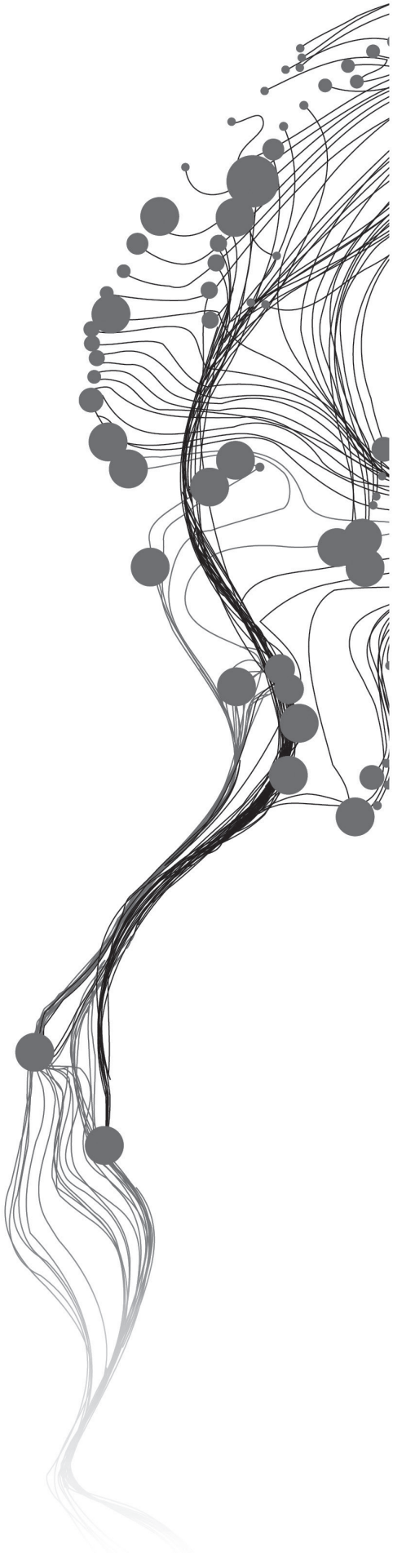
AUTOMATIC MODELING AND MAPPING OF PARTICULAR MATTER (PM10)

MONA ESHRAGHI
February, 2013

SUPERVISORS:

Dr. N. A. S. Hamm
Dr. I. Ivanova



AUTOMATIC MODELING AND MAPPING OF PARTICULAR MATTER (PM10)

MONA ESHRAGHI
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. N. A. S. Hamm
Dr. I. Ivanova

THESIS ASSESSMENT BOARD:

Prof. Dr. Ir. A. Stein (chair)
Dr. Ir. G. B. M. Heuvelink (external examiner)

ABSTRACT

World health organization (WHO) reported that outdoor air pollution caused 1.3 million death world wild (WHO, 2011) so it is a must to control the air quality and keep the pollution level under the thresholds defined by health organizations. It is also important that public be aware of near real time air quality situation where they live so they can make proper decision for their outdoor activity.

Air quality maps are one way to inform governments and public about different pollutants concentration. Daily recorded particular matter with a diameter of 10 micrometers or less (PM10) is the pollutant considered for modeling and mapping in this study. For automatic near real time modeling PM10, we used kriging with external drift (KED) using LOTOS-EUROS model output as a covariate and to make this map available for the users in real time, we used web processing services (WPS)and web map services (WMS).

The WPS used in this study is WPS4R produced by 52North which is capable of creating processes via R scripts. It enabled us to use R environment which is a very strong language for statistical data analysis.

Two maps are produced for each request from users, PM10 concentration and uncertainty maps. To improve the usability of the automatic air pollution mapping system, two groups of users are considered who those are familiar with geostatistical concepts and, the other group, who are not. First group are allowed for selecting different automatic interpolation methods which are ordinary kriging, universal kriging and kriging with external drift so they can compare different methods' result. For second group PM10 maps are produced using KED method as it was evident, in most of the days we analyzed, it gave better result in term of RMSE, ME and MSDR.

Keywords

air pollution, geostatistics, automatic spatial interpolation, web processing services

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God, the Almighty, for having made everything possible for me, by giving strength and courage to do this work.

My deepest gratitude to my first supervisor, Dr. Nicholas Hamm for his guidance, support and patience he demonstrated throughout. I learned a lot from him during the courses and the thesis period because of which I could complete my work. Thank you Dr. Hamm.

I would like to gratefully acknowledge my second supervisor, Dr. Ivana Ivanova for her encouragement, guidance and support for the whole thesis time. Thank you Dr. Ivanova.

I am really thankful to all my friends whose friendship made it so easy for me to endure pressures and being far from home and my family.

Last but not the least, I am indebt to my family, especially my mother for their support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Introduction

1.1 MOTIVATION AND PROBLEM STATEMENT

Air pollution is one of the major concerns in most countries in the world. A report by the Committee on the Medical Effects of Air Pollutants (COMEAP) indicates that in a western European population, a modest 10 μgm^{-3} increase in the ambient annual average level of particular matter (PM) is associated with a 6% increase in the death rate(COMEAP, 2009). It is government responsibility to control the air quality and keep pollutants lowers than defined thresholds by health providers. Requirements for this control are the knowledge about pollutants concentration in every location, every time as well as the uncertainty associated with predicted pollutant concentration. There are several stations installed all over the world recording amount of different air pollutants. For instance in Europe there is more that 4000 stations recording air pollution concentration. These numbers of stations do not satisfy the need of information on every location because they cannot cover the whole area. To overcome this problem there are different interpolation methods that model air pollution throughout area base on the values monitored by the stations to retrieve air pollution values in unmonitored locations.

Base on the Tobler first law of geography values of points more closer to each other have more similarity than those in more distance (Tobler, 1970). This is a fact that is used in spatial interpolation methods. The correlation between recorded values by the stations is used to reduce the prediction uncertainty of unmonitored points.

Because the interpolated values are retrieved from models and are not true values, they contain uncertainties. Between different spatial interpolation methods, geostatistical methods such as kriging are capable of calculating uncertainty of prediction in every predicted point. The result can be shown in a map which is called "Uncertainty map". It is valuable knowledge as it aids governments to decide on the placement of monitoring sites and it also can be used as information in risk assessment (Denby et al., 2007). In locations that prediction uncertainty is high, it is possible that pollutant concentrations exceed the limits but predicted amounts do not show exceeding (Senaratne et al., 2012).

Decision makers need air pollution behavior everyday if any event associated with air pollution (pollutant exceedance) took place in their region. Base on this reason that the near real time results (air pollution map and uncertainty map) are needed, automating the process is obligatory. Dubois and Galmarini (2005) defined proper geostatistical method for automation is the one that generate results:

1. In a minimum amount of time.
2. Without any human intervention, in the sense that only request for a map from the client is allowed.
3. That is "reasonable", which means low uncertainty of predictions. For example the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) has to be kept as low as possible.

4. That can handle outliers of input data (air pollution peak). Users that are considered to use the air pollution map and uncertainty map are governments, health providers, decision makers.

1.2 RESEARCH OBJECTIVE AND QUESTIONS

The main objective of this research is to automate the process of modeling and mapping PM10 considering results (PM10 concentration map and associated uncertainty) should be generated in near real time. The following are objectives (RO) and question (RQ) that have been answered during the research.

- **RO1** Find and apply the proper interpolation method for automatic PM10 modeling .
  - **RQ1.1** What are the suitable covariates to model PM10 concentration?
  - **RQ1.2** Which geostatistical interpolation method is more suitable to automate PM10 modeling?
- **RO2** Automate the PM10 concentration modeling and mapping and associated uncertainty by use of existing web services.
  - **RQ2.1** What are the rules for automatic modeling?
  - **RQ2.2** What are the steps to put the selected model on the Web?
- **RO3** Mapping of PM concentration and associated uncertainty by use of existing web services.
  - **RQ3.1** What are the steps to visualize the interpolated PM10 values and associated uncertainty by use of existing Web services?
  - **RQ3.2** What should be the resolution of the output map?

1.3 INNOVATION AIMED AT

The novelty of this research is the development of automatic modeling and mapping of air quality by applying the method that is fitted with automation.

1.4 THESIS STRUCTURE

This thesis has 8 chapters. Chapter 1 describes motivation and problem statement as well as research objectives and questions. Chapter 2 mentions related work that has been done in geostatistical modeling of air pollution, existing automatic interpolation services and visualizing methods for air quality and interpolation errors. Chapter 3 informs study area and data used in this research. Chapter 4 describes the adopted methodology for modeling air pollution and chapter 5 is about the prototype design and implementation . Chapter 6 shows the result, chapter 7 is discussion about the achieved result and chapter 8 is concludes and recommends.

Chapter 2
Related work

2.1 GEOSTATISTICS AND MODELING OF ENVIRONMENTAL VARIABLE

"Geostatistics is the area of spatial statistics that addresses prediction of unknown values at specified locations or aggregations of locations" (Holland et al., 2003). The advantage of geostatistical methods is modeling large scale variability (mean) with function of coordinates or external correlated variable and also modeling small scale variability, residuals, with spatial covariance function. Considering spatial correlation of variables results in decreasing prediction uncertainty (Holland et al., 2003). These methods generally refer to kriging and widely used by different scientist as spatial interpolation technique for modeling environmental data (e.g. soil moisture, air pollutants) for different purposes such as predicting the efficacy of emission control programs or designing of monitoring networks (EPA, 2004).

Hudson and Wackernagel (1994) applied kriging with external drift for mapping temperature in Scotland. They used elevation data as covariate (auxiliary variable). Kebaili Bargaoui and Chebbi (2009) compared ordinary kriging and kriging with external drift to model rainfall. They suggest 3-D estimation of the variogram rather than classical 2-D. In case of 3-D variogram kriging with external drift gave better result than ordinary kriging. They employed elevation to model the drift. In very admirable job by Zimmerman et al. (1999), they compared ordinary kriging, universal kriging and inverse distance weighting by use of synthetic data. They conclude that planar surface results in less uncertainty and also they mention that for some type of surfaces it is better to ignore the modeling of trend than to model it inappropriately . Lee et al. (2011) formed a linear mixed model between daily PM2.5 and calibrated AOT data which allows day-to-day variability in relationship between AOT and PM2.5. They used normal 10 km by 10 km grid for prediction over the study area. Kloog et al. (2011) extended and improved Lee et al. (2011) work by incorporating land use regression model and meteorological variables such as temperature, wind speed and elevation. Denby et al. (2010) applied log normal residual kriging with multiple linear regression to produce annual ozone and SO2 maps.

There are three ways for air quality mapping. One way is using accurate measurements of pollutants and apply some interpolation methods. The limitation is that in this method, no information about physical and chemical processes is considered . An alternative way is to use chemical transport model in which by use of physical laws and empirical relationship they describe chemical and transport process (Kassteele, 2006). Example of these models are LOTOS-EUROS, operational priority substances (OPS). LOTOS-EUROS model is a chemical transport model (CTM), also called simulator, that predict pollutants concentration, however, it severely underestimates the measured concentration, as with many other models (Denby et al., 2008). Although the outputs of these models are not accurate, they can provide important spatial information. Third way is to combine these two methods to do prediction on unmeasured points. It has been evident that combining theses two methods resulted in more accurate predictions (Kassteele, 2006).

Denby et al. (2008) and Denby et al. (2010) use LOTOS-EUROS model output and combined them with in-situ measurements to improve the predictions of PM10, ozone and SO2. What they

implemented is ordinary kriging on residuals after regression modeling of daily model concentrations and in situ measurements. Konovalov et al. (2009) and Honoré et al. (2008) use deterministic and statistical model for forecasting PM10 for one day ahead. They use statistical model in prediction correction made by a deterministic model known as model output statistics (MOS). They significantly improved the CTM output accuracy. Wackernagel et al. (2004) used kriging with external drift employing CTM outputs as covariate in modeling ozone concentration for risk management purposes. Wong et al. (2004) applied 4 different spatial interpolation on PM10 including spatial averaging, nearest neighbor, inverse distance weighting, ordinary kriging on a same study area. All the methods gave almost the same result (RMSE and MAE).

2.2 AUTOMATIC SPATIAL INTERPOLATION

When it comes to automatic modeling and mapping, not every interpolation methods can be useful. Lots of environmental monitoring networks collects data in real time but only few of them generates map automatically. This existing gap from recording data to displaying them on a map in away that is useful for users, indicates the difficulties in automatic modeling and mapping of environmental data (Brenning & Dubois, 2008). Webster and Oliver (2001) suggest that variogram modeling should be examined visually. Researchers, in order to benefit from strength of kriging methods in automatic modeling, have tried to solve the problem of automatic modeling of spatial correlation. Modeling of spatial correlation requires estimation of the so-called " variogram" parameters. In this subject INTAMAP took a step and provided a generic framework for automatic modeling of the spatial correlation. Papers by Skoien et al. (2008), Pebesma et al. (2011), De Jesus et al. (2008) describe the way that they took to tackle the problem of variogram modeling and also the improvement that has been made in this area.

"INTAMAP[1] is an interoperable framework for real time automatic mapping of critical environmental variables like air pollution by extending spatial statistical methods and employing open, web-based, data exchange and visualization" (Pebesma et al., 2011).The aim of the INTAMAP project is to provide interpolation without requiring any specified skills (Pebesma et al., 2011). In this project different generic spatial interpolation methods are defined which are ordinary kriging, copulas and trans-Gaussian kriging. These methods are suppose to work for any environmental variable so they can not support specific conditions such as including auxiliary data to model spatial variability. The INTAMAP system supports R statistical environment as an environment that supports user defined interpolation methods.

Some example of studies in automatic spatial interpolation are research by Abraham and Comrie (2004) in real time ozone mapping. What they implemented is regression interpolation hybrid approach in a way that they first subtracted mean from observed values and then ordinary kriging on the residuals. For modeling the spatial correlation, an empirically uniformed (i.e.,default) variogram model is used. Another example of attempt in automatic spatial interpolation is a large exercise which had dealt with mapping of radioactivity.The result and summary of this exercise is gathered in a document by Dubois and Galmarini (2005). Geostatistical modeling methods were successful in this exercise, however, they were not the best. The best result belongs to support vector machine algorithm.

For real time automatic mapping, other than automatic modeling, it is necessary that client can get maps on his/her platform so interopability is a main characteristics that real time automatic mapping system should have. Using web services e.g. web map services(WMS) and web processing services(WPS), that follows open geospatial consortium (OGC) standards assures this characteristic of a automatic mapping system. INTAMAP project uses WPS developed by 52North.

[1] `http:/www.intamap.org`

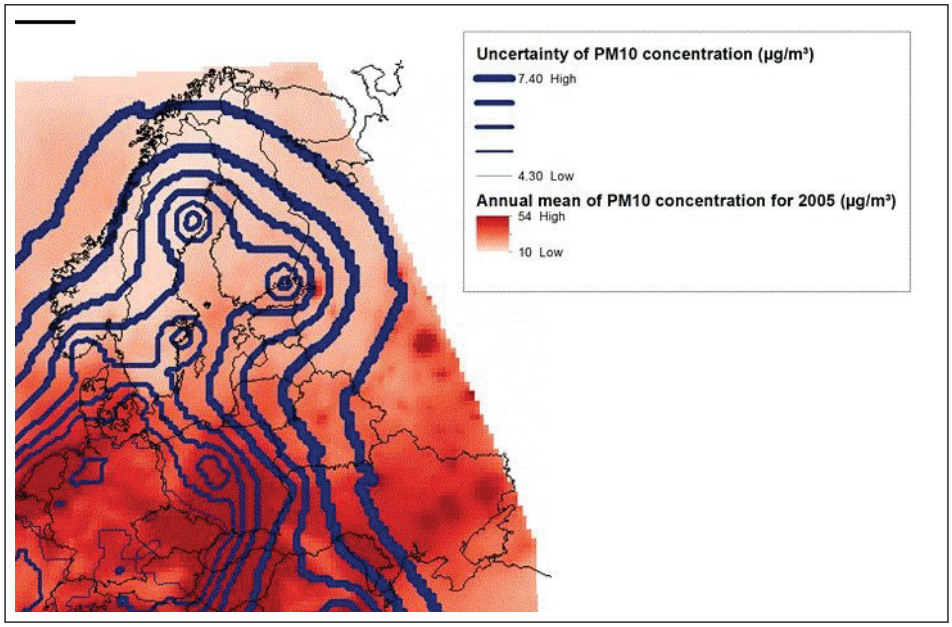


Figure 2.1: Contouring method. Pollutant concentration data is presented in the background and uncertainty in the data is represented through contours in the foreground taken from (Senaratne et al., 2012).

It also uses UncertML (a markup language to encode uncertain data) to encoding probability distribution of interpolation errors and web mapping services for visualization of interpolated values. The web client that INTAMAP provides for air quality data accepts data in form of XML(extensible markup language) and it returns interpolated values in form of XML.

