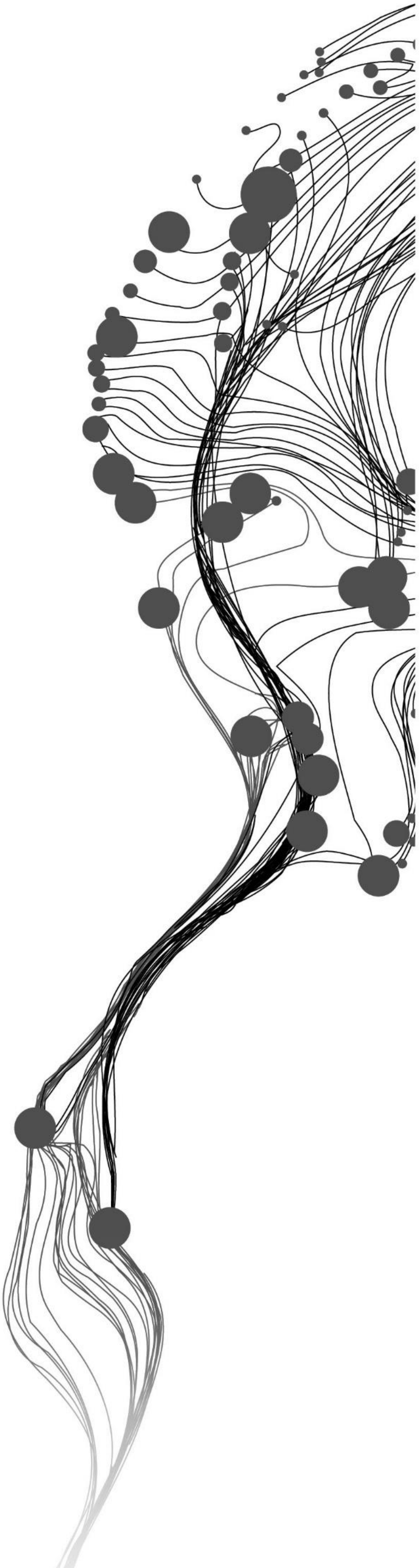


# Mapping Multiple Pollutants at a Continental Scale

INGRID MARISOL AMADOR FIGUEROA

February, 2015

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

THESIS ASSESMENT BOARD:
Prof. Dr. Ir. M.G. George Vosselman (Chair)
Dr. Ir. Gerard Hoek (University of Utrecht, Institute for Risk Assessment Sciences)

DISCLAIMER

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

# ABSTRACT

Air pollution generates different problems on ecosystems and human health. In order to improve the state of air quality, the European Union (EU) has established certain limits for all EU countries which must be accomplished. These limit target values have been established according to each individual pollutant. Particulate matter and ozone are Europe's most problematic pollutants in terms of harm to human health (EEA., 2013)).

Particulate matter is a type of pollutant that originates from primary and secondary particles generated in anthropogenic processes and natural emissions. Particulate matter is classified according to the size of the particle. Particulate matter equal or less than 1 micrometre is designated as $PM_{1,}$ $PM_{2.5}$ corresponds to the fine fraction of 2.5 micrometre or less in diameter and $PM_{10}$ has the size of 10 micrometres or less.

In order to evaluate the state of air pollution, accurate maps of interpolated data are needed. Different techniques have been used to obtain more accurate predictions. In this thesis, two methods, compositional kriging and cokriging are used to generate predictions for the spatial distribution of the pollutants. European data, taken from a freely available database is used as base data. Several combinations of covariates are used, until the best choice is found. The full comparison between cokriging and compositional kriging was not possible, as compositional kriging only provides data for the ratio between pollutants.

# ACKNOWLEDGEMENTS

To my parents:
Ingrid Figueroa for her example, love and unconditional support and reminding me the real important things in life and Domingo Amador for spreading his love for science and research as a tool for development.

To my family:
For their love and support through this process and for always motivate me to have the heart in the right place.

To my supervisors:
Dr. Nicholas Hamm for providing the data used during this study and for his professional advice during the development of this research.
Prof. Dr. Ir. Alfred Stein for sharing his great experience in this field.

To the Niche-Nuffic project:
For providing me the necessary sponsorship that allowed me to develop this research and to MSc. Lyande Eelderinck, Tonny Boeve and Dr. Javier Morales for their support and advice before and during my staying in the Netherlands.

To my friends:
Far away and close, new and old, for teaching me their culture, sharing their thoughts and their time with me and creating this an experience I will never forget. Special thanks to those who helped me in different ways during the six months of the research, giving me advices, helping me to solve technical problems or cheered me up with their positive attitude: Ana, Diana, Rodolfo, Manuel, Gustavo, Alfredo, Riddhi and Siddhi, Bernard, Emmanuel, Berihu, Christopher, Saad, Koen, Alicia, Eduardo and everyone else that has touched my life.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| clr | Centered Log-Ratio transformation |
| CTM | Chemical Transport Model |
| EEA | European Environmental Agency |
| ETRS89 | European Terrestrial Reference System |
| EU | European Union |
| ilr | Isometric Log-Ratio Transformation |
| PM | Particulate Matter |
| $PM_{10}$ | Particulate Matter less than 10μm |
| $PM_{2.5}$ | Particulate Matter less than 2.5 μm |
| $PM_{coarse}$ | Difference between $PM_{10}$ and $PM_{2.5}$ |

# 1. INTRODUCTION

## 1.1. Motivation and Problem statement

The air quality in any place on the Earth depends upon many factors, including topography, climatic conditions, social conditions, land use and the dispersion of the pollutants according to their physical and chemical properties. The variations of air quality levels in space can be determined as a function of these factors.

Air pollutants such as particulate matter (PM), ozone ($O_3$), nitrogen and organic compounds still represent a significant threat for human health and the environment in Europe, even after trying to improve the state of air pollution (EEA., 2013). Health effects of air pollution are dominated by particulate matter both $PM_{10}$ and $PM_{2.5}$, which is the inhalable size fraction of PM (Schaap *et al.* (2001)). Schaap *et al.,* (2001) also identified seven source groups for these two types of particulate matter: 1) nitrate-rich secondary aerosol, 2) sulphate-rich secondary aerosol, 3) traffic and re-suspended road dust, 4) metal from industrial activity/ incineration, 5) sea salt, 6) mineral dust, and 7) particles from residual oil combustion.

Air pollution can be measured from different techniques, viz. a) ground data, which are in situ measurements taken from stations, and b) remote sensing products. By using a combination of ground data and remotely sensed data and considering the composition of the pollutants, it is possible to predict levels of pollution. As long as the data is reliable (and corrected), it is likely that gaps and outliers, such as maximum and minimum values, are identified in the process of prediction.

A statistical method is required if predictions are to be made. The pollutants on air at any moment in time behave as a spatial phenomenon for which it is impossible to obtain data at every specific location. In order to use the data and identify the values where no measure has been taken it is helpful to use geostatistics. Geostatistics can help predict data taking into account the spatial correlation. Using geostatistical methods will enable us to make estimations or predictions without bias and with minimum error, and allows us to deal with varying properties at all spatial scales (Webster & Oliver (2007)).

Geostatistical prediction depends chiefly upon statistical linear models. Predictions are made by means of different techniques. The selection of an optimal model, with the lowest error can drastically improve the quality of the predictions. In this thesis, we focus in kriging. Kriging is regarded as the best linear unbiased estimator (BLUE) because it minimizes the prediction error variance. Different types of kriging are distinguished depending on the chosen model for the trend of the random function (Moral, Álvarez, & Canito, 2006). Cokriging and compositional kriging are two extensively validated methods for the prediction of air pollutants. Cokriging allows us to jointly estimate values of a coregionalization with spatially correlated components (Pawlowsky, 1989). It thus depends not only on the distance of the data, but also on the direction and orientation of the neighbouring data. On the other hand, compositional kriging considers all components simultaneously by minimizing the sum of their prediction error variances and by taking the unbiasedness, nonnegativity, and constant sum constrains into account (Walvoort & Gruijter, (2001)).

In this work, current values of particulate matter using multiple pollutants were predicted by combining the in situ data of Europe with data obtained from a Chemical Transport Model CTM to derive predictions on a map. These predictions were made by compositional kriging and cokriging, considering the two types of particulate matter, $PM_{2.5}$ and $PM_{10}$, as the components.

## 1.2. Research objectives

### 1.2.1. General Objective

Develop a geostatistical model for multiple pollutants and at a continental scale using compositional kriging and cokriging.

### 1.2.2. Specific Objectives

1. Perform an exploratory analysis over a selection of European ground and remotely sensed data for two types of particulate matter.
2. Predict values and uncertainties of particulate matter $PM_{2.5}$ using multivariable prediction.
3. Provide a general assessment of the data quality in terms of accuracy of the results obtained in the prediction of the composition.

### 1.2.3. Research Questions

**Objective 1: Exploratory Analysis**

1. Which covariates generate a model with lowest error?
2. What is the spatial distribution of the pollutants using these covariates?
3. What is the spatial variation of the measurements and how can it be interpreted?

**Objective 2: Prediction**

4. What are the results of the predictions with respect to the composition of $PM_{2.5}$?
5. What prediction method produces the lowest error: compositonal kriging or cokriging?

**Objective 3: Data quality**

6. What is the maximum spatial extent in which can be obtained accurate prediction results?
7. Which method is suitable for measuring the accuracy of the output?
8. What is the accuracy of the output?

## 1.3. Innovation

The novelty of this research resides in its aim to use two categories of particulate matter ($PM_{10}$ and $PM_{2.5}$) as components and also the use of multiple covariates to generate predictions. The prediction methods used are compositional kriging and cokriging, and their accuracy will be compared.

The second novel aspect of this work is the large spatial extent of the predictions. Some related work has already addressed the prediction of particulate matter concentrations in smaller areas with similar characteristics corresponding to single countries. The present research combines different, and larger, spatial extent areas and it is shown that reliable predictions can still be produced.

# 2. RELATED WORK

Denby *et al.* (2005) made a review on different techniques used for interpolation of air pollutants at a regional scale in Europe. Further research on air pollutants, specifically on particular matter ratios, was made by Leeuw & Horálek, (2009) who used $PM_{2.5}/PM_{10}$ ratios to prepare $PM_{2.5}$ concentration maps of Europe. On their research it is mentioned that on an annual basis the correlation between the co-located $PM_{2.5}$ and $PM_{10}$ daily averages should be at least 0.7. They found average ratios of 0.62 at rural stations and 0.65 urban stations, while the average ratio at traffic locations is 0.58 for the period 2004-2006.

Putaud *et al.* (2010) performed a synthesis of the data on particulate matter physical and chemical characteristics for 10 years. They state as one of their findings that there is no single ratio between $PM_{2.5}$ and $PM_{10}$ mass concentrations although fairly constant ratios that range from 0.5 to 0.9 are observed at most individual sites.

In the case of cokriging, Ver Hoef & Cressie, (1993), presented the best linear unbiased spatial prediction for multivariable data based on covariances or cross-variograms. Studies for air pollution prediction using co-kriging already exist in the literature. Singh, *et al.* (2011) implemented a cokriging technique using the results of a deterministic Chemical Transport Model (CTM) simulation as secondary variable. Bytnerowicz *et al.* (2002) in their research for ozone concentrations evaluated various interpolation methods, including both simple and geostatistical methods, for the interpolation of ozone measurements in the Carpathian Mountains of Eastern Europe. Results from that study indicate a number of appropriate interpolation methods. As a conclusion, they recommended the spherical model of cokriging with altitude as the secondary variable

In the case of compositional predictions, Pawlowsky (1989), analysed the use of cokriging to estimate regionalized compositions, making use of the additive-log-ratio transformation. Later, Walvoort & de Gruijter (2001), compared the performance of compositional kriging with that of the additive logratio-transform. In their work they describe a regionalized composition as a vector random function $z(x_i)$ located at a point $x_i$ in a spatial domain $D$ with $p$ components $z_k(x_i)$

$$z(x_i) = [z_1(x_i), z_2(x_i), \dots, z_p(x_i)\ ]^T$$

That are nonnegative and sum to a constant $c$ which usually equals 100(%) or 1.

Walvoort and de Guitjer suggest that prediction methods such as kriging and co-kriging do not satisfy the requirements for an appropriate spatial interpolation method for a regionalized composition and its constraints. Therefore compositional kriging is proposed, which is an unbiased predictor that minimizes prediction error variance.

Odeh, Todd, & Triantafilis, (2003) applied the additive and modified log-ratio transformation to the particulate matter data. The performance of the transformed data using ordinary kriging was compared with the prediction of the untransformed data using ordinary kriging, compositional kriging and cokriging. However, further research is required to understand the composition, spatial distribution and sources of particulate matter that could generate more accurate predictions. This may be achieved by using compositional kriging and assessing its results.

For completeness, we also note the book by Van den Boogaart & Tolosana-Delgado, (2013), which details different techniques to analyse compositional data on the R environment.

# 3. STUDY AREA AND DATA DESCRIPTION

## 3.1. Study Area

The area of study was determined by the countries that contributed with observations to the Airbase database (the database is described on the following section). The countries used for the prediction were selected according to the coherence of the measurements. The outliers, blank fields and unknown data were discarded. The countries which do not share land borders were also discarded, as it was considered that they could affect the final accuracy because the proximity within the stations is less, reducing the quality of the interpolation.

The countries selected for the analysis were the following:
Group A

North Western Europe: Belgium, Luxembourg, Netherlands and France (Latitude above or equal to 45 degrees).
Southern Europe: Italy, Spain, Portugal and France (Latitude below 45 degrees).
Central Eastern Europe: Austria, Switzerland, Czech Republic, Poland and Germany.

A second analysis was made with the following regions:
Group B

BENELUX: region formed by the countries Belgium, Netherlands and Luxemburg
GERPOL: Germany and Poland
SLOCZ: Czech Republic and Slovakia

The selection of the regions was defined according to the analysis of $PM_{2.5}/PM_{10}$ ratios and literature review of the work of Leeuw & Horálek, (2009), Hamm, *et al.* (2014) and Putaud *et al.* (2010).

## 3.2. Data Description

The data used for the prediction consists on $PM_{10}$ and $PM_{2.5}$ observations taken from stations across Europe of the database Airbase and measurements of the Chemical Transport Model (CTM) from LOTOS-EUROS.
The original data was provided in a Network Common Data Form (NetCDF) file extension and contains measurements for 3 years: 2007, 2008 and 2009. It is arranged in 4299 rows by 1096 columns. The columns correspond to the day number during the three years. The measurements of every day contain 15 variables with information of the stations and time in which the measurements were taken. The measurements selected correspond to the 5th of April of 2009. The 5th of April showed high air pollution levels and it was part of an air pollution event that occurred from the 2nd to the 7th of April, 2009. According to Hamm *et al.* (2014) "$PM_{10}$ concentrations in April were higher and fluctuated in both space and time (overall maximum: 185 $\mu g\ m^{-3}$: maximum daily median: 42 $\mu g\ m^{-3}$)".

After preliminary analysis of the data, the countries described in section 3.1 were retrieved.

### 3.2.1. Airbase Observations

AirBase (http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8#tab-european-data) is the air quality information system maintained by the European Environmental Agency (EEA) through the European topic centre on Air pollution and Climate Change mitigation. It contains air quality data delivered annually from an exchange of information measuring ambient air pollution within the Member States. (European Environmental Agency, 2014).

### 3.2.2. Covariates

In addition to the observed air quality data, it is possible to introduce supplementary data, with better spatial coverage, to improve the interpolation (Denby *et al.* (2005)). Typically the supplementary data should be representative of the data interpolated or should reflect correlation between the physical process that lead to the spatial distribution of the data interpolated. The additional data selected and tested in this case were the following: a) type of area (rural, suburban and urban), b) region and c) CTM.

#### 3.2.2.1. LOTUS-EUROS CTM

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 can simulate air quality on a regional and subregional scale for different components and includes data-assimilation (TNO, RIVM, KNIM, & PBL, 2015).

The formula used to calculate the model (LE) is the following:

$$LE\ formula\colon TPM_{2.5} |\ BC + PPM_{2.5} + SO_4 + NO_3 + NH_4 + Na_f * Na\ to\ seasalt + Dust_f$$

Where:
*TPM2.5* total particulate matter < 2.5 µg m$^{-3}$
*BC* Black Carbon
*$SO_4$* Sulfate
*$NO_3$* Nitrate
*$NH_4$* Ammonium
*$Na_f$* fine sodium
*Na* to seasalt Sodium to seasalt
*$Dust_f$* fine fraction of soil dust

The grid resolution is 0.50° longitude x 0.25° latitude. A plot from the grid can be found in Appendix 2.

# 4. METHODS

## 4.1. Preprocessing and Data Extraction

1. The NetCDF files that contained a matrix with the model measurements and the observations for $PM_{10}$ and $PM_{2.5}$ were downloaded. The content of the file was analysed.

2. The relevant station information was selected. Data frames for $PM_{10}$ observations, $PM_{10}$ model, $PM_{2.5}$ observations, and $PM_{2.5}$ model were created which contained the relevant information for the stations and the observations and model measurements for April 5th, 2009.

3. A projection and transformation of the data coordinates was made from latitude/longitude to the European Terrestrial Reference System ETRS89. The result obtained in meters was then transformed to kilometres.

4. The location of the countries was visually assessed from the plot of the stations to determine which countries do not share a boundary and are not affected or affect the emissions of other European countries. The stations located in islands that belong to European countries were omitted.

5. The countries were joined into the groups A and B. Each of these groups contained 3 different regions as described in section 3.1. The incomplete and unknown data cases were removed.

6. An R file was created containing two data frames one for $PM_{10}$ and another for $PM_{2.5}$ measurements to work with cokriging.

7. The measurements were exctracted into a data frame for those stations that had observations of both, $PM_{10}$ and $PM_{2.5}$, for the 5th of April 2009 to perform compositional data analysis and compositional kriging.

