ measurements of PM10 concentrations, selected explanatory variables (CTM outcome and CORINE Land cover type) were used for the separable spatio-temporal variogram calculation. Result of indicator variogram plot is shown in Figure 5-24. Fitted exponential model parameters are estimated and those are spatial nugget: 0.58, partial sill: 0.42, spatial range: 300km, temporal nugget: 0.44, temporal partial sill: 0.56 and temporal range 7 days.

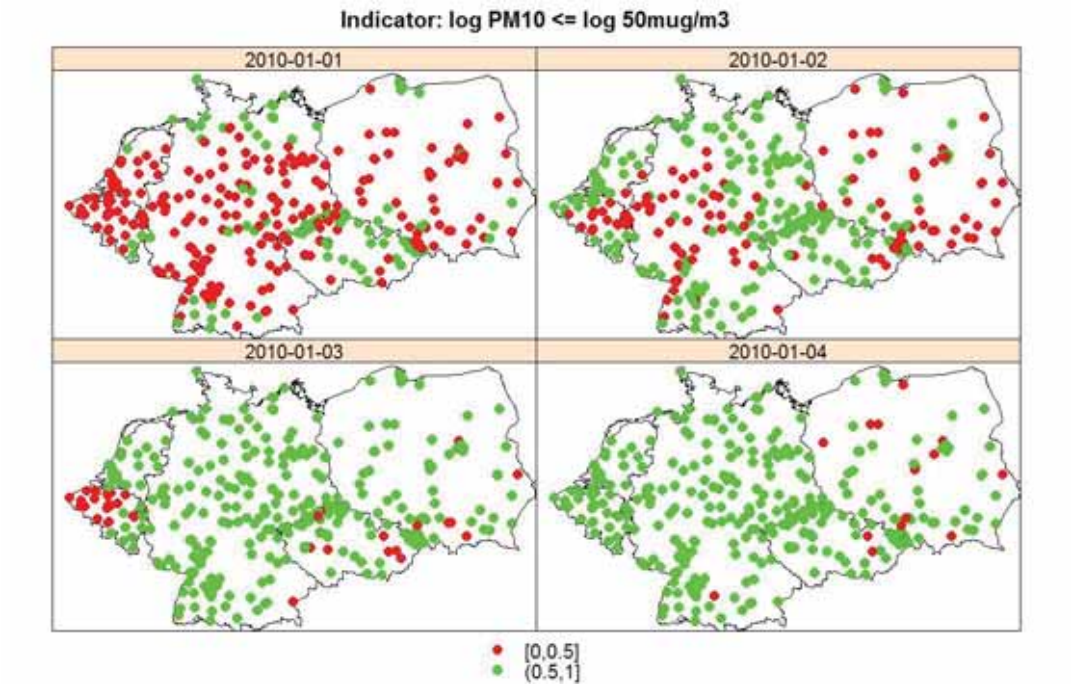


Figure 5-25 Indicator map of observed PM10 concentrations. Above/Below threshold is in red/green,

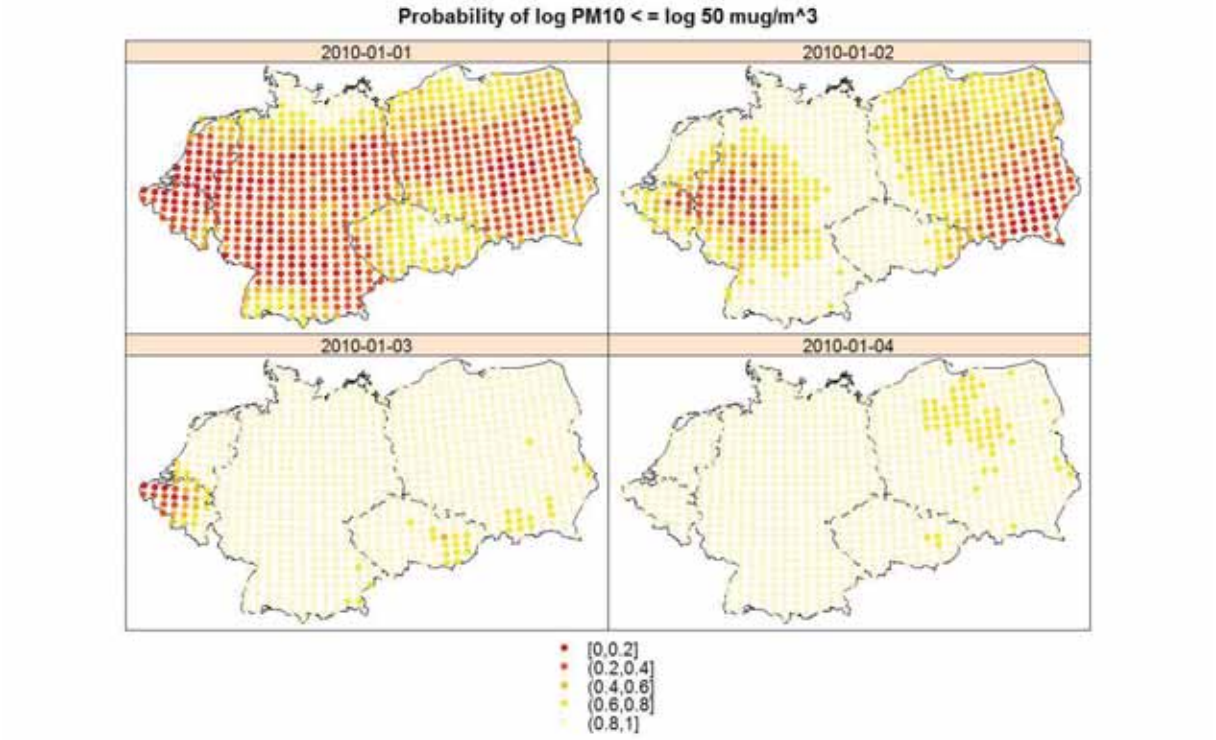


Figure 5-26 Probability map of exceedance at prediction locations. From yellow to red colour probability of exceedance increases

Using modelled indicator variogram, prediction of probability of exceedance at unsampled locations has been carried out. Range of predicted values of probability is [-0.033, 1.297]. In theory probability range should be in between 0 and 1. Therefore, these predicted probability values were limited to the theoretical range. Figure 5-26 shows probability of exceedance of PM10 concentrations over the study region in January 1st-4th. Spatial distribution of the probability varies depending on the distribution of the high observed concentrations.