2.3 VISUALIZATION

There are number of uncertainty visualization methods offered by different researcher. In general Senaratne and Gerharz (2011) categorized most spatio-temporal uncertainty visualization methods according to the parameters which are data type (continuous or categorical), data format (raster or vector), uncertainty type (positional uncertainty, temporal uncertainty or attribute uncertainty) and interaction type (static, dynamic, interactive). In this research, air pollution mapping, the uncertainty type that is considered is attribute uncertainty ,data type in continuous and data format is raster. For these parameters they offer different methods for visualization. For air quality, Senaratne et al. (2012) used two approaches (contouring method and adjustment map method) to visualize uncertainty. In contouring method 2.1, contours show uncertainty of predicted values of one pollutant throughout the region. The thickness of the lines indicates the uncertainty range . In adjustment map method 2.2, uncertainty of predicted values of one pollutant is presented using color sequences technique. In this method two side by side raster maps are used to visualize the value and the associated uncertainty side by side. Based on the result of a survey published by Senaratne et al. (2012), it was evident that contouring method and adjacent method had been better interpreted by participants.

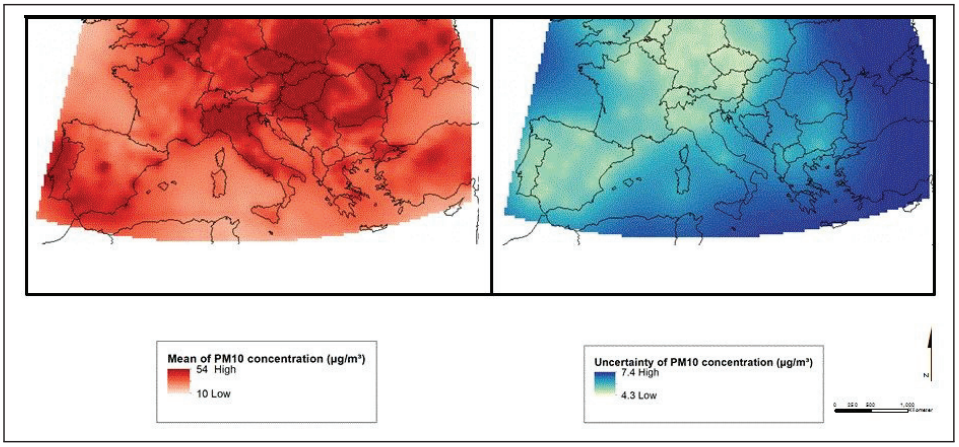


Figure 2.2: Adjacent Maps method. Pollutant concentration data (left) and uncertainty (standard deviation) of the pollutant data (right) over Europe are represented on two side by side maps taken from (Senaratne et al., 2012).

Chapter 3
Study area and data description

3.1 STUDY AREA

Study area of this research is the western part of the Europe including Germany, The Netherlands,Belgium and Luxembourg as is illustrated in figure 3.1. The area is extended from 2°E to 15°E and 47°N to 55°N.

3.2 DATA DESCRIPTION

There are three sources of data provided for this research:

- daily PM10 in situ measurements at various monitoring points;
- LOTOS-EUROS model output above monitoring points;
- station information including latitude, longitude, elevation, station code, country name, type of area i.e. rural, urban and suburban;
- LOTOS-EUROS model output for whole Europe with spatial resolution of 0.5° south to north and 0.25° west to east.

The daily in situ measurements are provided from AirBase database. PM10 concentration are recorded hourly but then are aggregated to give daily values. AirBase is the air quality database system of the European Economic Area (EEA). It contains air quality monitoring data and information submitted by the participating countries throughout Europe. Number of stations differs from one year to another. The reason is that some stations started recording during these three years and some stopped recording. In totally there are 967 stations in these four countries monitoring different pollutant concentration. However, only 292 of them recorded PM10 during 2010. Between these three years the year 2010 is selected because during this year more stations recorded PM10 in comparison with 2008 and 2009 and also the model output for the whole study area is available (see 3.1). The LOTOS-EUROS is a regional chemical transport model (CTM) designed for the assessment of gaseous and particulate air pollutants. The model is used for a wide range of scientific and regulatory supporting applications. LOTOS-EUROS [1] is built, maintained, improved by developers and researchers.

LOTOS-EUROS:

- simulates air quality over regional and sub-regional scale;
- covers an various number of pollutants (ozone, NO2, inorganic and organic PM2.5/PM10 , etc);

[1] http:/www.lotos-euros.nl/

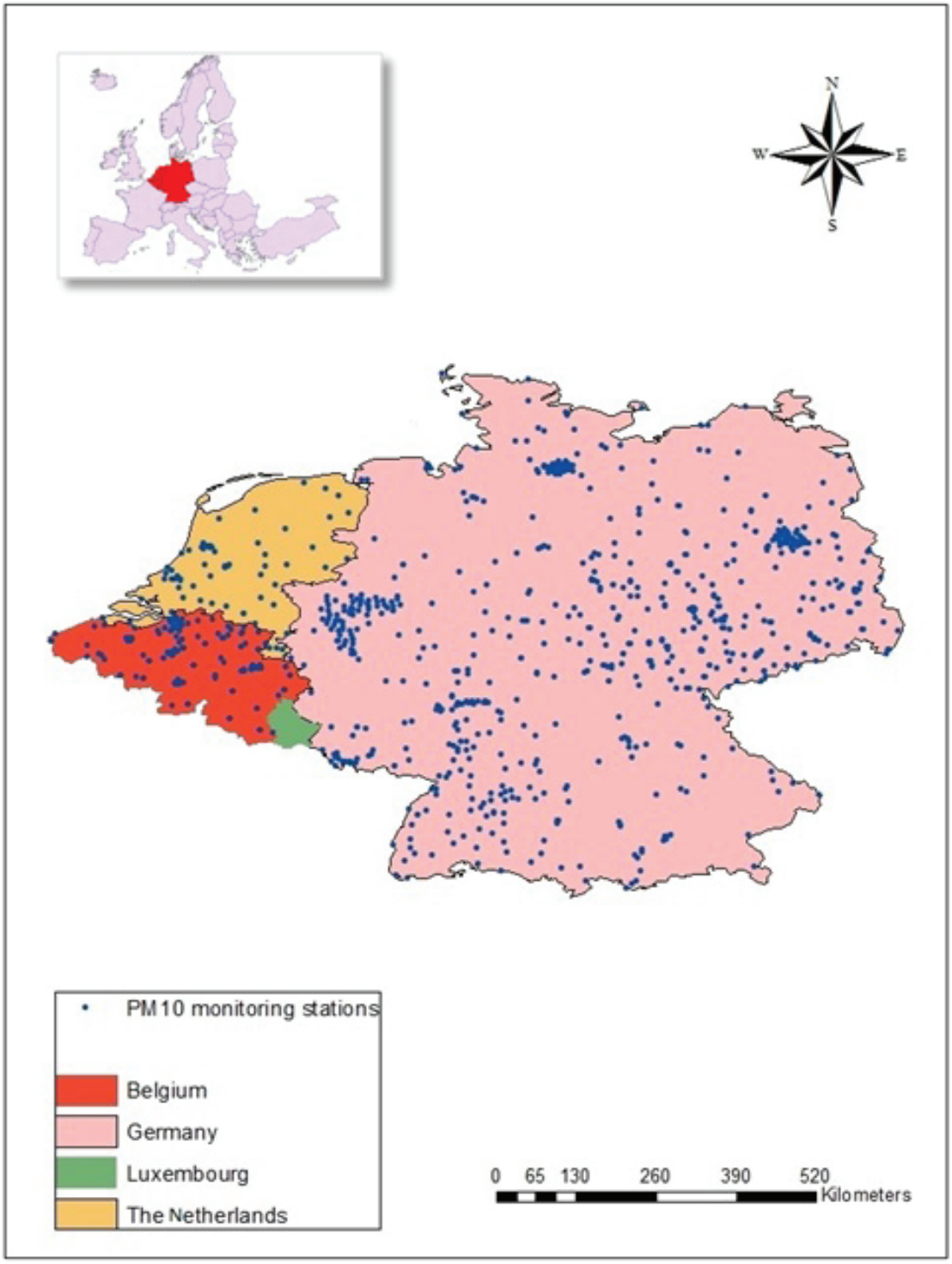


Figure 3.1: Study area and PM10 monitoring stations

Table 3.1: Summary of dataset used in the research

Country name	Number of stations recording PM10	Number of Stations in Rural Areas	Number of Stations in Urban and Suburban Areas
Belgium	40	14	26
Germany	225	61	164
The Netherlands	27	17	10
Luxembourg	0	0	0

Table 3.2: Number of stations recorded PM10 during different years

Year	**Number of stations recorded PM10**
2008	162
2009	274
2010	292

- includes data-assimilation (satellite data and ground measurements) (Schaap et al., 2008) .

It can be seen that there is no monitoring station in Luxembourg. Although there is no station, it is important to have knowledge about its air quality condition. This is done by modeling PM10 throughout the study area in which includes Luxembourg as well.In modeling PM10 all stations contribute not only those in rural areas. The reason is that the numbers of stations in rural areas are low, especially in Belgium and The Netherlands. Lots of information related to urban and suburban areas would be missed if only rural areas are considered.

In modeling PM10 all stations contribute not only those in rural areas. The reason is that the numbers of stations in rural areas are low, especially in Belgium and The Netherlands. Lots of information related to urban and suburban areas would be missed if only rural areas are considered.

Chapter 4
Methodology

4.1 DATA PREPROCESSING

4.1.1 Importing data

The provided datasets are in NetCDF (Network Common Data Form) format. The NetCDF data is a platform-independent, binary file which includes metadata exploring the contents and format of the data in the file. R has a package called "netcdf" which allows reading from, writing to, and creation of NetCDF files. “Panoply”software is used for reading, going through and understanding the data. In this research by use of "netcdf" package, data imported in to R, required data extracted and new tables created in order to be used in following analyses. The data extracted from the original file, are station latitude, station longitude, station elevation, country code, LOTOS-EUROS model output and daily PM10 value for a year 2010. Four tables are created, daily average PM10 in situ measurements, daily average LOTOS-EUROS model output in location of stations, daily average LOTOS-EUROS model output for whole study area and station information. Merging tables is done as needed during the interpolation process.

In this research, the datasets that are used are already stored in the local directory, and they are called via R scripts in the first step of the interpolation procedure. However, datasets can be imported from external servers using “sos4R” package. “sos4R” is a client for Sensor Observation Services (SOS) as specified by the Open Geospatial Consortium (OGC). It allows users to interactively create requests for near real-time observation data based on the available sensors and observations.

4.1.2 Data projection

In the original datasets station coordinates are in latitude and longitude. This coordinates system accounts for the curvature of the earth’s surface. However, statistical spatial interpolation techniques consider the linear distance between two points in space in order to define spatial correlation structure (EPA, 2004). These coordinates are transformed to European conventional. Terrestrial Reference System Lambert Azimuth Equal Area 1989 (ETRS LAEA 1989) projection reference system is selected as it is recommended to be used in Europe when statistical mapping at all scales and where true area representation is required (Annoni et al., 2001). Advantages of ETRS LAEA 1989 are:

- suitability for the whole Europe (25°W-45°E, 32°N-72°N);
- equal area representation of a given cell (pixel) throughout the raster (image/map);
- tolerable distortion of shape (Strobl et al., 2007).

4.1.3 Data cleaning

During the process, stations that recorded no value and stations with duplicate values are removed from the data in order to have proper data for modeling PM10. The reason is that the kriging interpolation can not be applied when duplicates are exist.

4.1.4 Exploratory data analysis

In this research histograms, Q-Q and bubble plots are used to summarize the main characteristics of the datasets in better understanding of the data. Last days of year 2010 from December 21 to December 31 (day 355 to 365) are selected and all the analysis applied for these days. By histograms and Q-Q plots we check if the data are normally distributed due to sensitivity of variogram calculation to highly skewed data distributions (Krige & Magri, 1982). If raw data do not follow a normal distribution, logarithm transformation should be applied to data and in the end, back transformation of the interpolated values and associated variance. Back transformation formulas are mentioned in Denby et al. (2008). To automate the process of logarithm transformation when is required , we used Shapiro-Wilk test. The Shapiro-Wilk test tests the null hypothesis that a sample came from a normally distributed population. The null hypothesis is that the sample is normally distributed. If the p-value is small (<0.01) the null hypothesis will be rejected in favor of the alternative.The Shapiro-Wilk test can be questionable due to not considering spatial correlation between observations (van de Kassteele et al., 2009). This test can be done in R. The code are mentioned in Appendix list.

In addition, average of PM10 concentration is calculated for each day to understand its variation for the whole year. In case that an unusual behavior, such as existing of extreme recorded values, is observed the corresponding day would be analyzed as well. For better understanding of the data and also to check if that was a correct decision to take both rural and urban areas in kriging procedure, the average of PM10 amount for rural and urban(urban and suburban) areas calculated separately.

4.2 DEFINE USERS TYPES AND USER'S REQUIREMENTS

It is important to know the users and their requirements from the automatic air pollution modeling and mapping system in order to design a system. In other words, it is essential to find out who are the potential users of the system before designing it to improve its usability (Senaratne et al., 2012) .

The potential users of the PM10 map and processing system are policy makers, health partitioner and public. In this research potential users are categorized in two groups. First are those who are familiar with geostatistics, modeling methods and the second groups are those who have little or no specific knowledge about these concepts. We call them advanced and normal users respectively. In addition , for user requirements, two possible scenarios are considered. First is when users upload their own air pollution data in Europe and ask for the associated map. We call this scenario as "general case". Second scenario is when user asks for a map in a same area of the research which is called "specific case ". In specific case, user either requests for PM10 map or may upload data for other pollutants such as SO2 in the same area of the study.

4.3 GEOSTATISTICAL MODELING OF PM10

4.3.1 Geostatistical modeling

In this part of the research, the objective is to apply different geostatistical interpolation methods to find out the appropriate and possible ones.

Deterministic models such as inverse distance weighting, nearest neighbor, have this limitation that they consider spatial variation among data as random variation. It means that they consider all the spatial variation has a predefined simple structure . However, stochastic models, behave with spatial variation differently in the way that they describe quantitatively how one variable value differs spatially. The general model is:

$$Z(u) = \mu(u) + S(u) + e \tag{4.1}$$

Where u denotes dependence on location, $\mu(u)$ is the deterministic trend, $S(u)$ is spatially correlated error and e is the spatially uncorrelated error. $\mu(u)$ is not dependent to location u if a constant mean can be assumed.

$$E[Z(u)] = \mu \tag{4.2}$$

What stochastic models apply is modeling the spatial correlation. Modeling of the spatial correlation is represented by variogram. Variogram values are retrieved by calculating half the expected squared difference:

$$\gamma(h) = \frac{1}{2}E[Z(u) - Z(u+h)]^2 \tag{4.3}$$

Where h is a distance, so-called 'lag', between each pair of random variables. Retrieved values from this equation results in cloud variogram. To generate an empirical variogram, partitioning the data on the distance between pairs of random variables (PM10) is necessary. Cressie (1993) suggests that number of bins should be large enough that spatial resolution retainment would be possible and yet low enough that the empirical variogram estimation be stable. Journel (1978) suggest that it should be at least 30 pairs in each bin. Another point that is considered in this research is isotropy assumption. In an isotropic field, the variation is the same in every direction. The covariance function only depends on the length of the distance vector, not on its direction. In contrast, in an anisotropic field the variation depends on both the length and on the direction of the distance vector. To check the isotropy assumption directional variogram is made in R software using "Gstat" package.

Empirical variograms do not provide all of the separation distances (h) and the corresponding semivariances needed by the kriging system. Hence, it is necessary to have a model that enables computing a variogram value for any possible separation distance.

After modeling of spatial variability, prediction on unrecorded locations is applied. Among different methods, geostatistical methods, kriging, are chosen because of their ability to produce uncertainty of predictions. The basic idea behind a kriging interpolation is to use the variogram to compute weights $\gamma(\mathbf{h})$, which minimize the variance in the estimated value. Kriging is the best linear unbiased estimator (Lark et al., 2005) because of its ability to minimize estimates of variance and mean absolute error. Following describes the different methods adopted for modeling PM10 throughout the study area.