## 4.2. Kriging of compositional Data

1. The coarse fraction of particulate matter was computed through:

$$PM_{coarse} = PM_{10} - PM_{2.5}$$

Figure 1 shows an explanation of the distribution of particulate matter sizes and PMcoarse

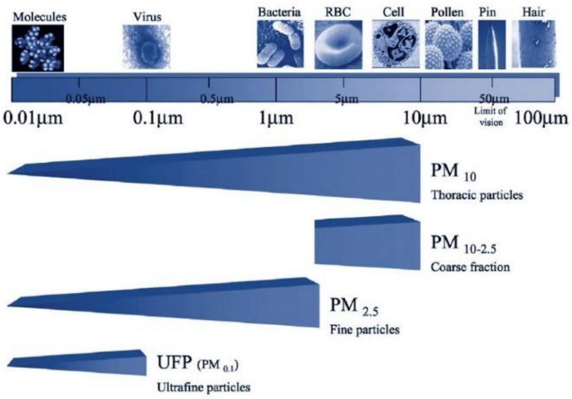


*Figure 1* Classification of particulate matter*

*Author: Linares, C & Díaz, J. (2008) El ecologista no. 58. Instituto de Salud Carlos III. Madrid, Spain. p. 6

2. The ratios $PM_{2.5}/PM_{10}$ and $PM_{coarse}/PM_{10}$ were computed for the countries of interest and the results were analysed.

3. The mean and standard deviation was calculated and analyzed for each of the groups. The values were grouped for the stations of:

   - The same country
   - The same region: Central-Eastern Europe, North-Western Europe and Southern Eruope and later for Belgium, Netherlands and Luxemburg (BENELUX), Germany and Poland, Czech Republic and Slovakia
   - The same type of area: rural, suburban and urban.

4. The data was treated as a composition by first "closing" the components to sum up to unity. To treat the data as a compositional data set it is necessary to assign a scale to the measurements. According to van den Boogaart & Tolosana-Delgado, (2013) from the scales one should be the preferred, as it is the one grounded on the mildest hypotheses: the Aitchinson geometry for compositions. Depending in the view of the problem other scales might be more appropriate. The scales are:

   **"rmult" Real Multivariate scale:** This is a scale to analyse multivariate vectors for which other multivariate methods such as cluster analysis and multivariate regression were created. In this case, the negative values are meaningful. The multivariate normal distribution is its central statistical model.
   **"rplus" Real interval restricted to the positive (plus) real orthant:** The rplus scale allows to perform the same process as with multivariate scale for positive data but it does not take into account some principles of compositional data analysis such as being scaling invariant. This leads to problems in the interpretation of the results, therefore it was not considered.
   **"aplus" Aitchison (i.e., ratio) geometry):** In this approach the amount of each component can be individually analysed as ratios of the components and distances computed based on the log transformed data. Therefore, the natural central distribution in this case should be the multivariate lognormal. This allows the data to be treated as a composition but many methods based on this approach are not scaling invariant.
   **"rcomp" Real (i.e, interval) compositional scale**: Treats the data as multivariate real datasets but does not considers the necessary constraints to treat compositional data. Not considered because of its similarities with "rplus" leading to difficulties on the interpretation. It does preserve mass.
   **"acomp" Aitchinson compositional scale:** It follows all the principles for the statistical analysis of compositional data: scaling invariance, perturbation invariance, permutation invariance, and subcompositional coherence. The reference distribution is the additive-logistic-normal distribution or the normal distribution on the simplex. The mathematical structure is a vector space structure in its own right isometrically equivalent to $\mathbb{R}^{D-1}$. An advantage of this equivalence is that it is possible to convert any statistical problem involving compositions of $D$ parts onto a classical multivariate problem involving real vectors of $D-1$ coordinates.
   **"ccomp" count compositional scale:** Used to treat count compositions. Not considered because the dataset to be used is not formed by counts but by true observations.

5. In order to translate real vectors to compositions we need to make a transformation from the simplex to the real space. Following Aitchison, (1999) there is equivalency between any $D$-part composition and its logratio vector. We can obtain an equivalence through an isometry which preserves angles and distances.

   The first type of transformation is the centred log ratio transformation (clr) which allows to treat the parts symmetrically:

$$clr(x) = \left(ln\frac{x_i}{g(x)}\right)_{i=1\dots D}$$

$$\text{With: } g(x) = \sqrt[D]{x_1 \cdots x_D}$$

   Where $x$ is the composition of $D$ number of parts (columns). In this case $D$ is equal to 2.

The log ratio of the vector is applied component-wise. The components sum up to zero. The image of the clr is a hyperplane of the real space orthogonal to the vector 1=[1,…,1], i.e., the bisector of the first orthant. This may be a source of problems when doing statistical analysis, as e.g., the variance matrix of a clr-transformed composition is singular. This transformation allows to use standard unconstrained multivariate methods (Aitchison, 2003).

The second is the Isometric Log-ratio Transformation (ilr). This is an isometric linear mapping between the Aitchison simplex and $\mathbb{R}^{D-1}$. The isometry is constructed by representing the result in the basis of the $(D-1)$ dimensional image space $\mathbb{H}$ of the clr transformation. It is possible to arrange the vectors $\{V_j *\}$ by columns in an $D \times (D-1)$ element matrix, denoted by $V$, with the following properties:

- It is a quasi-orthonormal matrix

$V^t \cdot V$ is an identity matrix of $D-1$ elements:

$$V^t \cdot V = I_{D-1}$$

$$V \cdot V^t = I_D - \frac{1}{D} 1_{D\times D}$$

Where $1_{D\times D}$ is a matrix full of ones.

- Its columns sum up to zero because they represent vectors of the clr-plane

$$ilr_v(x) = clr(x) \cdot V = \ln(x) \cdot V$$

$V$ is a matrix of $D$ rows and $D-1$ columns.

The transformation provides the coordinates of any composition with respect to a given orthonormal basis. The ilr transformation induces an isometric identification of $\mathbb{R}^{D-1}$ and $\mathbb{S}^D$. For measure and probability theory purposes, this induces an own measure for the simplex, called the Aitchinson measure on the simplex. (van den Boogaart & Tolosana-Delgado, 2013).

The philosophy of logratio analysis is the following:

1. Formulate the compositional problem in terms of the components of the composition.
2. Translate this formulation into terms of the logratio vector of the composition.
3. Transform the compositional data into logratio vectors.
4. Analyse the logratio data by an appropiate standard multivariate statistical method.
5. Translate back into terms of the compositions the inference obtained at the analysis (Aitchison, 2003).

6. The linear regression model was created for the case in which both, the dependent and independent variable, are a composition. The dependent variable was stated as the observations and the independent as the measurements of the model using the equation:

$$ilr(Y) = a + ilr(X) \cdot B + \varepsilon_i$$

Where:

$ilr(Y)$ and $ilr(X)$ are the isometric log ratio transformation of the dependent and independent compositional variable respectively.

$B$ is a square matrix representing a linear transformation between compositions in ilr space

$\varepsilon_i$ is the error

From this equation three models were obtained, one for each region. Later, a fourth model was computed testing two categorical covariates (the type of area and region) and a continuous variable (the CTM). The models where tested using the Analysis of variance (ANOVA) procedure.

7. Descriptive Statistics of the scaled dataframe:
The data was tested for normal disbrution using QQ-plots for compositional data with $\alpha = 0.05$.
The compositional mean of the dataset with $N$ number of observations and parts is the composition

$$\bar{x} = \frac{1}{N} \odot \oplus_{n=1}^{N} x_n = clr^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} clr(x_n)\right) =* \, lr^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} * \, lr(x_n)\right)$$

Where:

$\odot$ Denotes power transformation (geometrically equivalent to scaling)

$\oplus$ Denotes perturbation (translation of a composition)

$x_n$ Is the n$^{th}$ observation of the n$^{th}$ row of the compositional dataset $X$

$* \, lr$ Represents one of the log-ratio transformations.

The spread of the data was calculated through three methods:
The first method is the metric variance which is defined as the average distance between the composition and the mean and is equivalent to the average variation:

$$mvar(X) = \frac{1}{N-1}\sum_{n=1}^{N} d_A^2(x_n, \bar{x})$$

where $d$ is the distance from the centre to the dataset. To calculate the metric variance, the average of the squared distance is calculated and divided into the corrected degrees of freedom.

The second method was the metric standard deviation:

$$msd(X) = \sqrt{\frac{1}{D-1} mvar(X)}$$

This is the square root of the metric variance divided by the number of dimensions minus one $(D - 1)$. It can be interpreted as an average spread when the variance is not the same in all directions. When the average spread is the same in all direction this is the radial standard deviation on a log scale.

The codependence of the components needs a special treatment to avoid obtaining spurious effects. The equation of the metric variance does not give as a result any information about the codependence of components. To tackle this problems, Aitchison, (2003) suggests the variation matrix to replace the correlation. The variation matrix is symmetric and has $D^2$ components calculated with the following equation:

$$\tau_{ij} = var\left(ln\frac{x_i}{x_j}\right)$$

And estimated by

$$\hat{\tau}_{ij} = \frac{1}{N-1}\sum_{n=1}^{N} ln^2\,\frac{x_{ni}}{x_{nj}} - ln^2\,\frac{\bar{x}_i}{\bar{x}_j}$$

where $N$ is the number of observations. In other words, each component of the matrix is calculated as the variance of the natural logarithm of the *i*th part of the composition $x_i$ divided by the *j*th part $x_j$.

The variation matrix Ʈ has zero diagonal elements, and cannot be expressed as the standard covariance matrix of some vector. The smaller a variation element is, the better the proportionality between the two components (van den Boogaart & Tolosana-Delgado, 2013).

8. A variogram was created in order to analyse second-order moment descriptions of the form:

$$\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{(i,j) \in N(h)} \left(Z(x_i) - Z(x_j)\right)^2$$

$$N(h) = \{(i,j): \|x_i - x_j\| \approx h\}$$

Which shows the variance of the squared differences of $Z(x_j)$ and $Z(x_j)$at their respective location, separated by a distance $h$.
Later, an appropriate variogram model was fitted to the empirical variogram.

9. The values were predicted using compositional kriging which, according to Walvoort & Gruijter, (2001), is an extension of ordinary kriging. It considers all components simultaneously by minimizing the sum of their prediction error variances and by taking the unbiasedness, nonnegativity, and constant sum constraints into account. The equation is the following:

$$min_{w_k} \sum_{k=1}^{p} (\sigma_k^2 + \boldsymbol{w}_k^T \boldsymbol{C}_k \boldsymbol{w}_k - 2\boldsymbol{w}_k^T \boldsymbol{d}_k)$$

$$\boldsymbol{W}^T 1^{(n)} = 1^p$$

$$\boldsymbol{w}_k^T \boldsymbol{z}_k \geq 0 \; for \; k = 1, \ldots, p$$

$$tr(\boldsymbol{W}^T \boldsymbol{Z}) = 1$$

Where:
$\sigma^2$ is the variance of the $k$th component of $z(x_i)$,
$\boldsymbol{w}_k$ is the $k$th column of $W$,
$\boldsymbol{C}_k$ is the $n\,x\,n$ matrix containing the covariances between the data points for component $k$,
$\boldsymbol{d}_k$ is the vector of dimension $n$ containing the covariances between the data points and the prediction point for component $k$,
$\boldsymbol{z}_k$ represents the $k$th column of $Z$,
$tr()$ gives the trace of its argument.

10. The data were backtransformed.

## 4.3. Cokriging of $PM_{2.5}$ and $PM_{10}$

1. The complete extracted data set for $PM_{10}$ and $PM_{2.5}$ was used. In this case it is possible to use all the measurements since cokriging allows to use a sparsely sampled data as a primary variable ($PM_{2.5}$) in combination with abundant secondary information ($PM_{10}$).

2. The values to create data frames for the validation and prediction were retrieved. The validation data extracted was 25% of the total number of measurements. The remaining 75% was used for prediction. The validation and prediction data was plotted to see the spatial distribution of the stations used in each process.

3. The $PM_{10}$ and $PM_{2.5}$ data was inspected with descriptive statistics. The mean, median, first and third quartile and the minimum and maximum value were examined from the summary statistics. Both the histogram and Q-Q graph were plotted to check the probability distribution of the values.

4. The correlations between $PM_{10}$ and $PM_{2.5}$ measurements were computed and compared with the correlation between the log transformations of the two types of particulate matter.

5. A linear model with $PM_{2.5}$ as the primary variable and $PM_{10}$ as the secondary variable was created. Different models using different covariates were tested using ancillary and surrogate data. The best model according to the available data was selected using the criteria of the $R^2$ value.

6. The variograms were computed in order to later interpolate the data. According to Singh et al., (2011). The weights are estimated on the base of two semi-variograms and the cross variogram which describes the correlation between the two variables:

$$\gamma ZY(h) = \frac{1}{2N(h)} \sum_{(i,j)} [Z(x_i) - Z(x_j)\big(Y(v_i) - Y(v_j)\big)]$$

   where $N(h)$ is the number of station pairs $(x_i, x_j)$ separated by $h$.

7. The prediction was made by including the correlation of the variables $PM_{10}$ and $PM_{2.5}$ using cokring. The general equation is the following:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(X_i) + \sum_{j=1}^{m} \eta_j Y(X_j)$$

   Where
   $Z(s_i)$ are the primary data at a measurement point ( in this case $PM_{2.5}$)
   $Y(s_i)$ are the secondary data (in this case $PM_{10}$)
   $\lambda_i$ and $\eta_j$ are the weights, which are based on knowledge of the variograms and the crossvariogram

   According to Denby, *et al.* (2005) the crossvariogram is determined using the covariance of the two quantities $Z$and $Y$in a similar manner to the use of the variance to determine the semivariogram.

8. The mean error and root mean squared error was calculated using the following equations:

   Mean error

$$\varepsilon = \frac{1}{N} \sum_{i=1}^{N} [Z(x_i) - \hat{Z}(x_i)]$$

   Root mean squared error

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} [Z(x_i) - \hat{Z}(x_i)]^2}$$

where:
$N$ is the number of observations
$Z(x_i)$ is the measured concentration
$\hat{Z}(x_i)$ is the estimated concentration

# 5. RESULTS AND ANALYSIS

## 5.1. Preprocessing and Data Extraction

A visualization of the stations shown in figure 2 still shows some stations that do not share boundaries with other countries, even after modifying the bounding box to remove extreme locations.

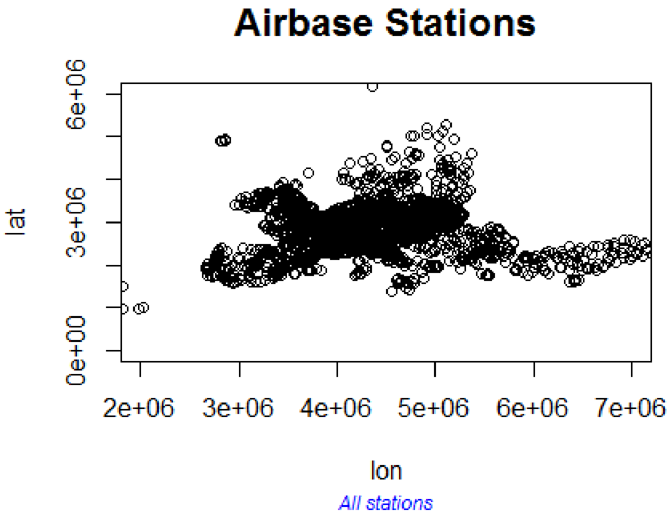


*Figure 2* Airbase stations locations (latitude and longitude) in Europe. Units: m.

These stations were removed and the countries with a greater number of stations and coherent data where retrieved. The countries that had missing values or unknown data where discarded from the analysis.

The following plots show the remaining stations for $PM_{2.5}$ (left plot) and $PM_{10}$ (right) for region A:

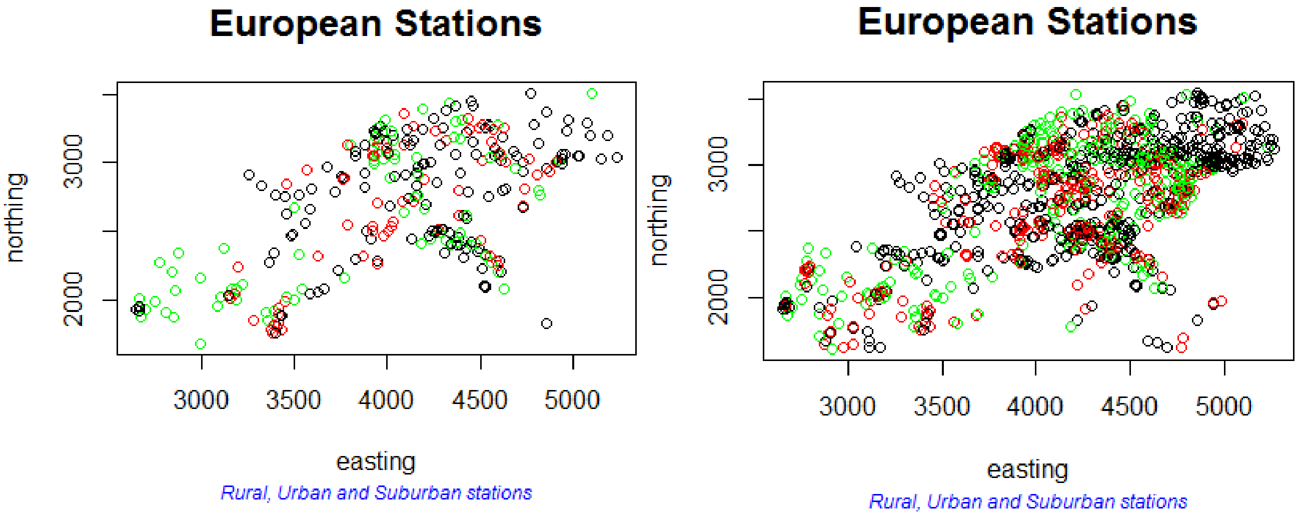


*Figure 3* Distribution of $PM_{2.5}$ and $PM_{10}$ European stations of group A in Cartesian coordinates (ETRS89). Units: km.

Rural stations are represented in green, suburban stations in red and urban stations in black.