Based on the probability threshold exceedance binary map has been created (Figure 5-27).

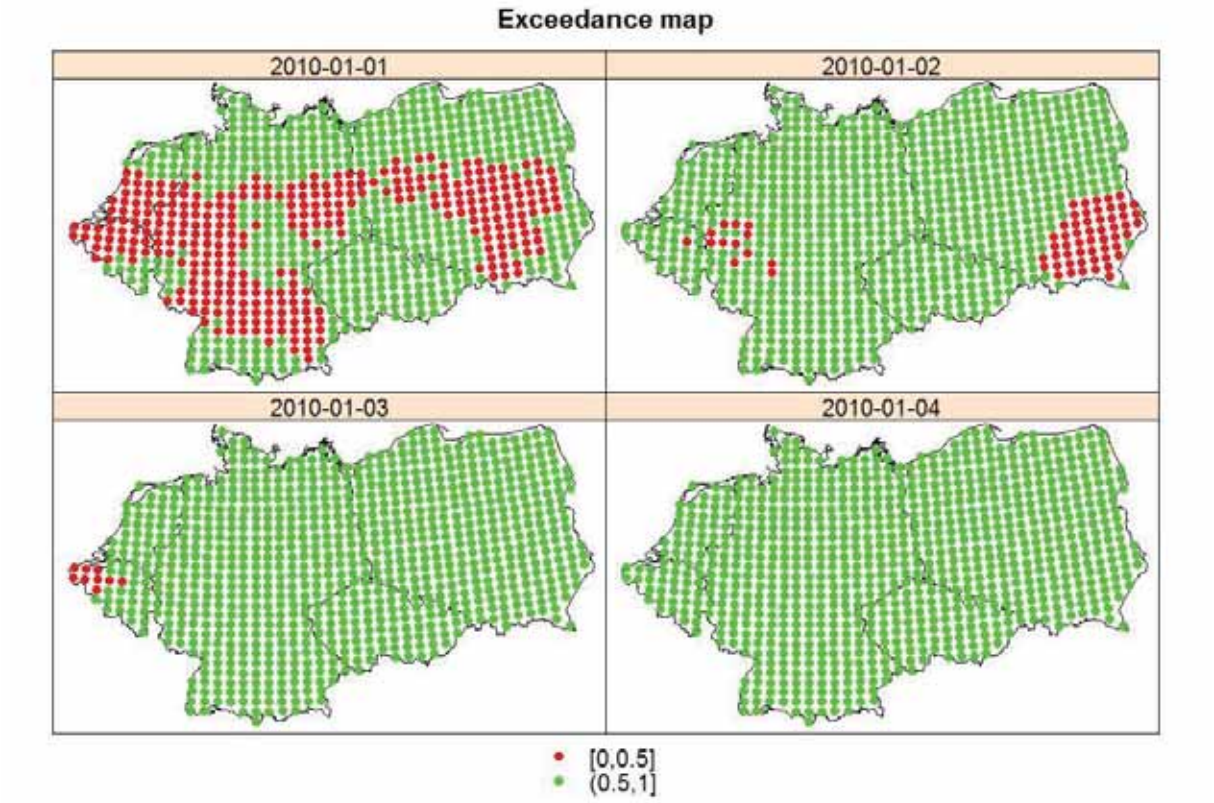


Figure 5-27 Predicted exceedance binary map. Red colour represents exceeded PM10 concentrations and green colour indicates below the defined thresholds.

6. DISCUSSION

6.1. Discussion

The aim of this research was to model and to map air pollution data by applying spatio-temporal geostatistical methods using secondary information.

Data pre-processing and combining all available dataset is essential and have been done prior to the implementation of spatio-temporal modelling. Besides primary in-situ measurements of PM10 concentrations, three different types of dataset were used. There were grid data of CTM outcome, and raster dataset of DEM and land cover map. All come with different formats and resolutions. In order to eliminate uncertainties come from resampling, spatial point based dataset was prepared based on given grid data of CTM outcome.

Overall, 580 monitoring stations were used in the spatio-temporal modelling. Referring to Figure 5-2, observed PM10 concentrations are positively skewed. This is due to the higher PM10 concentrations in the polluted atmosphere. Hence to achieve the normality, logarithmically transformed PM10 concentrations were used in the analysis (Figure 5-3). Observed PM10 concentrations vary in space (Figure 5-5) and time (Figure 5-6). And it's interesting to see that these variations are different in monthly-wise. During the months of the warm seasons (May, June, July and August), concentrations are relatively lower, and in colder seasonal times, extreme high concentration were observed. This could be explained by factors that influence the air pollution become active in cold seasonal days. On the other hand, PM10 concentrations in April are observed to be high concentrations and even their CTM outcomes were high (Figure 5-6 and Figure 5-10). Reason of this was explained by huge clouds of ash in the atmosphere due to the volcanic eruption in northern Iceland in between end of March and April. This kind of natural phenomena which affect air quality could be emerged at any time of the year. Therefore temporal aggregation like daily, monthly, seasonal or annual should be determined based on the exploratory analysis of the provided data.

Comparing summary statistics of in-situ measurements and CTM outcomes (Table 5-1 and Table 5-2), observed PM concentrations are higher than model outcome. According to LOTOS-EUROS (2011b), CTM underestimates the PM10 concentrations due to the large uncertainty in the modelling of secondary organic aerosols and crustal matter components. However, basic form of variation remains similar (for example, Figure 5-13).

In regards to data quality of in-situ measurements of PM10 concentrations, within the study region 62.5 per cent of given data had no observations at all during the year of 2010. And the remaining number of stations which are used in the analysis contains 9.7 per cent of missing values. Few numbers of daily observations have values less or equal than 0. Since it is a measurement of concentration, it cannot be negative. To avoid those, negative and 0 concentrations were replaced as missing values. In this research, data used as it is, however, presence of missing values and their cause should be examined at data exploration stage. And pre-interpolation could be useful to fill the gaps before the modelling and mapping.

Multiple regression analysis was carried out in order to choose significant explanatory variables as covariates. Regards to results of the regression analysis (Table 5-5), elevation gave insignificant

relationship with the observed PM10 concentrations. Land cover types from CORINE showed improved linear relationship. Although the regression analysis has been done in each month in order to study spatial relationship between dependent variables and explanatory variables, temporal aspect of the variables was disregarded.