- **Ordinary kriging**

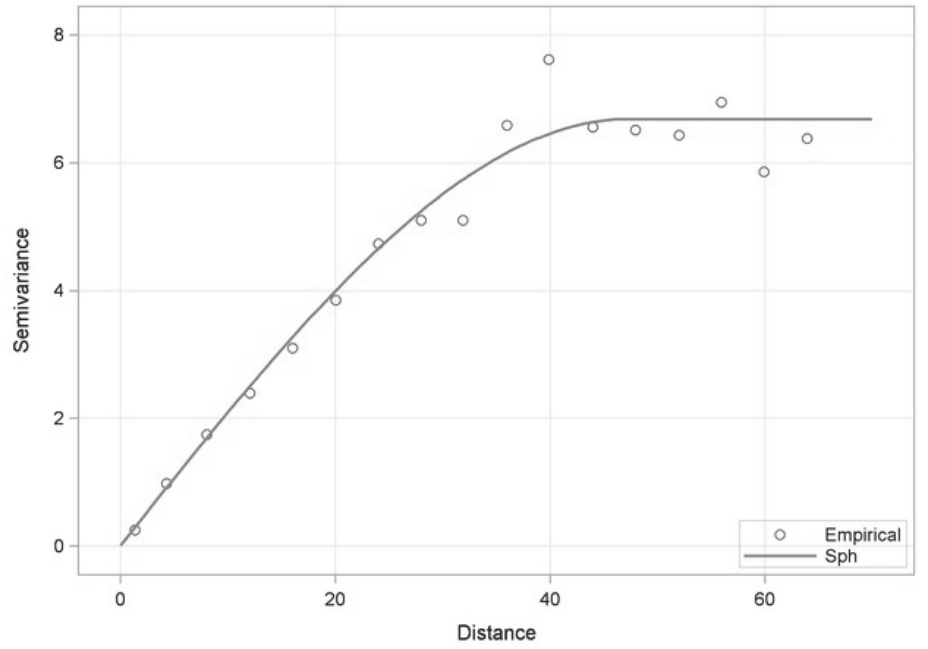


Figure 4.1: Example of empirical variogram and model variogram

In practice, Ordinary Kriging (OK) is the most widely-used interpolation method for environmental data (Webster & Oliver, 2007). In case of ordinary kriging, mean is assumed to be constant over study area so the general model in Equation (4.1) transform to:

$$Z(u) = \mu + S(u) + e \tag{4.4}$$

Where μ is not dependent on location u. The parameters that should be estimated are mean (μ) which is considered to be constant and variogram model parameters, which are nugget, sill and range. The ordinary kriging predictor is calculated:

$$\hat{T_0} = \hat{\mu} + c_0^T \, C^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}) \tag{4.5}$$

Where: **y** is the vector of observations of length n, c_0 is a vector of length n giving the covariance between all observations and the unsampled location, C is the $n \times n$ covariance matrix and **1** is a vector of 1' of length n. Ordinary kriging variance is retrieved from.

$$\hat{\sigma^2}(\hat{T} - T) = C(0) - c_0^T \, C^{-1} \, c_0 + (1 - c_0^T \, C^{-1}1)^T \, (1^T \, C^{-1}1)^{-1} \, (1 - c_0^T \, C^{-1}1) \tag{4.6}$$

First term in equation (4.6) indicates the covariance in lag zero. Second term reduces uncertainty in prediction points by use of correlation with neighboring points, and the third term increases the uncertainty because of the uncertainty in estimating the mean.

- **Hybrid Kriging**

Universal kriging and kriging with external drift (KED) are also called "hybrid"; non stationary(of the mean) geostatistical approaches (McBratney et al., 2000).

UK and KED are special case of kriging in which the drift defined through some auxiliary variables (covariate). They can be considered as a bivariate regression taking into account the spatial autocorrelation of the dependent variable. Hybrid kriging may be considered when there

is a former knowledge that a large-scale relationship exists in the target area (EPA, 2004). In UK, coordinates are auxiliary variable, so mean (large-scale variability) is modeled as a function of coordinates, and in KED, is modeled externally through some auxiliary variable (Hengl et al., 2003). The mean term in equation 4.1 is modeled by covariates so:

$$Z(u) = X\beta + S(u) + e \tag{4.7}$$

Where:

X is covariates' value and β is regression coefficients. The parameters to be estimated are trend coefficients and variogram parameters.

Regression analysis is done to find out the suitable auxiliary data which can be used in kriging procedure. In provided datasets, LOTOS-EUROS model output and coordinates are the data that are considered as covariates. The LOTOS-EUROS model has a spatial resolution of 35 km by 25 km. The selected grid size for mapping is 5 km by 5 km. The covariates values are required for every prediction points in order to apply KED, so we resampled LOTOS-EUROS model output into 5km by 5 km grids.

4.3.2 Automatic variogram modeling

Automation of the fitting variogram model to empirical variogram was the main obstacle for researchers to reach to the automatic modeling and mapping of environmental data. As mentioned earlier, user possible requirements categorized to general and specific case.

- Specific case

For a specific case, because the study area is fixed and spatial distribution and location of the stations that are recording PM10 are almost the same, the defined variogram modeling provided by "Gstat" package in R is used. By use of "Gstat" package, calculating empirical variogram and fitting a model to it can be applied automatically. Different cuts- off are examined to find out which one results in better variogram in the sense that it reaches to clear sill and also the minimum number of pairs in be more than 30 in each lag as Journel (1978) suggests. For fitting a model to empirical variogram, initial values for model parameters (nugget, sill, and range) are required. The procedure of variogram modeling in "Gstat" is done by iterative reweighted least squares so called Gauss-Newton fitting (Cressie, 1993). In this method For fitting a model to empirical variogram, initial values for model parameters (nugget, sill, and range) are required. Eleven days of data are analyzed, and empirical variograms are formed to find out the possible initial values for modeling. After fitting a model to each day's empirical variogram, initial values are selected and these values are considered for all remaining days of the year 2010. In addition, three common models; Spherical, exponential and Gaussian are applied . After analyzing the empirical variogram and fitted model the one with the minimum error sum of squares for 10 days is chosen. Choosing "Gstat package" gives us the ability to include auxiliary variable into the prediction process.

- General case

For the general case, in which users import their own data, the "automap" package, which apply automatic interpolation, is used. This package forms empirical variogram, fits a model to it and makes prediction in a automatic way. The method that "automap" uses for prediction is ordinary kriging, how ever it can enlist auxiliary variables for applying kriging with external drift and universal kriging. The reason behind selecting "automap" is the limitations that come with the

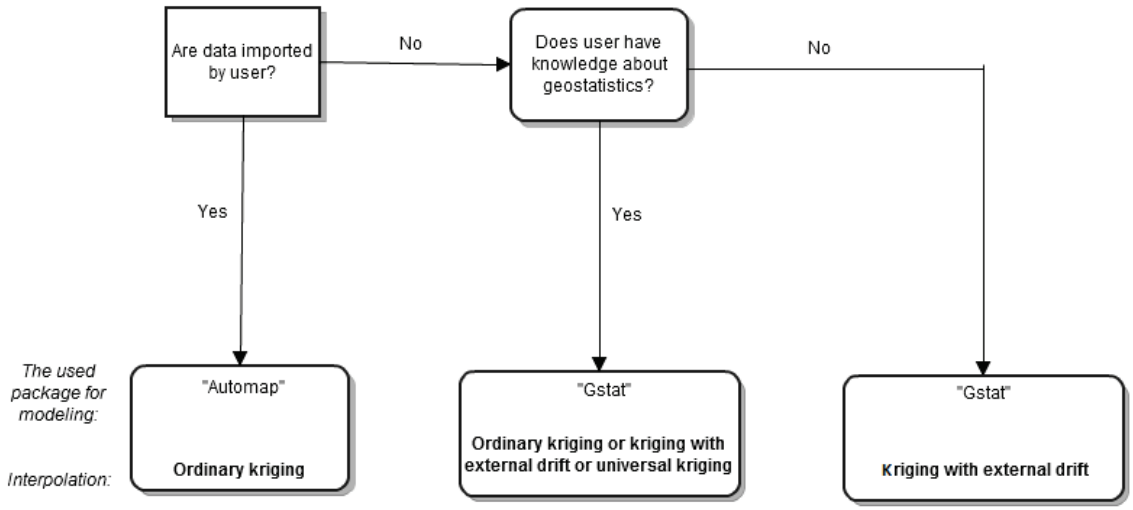


Figure 4.2: Decision tree for variogram modeling and interpolation methods choices that take place in R

uploaded data. Different data from various stations definitely have significantly different spatial characteristics so the suitable model variogram for them differs. Based on the spatial distribution and variogram values, “automap” fits a model to empirical variogram. For Automatic, omnidirectional variogram model fitting “automap” makes following choices:

- Cut off : %35 maximum distance

initial values for the fit :

- Sill: mean of the maximum and median value of the variogram values
- Nugget: minimum value of the variogram values
- Range: cutoff/3.5
- Model : Exponential, Spherical, Gaussian , the one with the minimum error sum of squares is selected

4.3.3 Spatial aggregation

Air pollution stations’ measurements are considered as point data. The spatial representativeness of each station is different from another due to their locations. Stations located in rural areas are representative of larger area than those are in urban or suburban areas. The spatial support of each station depends on a emission sources surrounding it (Spangl et al., 2007). Users, such as decision makers, may need to know pollutants’ (e.g.PM10) concentration for large areas. Changing scale from point to a area , from finer resolution to coarser resolution, is called upscaling (Van Bodegom et al., 2002). Block kriging is method to do this upscaling (Stein et al., 2001) to retrieve values from a point to a block. Block kriging is a specific case of kriging which make prediction on larger

areas by averaging. Besides the ability of this method to predict on blocks, it has been proven that in many environmental applications, block kriging usually exhibits smaller prediction errors than punctual kriging (Bivand et al., 2008). Block kriging can be easily applied by defining the block size in prediction command of "Gstat" package. What should be a block size is another issue. Apart from user requirements, selected grid size should not be computationally demanding (Hengl, 2006). Based on a report by Kuhlbusch et al. (2006), the resolution of air quality map should not be less than 5 km by 5 km as it may be needed for further health analyses. However, they suggest the resolution of 1 km by 1 km but grid size of 1 km by 1km in this study is time consuming. The prediction time in case of ordinary kriging is about 5 minutes. Hence, the grid resolution or block size considered in this study is 5 km by 5 km.

4.3.4 Accuracy assessment

Validating the models is required to check its quality. Cross validation is a technique that is applied in this study for accuracy assessment, as it can easily be applied automatically by use of "Gstat" package in R. Through cross-validation, a value for an observed point, based on all other data except that point, is predicted and then the predicted value is compared to the measured value. This procedure is applied for all the monitored points. The result of cross validation In R is a "spatial point data frame" in which stations coordinate, associated residual, prediction on monitored points and associated uncertainty are stored. Root mean square error(RMSE), Mean Error(ME) and Mean Square Deviation Ratio(MSDR) of the residuals to the prediction errors, can be calculated from the cross validation's result from following formulas:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z^*(u_i) - Z(u_i))^2}; \tag{4.8}$$

$$ME = \frac{1}{N}\sum_{i=1}^{N} Z^*(u_i) - Z(u_i); \tag{4.9}$$

and

$$MSDR = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(Z^*(u_i) - Z(u_i))^2}{\hat{\sigma^2(u_i)}}} \tag{4.10}$$

Where:

$\sigma^2(u_i)$ is kriging variance. These are the diagnostic measures for models evaluation. MSDR ideally should be 1 because the residuals from cross-validation should equal to prediction errors at each point that was cross out. ME should be ideally equal to 0 because kriging is the best linear unbiased predictor,however, ME is a weak diagnostic as it is not sensitive to choice of variogram modeling (Webster & Oliver, 2007).

Chapter 5
Prototype design and implementation

5.1 TECHNICAL SET-UP

One of the objectives for automatic modeling and mapping is to make these maps available for those who needs them but they have no specific knowledge about the procedures that results in map such as variogram modeling and kriging techniques (Pebesma et al., 2011). One of the other requirements is that, as mentioned earlier, users should be able to retrieve an air pollution map in real time, so they can make decisions on time. The Internet is an efficient way to enable users to access the information they need at their fingertips (Sheng et al., 2008). It is required for a web-based automatic interpolation service to be applicable in every platform so users do not need to have any specific software installed on their systems. This characteristic of a system is called interoperability.

If a system follows a service oriented architecture(SOA), it guaranteed its interoperability. A SOA is a collection of services that allows a web client to access and run a service and receive results from it (Izza et al., 2008). Therefore, geostatistical modeling can be considered client-service interaction that a client requests a map and the request goes through interpolation service located on a server and the result in the form of a map will appear on the client interface (De Jesus et al., 2008).

This study has three main stages including modeling PM10 which is done in R, automation of mapping procedure and automatic visualization of output. Geostatistical modeling and producing output take place in R environment. The output can be exported in raster(GeoTIFF) or vector(lines or GML) format.

The whole process of automatic mapping PM10 is displayed in figure 5.1 . This process starts with a request of PM10 map on a web-client user interface, and it ends with a map displaying on the client interface. Users select the date, desirable output format, e.g. GML, tiff, jpeg, etc., the specific location in which they want to know PM10 concentration and also the interpolation method that he/she desires such as ordinary kriging, universal kriging or kriging with external drift.

R codes should be placed in a server in order to produce interpolated values and associated uncertainty in a moment that user send a request. Web processing services (WPS) make this connection between user request and the algorithms. WPS is a web service for standardized processing of geodata which allows a client to benefit from preprogrammed calculations and computation models that operate on spatially referenced data over a net. The WPS used in this research is an open-source WPS provided by 52North[1]. The implementation of it is based on the current OpenGIS specification defined by the Open Geospatial Consortium (OGC)[2].

52North prepared a module so called “WPS4R” which allows WPS process creation via the R-scripts. The connection between R and WPS is made by “Rserve”. Rserve is an independent Transmission Control Protocol/Internet Protocol (TCP / IP) Server, available as an R package. R

[1] http:/52north.org/

[2] http:/www.opengeospatial.org

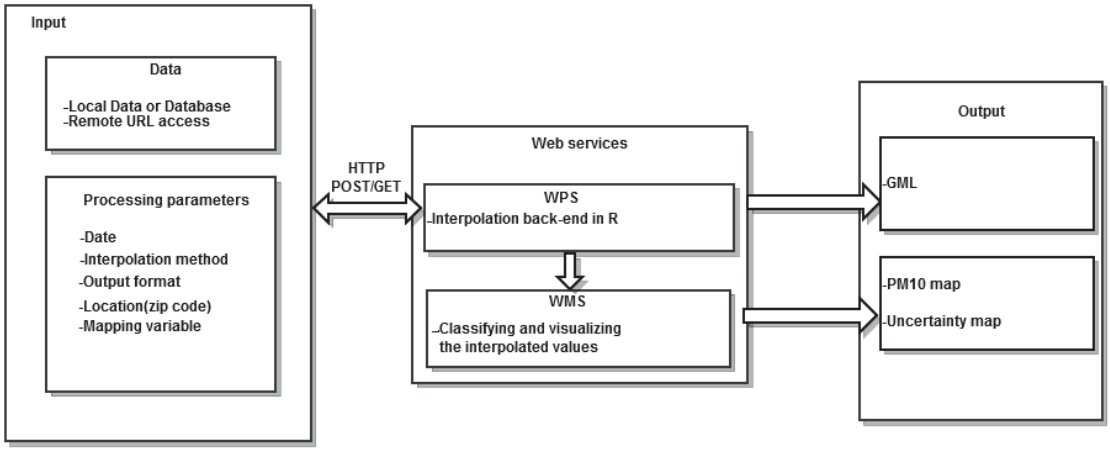


Figure 5.1: Technical set-up of the automatic interpolation service

scripts contain annotations including supplies process description, input and output information. The process description ,input and output format are coded like comments in R environment. The example of annotated R script is mentioned in Appendix. After the R scripts annotated properly, they should be uploaded to WPS (see figure 5.2).

We can get interpolated values directly from WPS in different formats such as geography markup language (GML) or GeoTIFF (see figure 5.4). Output in GML format maybe understandable for some users but not for all (see figure 5.3). GML output includes 2 bands of data; interpolated values and uncertainty of predictions. To visualize the output in a proper way, the GeoTIFF file produced through WPS is served to web map service (WMS). WMS, another web service used in this research, is a standard protocol for serving georeferenced map images over the Internet . These maps are generated by a map server. The specification was developed and published by the OGC . The WMS is used in this study is MapServer[3] which is an Open Source platform for publishing spatial data and interactive mapping applications to the web. By use of MapServer the GeoTIFF file including PM10 interpolated values and associated uncertainty are classified to different classes and styled.

5.2 DESIGN OF USER INTERFACE

The objective of the research is providing PM10 map in real time for a user through a web- client interface while the modeling and mapping procedure happens in back-end. The user interface should have the ability to serve the two groups described in this study(advance and normal group). To do so, two interface are designed one for general group and one for specific group. In figure 5.5 the possible user interface for advanced group is displayed. For general users the facilities of interface are limited to date, mapping variable and insert location. The design of user interface is done by HTML and Java script coding.