The locations of the extracted values for the $PM_{2.5}$ (left) and $PM_{10}$ (right) of the selected countries for region B (Belgium, Netherlands and Luxemburg) are shown on the following figure:

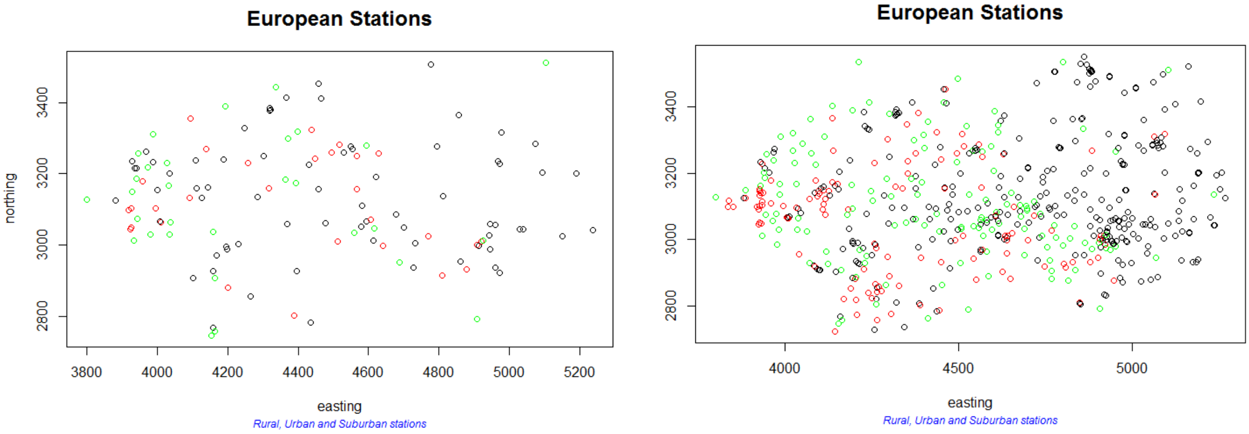


*Figure 4* Distribution of $PM_{2.5}$ and $PM_{10}$ European stations of group B in Cartesian coordinates. Units: km. Green represents rural stations, red suburban and black urban stations.

The number of stations of $PM_{2.5}$ measurements is smaller than for $PM_{10}$ on both groups of regions.

## 5.2. Kriging of compositional Data

### 5.2.1. Compositional Analysis and Prediction for the group of regions A: Central Eastern Europe (CEE), North-Western Europe (NWE) and Southern Europe (SE)

1. Computing of the coarse fraction

After joining the tables the number of stations available for the analysis was 255. The table presented repeated columns which were removed. The coarse measurements of the model and observations were computed.

2. Computing of ratios

The descriptive statistics for $PM_{2.5}$ and $PM_{coarse}$ measurements and ratios were computed. During the analysis the $PM_{coarse}$ fraction showed for some stations negative and zero values, indicating that the $PM_{2.5}$ value was higher than $PM_{10}$. This result did not seemed reasonable since $PM_{2.5}$ measurements are a fraction of $PM_{10}$ and therefore these stations where removed from the data frame.

*Table 1* Statistics of the compositional data for the group A

| **Observation ($\mu g \cdot m^{-3}$)** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
|---|---|---|---|---|---|---|
| **$PM_{2.5}$** | 3.75 | 17 | 29.61 | 32.57 | 44.53 | 104.60 |
| **$PM_{Coarse}$** | -29.64 | 5 | 9 | 9.96 | 13.36 | 50.333 |
| **$PM_{10}$** | 8 | 25.70 | 38 | 42.52 | 55.57 | 113.42 |
| **Ratios $_{2.5/10}$** | 0.23 | 0.63 | 0.77 | 0.75 | 0.86 | 1.70 |
| **Ratios $_{coarse/10}$** | -0.7 | 0.14 | 0.23 | 0.25 | 0.37 | 0.77 |
| **Model** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
| **$PM_{2.5}$** | 2.28 | 12.97 | 21.55 | 21.17 | 26.95 | 56.04 |
| **$PM_{Coarse}$** | 1.02 | 4.16 | 5.76 | 5.90 | 7.36 | 13.31 |
| **$PM_{10}$** | 4.58 | 18.34 | 27.50 | 27.06 | 33.99 | 69.35 |
| **Ratios $_{2.5/10}$** | 0.45 | 0.70 | 0.78 | 0.76 | 0.84 | 0.93 |
| **Ratios $_{coarse/10}$** | 0.07 | 0.16 | 0.22 | 0.24 | 0.29 | 0.54 |

The following table shows the results of the statistics for the 238 remaining stations:

*Table 2* Statistics of the compositional data without negative values for the group A

| **Observation ($\mu g \cdot m^{-3}$)** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
|---|---|---|---|---|---|---|
| **$PM_{2.5}$** | 3.75 | 17.19 | 29 | 31.62 | 44.0 | 81.25 |
| **$PM_{coarse}$** | 0.29 | 6 | 9.52 | 11.23 | 13.81 | 50.33 |
| **$PM_{10}$** | 8 | 26.10 | 38.00 | 42.86 | 55.40 | 113.42 |
| **Ratios $_{2.5/10}$** | 0.23 | 0.62 | 0.76 | 0.72 | 0.84 | 0.99 |
| **Ratios $_{coarse/10}$** | 0.01 | 0.16 | 0.24 | 0.28 | 0.38 | 0.77 |
| **Model** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
| **$PM_{2.5}$** | 2.28 | 13 | 21.53 | 20.84 | 26.92 | 46.49 |
| **$PM_{coarse}$** | 1.63 | 4.18 | 5.74 | 5.79 | 7.20 | 12.30 |
| **$PM_{10}$** | 4.58 | 18.50 | 27.47 | 26.63 | 33.76 | 54.20 |
| **Ratios $_{2.5/10}$** | 0.45 | 0.71 | 0.78 | 0.76 | 0.84 | 0.93 |
| **Ratios $_{coarse/10}$** | 0.07 | 0.16 | 0.22 | 0.24 | 0.29 | 0.54 |

The statistics of the stations seem to be more reasonable without negative values.

The stations left after the join (890 for $PM_{10}$ and 38 for $PM_{2.5}$) were used for the validation.

3. Analysis of the descriptive statistics for different groups

The data was grouped according to certain characteristics to identify if there was a pattern or relation between measurements with similar spatial context. The groups were made according to: the type of station (rural, suburban, urban), country, and region (Central Eastern Europe, North-Western Europe, Southern Europe).

The following tables show the number of stations, the average of the $PM_{2.5}/PM_{10}$ ratio and the standard deviation for the data grouped by type of station, country and region:

*Table 3* Mean and standard deviation of the ratio for the type of area, group A

| Type of area | Number of stations | Mean of the ratio | Standard Deviation of the ratio |
|---|---|---|---|
| **Rural** | 64 | 0.68 | 0.18 |
| **Suburban** | 60 | 0.75 | 0.15 |
| **Urban** | 114 | 0.72 | 0.16 |

*Table 4* Mean and standard deviation of the ratio for the type of area and country, group A

| Country | No. of stations | | | | Rural | | Suburban | | Urban | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R | S | U | T | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| **Austria** | *2* | *2* | *4* | ***8*** | 0.81 | 0.12 | 0.64 | 0.16 | 0.66 | 0.07 |
| **Belgium** | *6* | *6* | *2* | ***14*** | 0.79 | 0.04 | 0.79 | 0.09 | 0.76 | 0.04 |
| **Switzerland** | *2* | | *1* | ***3*** | 0.82 | 0.17 | | | 0.88 | NA |
| **Czech Republic** | *4* | *8* | *10* | ***22*** | 0.78 | 0.16 | 0.82 | 0.10 | 0.70 | 0.15 |
| **Germany** | *8* | *12* | *23* | ***43*** | 0.81 | 0.12 | 0.80 | 0.10 | 0.81 | 0.06 |
| **Spain** | *17* | *9* | *5* | ***31*** | 0.59 | 0.18 | 0.61 | 0.24 | 0.77 | 0.12 |
| **France** | *4* | *17* | *24* | ***45*** | 0.67 | 0.23 | 0.79 | 0.10 | 0.67 | 0.15 |
| **Italy** | *8* | *5* | *24* | ***37*** | 0.71 | 0.13 | 0.67 | 0.09 | 0.75 | 0.15 |
| **Netherlands** | *5* | *1* | *3* | ***9*** | 0.80 | 0.12 | 0.82 | NA | 0.91 | 0.12 |
| **Poland** | *1* | | *14* | ***15*** | 0.78 | NA | | | 0.66 | 0.15 |
| **Portugal** | *7* | | *4* | ***11*** | 0.46 | 0.09 | | | 0.35 | 0.11 |

*Table 5* Mean and standard deviation of the ratio for the type of area and region, group A

| Region | No. of stations | | | | Rural | | Suburban | | Urban | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R | S | U | T | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| **Central Eastern Europe** | *17* | *22* | *52* | ***91*** | 0.80 | 0.12 | 0.79 | 0.11 | 0.74 | 0.13 |
| **North-Western Europe** | *14* | *21* | *23* | ***58*** | 0.79 | 0.09 | 0.79 | 0.10 | 0.72 | 0.15 |
| **Southern Europe** | *33* | *17* | *39* | ***89*** | 0.58 | 0.17 | 0.66 | 0.20 | 0.69 | 0.19 |

The mean of the ratios can give information about those countries or regions for which $PM_{2.5}$ has a great local contribution on the $PM_{10}$ observations. The greater the value of the mean ratio is, the higher the contribution of $PM_{2.5}$. Since the particulate matter smaller than 2.5 µm are mainly generated from combustion of fuels and anthropogenic sources, while the $PM_{10}$ has an important explanation on natural phenomena, we expect higher ratios in areas where this type of human activity occurs. On table 5 we can see that the mean ratio decreases from rural to urban for Central Eastern Europe. From this region we can see that the Netherlands has very high mean ratios, particularly for the urban area which could have an explanation on shipping activities and the high population density. Regarding the North-western region the mean ratios are similar for the three types of area and the dispersion from the mean value is also low, but the total number of stations is the lowest. The ratios for South Europe are low compared to the other two regions, showing a higher contribution of the coarse fraction. The low rural ratio could have an explanation on the contribution of mineral Sahara dust (Leeuw & Horálek, 2009). In general the suburban region seems to contribute more with $PM_{2.5.}$

There does not seem to be a clear pattern of the weight of $PM_{2.5}$ over $PM_{10}$. Using data from all stations it becomes clear that in the rural area the fraction of $PM_{2.5}$ is lower but this differ according to each country.

The distribution of the stations is also important on the analysis of the ratios. When understanding the mean value and the standard deviation it is important to consider that some countries have a greater number of stations for one group compared with another. For example, for the type of area, there are about double the amount of urban stations (114) as there are rural (64) or suburban stations (60). However, in the aggregation per country Spain has a greater number of rural stations (17) compared to suburban (9) and urban (5), so care must be taken. The previous statement can be reaffirmed by looking at the distribution of the stations by type of area showed on figure 5.

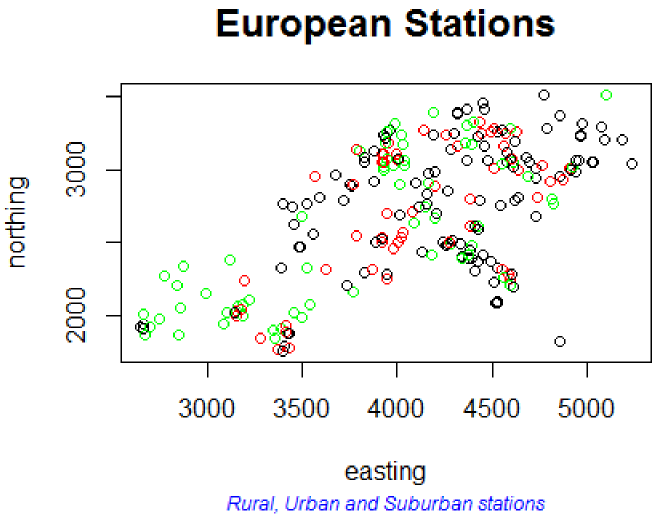


*Figure 5* Spatial distribution of the retrieved stations for the compositional analysis of group A in Cartesian coordinates (ETRS89). Units: km. The green dots represents rural stations, red suburban and black urban stations

Figure 5 shows the distribution of the stations represented by green for rural, red for suburban and black for urban. It is clear that in the South-east part of the continent (especially in Spain) the number of rural stations is higher than the urban and suburban number of stations. In France, Italy and Germany the number of urban stations is evidently higher than the number of rural and suburban. This can be corroborated with the data on table 4.

Some of the countries do not have enough stations for the analysis. That is the case of Switzerland, which has only 3 stations for the country. Given the problem of the concentration of the type of station on certain countries and the low number of stations in each country, the analysis by country was discarded.

Once the ratios were analysed it was decided that the further statistical analysis would be performed to the stations that share the same region because they seem to have a relatively similar value of standard deviation (less than 0.2 for the three types) with respect to the country and region. In addition to this, they have enough number of stations to work with, later on the variogram and compositional kriging.

4. Closing the composition and giving scale to the measurements

After selecting the group to use, the next step was to treat the data as compositions. For this the package "compositions" (van den Boogaart & Tolosana-Delgado, 2008) was used to close the parts of the compositions to sum up to one, which is the same principle as calculating the ratios, and to give scale to the measurements. The scale selected as appropriate for the dataset was "acomp" which is described on section 4.2.

5. Isometric transformation of the data

Figure 6 and figure 7 show the results obtained for the scaled measurements, Centered log-ratio (clr) transformation and Isometric log-ratio (ilr) transformation of each type of station:

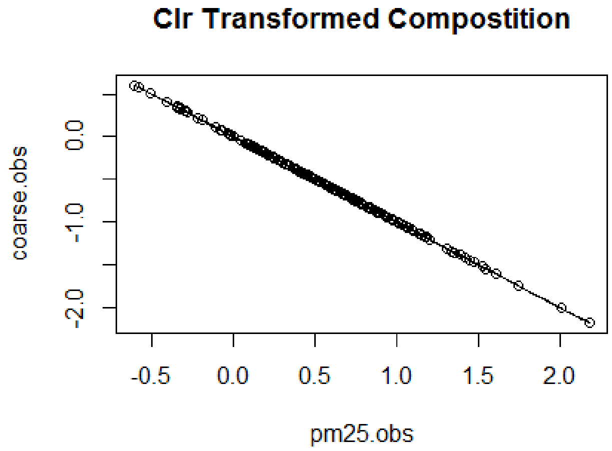


*Figure 6* Centred log-ratio transformation of the composition for group A

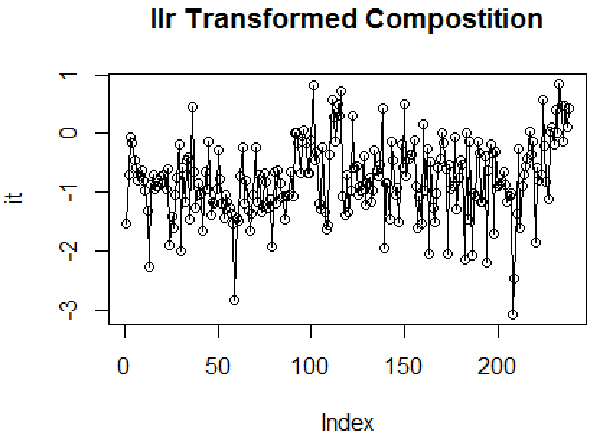


*Figure 7* Isometric log-ratio transformation of the composition for group A

6. Construction of the linear model

The model is intended to include the covariates that would improve the results of the prediction. The performance of three different covariates was tested. The continuous covariate was the CTM and the categorical covariates were the type of area and the region of the station. The results of the combination of different covariates with the response variable (compositional data with 238 observations) are shown in table 6.

*Table 6* Estimates for the regression models of compositional data group A

| Estimate/Model covariates | CTM, Type of area and Region | CTM and region | CTM and type of area | Region and type of area | CTM |
|---|---|---|---|---|---|
| **Residual Standard Error** | 0.5807 | 0.5854 | 0.6064 | 0.6104 | 0.6136 |
| **Degrees of Freedom** | 220 | 232 | 232 | 229 | 236 |
| **Adjusted R$^2$** | -0.07727 | -0.02155 | -0.02155 | -0.03493 | -0.004237 |

The model with the lowest residual standard error is the one which includes all the covariates: the Chemical Transport Model, type of area and region. Nevertheless, the summary statistics for this model indicates that none of the covariates are significant showing probabilities Pr(>|t|) greater than 0.1 for all the covariates. The R$^2$ yields negative values for all the models. It is important to mention that, according to van den Boogaart & Tolosana-Delgado, (2013) it is not recommended to work with the standard summary of linear models involving composition since it gives to the (arbitrary chosen) ilr transformation, an excessive importance. This is why the Analysis of Variance is also presented for the different models.

After testing the different combinations of the covariates, the chosen model for this analysis was the following:

lm(formula = ilr(y) ~ ilr(x1) * x3)

Where: ilr(y) is the response variable, the composition transformed by an isometric log-ratio transformation.
ilr(x1) Is an exploratory variable. It is the isometric log-ratio transformation of the CTM as a composition.
x3 the region of the station as a factor

This model shows a low residual standard error and also significant covariates. The coefficients found for this model are shown on tables 7 and 8:

*Table 7* Summary statistics of the selected compositional model for group A

| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
|---|---|---|---|---|
| -2.20307 | -0.33295 | 0.07138 | 0.38760 | 1.43911 |
| **Coefficients** | **Estimate** | **Standard Error** | **t value** | **Pr(>\|t\|)** |
| **(Intercept)** | -0.77203 | 0.20831 | -3.706 | 0.000263 *** |
| **ilr(x1)** | 0.16659 | 0.20569 | 0.810 | 0.418832 |
| **x3NWE** | 0.04813 | 0.30581 | 0.157 | 0.875072 |
| **x3SE** | 0.77327 | 0.23729 | 3.259 | 0.001287 ** |
| **ilr(x1):x3NWE** | 0.10796 | 0.33814 | 0.319 | 0.749796 |
| **ilr(x1):x3SE** | 0.49108 | 0.23779 | 2.065 | 0.040019 * |

*Table 8* Analysis of variance of the selected compositional model for the group A

| Coefficients | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **ilr(x1)** | 1 | 11.081 | 11.0812 | 32.3386 | 3.877e-08 *** |
| **x3** | 2 | 7.629 | 3.8145 | 11.1318 | 2.419e-05 *** |
| **ilr(x1):x3** | 2 | 1.740 | 0.8702 | 2.5396 | 0.08109 . |
| **Residuals** | 232 | 79.498 | 0.3427 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The analysis of variance shows that the CTM and region are significant on the model. The combination on the CTM and the type of area is significant at alpha= 0.01. We can state that with 99% confidence Y is dependent on the combination of the isometric log-ratio transformation of the model combined with the region.
The linear models were computed for the compositional data of the observations regressed on the CTM of each type of region (Central Eastern Europe, North-Western Europe and Southern Europe). The results of the model are shown in the following table:

*Table 9* Estimates of regression model (ilr of observations vs ilr of CTM) for each region, group A

| | **Region 1**<br>**Central Eastern Europe** | **Region 2**<br>**North-western Europe** | **Region 3**<br>**Southern Europe** |
|---|---|---|---|
| **Residual standard error** | 0.5432 | 0.5935 | 0.6262 |
| **Degrees of freedom** | 89 | 58 | 112 |
| **Adjusted R-squared** | -0.01124 | -0.01724 | -0.008929 |

The $PM_{2.5}$ and $PM_{coarse}$ model was plotted against the observations. The results for the centered log ratio (ilr) transformation of the region 1 are shown in figure 8.