It should be noted that used CORINE Land cover dataset has been created using satellite images that are acquired in 2006 with time consistency ± 1 year (Büttner *et al.*, 2012), and in-situ measurements of daily PM10 concentrations are observations in 2010. Four years difference in the two datasets could add uncertainty in the modelling.

Unlike the other studies (Beelen *et al.*, 2009; Desta, 2012), elevation did not give significant result in relation with the observed PM10 concentrations. Including elevation into the model is problematic since their values can be negative as well. Logarithm data transformation produces null values because of these negative values. And its distribution does not follow normality. Alternative solution could be replacing elevation values into factor for instance low, medium and high elevated.

Spatial and temporal correlations were examined separately using spatial variogram and auto and cross correlation functions. Figure 5-14 showed that variograms and their fitted exponential models of each month. Where variation of PM10 concentration is low, variogram tends to reach the sill at longer range. On the other hand, where the observed concentrations are relatively higher months, nuggets are estimated higher than others and variogram reaches sill at shorter range. Referring to the Table 5-6, variogram used secondary information reduced nugget effect. This could be concluded that non-spatial variability which cannot be conveyed by the data is reduced by using secondary information like CTM outcome and land cover types. As considering the data is multivariate time series, auto and cross correlation were computed. Correlation is dependent on time separation. In overall, cross correlations between two stations are significantly high up to four days. And as time lag increases, correlation decreases.

For the implementing spatio-temporal modelling, separable spatio-temporal variogram was used and exponential model was fitted. Referring to Figure 5-20, at 0 time and spatial lag, semi-variance could not be calculated. It is draw attention that at 0 spatial lag, semi-variances are relatively lower than other consecutive semi-variances. And these are increased abruptly at second spatial lag (20km). This could be explained by the no computed semi-variance at 0 time and spatial lag, and this variogram at 0 spatial lag is the purely temporal variogram which does not account for spatial structure in the computation. This separable model could be improved by extending the separable spatio-temporal model into product-sum model which has been actively studied and discussed in the De Iaco *et al.* (2002), De Iaco *et al.* (2011), De Iaco and Posa (2012) and Gräler *et al.* (2012).

It was assumed for the modelling that spatial structure of the daily PM10 concentrations is constant over a year, but unknown (second-order stationarity). However, data showed that PM10 concentrations vary both space and time (Figure 5-5). Thus, in order to improve the modelling, non-stationarity in the spatio-temporal process is needed to be taken into account.

Prediction map was created by spatio-temporal universal kriging based on the fitted separable spatio-temporal model at unsampled locations. Spatio-temporal kriging requires full dataset without any missing values. So for this reason, the stations that have continuous observation during whole month (more than 55 per cent of the data) were selected for the prediction. As for the remaining stations, there were used for the validation. This could lead to the lack of spatial support in the kriging when number of stations is decreased by around 40 per cent. Therefore, as mentioned before the cause and handling of missing values

should be carried out. As shown in the Figure 5-23, linear relationship between predicted versus observed values agreeably fitted to the model line with adjusted $R^2 = 0.63$.

Back transformation of the predicted values was done as explained in the Denby *et al.* (2008), however prediction result tends to smooth, large deviation between predicted and observed measurements at the same locations, and large prediction variances were obtained. This is because the equations provided in the paper are for the back transformation of the estimates from simple kriging (Webster & Oliver, 2008). Mean is known for simple kriging, while for the ordinary kriging the mean is unknown and only predicted values can be obtained. Kriging is based on weighted sum of observed values (Webster & Oliver, 2008), and according to identity of the logarithm, sum of logarithm is the logarithm of a product (Huntington, 1916). Hence weighting system becomes different than original. Back transformation equation for ordinary kriging is given in the Webster and Oliver (2008) with the additional term LaGrange multiplier. However, in this research universal kriging is used, meaning that observations are conditioned on CTM outcome and land cover types. Moreover back transformation would be differing than that for ordinary kriging. Furthermore, this back-transformation for universal kriging is worth studying in the further study.

Probability map of exceedance has been achieved by spatio-temporal indicator kriging based on given threshold value of daily PM10 concentrations. And exceedance map has been created by defining probability threshold as 0.697 in January. However, this value is subjective to the observations. When probability of exceedance is different in each day (Figure 5-25), the probability of threshold value should be set for each day. It is interesting to see that the patterns of the probability map. Lots of fireworks were occurred the end of December in previous year and this could explain the high probability of exceedance on 1st of January (Friday). To relate with the daily human activity for example working days and traffic, on Monday, high probability may occur and is gradually decreased until Sunday. But there are also several factors that affect the higher PM10 concentrations at any time in the atmosphere, for example, that volcano eruption in April, 2010 in Iceland.

Indicator kriging produced unrealistic prediction of probability, [-0.033, 1.297]. This is the one of the drawbacks of the indicator kriging which is that indicator kriging does not account for monotonic cdf (cumulative distribution function) property (Chiles & Delfiner, 1999; Christakos, 2000). Therefore, the probability map of exceedance could be improved by using other methods like simulation or disjunctive kriging approaches.

6.2. Limitation of the research

The research has encountered two main limitations and there were as follows,

- Limited to the given data and their characteristics;
- Limited to the available software environment and existing packages which are still under development.

7. CONCLUSION AND RECOMMENDATION

7.1. Conclusion

The conclusion has been made by answering the research questions.

1. *What is the spatial distribution of the PM concentrations over study area?*

Over the study area, spatial distribution of the PM10 concentrations varies. Based on the data exploratory analysis, distribution of the PM10 concentrations is positively skewed due to the higher concentrations then limited values are occurred in the some part of the region. Range of PM10 concentrations is between $0.20 \mu g/m^3$ and $455.0 \mu g/m^3$.

2. *What is the spatial relationship between in-situ measurements, CTM and additional sources such as elevation, and land cover over the regions in different temporal scales?*

The additional sources for the research were DEM, CTM outcome grid dataset and CORINE Land cover map. Referring to Table 5-5, DEM did not have significant correlation with observed PM10 concentrations and CORINE land cover has significant relationship with the observed PM10 concentrations. Referring to Figure 5-14, understanding the spatial structure was improved by using CORINE land cover types and CTM outcome.

3. *What is the temporal correlation between in-situ measurements, CTM and additional sources such as elevation, and land cover over the regions in different temporal scales?*

Temporal correlation has been examined only on in-situ measurements of PM10 observations since elevation and land cover types are temporally static compared to the daily variation of PM10 concentrations. There is around 4 successive days significant correlations were observed. And cross correlation between stations are relatively lower than auto-correlation.