[3] http:/mapserver.org/

Figure 5.2: Deploying a WPS4R process

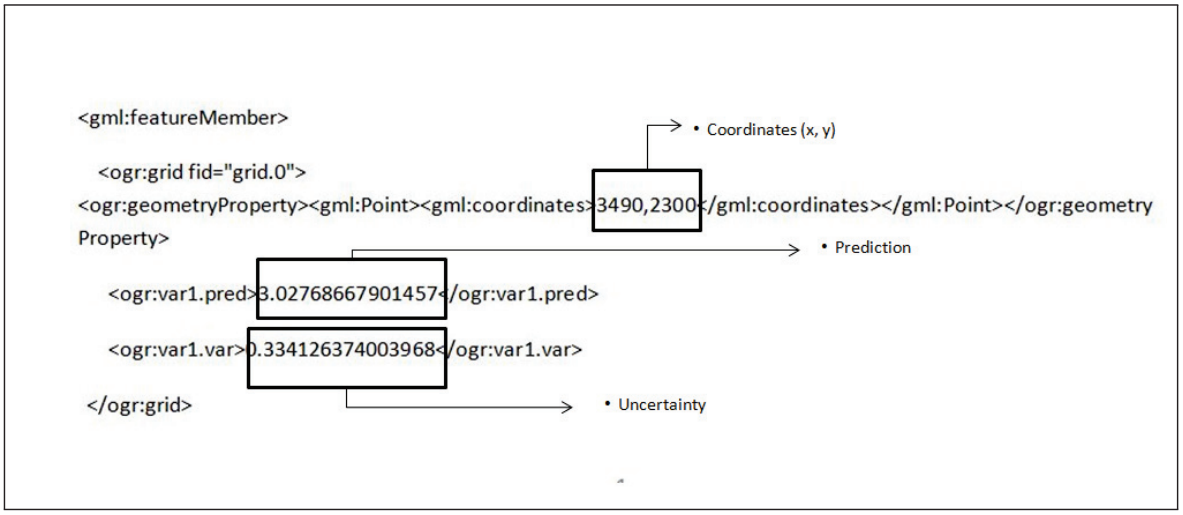


Figure 5.3: Output of WPS in GML format for advanced users.

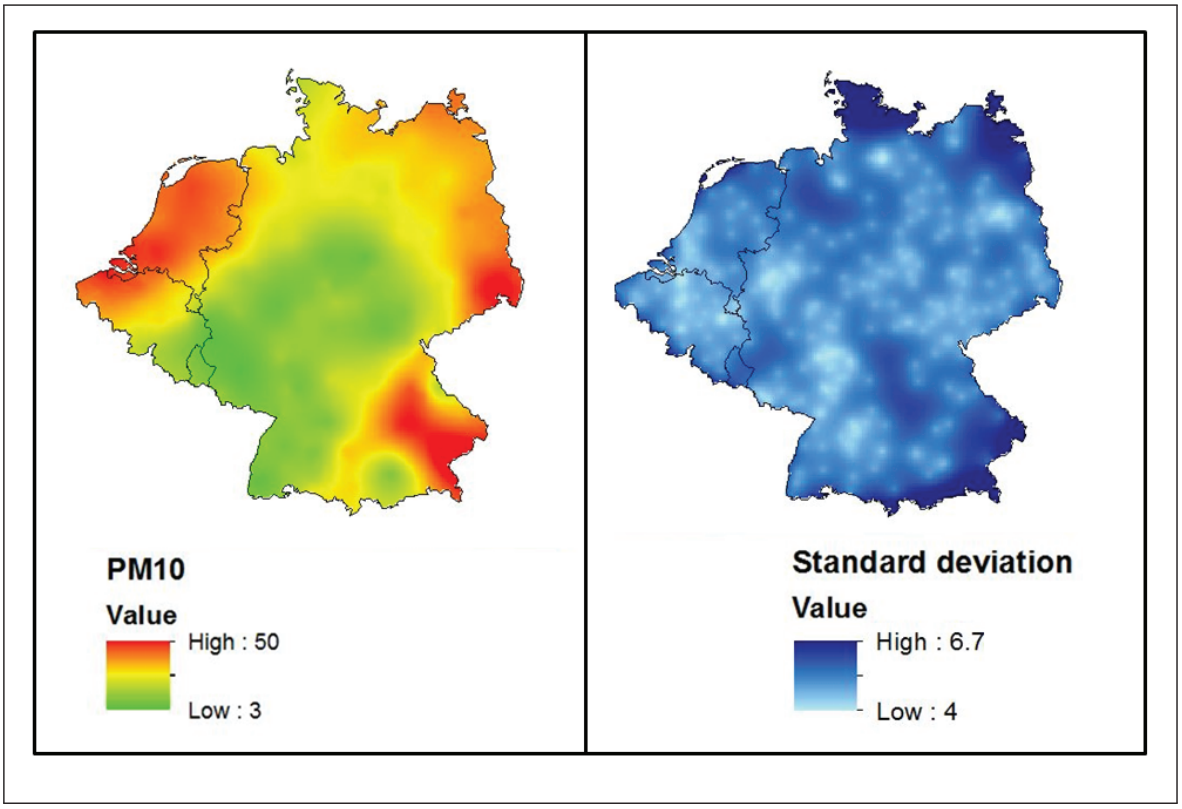


Figure 5.4: Adjacent maps map method to show the uncertainty

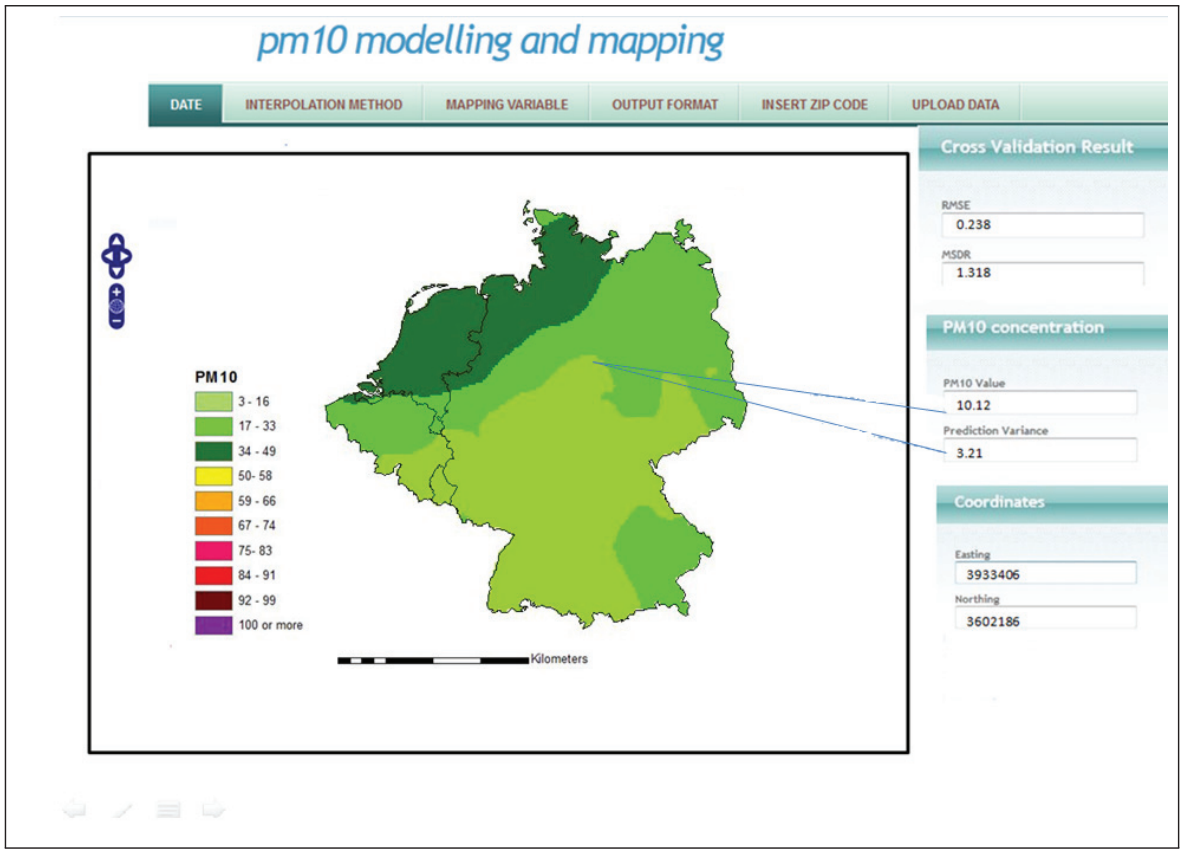


Figure 5.5: User Interface for advanced group

Table 5.1: Boundaries Between Index Points for PM10 taken from united kingdom department of environment and rural affairs website [5]

Index	1	2	3	4	5	6	7	8	9	10
Band	Low	Low	Low	Moderate	Moderate	Moderate	High	High	High	Very high
μm^{-3}	0-16	17-33	34-49	50-58	59-66	67-74	75-83	84-91	92-99	100 or more

Table 5.2: Health advice to accompany the Daily Air Quality Index taken from united kingdom department of environment and rural affairs website (`http://uk-air.defra.gov.uk/`)

Air pollution banding	Index	Health message for At-risk individuals	Health message for general Population
Low	1-3	Enjoy your usual outdoor activities.	Enjoy your usual outdoor activities.
Moderate	4-6	Adults and children with lung problems, and adults with heart problems, who experience symptoms, should consider reducing strenuous physical activity, particularly outdoors.	Enjoy your usual outdoor activities.
High	7-9	Adults and children with lung problems, and adults with heart problems, should reduce strenuous physical exertion, particularly outdoors, and particularly if they experience symptoms. People with asthma may find they need to use their reliever inhaler more often. Older people should also reduce physical exertion.	Anyone experiencing discomfort such as sore eyes, cough or sore throat should consider reducing activity, particularly outdoors.
Very high	10	Adults and children with lung problems, adults with heart problems, and older people, should avoid strenuous physical activity. People with asthma may find they need to use their reliever inhaler more often.	Reduce physical exertion, particularly outdoors, especially if you experience symptoms such as cough or sore throat.

5.3 VISUALIZATION

We use colors to display different PM10 concentration values when only PM10 map is required by the client. Colors are defined in a same way as suggested by (COMEAP, 2009). This index has ten points categorized to 4 levels which are low, moderate, high and very high. For each of the levels there is one advice for at-risk groups and one advice for the general population. At risk individuals are adults and children with heart or lung problems(see tables 5.1and 5.2).

When user asks for PM10 map and associated uncertainty, We use adjacent maps to visualize uncertainties as suggested by Senaratne et al. (2012). In the uncertainty map (right), uncertainty is represented through shades of blue. Darker blue represents higher uncertainties and lighter blue represents lower uncertainties(see figure 5.4).

Chapter 6
Result

In this chapter the result from the adopted method are displayed.

6.1 EXPLORATORY DATA RESULT

In this section the result of quantitative and descriptive data analysis are illustrated. The figure 6.1 displays the average of PM10 amount recorded by all the stations over the study area. It is obvious that day 1, 1/1/2010, has unusual behavior in comparison with other days. The bubble plots in figure 6.2 (a) also indicate this. There are some stations in this day that the recorded PM10 value is more than 200 and in one case even 300. Figures 6.2(b) and 6.3(a , b) demonstrate the variation of PM10 throughout the study area for day 355, 356 and 357. Corresponding histograms and normal Q-Q plots are displayed in figures 6.4 and 6.5. Since the data are skewed, the logarithmic transformation is applied to the raw data. The histograms and Q-Q plots of logarithm-transformed data are represented in figures 6.4 and 6.5.

Table 6.1 shows the minimum, maximum and mean of PM10 concentration in rural and non rural(urban and suburban) areas. It is inferred from this table that the difference between PM10 values in rural and urban areas are low in December day 355 ,356 and 357 and but it is high in day 1. Figure 6.6 displays the histograms of rural and urban (urban and suburban) areas separately and for all stations in 1/1/2010. It is also displays the Q-Q plot for raw and logarithm transformed PM10 concentration.

6.2 GEOSTATISTICAL MODELING

6.2.1 General case

In general case, where users upload their own data, automatic variogram modeling is applied by use of “automap” package. Figure 6.7 is the illustration of the produced empirical and model variograms for the four days of the data. Table 6.2 represents the model variogram parameters estimation by “automap” and the cross validation result for each day (day1,355,356 and 357 of 2010).

Table 6.1: Variation of PM10 recorded values in rural and urban areas

Day	Rural Min	Rural Mean	Rural Max	Urban Min	Urban Mean	Urban Max
Day 1	2	67	229	19	102	301
Day 354	7	27	63	13	32	76
Day 355	1	16	36	8	19	47
Day 356	1	18	60	5	19	44
Day 357	2	17	36	6	19	41

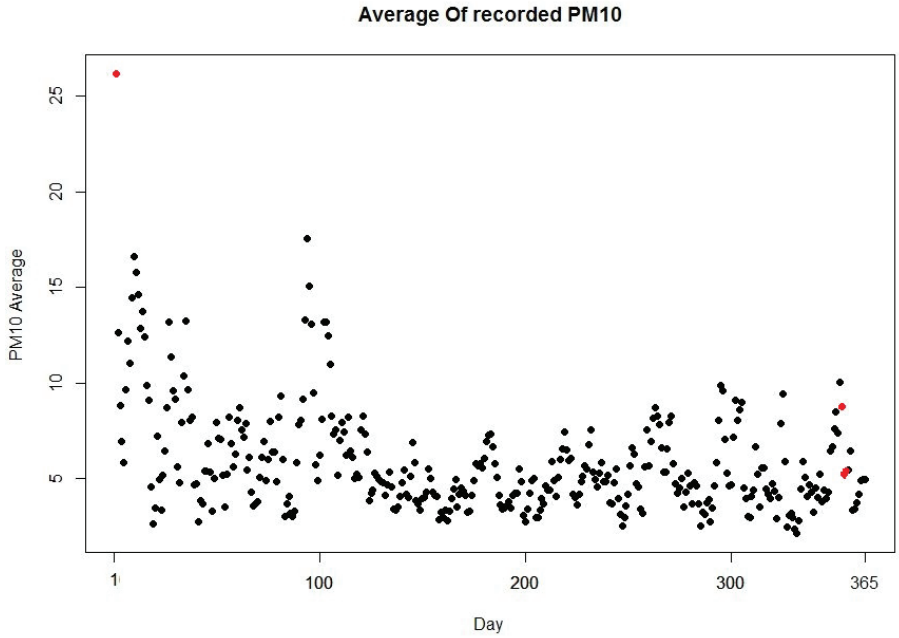


Figure 6.1: Scatter plot showing the average of the all stations' recorded PM10 for each day of the year 2010.