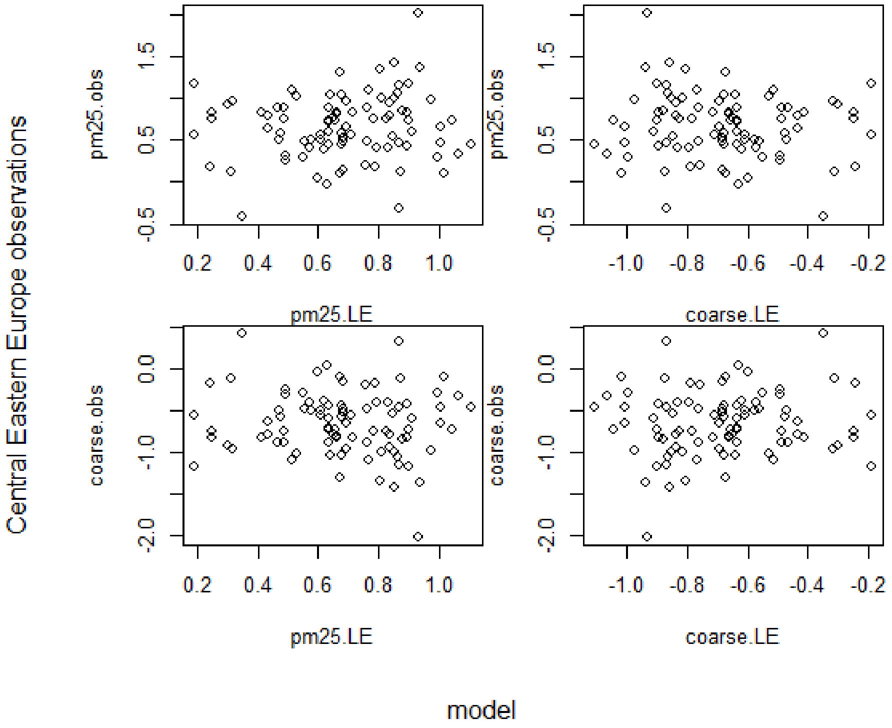


*Figure 8* Central Eastern Europe clr transformed observations vs. model

We can observe that there is a slight linear tendency within both the model and the observations. The region of Central Eastern Europe shows more correlation with the model than the other two regions.

The formula used for the linear model was the following:

lm(formula = ilr(compY1) ~ ilr(compX1))

where:
compY1 is the composition of the observations and compX1 is the composition of the CTM for Central Eastern Europe.

The dependent and the independent variable are transformed into an isometric log-ratio transformation (ilr). The coefficients for the intercept and the ilr transformed CTM data are the following:

*Table 10* Summary statistics for the linear regression model of Central Eastern Europe

| **Residuals:** | | | | |
|---|---|---|---|---|
| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
| -1.85 | -0.36 | -0.01 | 0.34 | 1.44 |
| **Coefficients:** | **Estimate** | **Standard Error** | **t value** | **Pr(>\|t\|)** |
| **(Intercept)** | -0.77 | 0.19 | -3.99 | 0.000134*** |
| **ilr(compX1)** | 0.17 | 0.19 | 0.87 | 0.385150 |

Significant Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The summary statistics show that there is no dependence on Y over X1 according to the probability value above 0.05. The analysis of variance showed the following results:

*Table 11* Analysis of variance for the regression model of Central Eastern Europe stations

| | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
|---|---|---|---|---|---|
| **ilr(compX1)** | 1 | 0.02 | 0.22 | 0.76 | 0.3851 |
| **Residuals** | 89 | 26.26 | 0.29 | | |

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The analysis of variance shows that the independent variable is not significant. The model is not improved by including this covariate.

The clr transformation of the observations against the model was plotted for region 2. For the North-Western Europe data the linear tendency is even less obvious.

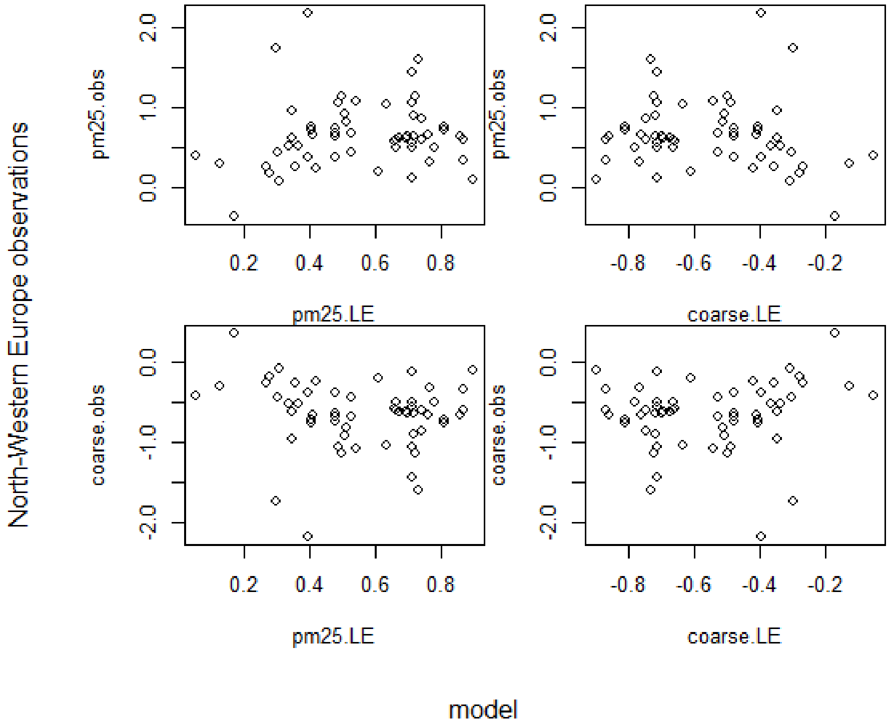


*Figure 9* North-Western Europe clr transformed observations vs. model

The formula used to calculate the coefficients was the following:

lm(formula = ilr(compY2) ~ ilr(compX2))

where compY2 is the composition of the observations and compX2 is the composition of the CTM model for North-Western Europe data. The coefficients obtained for the region 2 (NWE) regression model are the following:

*Table 12* Summary statistics for the linear regression model of the North-Western Europe stations

| **Residuals:** | | | | |
|---|---|---|---|---|
| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
| -2.20 | -0.21 | -0.11 | 0.31 | 1.28 |
| **Coefficients:** | **Estimate** | **Standard Error** | **t value** | **Pr(>\|t\|)** |
| **(Intercept)** | -0.72 | 0.22 | -3.18 | 0.00237** |
| **ilr(compX1)** | 0.27 | 0.27 | 1.01 | 0.31807 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

*Table 13* Analysis of variance for the linear regression model of North-Western Europe stations

| | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
|---|---|---|---|---|---|
| **ilr(compX2)** | 1 | 0.36 | 0.36 | 1.01 | 0.3181 |
| **Residuals** | 56 | 19.79 | 0.35 | | |

From the analysis of variance we can determine that the covariate is not significant in this case because the value of the probability obtained is higher than the F value.

In the case of the Southern Europe data, the plot shows a linear relation between the clr transformation of the model and the observations.

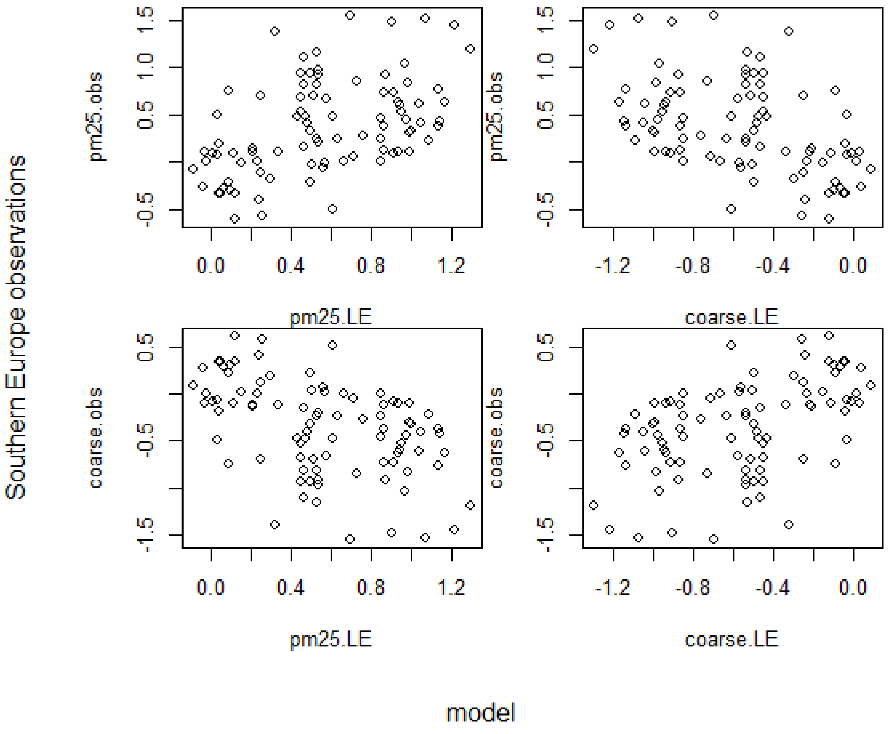


*Figure 10* Southern Europe clr transformed observations vs. model

The formula of the model was:

lm(formula = ilr(compY3) ~ ilr(compX3))

where compY3 is the composition of the observations and compX3 is the composition of the CTM model for Southern Europe data.

*Table 14* Summary statistics for the linear regression model of Southern Europe stations

| **Residuals:** | | | | |
|---|---|---|---|---|
| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
| -1.665 | -0.414 | 0.101 | 0.496 | 1.275 |
| **Coefficients:** | **Estimate** | **Standard Error** | **t value** | **Pr(>\|t\|)** |
| **(Intercept)** | 0.001 | 0.120 | 0.010 | 0.992 |
| **ilr(compX3)** | 0.66 | 0.126 | 5.203 | 1.29e-06*** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

*Table 15* Analysis of variance for the linear regression model Southern Europe stations

| | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
|---|---|---|---|---|---|
| **ilr(compX3)** | 1 | 10.410 | 10.410 | 27.076 | 1.288e-06*** |
| **Residuals** | 87 | 33.449 | 0.384 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The Southern Europe data shows a value for probability in the ANOVA test of 1.288e-06 and it is significant at a level of alpha= 0.

From the results above we can take the dependence of the Y on X as credible.

7. Analysis of the descriptive statistics of the compositional data

***Descriptive Statistics for Central Eastern Europe***

According to van den Boogaart & Tolosana-Delgado, (2013) it is not sufficient to accept joint normality only knowing that the marginals have a normal distribution. To determine a joint normal distribution the projection should be normal onto all directions. The quantile-quantile (Q-Q) plot was displayed to test the compositional normality. The plot for Central Eastern Europe stations is show in figure 11:

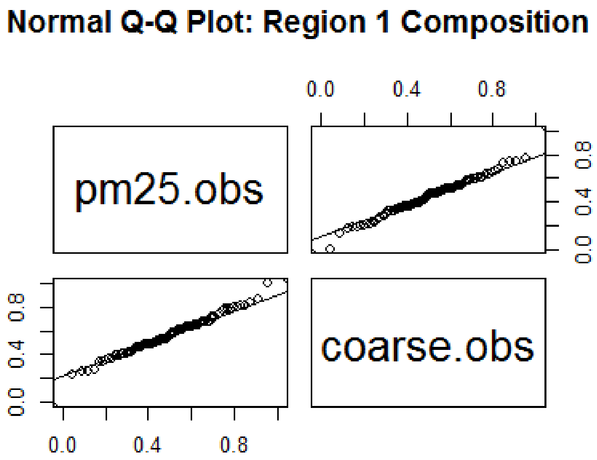


*Figure 11* Q-Q plot for Central Eastern Europe compositional data

The data seems to be normally distributed for both components with alpha=0.05. This joint normality suggest normality in all displayed marginal distributions.

The analysis of the descriptive statistics and the measures of dispersion were calculated for the regional compositions. The results are displayed in table 16:

*Table 16* Descriptive Statistics for Central Eastern Europe stations

| Component | Mean of the composition | Variation Matrix | Metric Variance | Metric Standard Deviation |
|---|---|---|---|---|
| **$PM_{2.5}$** | 0.789 | 0.589 | 0.294 | 0.542 |
| **$PM_{coarse}$** | 0.211 | | | |

The metric standard deviation (msd) is not the square root of the metric variance, but the square root of the mean of the eigenvalues of the variance matrix. "In this way it can be interpreted in units of the original natural geometry, as the radius of a spherical ball around the mean with the same volume as the 1-sigma ellipsoid of the data set" (van den Boogaart & Tolosana-Delgado, 2013). In this case the metric standard deviation can be interpreted as a 0.542 of average spread.

The variation matrix shows a relation in which the smaller a variation element is, the better the proportionality. In this case we can identify a medium variation (0.589) element which make us infer there is some proportionality between the two elements.

***Descriptive Statistics for North-Western Europe***

The North-Western Europe stations (region 2) show a normal distribution on the Q-Q plot with some outliers on the extreme values. The proportionality between the measurements is very low, with a result for the variation matrix of 0.7.

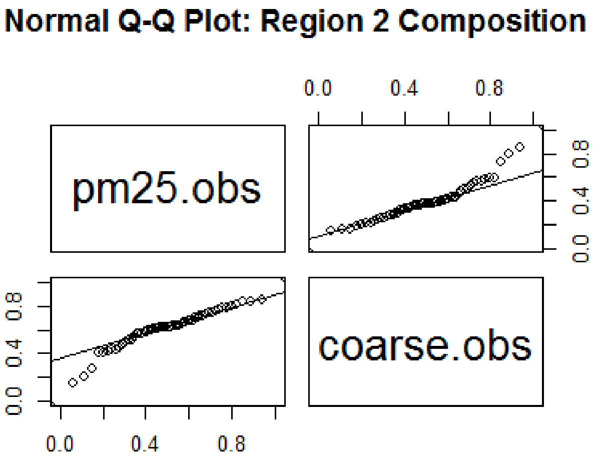


*Figure 12* Q-Q plot for North-Western Europe compositional data

*Table 17* Descriptive Statistics for North-Western Europe stations

| Component | Mean of the composition | Variation matrix | Metric Variance | Metric Standard Deviation |
|---|---|---|---|---|
| **$PM_{2.5}$** | 0.790 | 0.707 | 0.353 | 0.594 |
| **$PM_{coarse}$** | 0.209 | | | |

***Descriptive Statistics for Southern Europe***

The Q-Q plot of the southern European countries data shows a linear pattern in figure 13 indicating that the data is close to normality.

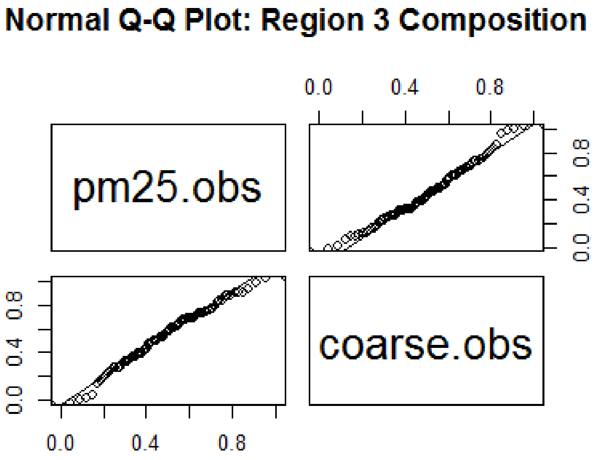


*Figure 13* Q-Q plot for Southern Europe compositional data

*Table 18* Descriptive Statistics for Southern Europe Stations

| Component | Mean of the composition | Variation Matrix | Metric Variance | Metric Standard Deviation |
|---|---|---|---|---|
| **PM$_{2.5}$** | 0.677 | 0.996 | 0.498 | 0.706 |
| **PM$_{coarse}$** | 0.323 | | | |

The proportionality between PM$_{10}$ and PM$_{coarse}$ for this area is very low.

8. Variogram

The empirical variogram for the prediction was calculated for the compositional data. The variogram and model variogram are shown on figure 14, 15 and 16:

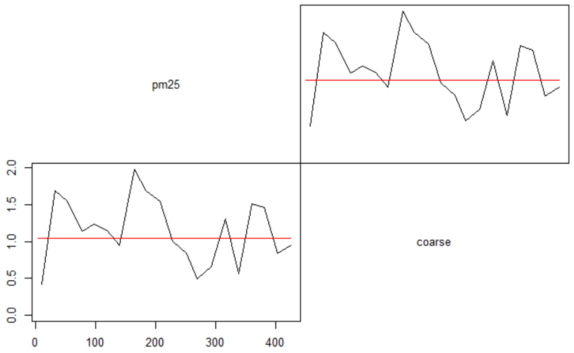


*Figure 14* Empirical Variogram (black line) and Model Variogram (red line) for Central Eastern Europe. Horizontal axis: distance in km, vertical axis: semi-variance

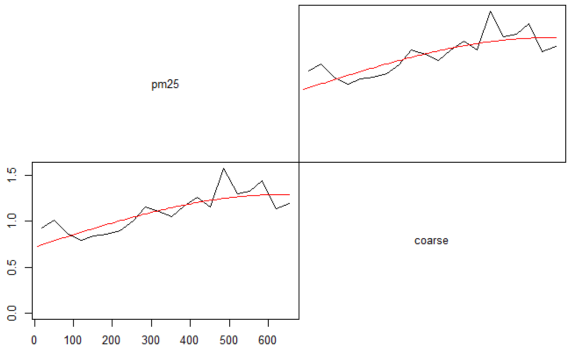


*Figure 15* Empirical Variogram (black line) and Model Variogram (red line) for North-Western Europe. Horizontal axis: distance in km, vertical axis: semi-variance

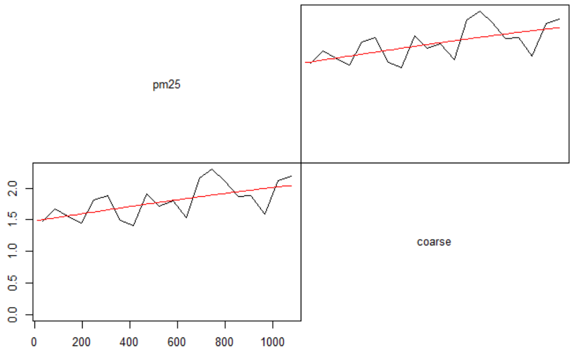


*Figure 16* Empirical Variogram (black line) and Model Variogram (red line) for Southern Europe. Horizontal axis: distance in km, vertical axis: semi-variance

The variogram for Central Eastern Europe shows a trend along the vertical direction, the spatial correlations decreases with the distance, the range is reached at 300 km. There is a cyclic behaviour noticeable on the North-Western Europe variogram as well as for the Southern Europe variogram. The third region also shows a vertical trend. The three variograms clearly show a nugget effect. According to Gringarten & Deutsch, (2001) the nugget effect can appear due to measurement errors or correlation ranges shorter than the sampling resolution. There is a lack of spatial correlation in the variograms, this effect could be linked to the limited number of data and the large distance between the measurements.