4. *How to integrate data with different spatial resolution and different qualities?*

Data integration is a big concern when it comes to handling different types of dataset from different sources. Research handled remotely sensed products: DEM and land cover map, in-situ measurements and CTM outcome grid dataset. To avoid the uncertainty comes from resampling remotely sensed products, representing the CTM grid data as a point and corresponding elevation and land cover information were assigned to the point based grid data. However, one could argue that this method is very naïve thing to do which does not consider scaling issues in the integration.

5. *How can we develop space-time models of PM concentrations?*

Separable spatio-temporal model of PM10 concentrations has been developed by adding spatial and temporal variogram functions. CTM outcome and CORINE land cover were included in the modelling as explanatory variables. Spatio-temporal variogram function is defined as a function of spatial and temporal distance h_s and h_t respectively. Kriging approach has been performed based on this variogram estimation. However, for this modelling, kriging requires complete time series at each location.

6. *How can exceedance map of PM concentration be made based on the defined threshold value?*

Exceedance map of PM10 concentrations has been created using indicator kriging. Observations below given threshold are set to 1 and exceeded observations are set to zero. Based on modeled spatio-temporal indicator variogram, prediction values of probability were obtained. Importantly, calculated probability threshold value (0.697) for creating daily exceedance map in January is critical, since this threshold values was computed from all data in January.

7. What area of Europe has a high probability of exceeding PM concentrations?

Area that has high probability of exceeding PM10 concentrations is different from region to region. However, probability map shows that probability of exceeding is likely to be higher in the eastern part of the study area, and to relate with the land cover, where area that covered by artificial surfaces has higher probability than others. For example, northern part of Belgium and southern part of The Netherlands are high probability of exceedance (Figure 3-4 and Figure 5-26).

7.2. Recommendation

Based on the outcomes of the research the following are recommended for further geostatistical mapping,

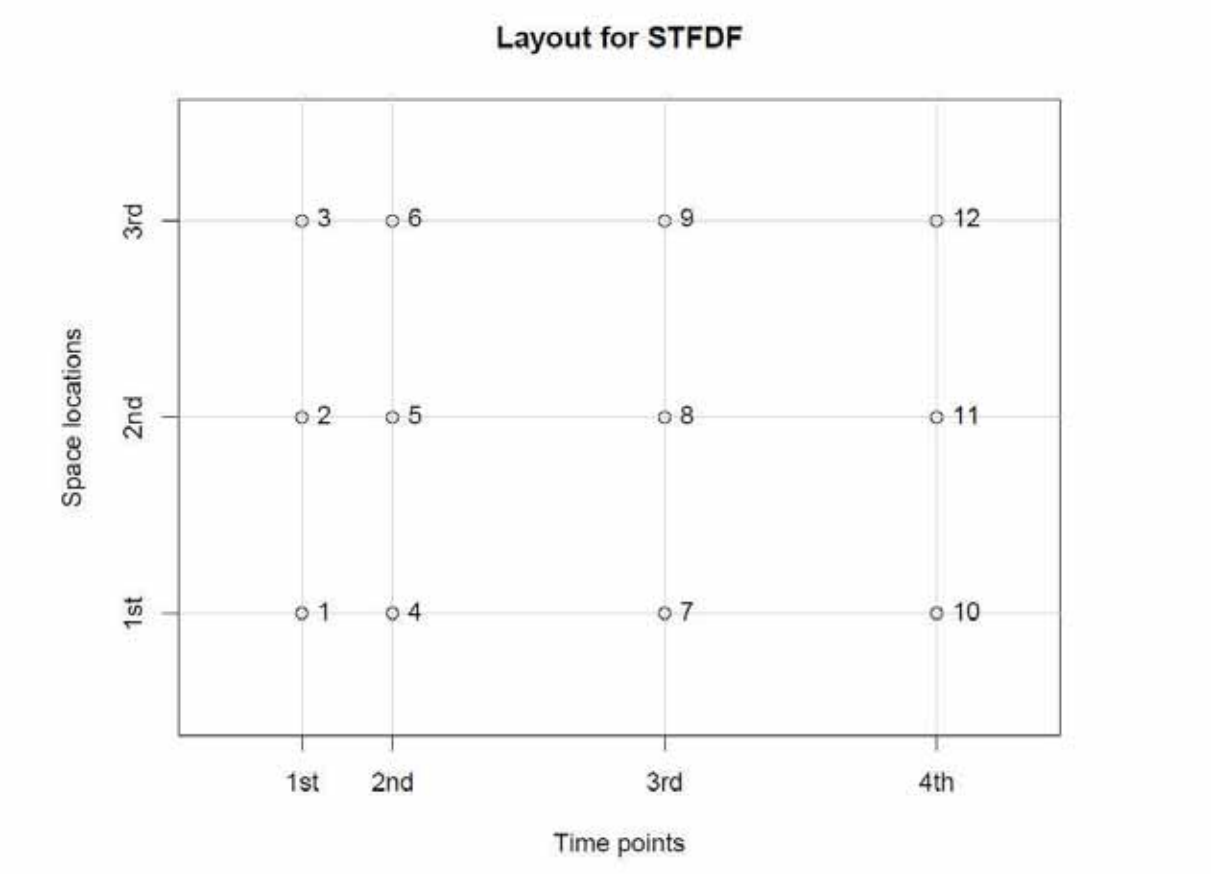
- Investigate the cause of the missing values and fill the possible gaps by linear interpolation beforehand;
- Include the DEM information as a factor into the model and check whether there is significance of elevation category;
- Separable spatio-temporal model could be improved by extending the covariance function (product-sum model);
- Extend the work by using dataset that has large temporal observations. This way seasonal variation could be modelled better;
- Examine the back transformation formula for universal kriging and
- Study the effect of defining the probability threshold value on daily basis and probability map of exceedance could be improved by other approaches like simulation and disjunctive kriging.