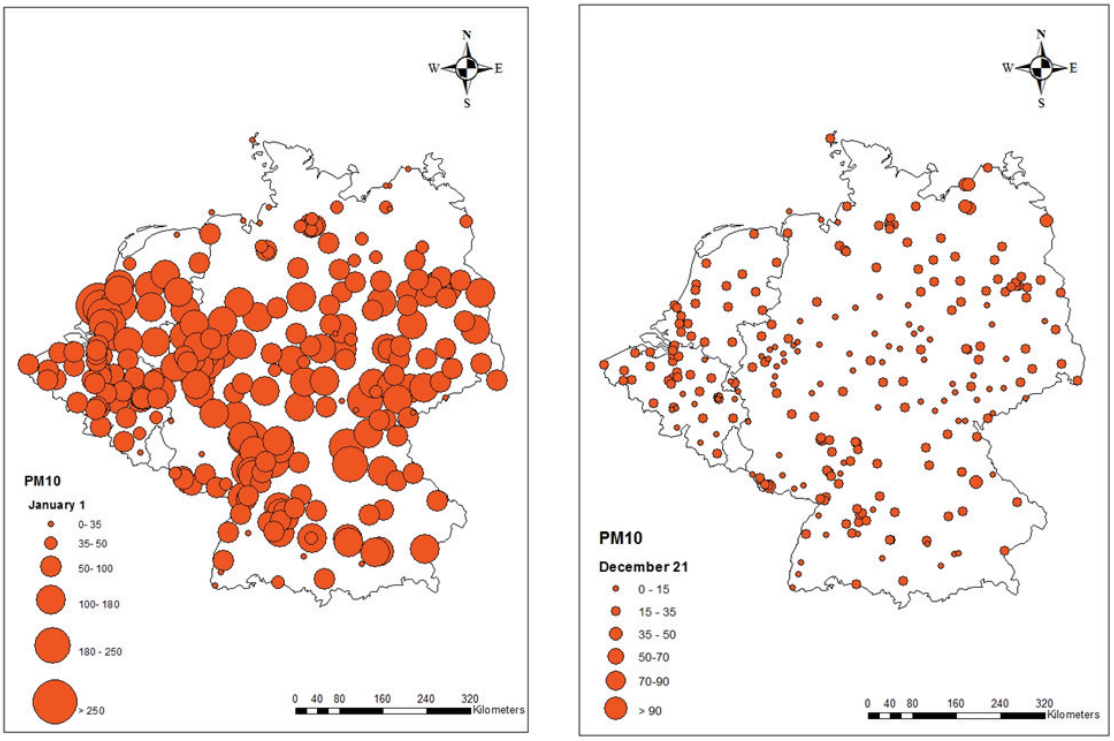


Figure 6.2: PM10 variation over the study area a) 1 January (day1) and b) 21 December (day355)

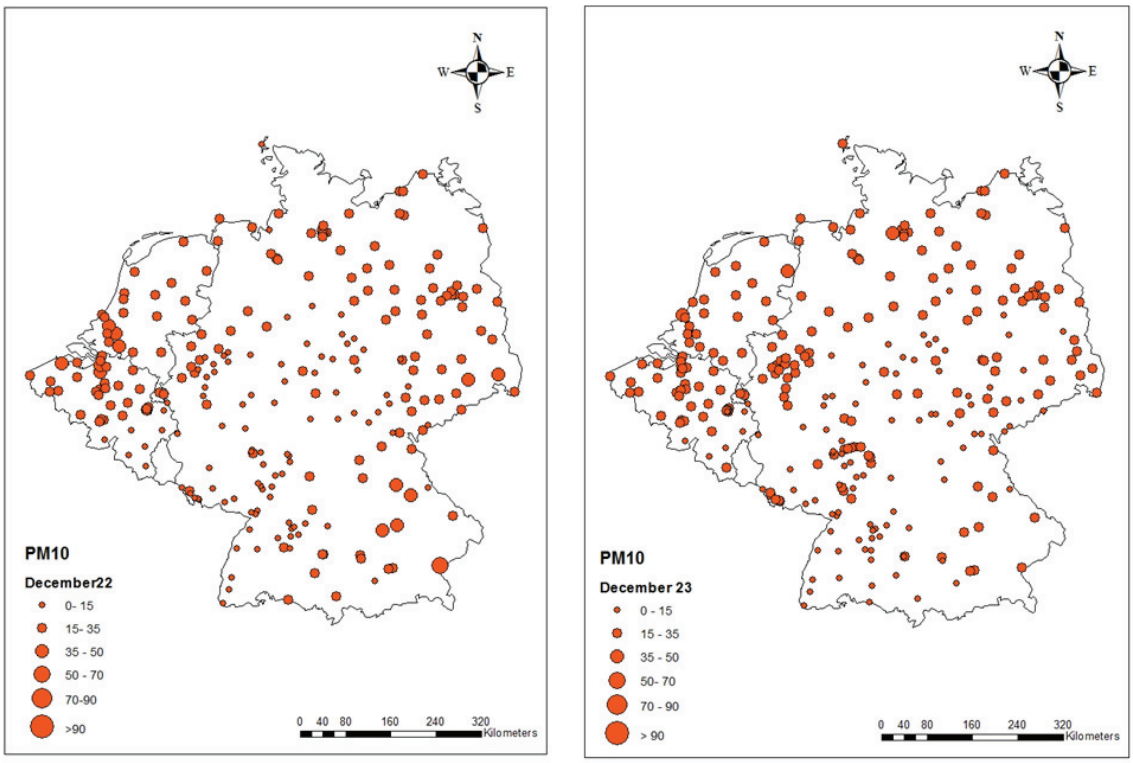


Figure 6.3: PM10 variation over the study area a) 22 December (day356)and b) 23 December (day357)

Table 6.2: Summary Of “automap” automatic variogram modeling. Psill and MinNP stand for Partial sill and minimum number of pairs among different lags respectively.

year 2010	SSErr	RMSE	ME	MinNP	Nugget	Psill	Range	Model
Day 1	0.004	0.434	-0.008	46	0.0416	1.52	7840	Exponential
Day 355	0.00077	0.352	-0.001	50	0.202	0.202	121	Exponential
Day 356	0.00037	0.398	-0.003	49	0.044	0.7	536	Exponential
Day 357	0.0015	0.386	-0.001	52	0.019	0.274	200	Exponential

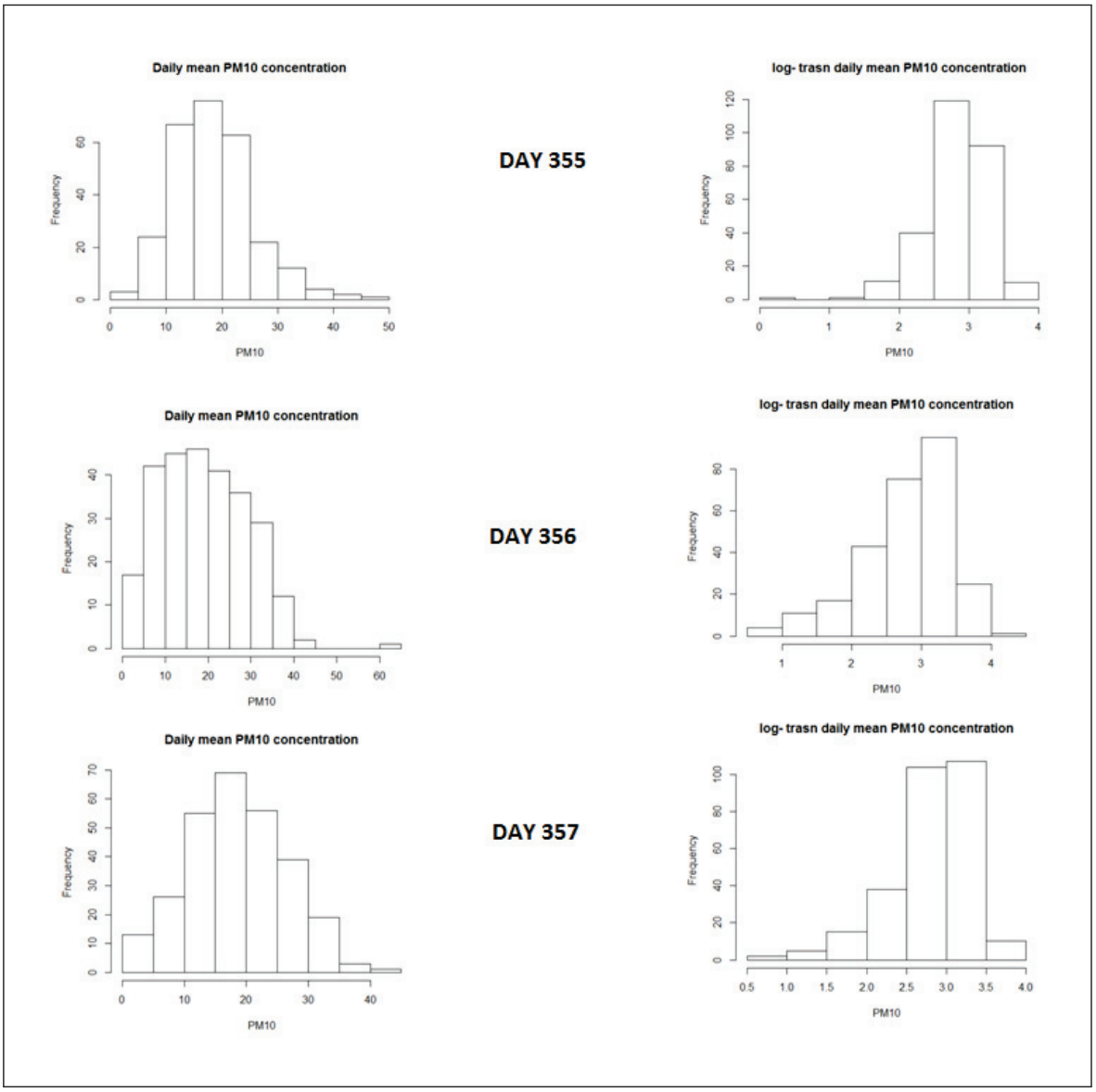


Figure 6.4: Histogram of Raw and Logarithm transformed PM10 for day 355,356 and 357

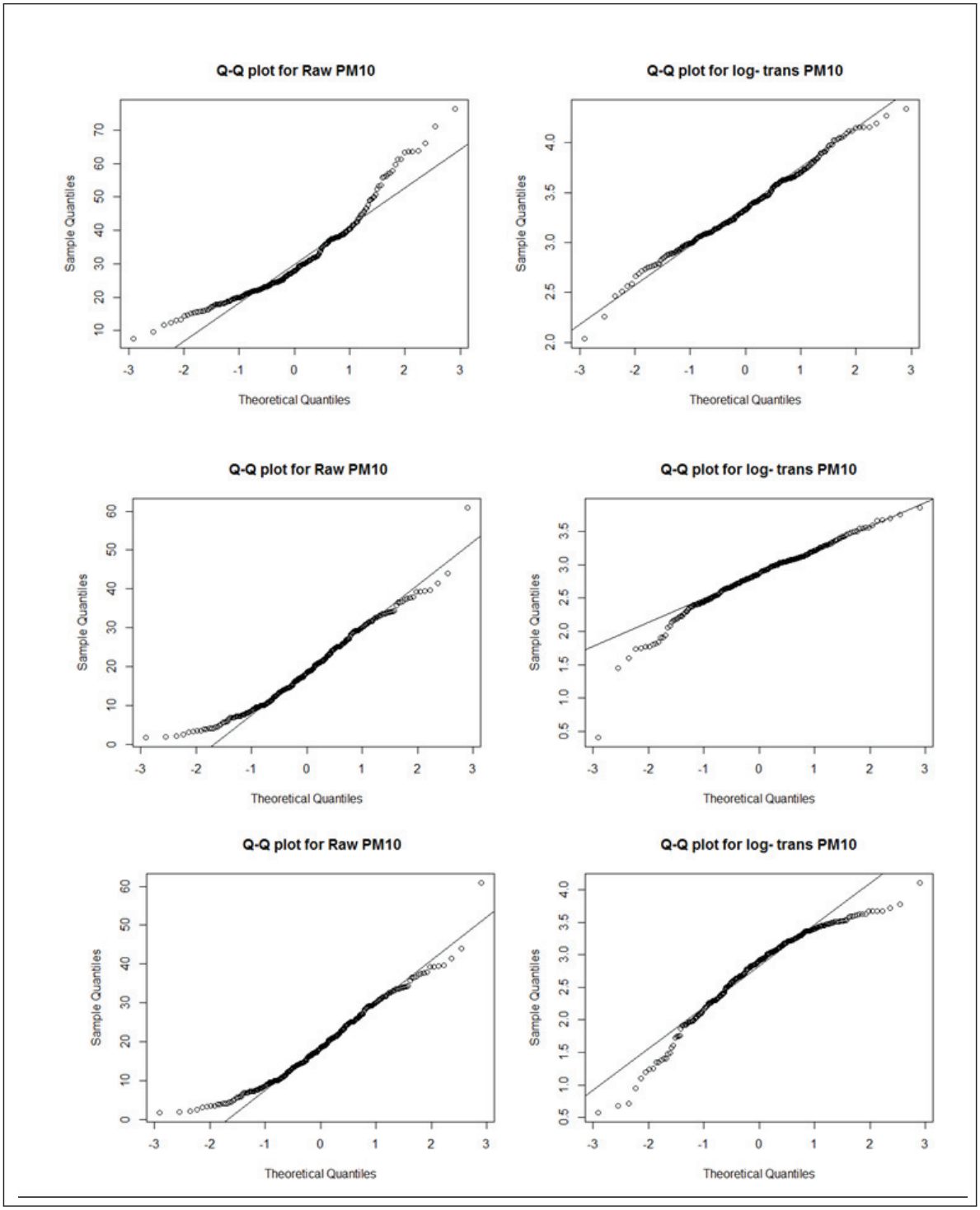


Figure 6.5: Q-Q plots of Raw and Logarithm transformed PM10 for day 355,356 aaadn 357

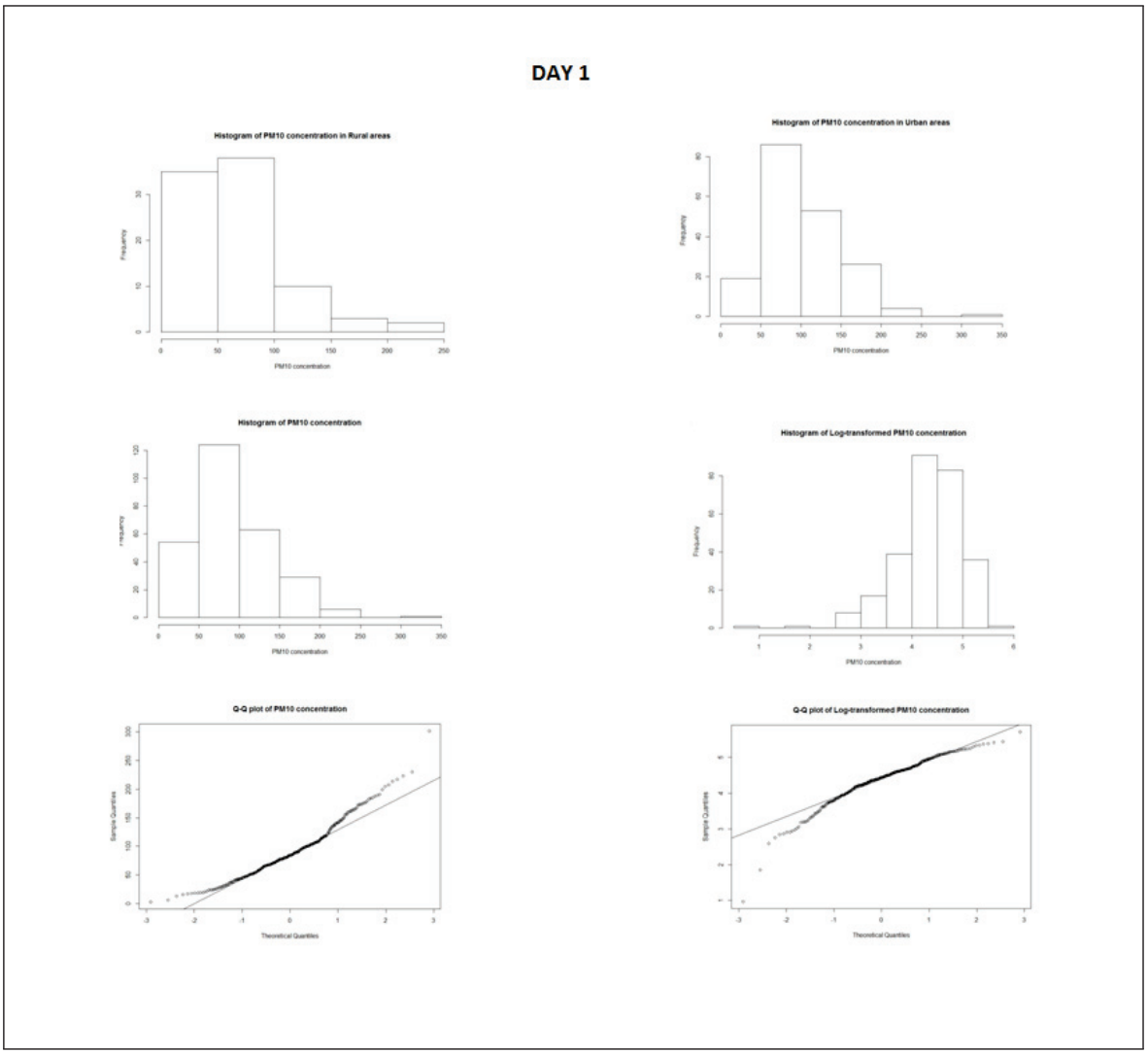


Figure 6.6: Histogram of PM10 concentration in a)Rural areas ,b)urban areas c)for all the stations Raw data,d)logarithm transformed of all the stations recorded PM10 value.e) is Q-Q plots of(c) and f) is Q-Q plots of(d).