9. Predicted values

The predicted values obtained for the group of regions A, are ratios of the measurements for $PM_{2.5}$. We can infer from these values the percentage of $PM_{2.5}$ over $PM_{10}$ and compare the values obtained from the analysis of the ratios made and analyzed in steps 2 and 3. This is in agreement with the ratios obtained showing lower values in Spain, Portugal and Southern France. The highest ratios appear to be in the Central Eastern Europe countries.

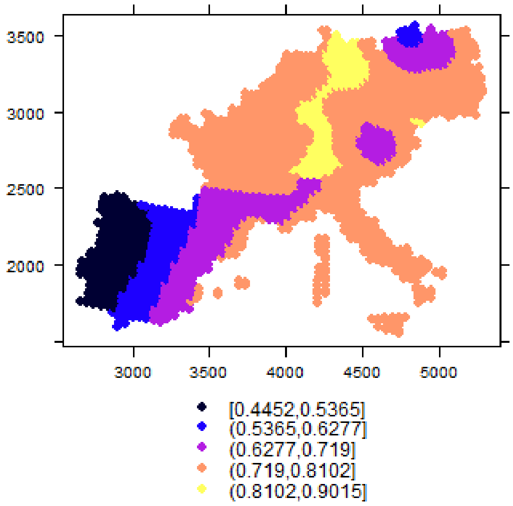


*Figure 17* Predicted values (compositional kriging) of $PM_{2.5}$ ratios for group A for April 5, 2009. Units: scaled $PM_{2.5}/PM_{10}$ ratios using Aitchison compositional scale.

This data can only show the predicted ratios. To evaluate the actual predicted values obtained it is necessary to backtransform the data into terms of µm$^{-3}$. However, this case is difficult since we do not have the total $PM_{10}$

observations for the predicted values. Nevertheless, the data was backtransformed using the total $PM_{10}$ from the Chemical Transport Model grid to obtain values of $PM_{2.5}$.

10. Backtransformed data

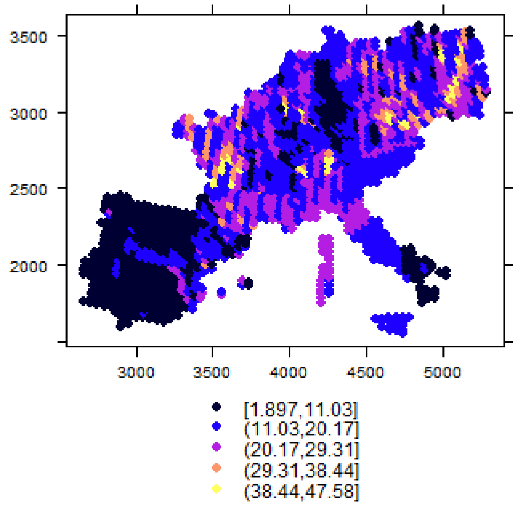


*Figure 18* Backtransformed values (compositional kriging) of $PM_{2.5}$ for group A for April 5, 2009. Units: µg·m$^{-3}$

### 5.2.2. Compositional Analysis and Prediction for the data grouped by the regions: Belgium, Netherlands and Luxemburg (BENELUX), Germany and Poland (GERPOL), Czech Republic and Slovakia (SLOCZ)

1. Computing of the coarse fraction

The stations necessary for the compositional analysis are those which measure both components. A more detailed description of the number of stations per group of countries can be found in table 19:

*Table 19* Number of stations per group for compositional analysis of group B

| Group | PM2.5 | PM10 |
|---|---|---|
| **Rural** | 30 | 62 |
| **Suburban** | 29 | 120 |
| **Urban** | 71 | 299 |
| **BENELUX** | 31 | 132 |
| **GERPOL** | 74 | 374 |
| **SLOCZ** | 25 | 115 |

2. Computing of the ratios

The correlation between the extracted observations of $PM_{10}$ and $PM_{2.5}$ is 0.780. After computing the coarse fraction and the ratios, some stations showed zero and negative values. As before this ratios were assumed to be incorrect measurements and removed from the dataset.

*Table 20* Descriptive statistics after removing negative and zero values for group B

| **Observation ($\mu g \cdot m^{-3}$)** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
|---|---|---|---|---|---|---|
| $PM_{2.5}$ | 12.84 | 35.13 | 44.08 | 45.51 | 55.60 | 81.25 |
| $PM_{Coarse}$ | 0.29 | 7.02 | 10.93 | 14.13 | 18.94 | 50.33 |
| $PM_{10}$ | 20.30 | 45.07 | 55.82 | 59.64 | 74.78 | 113.42 |
| Ratios $_{2.5/10}$ | 0.31 | 0.73 | 0.79 | 0.78 | 0.85 | 0.99 |
| Ratios $_{coarse/10}$ | 0.01 | 0.14 | 0.20 | 0.22 | 0.27 | 0.69 |
| **Model** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
| $PM_{2.5}$ | 8.46 | 20.01 | 25.16 | 26.18 | 30.79 | 46.49 |
| $PM_{Coarse}$ | 3.1 | 5.83 | 6.78 | 6.75 | 7.43 | 10.61 |
| $PM_{10}$ | 13.33 | 25.13 | 32.90 | 32.93 | 37.09 | 54.20 |
| Ratios $_{2.5/10}$ | 0.59 | 0.74 | 0.79 | 0.78 | 0.84 | 0.9 |
| Ratios $_{coarse/10}$ | 0.1 | 0.16 | 0.21 | 0.22 | 0.26 | 0.40 |

The ratios between $PM_{2.5}/PM_{10}$ vary from 0.31 to 0.99 for the observations and from 0.59 to 0.9 for the model. The mean of the ratios is the same (0.78) for the observations and model.

The descriptive statistics of the ratios are very similar between the observations and the model but they are also close to the values obtained for the analysis of the group of regions A described in section 5.2.2. For that case the mean value obtained for $PM_{2.5}/PM_{10}$ and $PM_{coarse}/PM_{10}$ was 0.72 and 0.22 respectively.

3. Analysis of the descriptive statistics for different groups

The mean and the standard deviation was computed for the ratios of different groups. The data was grouped by the type of area, country and regions. In this case, the regions were constructed with less countries and it was expected to show less dispersion. The results are shown on tables 21, 22 and 23:

*Table 21* Mean and standard deviation for the ratios by type of area, group B

| **Type of area** | **Number of stations** | **Mean of the ratio** | **Standard Deviation of the ratio** |
|---|---|---|---|
| **Rural** | 25 | 0.79 | 0.10 |
| **Suburban** | 27 | 0.80 | 0.09 |
| **Urban** | 54 | 0.75 | 0.13 |

*Table 22* Mean and standard deviation for the ratios by type of area and country, group B

| **Country** | **No. of stations** | | | | **Rural** | | **Suburban** | | **Urban** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **R** | **S** | **U** | **T** | ***Mean*** | ***SD*** | ***Mean*** | ***SD*** | ***Mean*** | ***SD*** |
| **Belgium** | *6* | *6* | *2* | ***14*** | 0.79 | 0.04 | 0.79 | 0.09 | 0.76 | 0.04 |
| **Czech Republic** | *4* | *8* | *10* | ***22*** | 0.78 | 0.16 | 0.82 | 0.10 | 0.67 | 0.15 |
| **Germany** | *8* | *12* | *23* | ***43*** | 0.81 | 0.12 | 0.80 | 0.10 | 0.81 | 0.06 |
| **Netherlands** | *5* | *1* | *3* | ***9*** | 0.80 | 0.12 | 0.82 | NA | 0.91 | 0.12 |
| **Poland** | *1* | | *14* | ***15*** | 0.78 | NA | | | 0.66 | 0.15 |
| **Slovakia** | *1* | | *2* | ***3*** | 0.78 | NA | | | 0.68 | 0.23 |

The stations for the individual countries by type of area show a low standard deviation. Although the value of the standard deviation is low, the analysis by country is not considered since some countries have very few stations; that is the case of Slovakia with 3 available stations from where is no suburban stations, and 2 urban and only 1 rural station.

*Table 23* Mean and standard deviation of the ratios by type of area and region, group B

| **Region** | **No. of stations** | | | | **Rural** | | **Suburban** | | **Urban** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **R** | **S** | **U** | **T** | ***Mean*** | ***SD*** | ***Mean*** | ***SD*** | ***Mean*** | ***SD*** |
| **Belgium, Netherlands, Luxemburg** | *11* | *7* | *5* | ***23*** | 0.79 | 0.08 | 0.79 | 0.08 | 0.85 | 0.12 |
| **Germany and Poland** | *9* | *12* | *37* | ***58*** | 0.80 | 0.11 | 0.80 | 0.10 | 0.75 | 0.12 |
| **Slovakia and Czech Republic** | *5* | *8* | *12* | ***25*** | 0.78 | 0.13 | 0.82 | 0.09 | 0.69 | 0.15 |

Germany and Poland show higher values for mean ratios, these countries are also part of the Central Eastern Europe region (for group A in the previous analysis), which also showed the highest mean ratios.

4. Closing the composition and giving scale to the measurements

The data was assigned with the class "acomp" as the group of regions A. The distribution of the stations by the type of area is showed on figure 19. It is important to consider the number of stations in the analysis of the standard deviation. The BENELUX region includes a higher number of rural stations, while the region of Germany and Poland and Slovakia and Czech Republic has a higher number of urban stations.

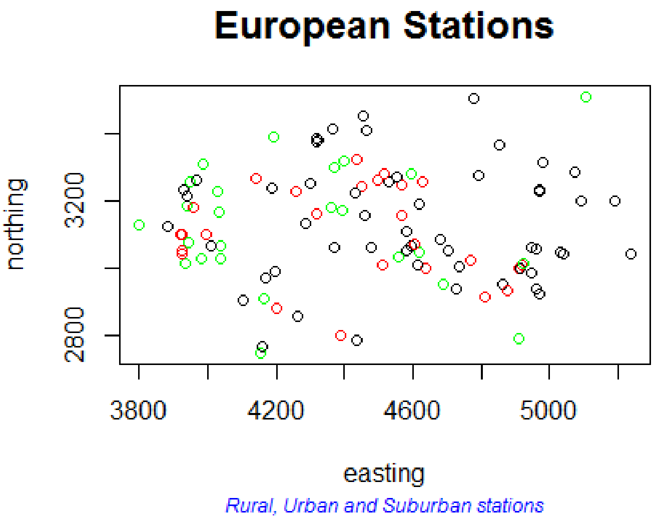


*Figure 19* Distribution of the retrieved stations for the compositional analysis of group B in Cartesian coordinates (ETRS89). Units: km. Rural stations in green, suburban in red and urban in black

5. Isometric transformation of the data

The data was transformed into clr transformation to visually inspect the relation between the measurements and CTM. The clr transformation allows to plot the measurements in $D$ number of dimension while for the computing of the estimates the ilr transformation was used which works with $D - 1$.

6. Construction of the linear model

*Table 24* Estimates for the regression models for the compositional data of group B

| **Estimate/Model covariates** | **CTM, Type of area and Region** | **CTM and region** | **CTM and type of area** | **Region and type of area** | **CTM** |
|---|---|---|---|---|---|
| **Residual Standard Error** | 0.5367 | 0.5633 | 0.5567 | 0.5495 | 0.5702 |
| **Degrees of Freedom** | 88 | 100 | 100 | 97 | 104 |
| **Adjusted R$^2$** | -0.1932 | -0.05 | -0.05 | -0.08247 | -0.009615 |

During the analysis of the second group of data the model including as covariates the ilr transformation of the Chemical Transport Model as a composition, type of area and region shows the lower residual standard error. The stations of the urban type of area are significant covariates at alpha=0.05 and the region "GERPOL" (Germany and Poland) at alpha=0.1. The combination of the isometric logratio transformation of the CTM and the region GERPOL is also significant at 0.1 with a probability of 0.0620. For the purpose of comparison, the model with the ilr transformation of CTM and region was further analyzed.

The clr transformation of the observations was plotted against the centered log ratio transformation of the CTM for the three regions. The clr measurements did not seem to have a linear relation on the graphs. The isometric logratio transformation of the measurements for this regions was also analyzed through the summary statistics and the ANOVA test of the model:

lm(formula = ilr(compY) ~ ilr(compX))

where:
compY is the $PM_{2.5}$ and $PM_{coarse}$ observation as a composition
compX is the $PM_{2.5}$ and $PM_{coarse}$ CTM measurements as a composition
ilr is the isometric log ratio transformation function

7. Analysis of the descriptive statistics

The data for the region formed by Belgium, Netherlands and Luxemburg does not seem to have a distribution close to the normal even after the clr or ilr transformation. Regions 2 and 3 formed by the countries Germany and Poland, and Slovakia and Czech Republic follow a distribution close to the normal. The descriptive statistics for the regions are shown on the following tables:

*Table 25* Descriptive Statistics for Belgium, Netherlands and Luxemburg Stations

| **Component** | **Mean of the composition** | **Variation Matrix** | **Metric Variance** | **Metric Standard Deviation** |
|---|---|---|---|---|
| **$PM_{2.5}$** | 0.83 | 0.84 | 0.42 | 0.65 |
| **$PM_{coarse}$** | 0.16 | | | |

*Table 26* Descriptive Statistics for Germany and Poland Stations

| Component | Mean of the composition | Variation Matrix | Metric Variance | Metric Standard Deviation |
|---|---|---|---|---|
| **PM$_{2.5}$** | 0.8 | 0.56 | 0.28 | 0.53 |
| **PM$_{coarse}$** | 0.20 | | | |

*Table 27* Descriptive Statistics for Slovakia and Czech Republic Stations

| Component | Mean of the composition | Variation Matrix | Metric Variance | Metric Standard Deviation |
|---|---|---|---|---|
| **PM$_{2.5}$** | 0.78 | 0.64 | 0.32 | 0.56 |
| **PM$_{coarse}$** | 0.22 | | | |

The region of Germany and Poland shows a higher relation between the components with the lower value obtained for the variation matrix (0.56). The metric variance and metric standard deviation for Germany and Poland is also the lowest compared to the other two regions. This is in agreement with the results of the evaluation of the models explained above in which the GERPOL region presented a significant probability.

8. Variogram

The variograms for Germany and Poland show that the spatial correlation decreases with the distance. For shorter distances, the spatial correlation between the measurements is higher. The variograms for Germany and Poland and Slovakia and Czech Republic display a nugget effect. The “BENELUX “and “SLOCZ” Region shows a cyclic behaviour with high and low correlation along the distance.

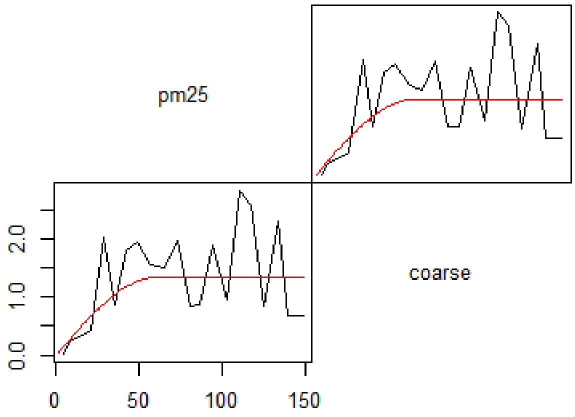


*Figure 20* Empirical Variogram (black line) and Model Variogram (red line) for Belgium, Netherlands and Luxemburg. Horizontal axis: distance in km, vertical axis: semi-variance

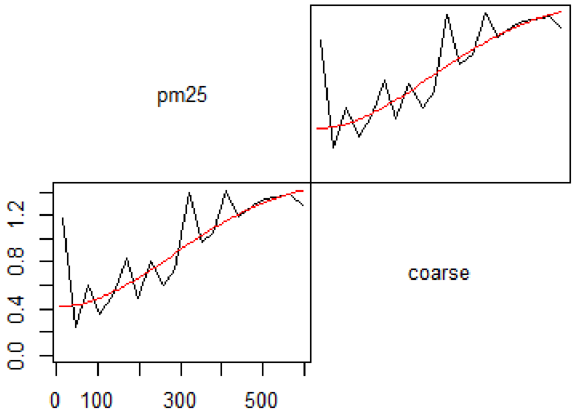


*Figure 21* Empirical Variogram (black line) and Model Variogram (red line) for Germany and Poland. Horizontal axis: distance in km, vertical axis: semi-variance

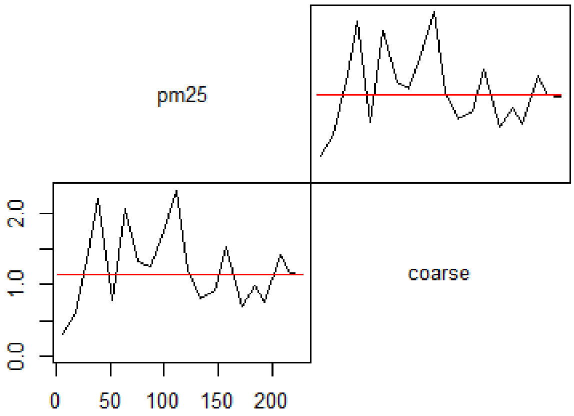


*Figure 22* Empirical Variogram (black line) and Model Variogram (red line) for Slovakia and Czech Republic. Horizontal axis: distance in km, vertical axis: semi-variance

9. Predicted values

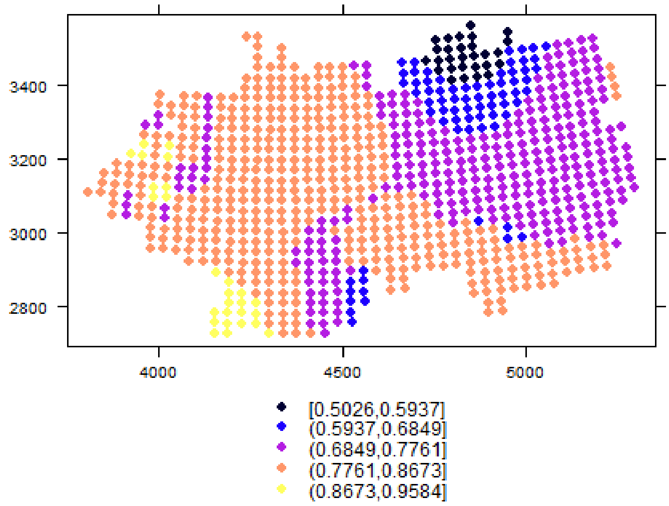


*Figure 23* Predicted values (compositional kriging) of the $PM_{2.5}$ ratios of group B for April 5, 2009. Units: scaled $PM_{2.5}/PM_{10}$ ratios using Aitchison compositional scale.