APPENDICES

Appendix-1 CORINE Land Cover Categories. Source from Büttner *et al.* (2012)

No.	CODE	LABEL1	LABEL2	LABEL3
1	111	Artificial surfaces	Urban fabric	Continuous urban fabric
2	112			Discontinuous urban fabric
3	121		Industrial, commercial and transport units	Industrial or commercial units
4	122			Road and rail networks and associated land
5	123			Port areas
6	124			Airports
7	131		Mine, dump and construction sites	Mineral extraction sites
8	132			Dump sites
9	133			Construction sites
10	141		Artificial, non-agricultural vegetated areas	Green urban areas
11	142			Sport and leisure facilities
12	211	Agricultural areas	Arable land	Non-irrigated arable land
13	212			Permanently irrigated land
14	213			Rice fields
15	221		Permanent crops	Vineyards
16	222			Fruit trees and berry plantations
17	223			Olive groves
18	231		Pastures	Pastures
19	241		Heterogeneous agricultural areas	Annual crops associated with permanent crops
20	242			Complex cultivation patterns
21	243			Land principally occupied by agriculture, with significant areas of natural vegetation
22	244			Agro-forestry areas
23	311	Forest and semi natural areas	Forests	Broad-leaved forest
24	312			Coniferous forest
25	313			Mixed forest
26	321		Scrub and/or herbaceous vegetation associations	Natural grasslands
27	322			Moors and heathland
28	323			Sclerophyllous vegetation
29	324			Transitional woodland-shrub
30	331		Open spaces with little or no vegetation	Beaches, dunes, sands
31	332			Bare rocks
32	333			Sparsely vegetated areas
33	334			Burnt areas
34	335			Glaciers and perpetual snow
35	411	Wetlands	Inland wetlands	Inland marshes
36	412			Peat bogs
37	421		Maritime wetlands	Salt marshes
38	422			Salines
39	423			Intertidal flats
40	511	Water bodies	Inland waters	Water courses
41	512			Water bodies
42	521		Marine waters	Coastal lagoons
43	522			Estuaries
44	523			Sea and ocean

Appendix-2 Structure of STFDF class in *spacetime* package, source from Pebesma (2012)



Appendix-3 Number of stations used in prediction and validation

	Total monitoring stations	Number of stations used for prediction	Number of stations used for validation
January	570	334	236
February	567	378	189
March	567	375	192
April	567	356	211
May	570	350	220
June	564	330	234
July	565	312	253
August	566	331	235
September	567	339	228
October	596	359	237
November	564	350	214
December	560	339	221

Appendix-4 Used[6] and implemented code in R

```
##################################################################################################
#### 1 LOAD REQUIRED LIBRARIES
##################################################################################################
library(gstat) # geostatistics
library(lattice) # lattice data
library(maptools) # map
library(rgdal) # handling geodata coordinate transformation
library(ncdf4) # extract NetCDF archived file
library(spacetime) #spacetime dat
library(gplots) # used for color scheme
##################################################################################################
#### 2 PREPARE DATA FROM ARCHIVE
##################################################################################################
LE  <- "D:\\MSc\\Module 16-23 MSc Research\\SPACETIME\\Data\\aqord__LE_eu__1d__pm10_mass__pm10.nc"
ap  <- "D:\\MSc\\Module 16-23 MSc Research\\SPACETIME\\Data\\aqord__obs__1d__pm10_mass__pm10.nc"
nc1  <- nc_open(ap)
nc2  <- nc_open(LE)
## Station information
lat  <- ncvar_get(nc1, "station_lat")
lon <- ncvar_get(nc1, "station_lon")
height  <- ncvar_get(nc1, "station_height")
st.code <- ncvar_get(nc1, "station_code")
st.name <- ncvar_get(nc1, "station_name")
st.type <- ncvar_get(nc1, "airbase_station_type")
st.type_area <- ncvar_get(nc1, "airbase_station_type_of_area")
# This dataframe contains the PM10 insitu observations
pm10.airbase  <- ncvar_get(nc1, "pm10_mass__pm10")
# This dataframe contains the PM10 LE simulator outputs
pm10.LE   <- ncvar_get(nc2, "pm10_mass__pm10")
# Close the links to the NetCDF files
nc_close(nc1)
nc_close(nc2)
tmp0  <- data.frame(lat=lat, lon=lon, z=1)
tmp1  <- tmp0
coordinates(tmp1) <- ~ lon + lat
#proj4string(tmp1) <- CRS("+proj=latlong")
proj4string(tmp1) <- CRS("+proj=longlat")
proj4string(tmp1)
# http://spatialreference.org/ref/epsg/3035/
tmp2  <- spTransform(tmp1, CRS("+init=epsg:3035"))
tmp3  <- as.data.frame(tmp2)
colnames(tmp3)  <- c(colnames(tmp3)[1], "x", "y")
tmp3  <- data.frame(tmp3, lat=lat, lon=lon)
# This data fram contains the station information.
st.info.airbase <- data.frame(lat=lat, lon=lon, easting=tmp3$x, northing=tmp3$y, height=height, code=st.code, name=st.name,
 type=st.type, type_area=st.type_area)
# Save to .Rdata files
save(st.info.airbase, pm10.airbase, pm10.LE, file="airbaseLE.Rdata")

# Delete all variables and then restore the key data
rm(list=ls())
load("airbaseLE.Rdata")
attach(st.info.airbase)
tmp0  <- data.frame(mNo=1:length(lat), lat=lat, lon=lon, easting=easting, northing=northing, height=height, st.code=code,
 type=type, type_area=type_area, country=substr(code, 1, 2))
detach(st.info.airbase)
head(tmp0)
# FOR attach CORINE information to the file
write.csv(tmp0, sep=";", file="D:\\MSc\\Module 16-23 MSc Research\\SPACETIME\\matrix.csv")
# some work in ArcGIS to get the Corine LC information!!!! export as .csv file and load to R as data frame

##################################################################################################
#### 3 LOAD THE PREPARED CSV FILE
##################################################################################################
st.info.airbase.lc <- as.data.frame(read.csv("all_station_dem_lc.csv"))  #metadata with corie lc
head(st.info.airbase.lc)
load("airbaseLE.Rdata")

##################################################################################################
#### 4 PREPARE STFDF FILE FOR INSITU (ALL STATIONS INCLUDED)
##################################################################################################
# SP data
row.names(st.info.airbase.lc)=paste(st.info.airbase.lc$st_code)
sp1=SpatialPointsDataFrame(data.frame(st.info.airbase.lc$easting/1000, st.info.airbase.lc$northing/1000), st.info.airbase.lc
)    # in km
sp2=SpatialPointsDataFrame(data.frame(st.info.airbase.lc$easting, st.info.airbase.lc$northing), st.info.airbase.lc)    # in m

# time
time365 = as.Date(as.POSIXct("2010-01-01")+3600*seq(24,8760, by=24)) #  prepare time value 365 days
timeIsInterval(time365)=TRUE
time365
# value
mydata1=pm10.airbase[,732:1096] # extracting 2010 daily measurements
mydata1[mydata1<=0]<-NA      # if value less or equal than Zero, then put NA (because when apply log trans, it produce
 infinite values)
#mydata1=log(mydata1)
mydata2=as.numeric(mydata1)
mydata3=data.frame(value=mydata2)
st.all.obs=STFDF(sp2, time365, mydata3)
st.all.obs@sp@proj4string <- CRS("+init=epsg:3035")
```