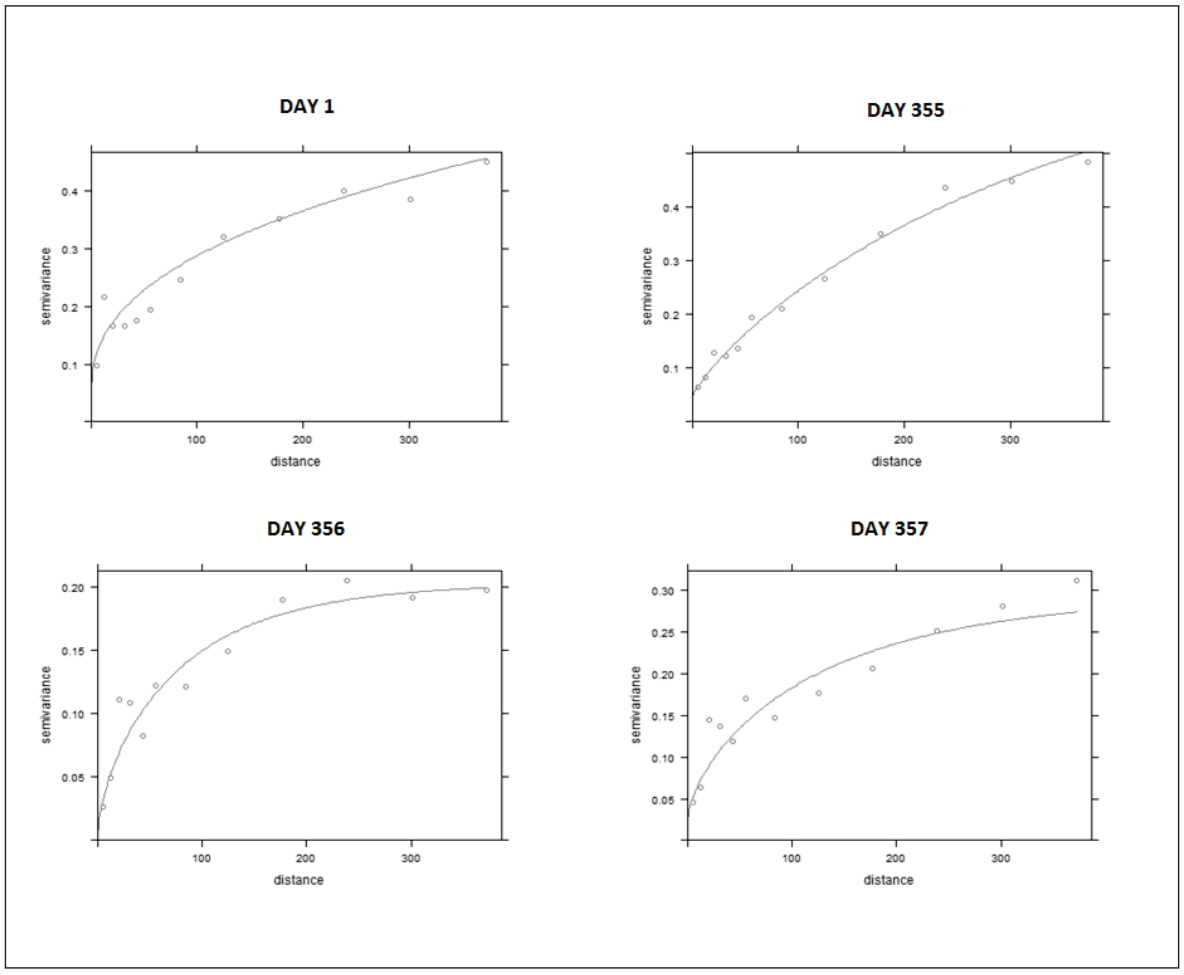


Figure 6.7: Automap ’s empirical variograms and fitted models

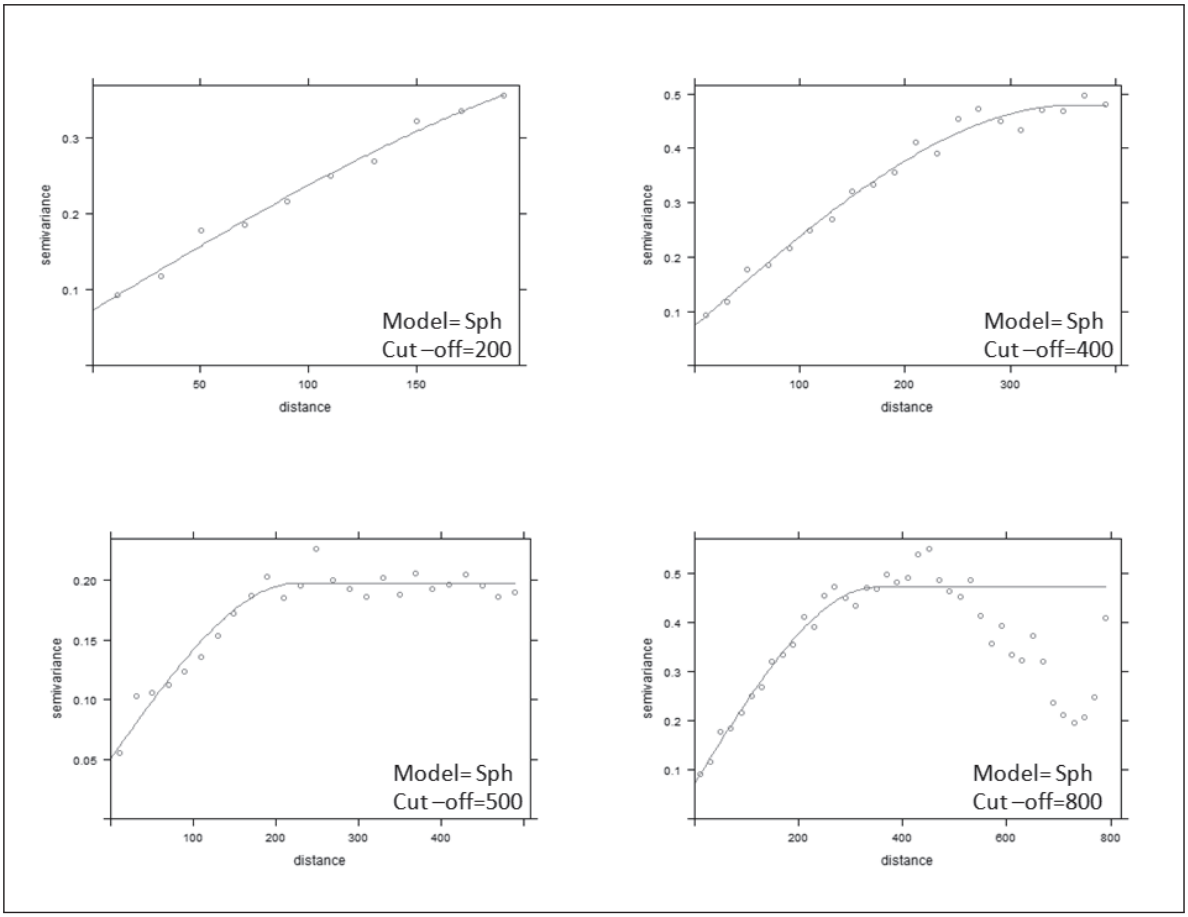


Figure 6.8: Empirical and Spherical(Sph) model variograms for different cut-offs

6.2.2 Specific case

In specific case, user requests for a PM10 map or any other pollutants in the predefined area (Germany, Belgium, The Netherlands and Luxembourg). Different cuts off (200 km, 400 km, 500 km and 800 km) are examined to find the one which reaches to clear sill and also the minimum number of pairs is checked to be more than 30. Spherical model is selected to be fitted to empirical variograms. Figure 6.8 shows the variograms of different cut offs for day 356 and table 6.3 is representation of the impact of this selection in value of RMSE, ME and MSDR and model parameters. The result for finding proper cut off in case of universal kriging and kriging with external drift are presented in Appendix.

What can be concluded from table 6.3 and figure 6.8 is that the cutoff equal to 500 km is the better choice as the fitted model reaches to clear sill and the minimum number of pairs is 168. In both cases, when cut off is 200 km and 400 km, model variogram do not reach to the clear sill and when cut off is 800 km the minimum number of pairs is less than 30. The RMSE resulted from various cut offs, are not significantly different, however, it is less in case of 500 km.

6.2.3 Choice of variogram model and initial values

The common models (spherical, exponential and Gaussian) are selected and fitted to empirical variogram to determine which one has less SSErr and MSDR. The table 6.4 shows the result for

Table 6.3: Summary for the impact of different cut-off selection

22 December (day 356)	SSErr	MSDR	RMSE	ME	MinNP	Nugget	sill	Range
cut-off=200/ width=20	0.0001	1.155	0.398	-0.002	160	0.07	0.834	451
cut-off=400/ width=20	0.0002	1.148	0.397	-0.002	160	0.05	0.629	310
cut-off=500/ width=20	0.0003	1.146	0.397	0.0008	168	0.050	0.147	226
cut-off=800/ width=20	0.0008	0.396	0.396	-0.002	26	0.061	0.478	201

Table 6.4: Summary for the SSErr retrieved from different models

Model	SSErr/ 21 December	SSErr/ 22 December	SSErr/ 23 December
Exponential	0.0002	0.0003	0.0007
Spherical	0.0003	0.0002	0.0009
Gaussian	0.0009	0.002	0.002

day 355, 356 and 357, and figure 6.9 illustrates the empirical variograms and fitted models.

Shape of empirical variograms does not suggest the Gaussian model as well as the result in tables 6.4, 6.5 and figure 6.9. From table 6.5, it can be concluded that Exponential model is a better choice due to less MSDR, RMSE and ME, hence, exponential model is chosen for variogram modeling.

Initial values are required to fit a non linear model to the empirical variogram. Usually these values are selected base on the shape of the empirical variogram. As long as the initial guess does not differ enormously from the true values the procedure of variogram modeling can be applied (the model variogram does not get singular). It means that selecting different initial values results in almost the same model variogram since the initial guesses are logical. To check this, the initial values used for 22/12/2010 also used for 1/1/2010 and the last 10 days of 2010 and the model could be fitted to each empirical variogram. Note that in 1/1/2010 the PM10 concentrations recorded by the stations are significantly different from other days, hence we conclude that initial values selected for day 356 can be used for all other days. The selected initial values for day 356 are :

- Nugget: 0.4
- Range: 250 km
- Partial sill :0.2

The estimated model parameters, based on the mentioned initial values, for day 1, 355 and 357 are presented in table 6.6.

6.2.4 Prediction on unmonitored locations

Three geostatistical interpolation methods are used to predict at unsampled locations; OK, UK and KED. Using hybrid approaches can be beneficiary if the target variable(PM10) and the aux-

Table 6.5: Summary for the MSDR retrieved from different models

Model	MSDR/ 21 December	MSDR/ 22 December	MSDR/ 23 December
Exponential	1.344	1.146	1.244
Spherical	1.393	1.168	1.248
Gaussian	1.538	1.441	1.350

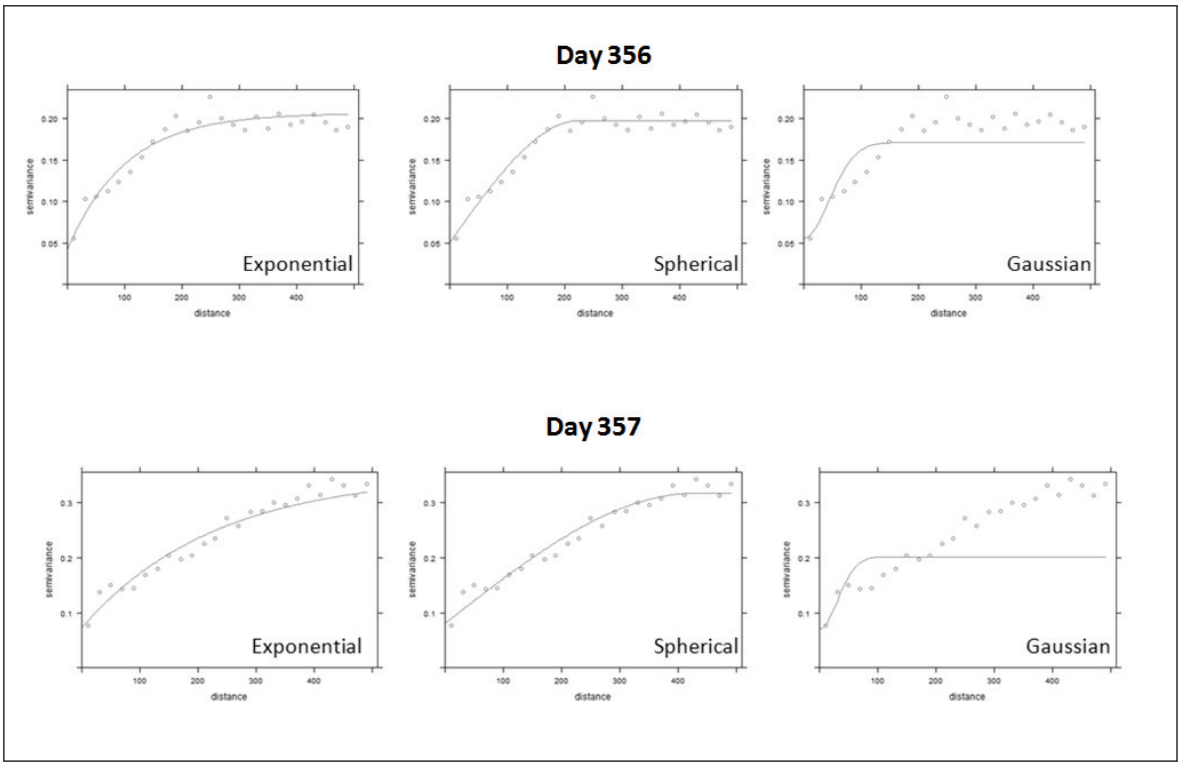


Figure 6.9: Fitted Exponential, Spherical and Gaussian model to the empirical variogram

Table 6.6: Model parameters' estimates

Day	Nugget	Partial sill	Range
Day 1	0.15	0.43	340
Day 355	0.02	0.16	396
Day 356	0.04	0.16	101

Table 6.7: Summary of regression analysis for PM10 and model output

Model output	P-value	R square	F statistic
Day1	«0.001	0.06	17
Day355	«0.001	0.10	32
Day356	«0.001	0.20	68
Day357	«0.001	0.31	130

Table 6.8: Summary of regression analysis for PM10 and coordinates

Coordinates	P-value	R square	F statistic
Day1	0.1	0.01	2.308
Day355	0.0001	0.061	8.86
Day356	«0.001	0.109	16.53
Day357	«0.001	0.442	110.2

iliary variable are correlated. Regression analyses result indicates if the correlation between variables is significance so they can be used as auxiliary variable in modeling . Every test of significance begins with a null hypothesis. We test if the regression coefficients are significantly different from zero. The test of significance is based on the t statistic. If this is a model that can be fitted to scatter plot of 2 variables:

$Y = \beta_0 + \beta_1 X$

It is tested if β_1 is significantly different from zero.

- H_0 : β_1 is zero
- H_1 : β_1 is not zero

If p-value is small (<0.01) so the null hypothesis is rejected in favor of the alternative. It means that the correlation between two variables are significant. In this research model output and coordinates are checked if they can be used as covariates. Apart from p-value, the R square and F also show the correlation.

From Tables 6.7 , it can be implied that the significant correlation between PM10, model output exists everyday that we analyzed. Table 6.8 shows that PM10 and coordinates are not correlated everyday. For instance, the significance correlation exists between coordinates and PM10 for day 356 but not for day 1.

The cross validation result for different kriging methods for day 1, 355, 356 and 357 are presented in tables 6.9, 6.10, 6.11 and 6.12. In general, KED gave better result.

This is also evident by uncertainty maps presented in figure 6.11. it is obvious that the kriging standard deviation is lower in case of KED in comparison with UK and OK in whole study area. Figure 6.10 shows the produced PM10 concentration maps from different methods.

6.3 DESIGN OF USER INTERFACE

The designed user interface is illustrated in figure 6.12. The defined user interface in chapter 5 can not be applied due to lack of time. This is a very simple interface that is made by use of HTML and Java script and open layers . The PM10 concentration map is displayed over base world map.