10. Backtransformed data

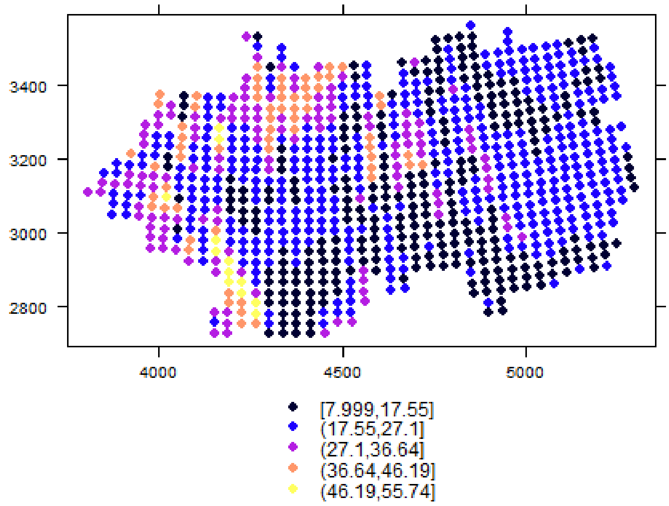


*Figure 24* Backtransformed values (compositional kriging) of $PM_{2.5}$ for group B for April 5, 2009. Units: μg·m$^{-3}$

## 5.3. Cokriging of $PM_{2.5}$ and $PM_{10}$

### 5.3.1. Analysis and prediction using cokriging for the data grouped by the regions: Central Eastern Europe (CEE), North-Western Europe (NWE) and Southern Europe (SE)

1. Analysis of the data set

The data contained data frames for $PM_{10}$ (1145 observations) and $PM_{2.5}$ (293 observations). The primary variable to use was the $PM_{2.5}$ since it is sparse and because $PM_{10}$ has a greater number of observations.

2. Retrieval of the prediction and validation data set

The validation data set from the main attribute $PM_{2.5}$ was extracted from the original data set. The validation data contained 25% of random rows which is 75 observations from the complete data set consisting of 293 observations. The remaining data, 218 observations, were assigned as the prediction dataset.

3. Analysis of the descriptive statistics

The results of the descriptive statistics for $PM_{10}$ and $PM_{2.5}$ observations and model are shown on table 28

*Table 28* Descriptive statistics $PM_{10}$ observations and CTM for group of regions A

| Statistic | $PM_{10}$ observations | $PM_{10}$ CTM | $PM_{2.5}$ observations | $PM_{2.5}$ CTM |
|---|---|---|---|---|
| **Minimum** | 1.79 | 4.58 | 3.75 | 2.28 |
| **1st quartile** | 25.8 | 19.86 | 18.0 | 13.52 |
| **Median** | 38.0 | 26.33 | 29.0 | 21.60 |
| **Mean** | 43.32 | 27.4 | 32.75 | 21.23 |
| **3rd quartile** | 56.0 | 33.8 | 44.58 | 26.85 |
| **Maximum** | 152.0 | 69.35 | 104.60 | 56.04 |

The values are within a range of 1.79 to 152 for the PM10 observations and from 4.58 to 69.35 µg $m^{-3}$ for the CTM. The values suggest that the model, in general, underestimates the concentrations. The same occurs between the observations and the model of PM2.5 measurements. Authors such as Schaap *et al.* (2001), Schaap *et al.* (2009) and Adams & Lükewilee, (2010) mention in their work the underestimation of the CTM due to the higher uncertainties in PM emission inventories. The mean values of $PM_{10}$ values are higher than $PM_{2.5}$ for both, the observations and model.

The histogram of the data for region A was plotted and showed a distribution skewed to the left. The data was log transformed and compared to the raw data. The plots are shown in figure 25:

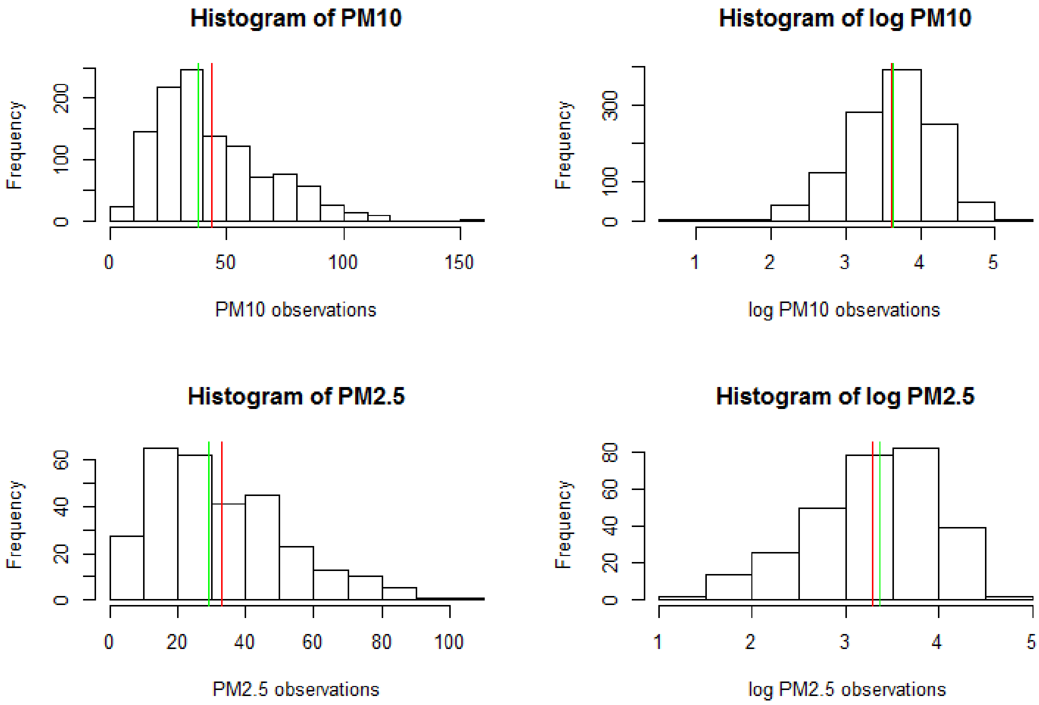


*Figure 25* Histograms for cokriging data, observations and log transformation of $PM_{10}$ and $PM_{2.5}$, of group A

Both observation measurements, $PM_{10}$ and $PM_{2.5}$, seem to have a distribution closer to the normal after the logarithmic transformation. This is also reiterated with the Q- Q plot, which shows in figure 26 that the distribution is closer to the normal for the logarithmic transformation.

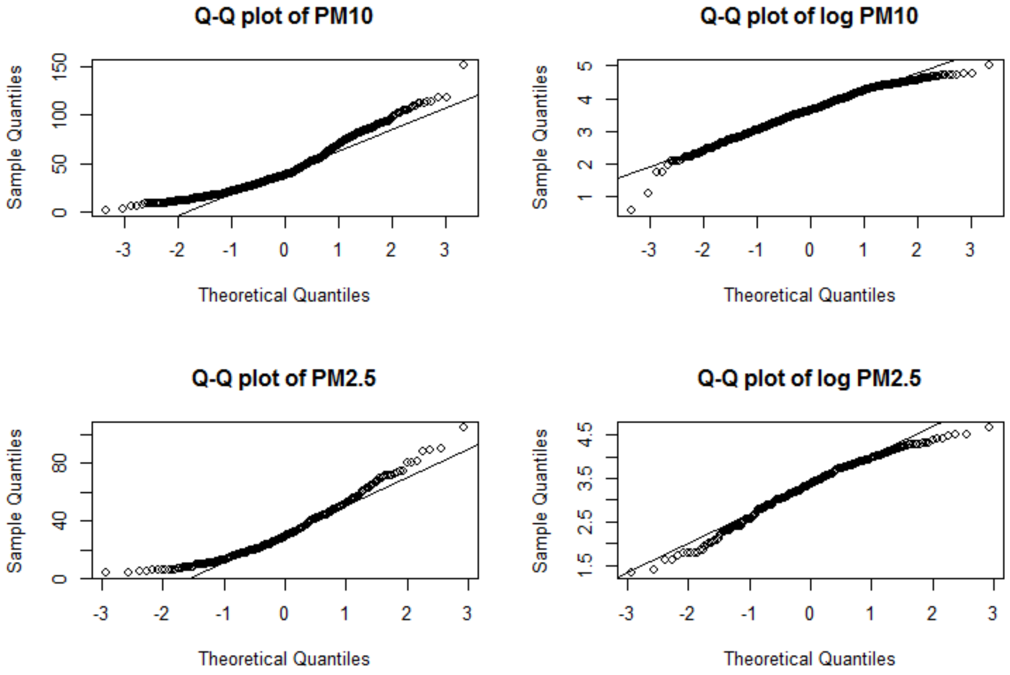


*Figure 26* Q-Q plot for cokriging data, $PM_{10}$ and $PM_{2.5}$ observations and log transformation, of group A

4. Correlations between the observations and the logarithmic transformation of the measurements

The correlation found between the PM2.5 and PM10 observations and the CTM is 0.74 and 0.69 for respectively. The correlation between the logarithm of the observations and the model is 0.77 for $PM_{2.5}$ and 0.68 for $PM_{10}$.

5. Model

Two models were constructed to evaluate the performance of the covariates for both, $PM_{2.5}$ and $PM_{10}$. The first model was prepared with the logarithm of $PM_{2.5}$ observations regressed on the logarithm of the CTM, type of area and region. The formula used was the following:

$$\log(pm.obs) \sim \log(pm.LE) + x1 + x2$$

where:
pm.obs are the observations extracted from the stations
pm.LE are the data extracted from the Chemical Transport Model –CTM-
x1 is the type of area
x2 is the type of region

In the case of $PM_{2.5}$ measurements, the log transformation of the CTM and the Southern Europe region are significant at 0.001 level with a probability of <2.2e-16 while the suburban type of area is significant at 0.05 with a value of 0.02017.

The urban type of area is the only covariate that may not improve the model because the probability (0.066) is higher than the tolerance (alpha=0.05). For the $PM_{10}$ model, the urban region has highest probability value (0.00148) but it is still below the tolerance level so it is possible to reject the null hypothesis that $\beta = 0$.

The type of area showed the highest probability values for the analysis of variance, therefore the second model was constructed with the covariates logarithmic transformation of the CTM and region, to asses if an improvement on the model could arise. The equation of the second model was:

$$\log(pm.obs) \sim \log(pm.LE) + x2$$

The following table shows the comparison of the fit of the models:

*Table 29* Comparison between the fit of the models for cokriging of group A

| Model | PM2.5 Log(CTM), type of area and region | PM2.5 Log(CTM) and region | PM10 Log(CTM), type of area and region | PM10 Log(CTM) and region |
|---|---|---|---|---|
| **Residual standard error** | 0.346 | 0.3491 | 0.3915 | 0.395 |
| **Degrees of freedom** | 212 | 214 | 952 | 954 |
| **Adjusted $R^2$** | 0.7349 | 0.7301 | 0.5485 | 0.5404 |

The first model for $PM_{10}$ and $PM_{2.5}$ which includes all the covariates show a slightly improvement over the second model. Considering the minor improvement of the type of region as a covariate (probability of 0.05 in the ANOVA test between model 1 and model 2) and the ambiguous relation of what is defined as a rural, urban or suburban type of area, the further analysis and prediction was made for the model 2.

In this way, there is also the possibility of comparing the methods, compositional kriging and cokriging, under the same conditions.

The summary statistics for the selected model are displayed on table 31 for the $PM_{2.5}$ and 32 for $PM_{10}$ in table 32.

*Table 30* Summary for the selected $PM_{2.5}$ regression model of group A

| **Residuals:** | | | | |
|---|---|---|---|---|
| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
| -1.15242 | -0.21221 | 0.00457 | 0.23430 | 0.87290 |
| **Coefficients:** | **Estimate** | **Standard Error** | **t value** | **Pr(>\|t\|)** |
| **Intercept** | 1.67 | 0.15 | 10.94 | < 2e-16 *** |
| **log(pm25.LE** | 0.66 | 0.04 | 14.43 | <2e-16 *** |
| **x2NWE** | -0.19 | 0.06 | -3.053 | 0.00255 ** |
| **x2SE** | -0.65 | 0.06 | -10.8 | < 2e-16 *** |

*Table 31* Summary for the selected $PM_{10}$ regression model of group A

| **Residuals:** | | | | |
|---|---|---|---|---|
| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
| -3.6469 | 0.2140 | 0.0265 | 0.2678 | 1.3396 |
| **Coefficients:** | **Estimate** | **Standard Error** | **t value** | **Pr(>\|t\|)** |
| **Intercept** | 1.75 | 0.12 | 15.001 | < 2e-16 *** |
| **log(pm25.LE** | 0.63 | 0.03 | 18.7 | <2e-16 *** |
| **x2NWE** | -0.18 | 0.03 | -5.47 | 5.72e-08 *** |
| **x2SE** | -0.49 | 0.03 | -14.09 | < 2e-16 *** |

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

6. Empirical Variogram, Model Variogram and Cross-Variogram

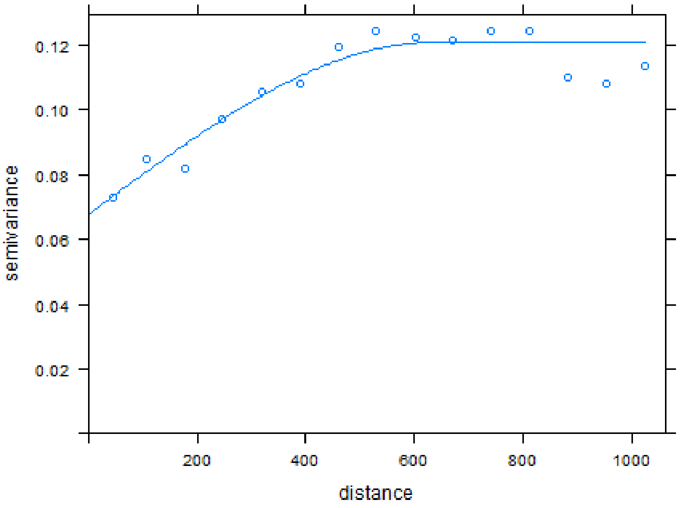


*Figure 27* Empirical and model variogram for $PM_{2.5}$ of group A. Distance in km.

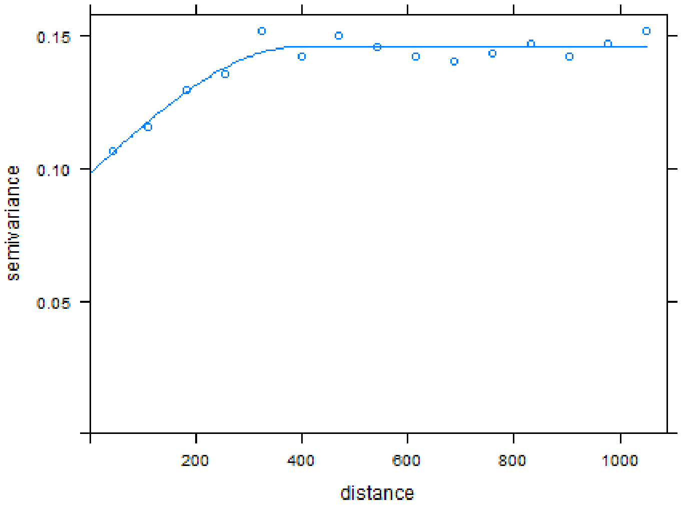


*Figure 28* Empirical and model variogram for $PM_{10}$ of group A. Distance in km

There is spatial correlation between the measurements. There is an extended correlation range in the horizontal direction for $PM_{10}$. The spherical variogram models were fitted to the empirical variograms of both measurements. The estimated parameters are shown in table 33.

*Table 32* Estimated Parameters of the variogram for CEE, NWE and SE for cokriging

| Pollutant | Model | Partial Sill | Nugget | Range | SSErr |
|---|---|---|---|---|---|
| **PM2.5** | Spherical | 0.0675 | 0.0533 | 637 | 2.67e-06 |
| **PM10** | Spherical | 0.0981 | 0.048 | 396 | 2.85e-05 |

The cross variogram shows that the changes in both pollutants are spatially similar. The spherical model was used for the cross variogram with a range of 400 km. The results are presented in figure 29:

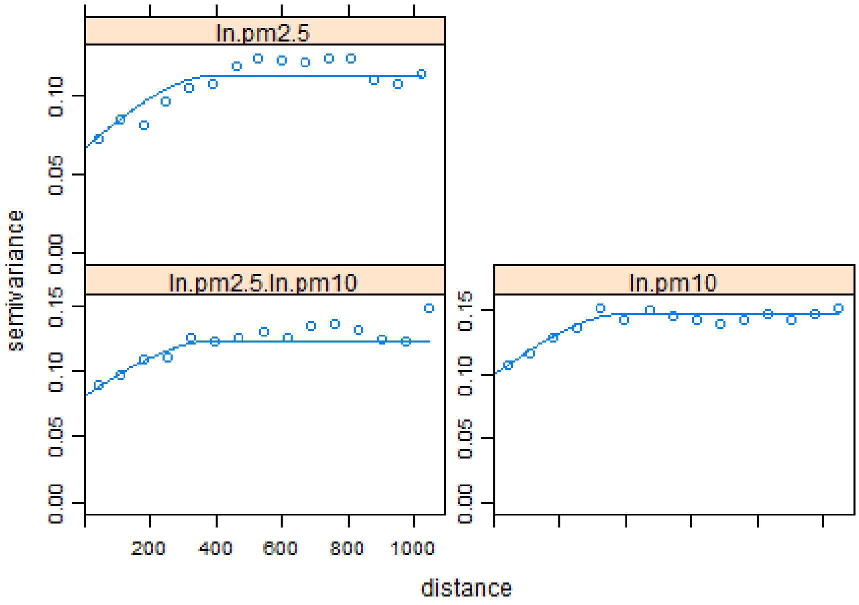


*Figure 29* Cross variograms and variogram model for group A. Distance in km.