[6] R code for extracting NetCDF4 archive file into data frame is acknowledged to Dr. Nicholas Hamm.

```
#####################################################################################
#### 6 SUMMARY STATISTICS
#####################################################################################
# BY COUNTRY
summary(as((r2010[which(r2010@sp@data$country == "BE")]), "STFDF"))
summary(as((r2010[which(r2010@sp@data$country == "CZ")]), "STFDF"))
summary(as((r2010[which(r2010@sp@data$country == "DE")]), "STFDF"))
summary(as((r2010[which(r2010@sp@data$country == "NL")]), "STFDF"))
summary(as((r2010[which(r2010@sp@data$country == "PL")]), "STFDF"))
# BY LAND COVER
summary(as((r2010.LE[which(r2010@sp@data$LABEL1 == "Agricultural areas")]), "STFDF"))
summary(as((r2010.LE[which(r2010@sp@data$LABEL1 == "Artificial surfaces")]), "STFDF"))
summary(as((r2010.LE[which(r2010@sp@data$LABEL1 == "Forest and semi natural areas")]), "STFDF"))
summary(as((r2010[which(r2010@sp@data$LABEL1 == "Water bodies")]), "STFDF"))
# summary(as((r2010.obs[which(r2010.obs@sp@data$LABEL1 == "Wetlands")]), "STFDF"))
summary(r2010)
library(gplots)
boxplot.n(r2010[,"2010-01-01"]$value~r2010@sp@data$LABEL1, top=T, horizontal=F)
abline(h=mean(r2010[,"2010-01-01"]$value), lwd=2, col="red")
boxplot(r2010@sp@data$LABEL1)
#####################################################################################
#### 7  histograms and QQ plots data transformation
#####################################################################################
# PM10
X11()
par(mfrow=c(2,2))
hist(na.omit(r2010[,1:31]@data$value), main="", xlab="PM10 Concentration", cex.lab=1.5, prob=T)
lines(density((r2010[,1:31]@data$value),na.rm=T, adjust=2), col="red")
qqnorm(r2010[,1:31]@data$value, main="",  cex.lab=1.5)
qqline(r2010@data$value, col="red")
r2010@endTime <-delta(r2010@time)
# logPM10
X11()
hist(log(r2010[,1:31]@data$value), main="", xlab="log PM10 Concentration", cex.lab=1.5, prob=T)#,col=as.numeric(r2010@sp
@data$LABEL1))
lines(density(log(r2010[,1:31]@data$value),na.rm=T, adjust=2), col="red")
#####################################################################################
# This code is for preparing CTM grd STFDF file
#####################################################################################
library(gstat) # geostatistics
library(lattice) # lattice data
library(maptools) # map
library(rgdal) # handling geodata coordinate transformation
library(ncdf4) # extract NetCDF archived file
library(spacetime) #spacetime dat
library(gplots)
######## Read NetCDF file from directory and bind all together as data.frame
LE  <- dir("D:\\MSc\\Module 16-23 MSc Research\\LE\\data", full.names=T)
nc <- nc_open(LE[1])
lat  <- ncvar_get(nc, "lat")
lon  <- ncvar_get(nc, "lon")
nc1 <- as.matrix(as.numeric(t(ncvar_get(nc, "tpm10"))))
####### SP object from grid data
att1=expand.grid(lat=lat, lon=lon)
coordinates(att1)=~lon+lat
proj4string(att1)=CRS("+proj=longlat")
p=SpatialPoints(att1)
# for spatial join in ArcGIS for getting Country names from europe boundary polygon file
write.csv(p, sep=";", file="D:\\MSc\\Module 16-23 MSc Research\\CTM\\ctm_center.csv")
# Using SpatialJoin to add country name attribute. All done in ArcGIS
ctm.all <- as.data.frame(read.csv("D:\\MSc\\Module 16-23 MSc Research\\CTM\\SHP\\ctm_all.csv"))  #metadata with corie lc
grd.att1=data.frame(p, name=ctm.all$NUTS_ID, LABEL1=ctm.all$LABEL1,dem=ctm.all$RASTERVALU)
coordinates(grd.att1)=~lon+lat
proj4string(grd.att1)=CRS("+proj=longlat")
grd.proj  <- spTransform(grd.att1, CRS("+init=epsg:3035"))
for(i in 2:length(LE)){
  nc2=nc_open(LE[i])
  nc1=cbind(nc1,as.matrix(as.numeric(t(ncvar_get(nc2, "tpm10")))))
  nc_close(nc2)
}
a=as.numeric(nc1) # for STFDF
time365 = as.Date(as.POSIXct("2010-01-01")+3600*seq(24,8760, by=24)) #  prepare time value 365 days
timeIsInterval(time365)=TRUE
time365
##### Creat STFDF
LE.ST=STFDF(grd.proj, time365, data.frame(LE=a*1000000000), as.POSIXct("2009-12-31")+3600*seq(24,8760, by=24))
summary(LE.ST)
stplot(LE.ST[,1:2],col.regions=colorpanel(20, "yellow", "blue", "red"),sp.layout=layout)
# only Study area
summary(LE.ST@sp@data$name)
tmp3=as((LE.ST[which(LE.ST@sp@data$name != "AD")]), "STFDF")
tmp3=as((tmp3[which(tmp3@sp@data$name != "AL")]), "STFDF")
tmp3=as((tmp3[which(tmp3@sp@data$name != "AT")]), "STFDF")
tmp3=as((tmp3[which(tmp3@sp@data$name != "BA")]), "STFDF")
tmp3=as((tmp3[which(tmp3@sp@data$name != "BG")]), "STFDF")
tmp3=as((tmp3[which(tmp3@sp@data$name != "CH")]), "STFDF")
tmp3=as((tmp3[which(tmp3@sp@data$name != "CY")]), "STFDF")
tmp3=as((tmp3[which(tmp3@sp@data$name != "DK")]), "STFDF")
tmp3=as((tmp3[which(tmp3@sp@data$name != "EE")]), "STFDF")
```

```
##############################################################################################
# Spatio temporal variogram, January
##############################################################################################
stplot(JAN, mode="xt",col.regions=bpy.colors())
r1.vv=variogramST(PM10~LE+LC, JAN, width=20000, cutoff=500000,tlags=0:6)
plot(r1.vv, ylab="time lag (days)", xlab="distance (km)")
plot(r1.vv, map=FALSE, ylab="time lag (days)", xlab="distance (km)")
# fit ST Variogram
sepVgm <- list(space=vgm(0.3, "Exp", 300000, 0.15),
               time= vgm(0.3, "Exp",    3, 0.15),
               sill=0.3, nugget=0.15, stModel="separable")
sepVgm<-fit.StVariogram(r1.vv, sepVgm, method="L-BFGS-B")$StVgmFit
X11()
plot(r1.vv, sepVgm,ylab="time lag (days)", xlab="distance (km)")
plot(r1.vv, sepVgm, map=F,ylab="time lag (days)", xlab="distance (km)")
plot(r1.vv, wireframe=T, pretty=T, col.regions=bpy.colors())

LE.grid.jan.1$LE <-log(LE.grid.jan.1$LE)
LE.grid.jan.1$LC <-rep(LE.grid.jan.1@sp$LABEL1,31)
LE.grid.jan.1$LC[which(LE.grid.jan.1$LC == "Wetlands")] <- "Agricultural areas"
LE.grid.jan.1$LC[which(LE.grid.jan.1$LC == "Water bodies")] <- "Artificial surfaces"
# kriged.pl<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="PL")], modelList
=sepVgm, fullCovariance=T)
# kriged.pl.cv<-gstat.cv(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="PL")], modelList
=sepVgm)
save(LE.grid.jan.1, file="JANGRID.Rdata")
##############################################################################################
# Spatio temporal kriging, January
##############################################################################################
kriged.pl.v1<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="PL")][1:100,],
 modelList=sepVgm, computeVar=T)
kriged.pl.v2<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="PL")][101:200,],
 modelList=sepVgm, computeVar=T)
kriged.pl.v3<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="PL")][201:322,],
 modelList=sepVgm, computeVar=T)
kriged.nl.v<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="NL")], modelList
=sepVgm, computeVar=T)
kriged.cz.v<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="CZ")], modelList
=sepVgm, computeVar=T)
kriged.de.v1<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="DE")][1:100,],
 modelList=sepVgm, computeVar=T)
kriged.de.v2<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="DE")][101:250,],
 modelList=sepVgm, computeVar=T)
kriged.de.v3<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="DE")][251:369,],
 modelList=sepVgm, computeVar=T)
kriged.be.v<-krige(PM10~LE+LC, locations=JAN.no, newdata=LE.grid.jan.1[which(LE.grid.jan.1@sp$name =="BE")], modelList
=sepVgm, computeVar=T)
str(kriged.be.v)
summary(kriged.be.v)

##########################################################################
#indicator kriging
##########################################################################
summary(iJAN)
iJAN=JAN[,,1:3]
iJAN$PM10i=ifelse(iJAN$PM10 <= log(50), 1, 0)
iJAN$iLE=ifelse(iJAN$LE <= log(50), 1, 0)
summary(iJAN)
r1.vv.i=variogramST(PM10i~LE+LC, iJAN, width=20000, cutoff=500000,tlags=0:6)
r1.vv.ci=variogramST(PM10i~LE+LC, iJAN, width=20000, cutoff=500000,tlags=0:6, cloud=T)

plot(r1.vv.i, ylab="time lag (days)", xlab="distance (km)", cloud=T)
plot(r1.vv.i, map=FALSE, ylab="time lag (days)", xlab="distance (km)")
# fit ST Variogram
isepVgm <- list(space=vgm(0.4, "Exp", 300000, 0.5),
               time= vgm(0.4, "Exp",    3, 0.5),
               sill=0.4, nugget=0.06, stModel="separable")
isepVgm<-fit.StVariogram(r1.vv.i, isepVgm, method="L-BFGS-B")$StVgmFit
```