Table 6.9: UK, KED and OK method validation result for day (1)

1 January (day 1)	KED	UK	OK
RMSE	0.483	0.485	0.485
ME	-0.003	-0.004	-0.0005
MSDR	1.093	1.106	1.112

Table 6.10: UK, KED and OK method validation result for day (355)

21 December (day 355)	KED	UK	OK
RMSE	0.343	0.346	0.346
ME	-0.003	-0.0002	-0.0002
MSDR	1.331	1.353	1.344

Table 6.11: UK, KED and OK method validation result for day (356)

22 December (day 356)	KED	UK	OK
RMSE	0.396	0.398	0.397
ME	-0.0002	-0.002	-0.002
MSDR	1.184	1.152	1.146

Table 6.12: UK, KED and OK method validation result for day (357)

23 December (day 357)	KED	UK	OK
RMSE	0.379	0.391	0.382
ME	-0.0003	-0.006	-0.002
MSDR	1.197	1.207	1.244

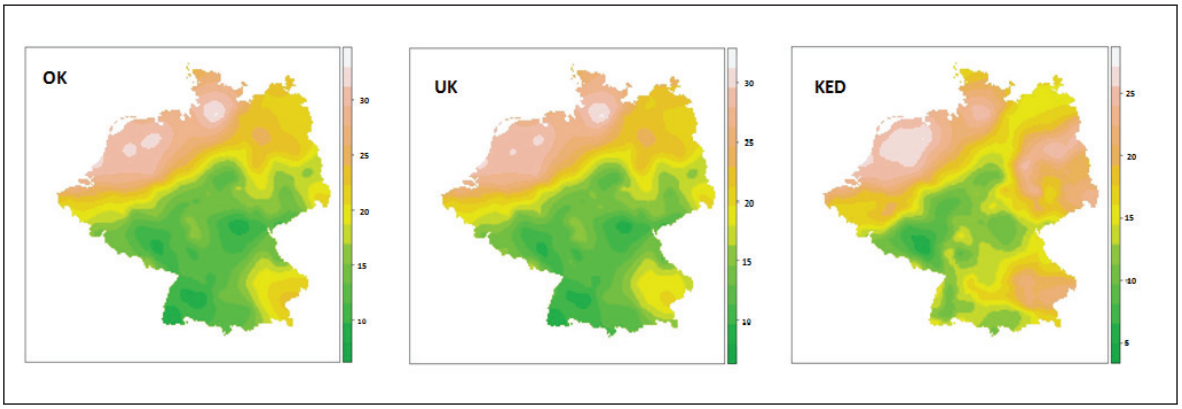


Figure 6.10: Retrieved PM10 concentration from different models (OK, UK, KED) for day (357)

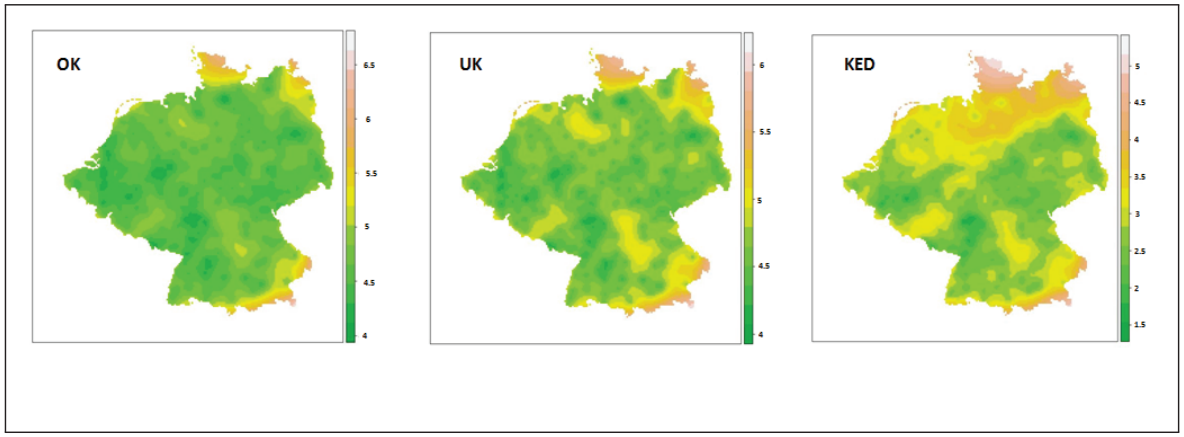


Figure 6.11: Uncertainty maps (standard deviation) from different models (OK, UK, KED) for day (357)

The PM10 concentration values are calculated in WPS and classified and visualized through WMS in the map file. The HTML and Java script codes the map file are presented in Appendix.

In concludes, all considered methods of interpolation the cut off 500 is the best choice and exponential model selected to fit to empirical variogram. Between different methods KED is most suitable one for automatic modeling and mapping because LOTOS-EUROS model out put in most of the days has a significant correlation with PM10 which consequently results in less prediction error.

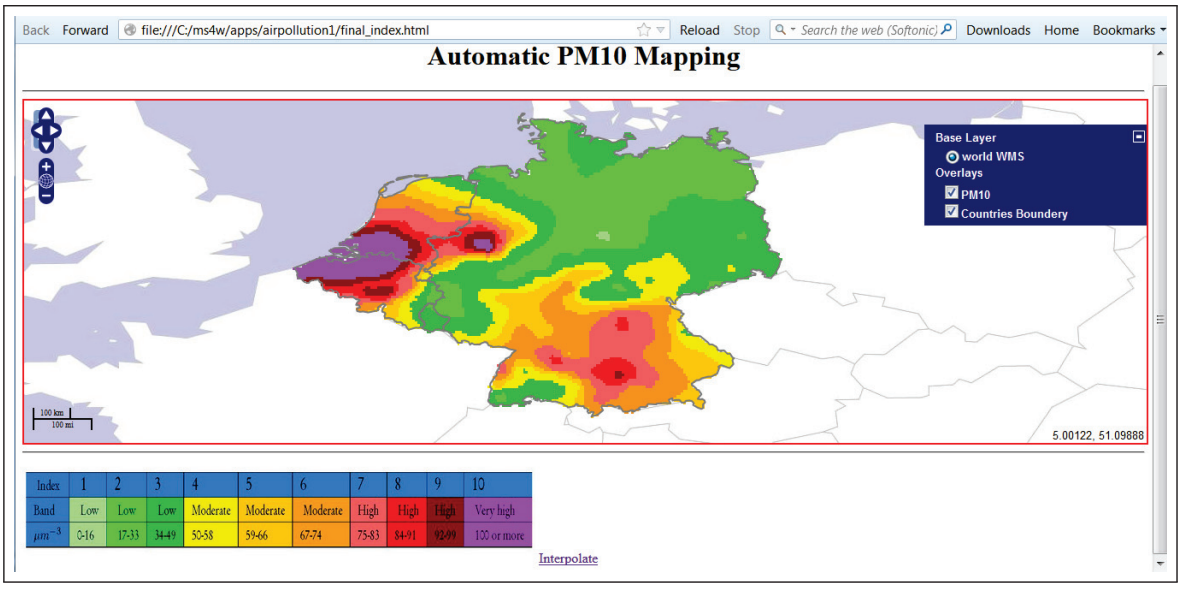


Figure 6.12: user interface. Base map and overlayed PM10 concentration map.

Chapter 7
Discussion

In this chapter the result archived and displayed in chapter6 are discussed.

7.1 EXPLORATORY DATA RESULTS

From the exploratory data results displayed in figures 6.2 and 6.3 , it can be concluded that the spatial distribution of PM10 varies day by day. However, the most extreme case is happened during the 2010 in day1 ,January first, because of the fireworks in new years Eve. In this day almost every stations shows that PM10 concentration exceeded the limit ($50\mu gm^{-3}$). We include this day into the analysis to check our model functionality to see if that works in the most extreme situation. From the histograms and Q-Q plots it can be inferred that data of some days like day 355 and 356 need to be logarithm transformed but not day 357. The result of Shapiro-Wilk test also demonstrate this(see appendix).

7.2 GEOSTATISTICAL MODELING

Two scenarios are considered in this study. First when users request for map in the same region of the study area and second scenario when users upload their data, not in the same region. The prior knowledge about the study area and spatial distribution of stations can be beneficial in the selection of cut off for empirical variogram calculation. The result shows that when proper cut off(the one which reaches to clear sill) is selected, it resulted in less kriging variance (see MSDR results in table 6.3). The model variograms by "automap" in figure 6.7 show that non of them reaches to the clear sill because cut off is one third of maximum distance by default.

As it is mentioned earlier, for specific case user can upload their data in the same region of the study area. These data can be other pollutants concentration that recorded by the same stations like ozone or SO2 but not PM2.5. The reason is that number of stations recorded PM2.5 are less than stations that are recording other pollutants. Hence, the selected initial values to fit a model to empirical variogram for PM10 can not be used for PM2.5. We can not consider in this case that there is prior knowledge about the spatial distribution of the monitoring points.

Three methods, OK, UK and KED selected for modeling PM10. From tables 6.7 - 6.12 it can be concluded that using auxiliary variable improve the result(RMSE and MSDR) when they are correlated with the target variable (PM10). In tables 6.7 and 6.8, It is displayed that in all days we analyzed the correlations between PM10-model output and PM10-coordinates are significant. It is evident from the global statistics and also generated uncertainty maps for day 357 that KED resulted in lower uncertainty in comparison with two other methods .

Another reason for selecting KED over OK is that using auxiliary variables can be helpful to describe some part of the spatial variation of the correlated observation (Kassteele, 2006). There are some days that there are lots of stations do not record PM10. In such cases using covariates which have a full coverage over the region can reduce the uncertainty of predictions. This is also concluded by (Kassteele, 2006) in a case study in The Netherlands.

One limitation of the LOTOS-EUROS is its spatial resolution (35 km by 25 km) in compare with the required resolution of PM10 (5 km by 5 km) concentration map. Using models with finer resolution can decrease the uncertainty in the predictions.

7.3 PROTOTYPE DESIGN AND IMPLEMENTATION

In figure 6.12 the simple designed interface is presented. This can be improved in many ways by careful consideration to find out what is suitable for users, or better to say what is ”user friendly” and also to find out who are the potential users.

PM10 interpolated values and uncertainty are produced in R back-end ,are restored in WMS directory and then recalled for being displayed in HTML file. There should be a system to delete produced GeoTIFF files in WMS directory time to time to prevent storage problem.

Chapter 8
Conclusion and Recommendations

8.1 CONCLUSION

In this research the objective was to automate the process of modeling and mapping of PM10 and associated uncertainty. To reach this objective several questions addressed in chapter one, section" research objective and questions", are answered.

RQ1.1 What are the suitable covariates to model PM10 concentration?

Based on the regression analysis done in the research and also available data model output and coordinates considered as suitable covariates. However, this should be noted because we automate the daily data, we cannot conclude that model output or coordinates are always correlated with target variable as it is displayed in table 6.7. But in case of model output most of the days, it was highly correlated with the PM10 concentration. Another point is that model output is a physical model which lot of meteorological data contains in its calculation so indirectly we include meteorological data in our modeling by using it as an auxiliary variable.

RQ1.2Which geostatistical interpolation method is more suitable for automating air pollution?

KED is more suitable for automating air pollution modeling.The reasons are First it can be applied with no human intervention, second; it can calculate the uncertainty in prediction points and third reason is that model out put and PM10 concentration demonstrate to have correlation in most of the days which results in less prediction uncertainty. In addition LOTOS-UEROS model has a full coverage over the European continent with spatial resolution of 25 km by 35 km. For those areas that the number of recording stations are low using covariate is beneficial.

RQ2.1 What are the rules for automatic modeling?

Automatic modeling in this research equals to automatic variogram modeling procedure. For two defined scenarios in this research two groups of rules are used. General case, where users import the data in a region other than the defined study area. The automation rules in "automap" are used which are:

- Cut off : %35 maximum distance

initial values for the fit :

- Sill: mean of the maximum and median value of the variogram values
- Nugget: minimum value of the variogram values
- Range: cutoff/3.5
- Model : Exponential, Spherical, Gaussian , the one with the minimum error sum of squares is selected

and for specific case, where users require air pollution map in a defined study area, we improve the variogram model by selecting the proper cut off and number of bins. The selected cut off is equal to 500 km and initial variogram parameters are:

- Nugget: 0.4
- Range: 250 km
- Partial sill :0.2

RQ2.2 What are the steps to put the selected model on the Web?

By use of WPS which follows SOA the model can be used over the net. To model PM10, we used R which is very powerful environment for any statistical computing. We benefited from WPS4R, a module provided by 52North, which allows WPS process creation via the R-scripts. The requirements is that the annotated R-scripts and also "Rserve" which connects WPS to R.

RQ3.1 What are the steps to visualize the interpolated PM10 values and associated uncertainty by use of existing Web services?

WMS is a standard protocol for serving georeferenced map files over the Internet. The GeoTIFF files including PM10 interpolated values and associated uncertainty produced in modeling step, are classified and styled trough WMS. OGC defined the standard ways of visualization such as specifying colors associated to each class. The adjacent maps are used to visualize uncertainties. The WMS used is in this study is "Mapserver". The GML format of predictions and uncertainties are produced for those users who select GML as an output format. The GML file is an output of WPS which directly displays in user interface.

RQ3.2 What should be the resolution of the output map?

The selected resolution of output map is 5km by 5km as it is suggested by Kuhlbusch et al. (2006). This is also a grid size that is not computationally demanding so it is suitable for automation.

8.2 RECOMMENDATIONS

This study was one step forward to real time automatic air pollution modelling and mapping. There are some limitations in every work such as time limit.

In this thesis, we used daily PM10 data, and we restored the data in the local directory and recalled it in to interpolation process via R- scripts. I suggest using "sos4r" package in order to equip the interpolation system with retrieving near real time monitoring data.

We made some assumption for modeling PM10. One of them was an isotropy assumption. "INTAMAP" already deal with anisotropy automatically. If the system finds out that data are anisotropy significance, it corrects for it. We can use this "INTAMAP" capability to develop our system.

We chose KED for automatic modeling although it was not always beneficial. In some days that correlation between LOTOS-EUROS model out put and PM10 were not significant, OK gave better result in terms of RMSE and MSDR. One further work would be to check the correlation between the target variable and covariate automatically by use of the result of regression analysis (P-value). If the correlation is significant, the interpolation system applies KED otherwise ordinary kriging.

We selected some visualization techniques to display PM10 map concentration and associated uncertainty. It is useful to create an interface that the user can choose different visualization tools in order to get what is best for him/her application.

We could not consider "user friendly" term in designing of the user interface due to lack of time. More work need to be done to understand user requirements and design an interface which fits to users' needs.

Chapter 9
References

Abraham, J., & Comrie, A. (2004). Real-time ozone mapping using a regression-interpolation hybrid approach, applied to Tucson, Arizona. *Journal of the Air & Waste Management Association*, *54*(8), 914–925.

Annoni, A., Luzet, C., Gubler, E., & Ihde, J. (2001). Map projections for europe. *Report EUR*, *20120*, 131.

Bivand, R., Pebesma, E., & Gómez-Rubio, V. (2008). *Applied spatial data analysis with r*. Springer.

Brenning, A., & Dubois, G. (2008). Towards generic real-time mapping algorithms for environmental monitoring and emergency detection. *Stochastic Environmental Research and Risk Assessment*, *22*(5), 601–611.

COMEAP. (2009). Committee on the medical effects of air pollutants. long-term exposure to air pollution: effect on mortality.

Cressie, N. (1993). *Statistics for spatial data, revised edition* (Vol. 928). Wiley, New York.

De Jesus, J., Dubois, G., & Hiemstra, P. (2008). Web-based geostatistics using WPS. *Proceedings of the 6th Geographic Information Days*, *32*, 199–218.

Denby, B., Costa, A., Monteiro, A., Dudek, A., & Erik, S. (2007). Uncertainty mapping for air quality modelling and data assimilation.

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.

Denby, B., Sundvor, I., Cassiani, M., de Smet, P., de Leeuw, F., & Horálek, J. (2010). Spatial mapping of ozone and SO2 trends in europe. *Science of the Total Environment*, *408*(20), 4795–4806.

Dubois, G., & Galmarini, S. (2005). Spatial interpolation comparison (sic): introduction to the exercise and overview of results. *Automatic mapping algorithms for routine and emergency monitoring data*, 150.

EPA. (2004). Developing spatially interpolated surfaces and estimating uncertainty.

Hengl, T. (2006). Finding the right pixel size. *Computers and Geosciences*, *32*(9), 1283–1298.

Hengl, T., Heuvelink, G., & Stein, A. (2003). Comparison of kriging with external drift and regression-kriging. *Technical note, ITC*.

Holland, D., Cox, W., Scheffe, R., Cimorelli, A., Nychka, D., & Hopke, P. (2003). Spatial prediction of air quality data. *EM-pittsbutgh-air and waste management association*, 31–35.

Honoré, C., Rouïl, L., Vautard, R., Beekmann, M., Bessagnet, B., Dufour, A., … others (2008). Predictability of european air quality: Assessment of 3 years of operational forecasts and analyses by the prevŠair system. *J. Geophys. Res*, *113*, D04301.

Hudson, G., & Wackernagel, H. (1994). Mapping temperature using kriging with external drift: theory and an example from scotland. *International journal of Climatology*, *14*(1), 77–91.

Izza, S., Vincent, L., & Burlat, P. (2008). Exploiting semantic web services in achieving flexible application integration in the microelectronics field. *Computers in industry*, *59*(7), 722–740.

Journel, A. (1978). Mining geostatistics. *Academic Press. London. Mathematical Geology*, *17*.

Kassteele, J. (2006). *Statistical air quality mapping*. [Sl: sn].

Kebaili Bargaoui, Z., & Chebbi, A. (2009). Comparison of two kriging interpolation methods applied to spatiotemporal rainfall. *Journal of Hydrology*, *365*(1), 56–73.