## 7. Prediction

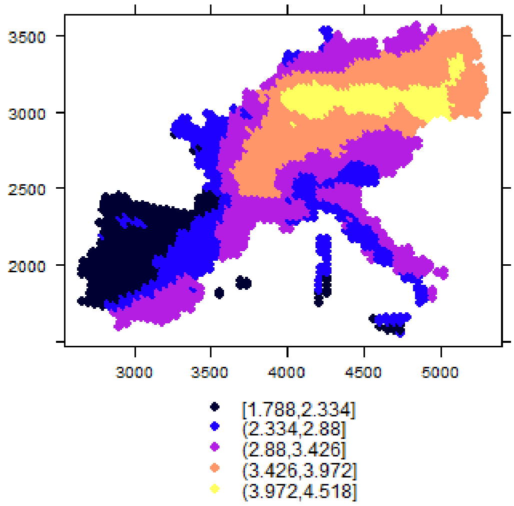


*Figure 30* Log transformed (cokriging) predictions of group A for April 5, 2009.

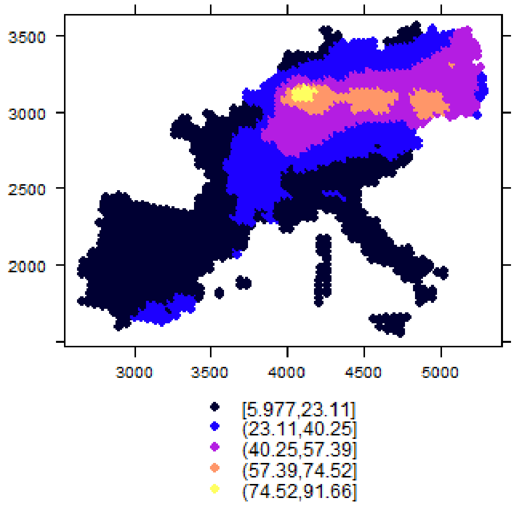


*Figure 31* Backtransformed prediction after cokriging of group A for April 5, 2009. Units: µg·m$^{-3}$.

## 8. Mean error and Root Mean Squared Error

*Table 33* Mean error and Root Mean Squared Error, cokriging group A

| Measure of error | Cokriging | Ordinary Kriging |
|---|---|---|
| **Mean Error** | 5.921189e-18 | -8.881784e-18 |
| **Root Mean Squared Error** | 1.1466e-16 | 5.773314e-17 |

### 5.3.2. Analysis and prediction using cokriging for the group of regions B: Belgium, Netherlands and Luxemburg (BENELUX), Germany and Poland (GERPOL) and Czech Republic and Slovakia (SLOCZ)

1. Analysis of the data set

The retrieved data for the three regions contained a data frame with 130 observations for $PM_{2.5}$ and a second with 551 observations for $PM_{10}$.

2. Retrieval of the prediction and validation data set

The validation data set was retrieved as 25% of random observations for $PM_{2.5}$ (33 observations). The remaining data was assigned as the prediction data set with a total number of observations equal to 97.

3. Analysis of the descriptive statistics

The mean and median for both types of particulate matter for the group of regions B are higher than those for group A. This could be caused by the countries included, for which is known to contribute with high values of particulate matter. According to Adams & Lükewilee, (2010) in rural areas, largely constant $NH_3$ emissions from agriculture have contributed to the formation of secondary particulate matter and prevented significant reductions of PM in, for example, the Netherlands and north-western Germany. Also, Hamm *et al.* (2014) presented the highest values of $PM_{10}$ for these countries on the 5th of April of 2009.

*Table 34* Descriptive statistics of cokriging data, $PM_{10}$ and $PM_{2.5}$ observations and CTM

| Statistic | $PM_{10}$ observations | $PM_{10}$ CTM | $PM_{2.5}$ observations | $PM_{2.5}$ CTM |
|---|---|---|---|---|
| **Minimum** | 1.79 | 8.88 | 12.84 | 8.46 |
| **1st quartile** | 41.19 | 24.18 | 35.13 | 19.50 |
| **Median** | 55.17 | 31.62 | 44.60 | 25.05 |
| **Mean** | 58.69 | 33.25 | 47.51 | 26.47 |
| **3rd quartile** | 75.35 | 39.19 | 57.13 | 31.92 |
| **Maximum** | 152.0 | 69.35 | 104.60 | 56.04 |

The raw data does not follow a normal distribution and is slightly skewed to the left. The logarithmic transformation was applied to the $PM_{10}$ and $PM_{2.5}$ data. After the transformation the distribution of the data was skewed to the right but closer to be symmetrical and to a normal distribution with a similar value for the mean (red in figure 32) and median (green in figure 32). The Q-Q plot shows an improvement on the distribution of the data.

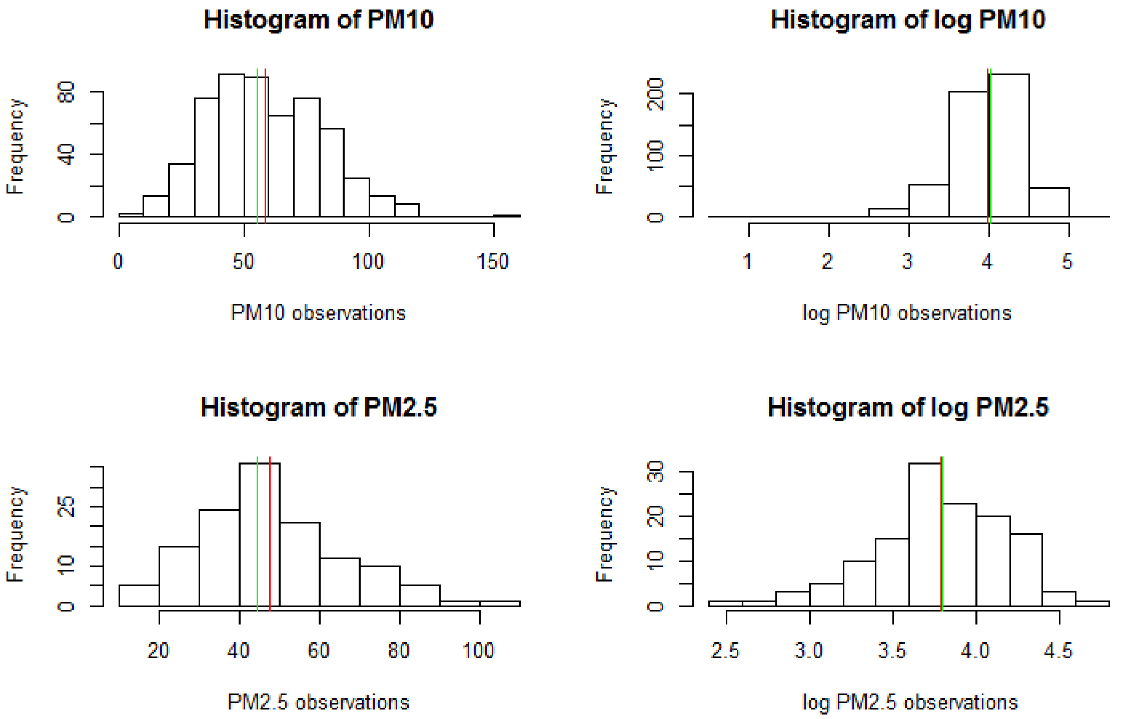


*Figure 32* Histograms for cokriging data, observations and log transformation of $PM_{10}$ and $PM_{2.5}$, of group B

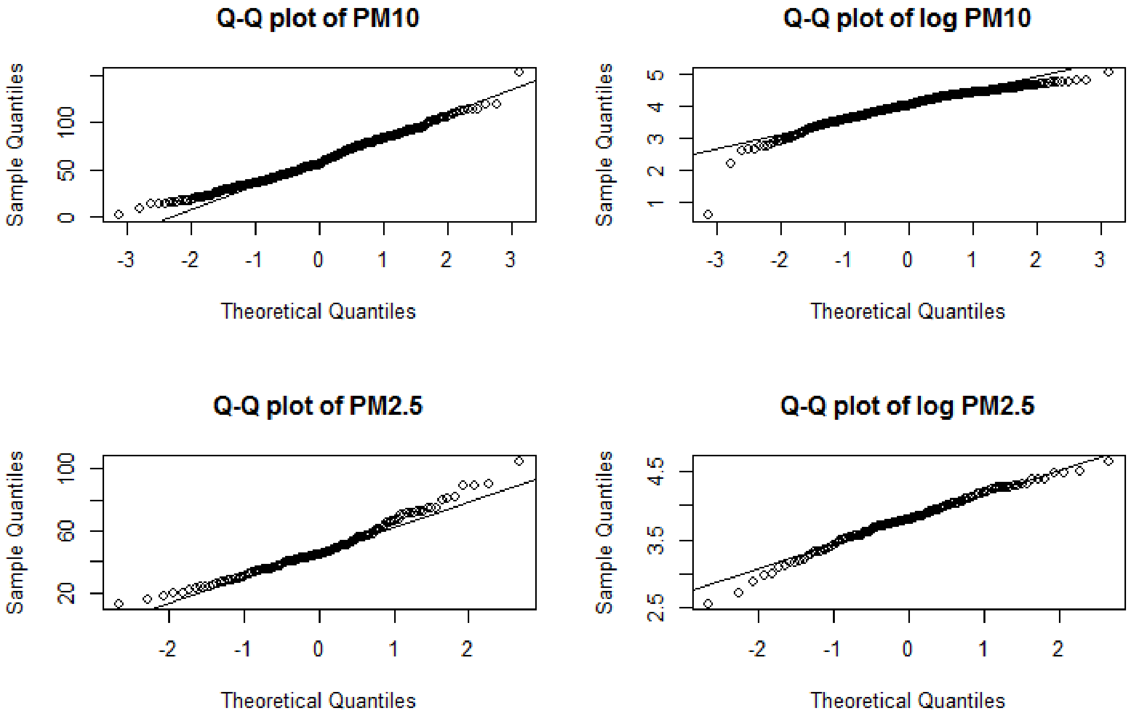


*Figure 33* Q-Q plot for cokriging data, $PM_{10}$ and $PM_{2.5}$ observations and log transformation, of group B

4. Correlations between the observations and the logarithmic transformation of the measurements

The correlation between the observatios and the Chemical Transport Model –CTM- is higher for the $PM_{2.5}$ measurements resulting on a value 0.73 than for the measurements of $PM_{10}$ in which the outcome obtained was 0.53.

5. Model

Four different models were analyzed the same way as for regions "A". The results in table 34 indicate that the best model for both categories of pollutants is obtained from including all the available covariates. Considering the fact that the $R^2$ increases with every predictor, the adjusted $R^2$ was evaluated to compare the models with different number of independent variables. As in the first analysis, the highest values of adjusted $R^2$ were achieved for the model using the three covariates.

*Table 35* Comparison between the fit of the models for cokriging of group B

| Coefficient/Model | PM2.5 Log(CTM), type of area and region | PM2.5 Log(CTM) and region | PM10 Log(CTM), type of area and region | PM10 Log(CTM) and region |
|---|---|---|---|---|
| **Residual standard error** | 0.2325 | 0.2433 | 0.3968 | 0.3982 |
| **Degrees of freedom** | 91 | 93 | 461 | 463 |
| **Adjusted $R^2$** | 0.6166 | 0.5803 | 0.2805 | 0.2753 |

The first model is to some extent an improvement over the second model. The difference between the coefficients is very low. As in the first cokriging analysis, it was considered that by adding the type of area as a covariate an error could be included on the prediction. The previous statement is because while the region used has universally defined boundaries, the criteria to define the type of area boundaries on the maps differs according to the author.

Finally the selected model included the logarithmic transformation of the CTM measurements and the type of region. The residuals plot of the model for $PM_{2.5}$ is showed on figure 34:

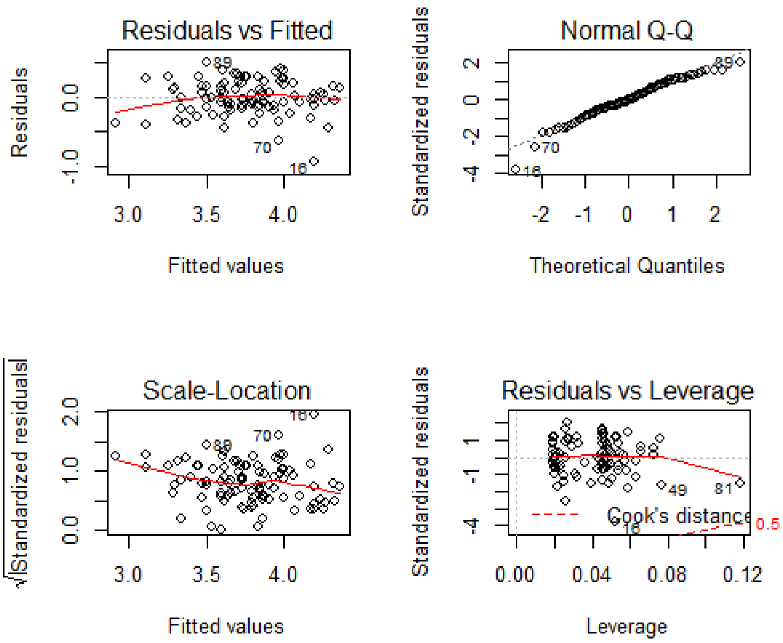


*Figure 34* Residuals plot for $PM_{2.5}$ model of group B

The analysis of variance and summary statistics (tables 35 and 36) show that the logarithmic transformation of the CTM and the region are significant predictors.

*Table 36* Analysis of variance for $PM_{2.5}$ model of group B

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **log(pm25.LE)** | 1 | 7.99 | 7.99 | 125.18 | <2e-16 *** |
| **x2** | 2 | 0.48 | 0.24 | 3.78 | 0.02633* |
| **Residuals** | 93 | 5.93 | 0.06 | | |

*Table 37* Summary statistics for $PM_{2.5}$ model of group B

| **Residuals:** | | | | |
|---|---|---|---|---|
| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
| -0.93870 | -0.13928 | -0.00143 | 0.19321 | 0.51011 |
| **Coefficients:** | **Estimate** | **Standard Error** | **t value** | **Pr(>\|t\|)** |
| **Intercept** | 1.40 | 0.22 | 6.43 | 5.47 e-09 *** |
| **log(pm25.LE)** | 0.71 | 0.07 | 10.57 | <2e-16 *** |
| **X2GERPOL** | 0.11 | 0.06 | 1.78 | 0.07766. |
| **x2SLOCZ** | 0.21 | 0.08 | 2.74 | 0.00732** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The $PM_{10}$ model residuals are close to a normal distribution and show a linear relationship between the variables. The plot shows some outliers that correspond to two urban stations in Poland (303 and 362) and a rural station in Czech Republic (95).

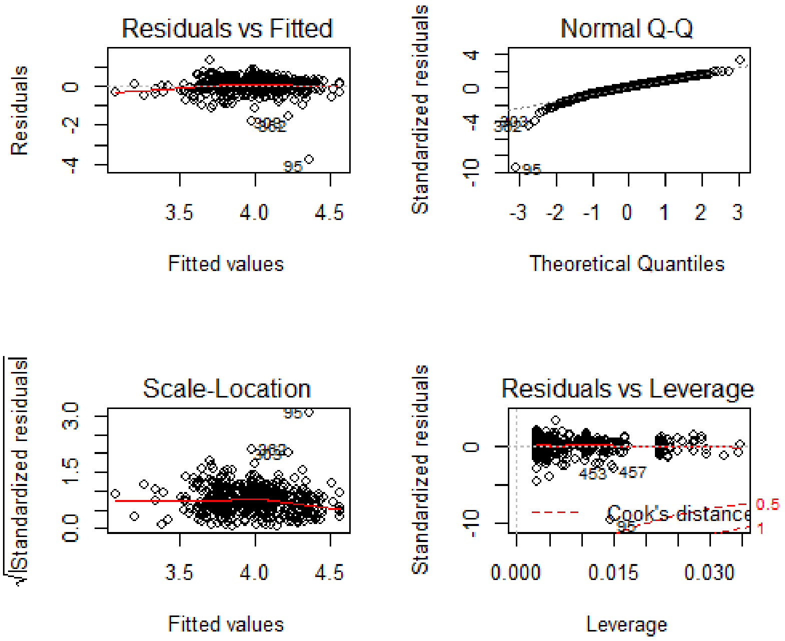


*Figure 35* Residuals plot for $PM_{10}$ model of group B

The analysis of variance and summary statistics report a high significance for the covariates included.