LIST OF REFERENCES

The 2010 Eruptions of Eyjafjallajökull. (2011). [The unofficial blog of the UPJ Geology & Planetary Sciences department]. Retrieved from http://mountaincatgeology.wordpress.com/2011/02/19/the-2010-eruptions-of-eyjafjallajokull/

Annoni, A., Luzet, C., Gubler, E., & editors. (2003). Map Projections for Europe (Vol. EUR 20120 EN): European Commission.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data* (Vol. 101). Boca Raton etc.: Chapman & Hall

CRC.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., & Briggs, D. J. (2009). Mapping of background air pollution at a fine spatial scale across the European Union. *Science of The Total Environment, 407*(6), 1852-1867.

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H.). *Spatial and Spatio-Temporal models with R-INLA*. Spatial and Spatio-temporal Epidemiology, (0).

Büttner, G., Kosztra, B., Maucha, G., & Pataki, R. (2012). Implementation and achievement of CLC2006.

Cesare, L. D., Myers, D. E., & Posa, D. (2001). Estimating and modeling space–time correlation structures. *Statistics & Probability Letters, 51*(1), 9-14.

Chiles, J. P., & Delfiner, P. (1999). *Geostatistics : modeling spatial uncertainty*. New York etc.: Wiley & Sons.

Christakos, G. (2000). *Modern spatiotemporal geostatistics*. GB: Oxford University Press.

Cocchi, D., Greco, F., & Trivisano, C. (2007). Hierarchical space-time modelling of PM10 pollution. *Atmospheric Environment, 41*(3), 532-542.

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio - temporal data*. Hoboken: Wiley & Sons.

Dadvand, P., Rankin, J., Rushton, S., & Pless-Mulloli, T. (2011). Association Between Maternal Exposure to Ambient Air Pollution and Congenital Heart Disease: A Register-based Spatiotemporal Analysis. *American Journal of Epidemiology, 173*(2), 171-182.

de Fouquet, C., Malherbe, L., & Ung, A. (2011). Geostatistical analysis of the temporal variability of ozone concentrations. Comparison between CHIMERE model and surface observations. *Atmospheric Environment, 45*(20), 3434-3446.

De Iaco, S., Myers, D. E., Palma, M., & Posa, D. (2010). FORTRAN programs for space–time multivariate modeling and prediction. *Computers & Geosciences, 36*(5), 636-646.

De Iaco, S., Myers, D. E., & Posa, D. (2001). Space–time analysis using a general product–sum model. *Statistics & Probability Letters, 52*(1), 21-28.