Kloog, I., Koutrakis, P., Coull, B., Lee, H., & 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.

Konovalov, I., 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.

Krige, D., & Magri, E. (1982). Studies of the effects of outliers and data transformation on variogram estimates for a base metal and a gold ore body. *Mathematical Geology*, *14*(6), 557–564.

Kuhlbusch, T., John, A., Hugo, A., Peters, A., von Klot, S., Cyrys, J., … Bruckmann, P. (2006). Analysis and design of local air quality measurements. *Final report to the European Commission. IUTA, Duisburg*.

Lark, R., Cullis, B., & Welham, S. (2005). On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (e-blup) with REML. *European Journal of Soil Science*, *57*(6), 787–799.

Lee, H., Liu, Y., Coull, B., Schwartz, J., & Koutrakis, P. (2011). A novel calibration approach of MODIS AOD data to predict PM 2.5 concentrations. *Atmos. Chem. Phys*, *11*, 7991–8002.

McBratney, A., Odeh, I., Bishop, T., Dunbar, M., & Shatar, T. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma*, *97*(3), 293–327.

Pebesma, E., Cornford, D., Dubois, G., Heuvelink, G., Hristopulos, D., Pilz, J., … Skøien, J. (2011). Intamap: the design and implementation of an interoperable automated interpolation web service. *Computers & Geosciences*, *37*(3), 343–352.

Schaap, M., Timmermans, R., Roemer, M., Boersen, G., Builtjes, P., Sauter, F., … Beck, J. (2008). The lotos? euros model: description, validation and latest developments. *International Journal of Environment and Pollution*, *32*(2), 270–290.

Schut, P. (2007). Opengis® web processing service 1.0. 0, ogc. document ogc 05-007r7.

Senaratne, H., & Gerharz, L. (2011). An assessment and categorisation of quantitative uncertainty visualisation methods for geospatial data.

Senaratne, H., Gerharz, L., Pebesma, E., & Schwering, A. (2012). Usability of spatio-temporal uncertainty visualisation methods. *Bridging the Geographic Information Sciences*, 3–23.

Sheng, G., Darka, M., Xiaolun, Y., Francois, A., Eddie, O., & David, C. (2008). Towards web-based representation and processing of health information. *International Journal of Health Geographics*.

Skoien, J., Pebesma, E., & Blöschl, G. (2008). Geostatistics for automatic estimation of environmental variablesŮsome simple solutions. *Georisk*, *2*(4), 259–272.

Spangl, W., Schneider, J., Moosmann, L., & Nagl, C. (2007). Representativeness and classification of air quality monitoring stations. *Umweltbundesamt report*.

Stein, A., Riley, J., & Halberg, N. (2001). Issues of scale for environmental indicators. *Agriculture, Ecosystems and Environment*, *87*(2), 215 - 232.

Strobl, P., Reithmaier, L., Soille, P., Mehl, W., & Bielski, C. (2007). Assembly of a remote sensed reference image database of europe at 25 m resolution. In *Proceedings of the 27th earsel symposium, geoinformation in europe, ma gomarasca (ed.). millpress, the netherlands* (pp. 515–522).

Van Bodegom, P., Verburg, P., tein, A., Adiningsih, S., & Denier van der Gon, H. (2002). Effects of interpolation and data resolution on methane emission estimates from rice paddies.

Environmental and Ecological Statistics, 5-26.

van de Kassteele, J., Stein, A., Dekkers, A., & Velders, G. (2009). External drift kriging of NO x concentrations with dispersion model output in a reduced air quality monitoring network. *Environmental and Ecological Statistics*, *16*(3), 321–339.

Wackernagel, H., Lajaunie, C., Blond, N., Roth, C., & Vautard, R. (2004). Geostatistical risk mapping with chemical transport model output and ozone station data. *Ecological Modelling*, *179*(2), 177 - 185. (Control of Distributed Systems and Environmental Applications)

Webster, R., & Oliver, M. (2001). Geostatistics for environmental scientistsjohn Wiley & Sons. *New York*.

Webster, R., & Oliver, M. (2007). *Geostatistics for environmental scientists*. Wiley.

WHO. (2011). World health organization. air quality and health: Fact sheet n 313. *Updated September*.

Wong, D., Yuan, L., & Perlin, S. (2004). Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science and Environmental Epidemiology*, *14*(5), 404–415.

Zimmerman, D., Pavlik, C., Ruggles, A., & Armstrong, M. (1999). An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology*, *31*(4), 375–390.

Appendix A

Code

A.1 MAPSERVER, MAP FILE

```
MAP
    NAME airpollution
    IMAGECOLOR 255 255 255
    IMAGETYPE PNG24 ## use AGG to for anti-aliassing
    OUTPUTFORMAT
    NAME 'AGG'
    DRIVER AGG/PNG
    MIMETYPE "image/png"
    IMAGEMODE RGB
    EXTENSION "png"
    END # outputformat
  PROJECTION
   "init=epsg:3035" #latlon on etrs 1989 laea
  END
  EXTENT 3700000 2600000 4700000 3600000 # meters extents of study area
    WEB
    IMAGEPATH "c:/tmp/ms_tmp/"
    IMAGEURL "/ms_tmp/"
    METADATA
  "ows_enable_request"   "*"
      "map" "c:/ms4w/Apache/htdocs/airpollution1/config.map"
      "ows_schemas_location" "http://schemas.opengeospatial.net"
      "ows_title" "Sample WMS"
  "ows_enable_request"   "*"
      "ows_onlineresource" "http://localhost:7070/
      cgi-bin/mapserv.exe?map=C:/ms4w/apps/airpollution1/config.map&"
      "ows_srs" "EPSG:3035 " #meter
      "wms_feature_info_mime_type" "text/plain"
      "wms_feature_info_mime_type" "text/html"
      "wms_server_version" "1.1.1"
      "wms_formatlist" "image/png,image/gif,image/jpeg, image/geotiff"
      "wms_format" "image/png"
    END #metadata
    END #web

LAYER
```

```
NAME "pm10"
DATA  "pm10_357.tif"
TYPE RASTER
STATUS ON
METADATA
  "ows_title" "pollution"
END #metadata
PROJECTION
  "init=epsg:3035"
    END #projection
CLASSITEM "[pixel]"
 # class using simple string comparison, equivelent to ([pixel] = 0)

 # class using an EXPRESSION using only [pixel].
CLASS
EXPRESSION ([pixel] >-100 AND [pixel] < 17)
STYLE
  COLOR 150 255 150
END
END
CLASS
EXPRESSION ([pixel] >= 17 AND [pixel] < 34)
STYLE
  COLOR 50 255 0
END
END
CLASS
EXPRESSION ([pixel] >= 34 AND [pixel] < 50)
STYLE
  COLOR 50 200 0
END
END
CLASS
EXPRESSION ([pixel] >= 50 AND [pixel] < 59)
STYLE
  COLOR 255 255 0
END
END
CLASS
EXPRESSION ([pixel] >= 59 AND [pixel] < 67)
STYLE
  COLOR 255 200 0
END
END
CLASS
EXPRESSION ([pixel] >= 67AND [pixel] < 75)
STYLE
  COLOR 255 150 0
```

```
END
END
CLASS
EXPRESSION ([pixel] >= 75AND [pixel] < 84)
STYLE
  COLOR 255 100 100
END
END
CLASS
EXPRESSION ([pixel] >= 84AND [pixel] < 92)
STYLE
  COLOR 255 0 0
END
END
CLASS
EXPRESSION ([pixel] >= 92AND [pixel] < 100)
STYLE
  COLOR 150 0 0
END
END
CLASS
EXPRESSION ([pixel] >= 100)
STYLE
 COLOR 200 50 255
END#style
END#class

END #layer pm10
LAYER
NAME "countrybouderies"
TYPE POLYGON
STATUS ON
DATA data/Belenux_ger_etrs
METADATA
  "ows_title" "Belenux_ger_etrs"
END #metadata
PROJECTION
  "init=epsg:3035"
END
CLASS
NAME "Belenux_ger_etrs"
STYLE
OUTLINECOLOR 127 127 127
WIDTH 2
END # style
END #class countries
END #layer countries
```

```
END #map
```

A.2 HTML FILE

```
<html>
<head>
<title > PM10 </title>

<!-- OpenLayers core js -->
<script type="text/javascript"
src="http://www.openlayers.org/dev/OpenLayers.js">
</script>
<!-- OpenStreetMap base layer js -->
<script type="text/javascript"
src="http://www.openstreetmap.org/openlayers/OpenStreetMap.js">
</script>


 <script type="text/javascript">
var myMap, myWMSBaseLayer, pm10,bounderies ;
var myCenter = new OpenLayers.LonLat(
10,51
);
function init() {
myMap = new OpenLayers.Map("mapDiv");
    myWMSBaseLayer = new OpenLayers.Layer.WMS(
"world WMS",
"http://geoserver.itc.nl/cgi-bin/mapserv.exe?map=
D:/Inetpub/geoserver/mapserver/config_world.map&",
{layers: "world"}
);

pm10 = new OpenLayers.Layer.WMS
("PM10", "http://localhost:7070/cgi-bin/mapserv.exe?map=
C:/ms4w/apps/airpollution1/config.map",
            {layers: "pm10",transparent: "true", format: "image/png"}
            );
bounderies= new OpenLayers.Layer.WMS
("Countries Boundery", "http://localhost:7070/cgi-bin/mapserv.exe?map=
C:/ms4w/apps/airpollution1/config.map",
            {layers: "countrybouderies",transparent: "true", format: "image/png"}
            );
myMap.addLayers([myWMSBaseLayer, pm10, bounderies]);
myMap.addControl(new OpenLayers.Control.MousePosition());
            myMap.addControl(new OpenLayers.Control.ScaleLine());
myMap.addControl(new OpenLayers.Control.Navigation());
myMap.addControl(new OpenLayers.Control.PanZoom());
```

```
      myMap.addControl(new OpenLayers.Control.LayerSwitcher());
myMap.setCenter(myCenter,6);

      request = OpenLayers.Request.POST({
url: "http://localhost:8080/wps/WebProcessingService?Request=Execute&Service=WPS&
Version=1.0.0&Identifier=org.n52.wps.server.r.OK6",
data: OpenLayers.Util.getParameterString(),
headers: {
"Content-Type": "application/x-www-form-urlencoded"
},
callback: handler
})
}

    </script>
  </head>
  <body onload="init()" bgcolor="#ffffff">
<center><h1> Automatic PM10 Mapping</h1></center>
<hr>
    <!-- div containers for map, lat/long coordinate
    output and transaction messages -->
    <div >
        <div id="mapDiv" style="bottom:100px;width:100;
        height:400px; border:2px solid red;">
           <!-- <div id="coordinates" style="z-index:999"></div>-->
        </div>

</div>
<hr>
<img border="0" src="Legend.JPG" alt="Pulpit rock" width="600" height="120">
<a href="http://localhost:8080/wps/WebProcessingService?Request=
Execute&Service=WPS&Version=1.0.0&Identifier=org.n52.wps.server.r.OK6">Interpolate</a>


</body>
</html>
```

A.3 R- SCRIPTS ANNOTATION

```
#wps.des: title = Ordinary kriging in R, ( description of the process)
#abstract = predicting PM10 on unmonitored points;
#wps.in: day, type = integer, abstract = Points for OK; (defining input data)
#wps.out: output ,type =integer, abstract =RMSE,ME and MSDR; (defining output data)
```

A.4 R- SCRIPTS FOR AUTOMATING THE PROCESS OF LOGARITHM TRANSFORMATION

```
b<- shapiro.test(data$PM10)
 if(b$p.value< 0.01) {
data$PM10<- log(data$PM10)
}
```

A.5 DIFFERENT CUT-OFFS FOR KED AND UK

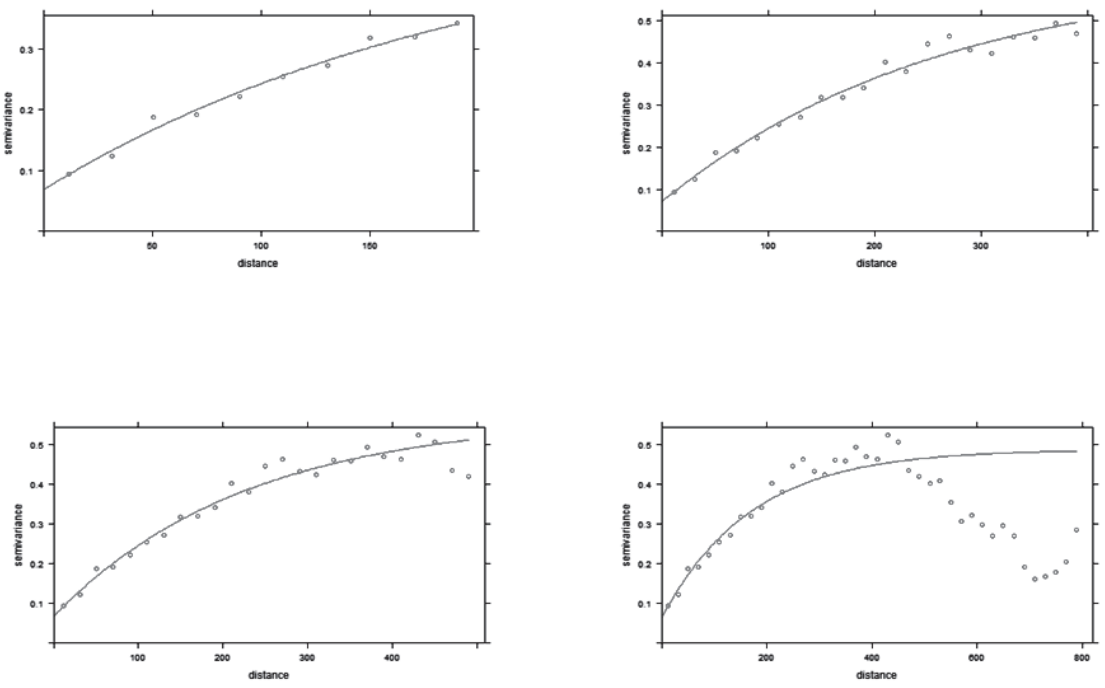


Figure A.1: Diffrent cut offs in case of KED

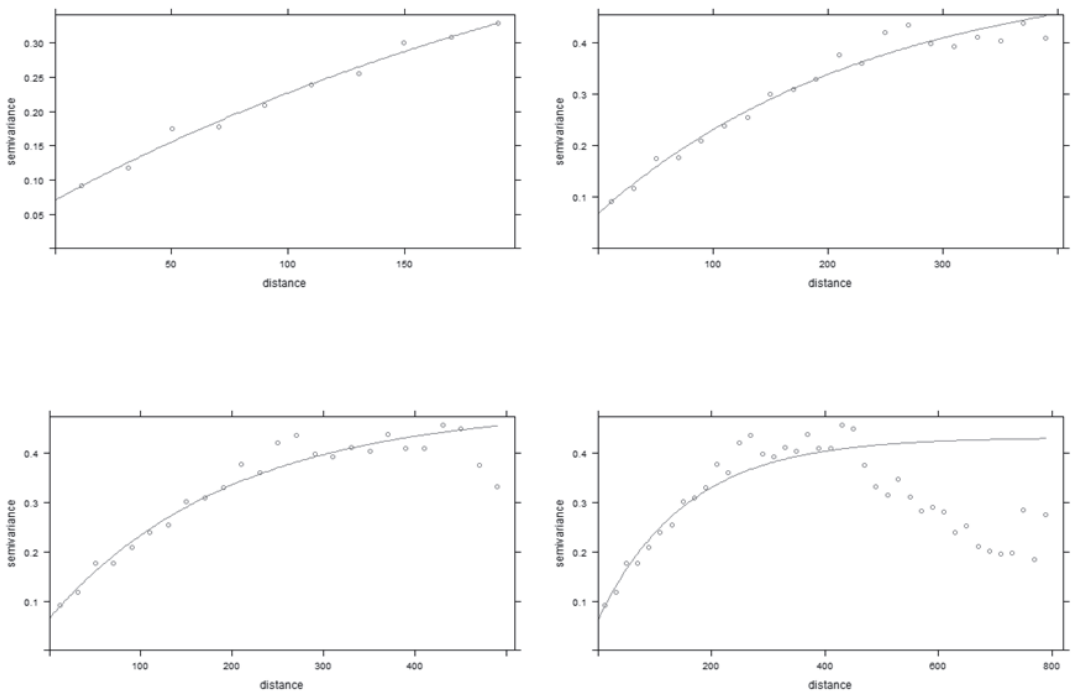


Figure A.2: Different cut offs in case of universal kriging

A.6 RESULT OF SHAPIRO-WILK TEST

Table A.1: Result of shapiro-wilk test

year 2010	p-value
day1	«0.001
day355	«0.001
day356	«0.001
day357	0.05