De Iaco, S., Myers, D. E., & Posa, D. (2002). Space–time variograms and a functional form for total air pollution measurements. *Computational Statistics & Data Analysis, 41*(2), 311-328.

De Iaco, S., Myers, D. E., & Posa, D. (2011). On strict positive definiteness of product and product–sum covariance models. *Journal of Statistical Planning and Inference, 141*(3), 1132-1140.

De Iaco, S., & Posa, D. (2012). Predicting spatio-temporal random fields: Some computational aspects. *Computers & Geosciences, 41*(0), 12-24.

Denby, B., Schaap, M., Segers, A., Builtjes, P., & Horálek, J. (2008). Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale. *Atmospheric Environment, 42*(30), 7122-7134.

Desta, F. S. (2012). *Non - stationary linear mixed modelling of air quality.* MSc Thesis, University of Twente Faculty of Geo-Information and Earth Observation ITC, Enschede. Retrieved from http://www.itc.nl/library/papers_2012/msc/gfm/desta.pdf

Dickey, J. H. (2000). Selected topics related to occupational exposures Part VII. Air pollution: Overview of sources and health effects. *Disease-a-Month, 46*(9), 566-589.

EEA. (2012). Air quality in Europe 2012 *Air quality in Europe* (pp. 104).

European Environment Agency. Retrieved August 1st, 2012, from http://www.eea.europa.eu/

Gerharz, L. E., Klemm, O., Broich, A. V., & Pebesma, E. (2013). Spatio-temporal modelling of individual exposure to air pollution and its uncertainty. *Atmospheric Environment, 64*(0), 56-65.

Gething, P. W., Atkinson, P. M., Noor, A. M., Gikandi, P. W., Hay, S. I., & Nixon, M. S. (2007). A local space–time kriging approach applied to a national outpatient malaria data set. *Computers & Geosciences, 33*(10), 1337-1350.

Goovaerts, P., Webster, R., & Dubois, J. P. (1997). Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental and Ecological Statistics, 4*(1), 49-64.

Gräler, B., Gerharz, L., & Pebesma, E. (2012, 30th January). Spatio-temporal analysis and interpolation of PM10 measurements in Europe Retrieved 30th July, 2012, from http://acm.eionet.europa.eu/reports/ETCACM_TP_2011_10_spatio-temp_AQinterpolation

Huntington, E. V. (1916). An Elementary Theory of the Exponential and Logarithmic Functions. *The American Mathematical Monthly, 23*(7), 241-246.

Jost, G., Heuvelink, G. B. M., & Papritz, A. (2005). Analysing the space–time distribution of soil water storage of a forest ecosystem using spatio-temporal kriging. *Geoderma, 128*(3–4), 258-273.

Kitagawa, G. (2010). *Introduction to time series modeling* (Vol. 114). Boca Raton: CRC.

Kloog, I., Koutrakis, P., Coull, B. A., Lee, H. J., & Schwartz, J. (2011). Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric Environment, 45*(35), 6267-6275.

Koelemeijer, R. B. A., Homan, C. D., & Matthijsen, J. (2006). Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmospheric Environment, 40*(27), 5304-5315.

Konovalov, I. B., Beekmann, M., Meleux, F., Dutot, A., & Foret, G. (2009). Combining deterministic and statistical approaches for PM10 forecasting in Europe. *Atmospheric Environment, 43*(40), 6425-6434.

Kyriakidis, P. C., & Journel, A. G. (1999). Geostatistical space-time models: A review. *Mathematical Geology, 31*(6), 651-684.

Lark, R. M., & Cullis, B. R. (2004). Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science, 55*(4), 799-813.

Lee, H. J., Liu, Y., Coull, B. A., Schwartz, J., & Koutrakis, P. (2011). A novel calibration approach of MODIS AOD data to predict PM2.5 concentrations. *Atmospheric Chemistry and Physics, 11*(15), 7991-8002.

LOTOS-EUROS. (2011a, 1st Dec). LOTOS-EUROS Retrieved 3rd Sep, 2012, from http://www.lotos-euros.nl/

LOTOS-EUROS. (2011b). Products, Quality and Background Information. In H. Elbern, V.-H. Peuch & L. Rouïl (Eds.).

Mwenda, L. P. (2011). *Geostatistical analysis of air pollution using models, in situ and remote sensed data.* MSc Thesis, University of Twente Faculty of Geo-Information and Earth Observation ITC, Enschede.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences, 30*, 683-691.

Pebesma, E. J. (2012). spacetime: Spatio-Temporal Data in R. *Journal of Statistical Software, 51*(7), 1-30.

Riccio, A., Barone, G., Chianese, E., & Giunta, G. (2006). A hierarchical Bayesian approach to the spatio-temporal modeling of air quality data. *Atmospheric Environment, 40*(3), 554-566.

Romanowicz, R., Young, P., Brown, P., & Diggle, P. (2006). A recursive estimation approach to the spatio-temporal analysis and modelling of air quality data. *Environmental Modelling & Software, 21*(6), 759-769.

Sahu, S. K., & Mardia, K. V. (2005). *Recent trends in modeling spatio-temporal data*. Paper presented at the Meeting of the Italian Statistical Society on Statistics and the Environment. http://eprints.soton.ac.uk/30048/

Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindstrom, J., & Kaufman, J. D. (2011). Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment, 45*(36), 6593-6606.

The New York Times. (2010). Iceland Volcano Eruption of 2010 (Eyjafjallajokull Volcano) Retrieved 2013-02-12, 2013, from http://topics.nytimes.com/top/news/international/countriesandterritories/iceland/eyjafjallajokull/index.html

van de Kassteele, J., Koelemeijer, R. B. A., Dekkers, A. L. M., Schaap, M., Homan, C. D., & 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 de Kassteele, J., Stein, A., & Dekkers, A. L. M. (2006). *Statistical air quality mapping.* PhD Thesis, Wageningen Universiteit, Wageningen.

van der Meer, F. D. (2012). Remote sensing image analysis and geostatistics. *JournalInternational journal of remote sensing, 33*(18), 5644-5676.

Webster, R., & Oliver, M. A. (2008). *Geostatistics for environmental scientists : e-book* (Second edition ed.): Wiley & Sons.

WHO. (2011, September 2011). Air quality and health Retrieved 8th August, 2012, from http://www.who.int/mediacentre/factsheets/fs313/en/

World Health Organization. (2006). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide - Global update 2005.