*Table 38* Analysis of variance for $PM_{10}$ model of group B

| | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
|---|---|---|---|---|---|
| **log(pm10.LE)** | 1 | 25.62 | 25.62 | 161.61 | <2.2e-16 *** |
| **x4** | 2 | 2.92 | 1.46 | 9.22 | 0.0001182 ** |
| **Residuals** | 463 | 73.41 | 0.16 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

*Table 39* Analysis of variance for $PM_{10}$ model of group B

| **Residuals:** | | | | |
|---|---|---|---|---|
| **Minimum** | **1st Quartile** | **Median** | **3rd Quartile** | **Maximum** |
| -3.7949 | -0.1872 | 0.0336 | 0.2252 | 1.3237 |
| **Coefficients:** | **Estimate** | **Standard Error** | **t value** | **Pr(>\|t\|)** |
| **Intercept** | 1.26 | 0.20 | 6.16 | 1.56e-09 *** |
| **log(pm10.LE)** | 0.72 | 0.06 | 12.75 | < 2e-16 *** |
| **x4GERPOL** | 0.24 | 0.06 | 3.83 | 0.000143 *** |
| **x4SLOCZ** | 0.3 | 0.07 | 4.19 | 3.36e-05 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

6. Empirical Variogram, Model Variogram and Cross Variogram

*Table 40* Model cross variogram coefficients

| **Pollutant** | **Model** | **Partial Sill** | | **Range** | **SSErr** |
|---|---|---|---|---|---|
| **PM2.5** | Exponential | 0.0291 | 0.0445 | 131 | 1.52e-06 |
| **PM10** | Spherical | 0.1351 | 0.0265 | 603 | 6.79e-06 |

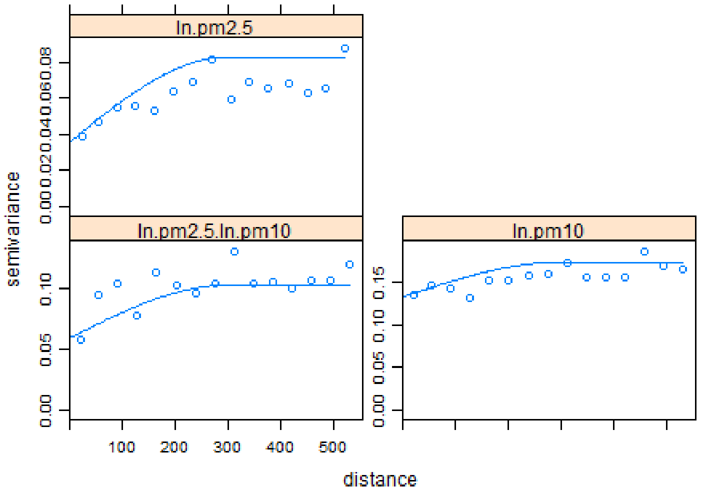


*Figure 36* Cross variogram and variogram model for group B. Distance in km

7. Prediction

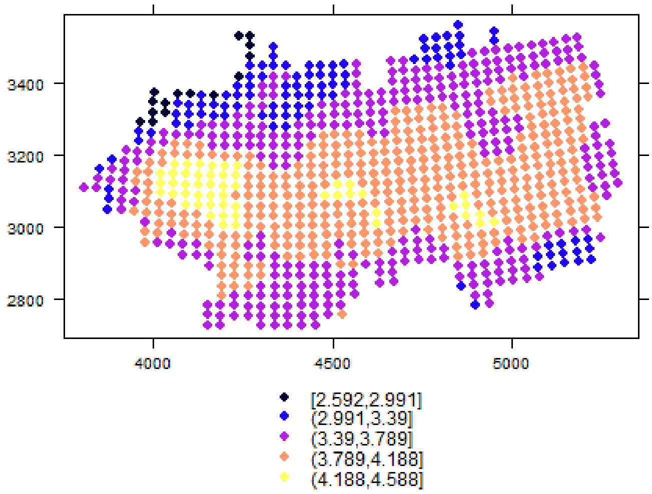


*Figure 37* Log transformed predictions (cokriging) of group B for April 5, 2009.

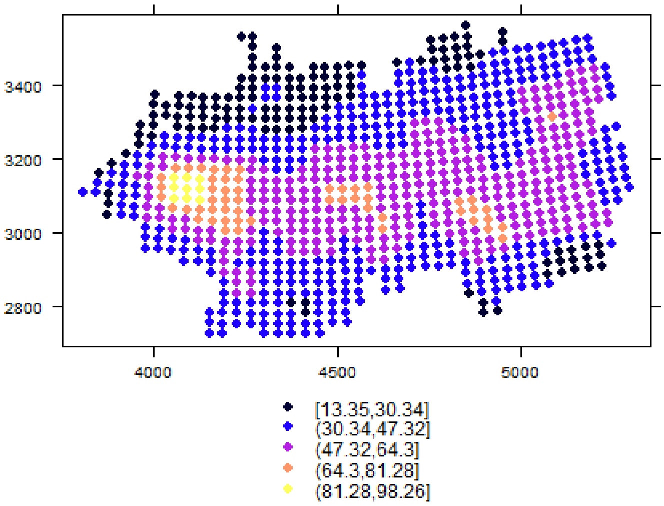


*Figure 38* Backtransformed prediction after cokriging for group B for April 5, 2009. Units: µg·m$^{-3}$

8. Mean Error and Root mean squared error

*Table 41* Mean error and Root mean squared error, cokriging of group B

| Measure of Error | Cokriging | Ordinary Kriging |
|---|---|---|
| **Mean Error** | 0.0129149 | 0.008370715 |
| **Root Mean Squared Error** | 0.110435 | 0.1149565 |

# 6. DISCUSSION

According to the nature of the data both methods, compositional kriging and cokriging, represent an appropriate way of prediction. The advantages in both methods are that they minimize the prediction error variance and include the concept of best linear unbiased predictor (BLUP).

For this case and in air pollution, cokriging is a good method considering that not all the stations in Europe measure the same type of pollutants, therefore, it is possible to include other particles such as the ultrafine particles ($PM_1$) as a sparsely sampled variable. For this study, the performance of compositional kriging was tested over a composition of two parts due to the lack of data, while according to Tolosana-Delgado in theory a composition of two parts that sums up to 1 or 100% is in fact only one part or variable (personal communication, Jan 19, 2015). The compositional data analysis and prediction lead to some errors that may be avoided by using a logistic transformation for this two parts. Further research is then possible in this sense by either including a greater number of parts or performing logistic transformation over the two parts.

The low number of observations for some countries, may have generated less accurate results. While the inclusion of multiple pollutants for this study could have generated results with greater accuracy for cokriging, it would have decreased the accuracy of the compositional prediction because it only makes sense to analyze the components measured at the same spatial location which reduces the number of stations for the analysis.

This research did not include the model of the covariate with the lowest error due to the problem of the difference between interpretations of the type of area. In order to generate a better prediction it is possible to test a coherent map between this areas and the measurements.

The improvement of cokriging over compositional kriging is not clear since the problem with cokriging is that there is no accounting for the correlation between predicted values (Ver Hoef & Cressie, (1993)). As mentioned by Aitchison, (2003), there could be a spurious effect on the correlation between the measurements of compositions in the same location. Being $PM_{2.5}$ a fraction of $PM_{10}$ it is possible that this problem arises in the analysis for close stations given that the variance within the measurements is less.

The main problem encountered during this study was the difficulty on comparing results across methods. The two methods work in a different way and produce different type of results. In this case even though, the data was transformed back to the original measurements on both methods, they are not comparable since one is being backtransformed from natural logarithm to the original input and the other one from ratio to the total measure $PM_{2.5}$ obtained using the values in the CTM.

The methods also allow to include multiple pollutants in the process but it is necessary to clarify that they do it in two different ways. Compositional kriging provides a way to predict multiple pollutants at the same time while cokriging can include multiple variables to predict a sparsely sampled variable. Nevertheless it is possible to work with a combination of this two methods: compositional cokriging. Compositional cokriging allows to jointly estimate values of a coregionalization with spatially correlated components (Pawlowsky, (1989)).

# 7. CONCLUSIONS AND RECOMMENDATIONS

## 7.1. Conclusions

**Objective 1: Exploratory Analysis**

1. Which covariates generate a model with lowest error?

   The comparison between the models present similar values of Residual Standard Error and Adjusted $R^2$ between the models using the three covariates (chemical transport model from LOTUS EUROS, the type of area and region) and the model using the CTM and region. For the compositional analysis of group A, the most adequate model was obtained using the CTM and region as covariates. The rest of the analysis obtained lower residual standard error and higher adjusted R2 values for the model including the three covariates. The difference between this two models was very low (in some cases not significant at $\alpha=0.05$ in the ANOVA). The available data to include the type of area as a covariate needed further preprocessing to ensure that the data is coherent when comparing the distinction between the classes, i.e. rural, urban and suburban. This process could generate further errors for the prediction that would not be shown in the error of the model. Therefore the final model chosen was the one including the region and the CTM as independent variables.

2. What is the spatial distribution of the pollutants using these covariates?

3. What is the spatial variation of the measurements and how can it be interpreted?

   The highest levels of pollution found for PM2.5 were located on Central Eastern Europe for the analysis of region A and for the countries of Germany and Poland in the analysis of region B so we can say in both cases spatial distribution of highest levels of pollution was located around the same places. The high values also correspond to West Germany and are concentrated on the location of the industrial agglomerations around the river Rhine.

   The lower values of PM2.5 were located on the south coast of region A northern-west coast of region B. The low values of PM2.5 of region can be explained on higher contributions of the coarse fraction for this areas than the particulate matter less than 2.5 μg.

   For compositional kriging the results can only be truly analysed in terms of ratios, the backtransformed values may include error and were only plotted for explanation purposes. Region A ranged from 5.98 to 91.66 while region B ranged from 13.35 to 98.26 for the predictions based on cokriging.

**Objective 2: Prediction**

4. What are the results of the predictions with respect to the composition of $PM_{2.5}$?

   As stated in the answer to question 3, the compositional prediction can only be discussed in terms of ratios. The back transformed data is only an example representing what the actual values could be if the total $PM_{10}$ was the measurement obtained for the LOTUS-EUROS Chemical Transport Model. However, the ratios also provide important information of the pollutants and can describe in which areas the highest concentrations of $PM_{2.5}$ occur. The areas with low values of this pollutant can be explained on higher contributions of total $PM_{10}$ of the coarse fraction. The ratios ranged from 0.44 to 0.90 for region A and from 0.50 to 0.96 in region B, the mean values are 0.67 and 0.73 respectively.

5. What prediction method produces the lowest error: compositonal kriging or cokriging?

   Comparing the cokriging approach with the compositional approach is difficult as the compositional approach provides ratios of PM while cokriging can produce actual values. In this analysis we are interested in the absolute values of PM2.5. If instead we were interested in the ratios a direct comparison between both approaches could be made as done by Odeh et al. (2003).

**Objective 3: Data quality**

6. What is the maximum spatial extent in which can be obtained accurate prediction results?

   The spatial extent show similar trends for the data, the difference of the results was mainly between the methods and not between the spatial extents. Nevertheless, the prediction for bigger areas show lower errors compared to the regions of smaller size. This could be because the amount of data is higher.

7. Which method is suitable for measuring the accuracy of the output?

   Following the studies by Odeh *et al.,* (2003), Denby *et al.,* (2005) and Hamm *et al.* (2014), the model validation diagnostics where calculated. The mean error and root mean squared error was computed between the prediction and validation dataset. The mean error was used to quantify the bias and the root mean squared error to quantify the dispersion of the error.

8. What is the accuracy of the output?

   The prediction based on cokriging generated a Mean Error of 5.92e-18 and 0.1104 for regions A and B respectively. The RMSE obtained for region A was 1.14e-16 compared to 0.1149 for region B. The prediction obtained for a larger spatial extent is more accurate.

## 7.2. Recommendations

1. The results of this study could be further analysed or improved by including other pollutants that are fractions of $PM_{10.}$ The challenge lays in that different stations measure different type of pollutants, in the case of cokriging this is a problem because there will be less number of stations to include in the prediction.

2. The model could be improved by testing the performance of a coherent land cover map as a predictor. Others covariates such as DEM could also be tested. It is not recommended to include more than 4 covariates for the model to avoid falling in the error of overfitting.

3. Other softwares and packages are available to treat compositional data, an example is CoDA work 2015.

# LIST OF REFERENCES

Adams, M., & Lükewilee, A. (2010). The european environment state and outlook 2010. *European Environment Agency.* Copenhagen. p 42.

Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology, 31*(5)*,* 563–580.

Aitchison, J. (2003). A concise guide to compositional data analysis. *Compositional Data Analyisis Workshop*; Girona, Italy. p. 97. Retrieved from http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf. Accessed the 12 of December 2014.

Bytnerowicz, A., Godzik, B., Frączek, W., Grodzińska, K., Krywult, M., Badea, O., Barancok, P., Blum, O., Cerny, M., Godzik, S., Mankovska, B., Manning, W., Moravcik, P., Musselman, R., Oszlanyi, J., Postelnicu, D., Szdzujh, J., Varsavova, M., Zota, M. (2002). Distribution of ozone and other air pollutants in forests of the Carpathian Mountains in central Europe. *Environmental Pollution, 116*(1)*,* 3–25.

Denby, B., Horálek, J., Walker, S. E., & Eben, K., & Fiala, J. (2005). Interpolation and assimilation methods for European scale air quality assessment and mapping - Part I: Review and recommendations. *European Topic Centre on Air and Climate Change Technical Paper 2005*/7. p. 51.

EEA. (2013). Air quality in Europe-2013 report. *European Environment Agency*. Copenhagen (2013). EEA report No. 9/2013.

Gringarten, E., & Deutsch, C. V. (2001). Teacher's Aide Variogram Interpretation and Modeling. *Mathematical Geology, 33*(4)*,* 507–534.

Hamm, N.A.S., Finley, A.O., Schaap, M., & Stein, A. (2015). A spatially varying coefficient model for mapping PM10 air quality at the European scale. *Atmospheric Environment, 102*(2015), 393-405.

Leeuw, F. De, & Horálek, J. (2009). Assessment of the health impacts of exposure to PM 2 . 5 at a European level. *European Topic Centre on Air and Climate Change Technical Paper 2009*/1. p. 21.

Moral, F. J., Álvarez, P., & Canito, J. L. (2006). Mapping and hazard assessment of atmospheric pollution in a medium sized urban area using the Rasch model and geostatistics techniques. *Atmospheric Environment, 40*(8)*,* 1408–1418.

Odeh, I. O. A., Todd, A. J., & Triantafilis, J. (2003). Spatial Prediction of Soil Particle-Size Fractions As Compositional Data. *Soil Science, 168*(7), 501–515.

Pawlowsky, V. (1989). Cokriging of regionalized compositions. *Mathematical Geology, 21*(5)*,* 513–521.

Putaud, J. P., Van Dingenen, R., Alastuey, A., Bauer, H., Birmili, W., Cyrys, J., Flentje, H., Fuzzi, S., Gehrig, R., Hansson, H. C., Harrison, R. M., Herrmann, H., Hitzenberger, R., Huglin, C., ¨ Jones, A. M., Kasper-Giebl, A., Kiss, G., Kousa, A., Kuhlbusch, T. A. J., Loschau, G., Maenhaut, W., Molnar, A., Moreno, T., ¨ Pekkanen, J., Perrino, C., Pitz, M., Puxbaum, H., Querol, X., Rodriguez, S., Salma, I., Schwarz, J., Smolik, J., Schneider, J., Spindler, G., ten Brink, H., Tursic, J., Viana, M., Wiedensohler, A., & Raes, F. (2010). A European aerosol phenomenology – 3: Physical and chemical characteristics of particulate matter from 60 rural, urban, and kerbside sites across Europe. *Atmospheric Environment, 44*(10), 1308–1320.

Schaap, M., Manders, A. M. M., Hendriks, E. C. J., Cnossen, J. M., Segers, A. J. S., Denier van der Gon, H. A. C., Jozwicka, M., Sauter, F.J., Velders, G. J. M., Matthijsen, J., Builtjes, P. J. H. (2009). Regional Modelling of Particulate Matter for the Netherlands. PBL Report 500099008, Bilthoven, The Netherlands. (p. 99).

Schaap, M., Weijers, E. ., Mooibroek, D., Nguyen, L., & Hoogerbrugge, R. (2001). Composition and origin of particulate matter in the Netherlands. RIVM Rapport *650010029* (p. 69). Bilthoven, The Netherlads. Retrieved from http://rivm.openrepository.com/rivm/handle/10029/257110

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

TNO, RIVM, KNIM, & PBL. (2015). LOTOS-EUROS model - Home. Retrieved February 14, 2015, from http://www.lotos-euros.nl/

Van den Boogaart, K. G., & Tolosana-Delgado, R. (2008). "Compositions": A unified R package to analyze compositional data. *Computers & Geosciences, 34*(4), 320–338.

Van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). Analyzing Compositional Data with R. Berlin: Springer.

Ver Hoef, J. M., & Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology, 25*(2), 219–240.

Walvoort, D., & Gruijter, J. de. (2001). Compositional kriging: a spatial interpolation method for compositional data. *Mathematical Geology, 33*(8), 951–966.

Webster, R., & Oliver, M. A. (2007). Geostatistics for environmental scientists, second ed. Chichister: John Wiley and Sons Ltd.

# APPENDICES

## Appendix1: R codes

```
# Code to import and process the PM data for group of regions A
# from Arjo Seger's NetCDF file

# Author: Dr Nicholas Hamm
# Modifications Marisol Amador

##########Load the necessary libraries
library(ncdf4) # Version 1.4 (1.3 doesn't seem to work)
library(gstat)
library(rgdal)

##########Extract the data
#Open the netcdf data file (Run it either for PM2.5 or PM10, example for PM2.5)
#Let's create first the PM2.5
ap       <- "C:/GFM MSc/Thesis/Preprocessing/airbase/airbase/aqord__obs__1d__pm25_mass__pm25.nc"
LE  <- "C:/GFM MSc/Thesis/Preprocessing/airbase/airbase/aqord__LE_eu__1d__pm25_mass__pm25.nc"

#For PM10
#LE  <- "C:/GFM MSc/Thesis/Preprocessing/airbase/airbase/aqord__LE_eu__1d__pm10_mass__pm10.nc"
#ap      <- "C:/GFM MSc/Thesis/Preprocessing/airbase/airbase/aqord__obs__1d__pm10_mass__pm10.nc"

nc1      <- nc_open(ap)
nc2      <- nc_open(LE)

# now you can inspect the contents of nc (by typing "nc")
nc1
nc2

# 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")

##########Create data frames
# PM2.5 data
# This dataframe contains the PM2.5 in situ observations
pm25.airbase       <- ncvar_get(nc1, "pm25_mass__pm25")
# This dataframe contains the PM2.5 LE simulation outputs
pm25.LE            <- ncvar_get(nc2, "pm25_mass__pm25")

# PM10 data
# This dataframe contains the PM10 in situ observations
#pm10.airbase  <- ncvar_get(nc1, "pm10_mass__pm10")
```

```
# This dataframe contains the PM10 LE simulation outputs
#pm10.LE          <- ncvar_get(nc2, "pm10_mass__pm10")

# Close the links to the NetCDF files
nc_close(nc1)
nc_close(nc2)

#Create a data frame with the coordinates for April 5 2009
tmp0  <- data.frame(lat=lat, lon=lon, height=height, st.code=st.code, type=st.type, type_area=st.type_area,
               coun=substr(st.code, 1, 2), pm25.obs=pm25.airbase[,826], pm25.LE=pm25.LE[,826])
#For PM10
#tmp0  <- data.frame(lat=lat, lon=lon, height=height, st.code=st.code, type=st.type, type_area=st.type_area,
               coun=substr(st.code, 1, 2), pm10.obs=pm10.airbase[,826], pm10.LE=pm10.LE[,826])

tmp1      <- tmp0
coordinates(tmp1) <- ~ lon + lat

##########Coordinates
#Projected coordinates WGS84
proj4string(tmp1) <- CRS("+proj=longlat")
proj4string(tmp1)

# Transform from Lat/Lon to ETRS89, see
# http://spatialreference.org/ref/epsg/3035/
tmp2      <- spTransform(tmp1, CRS("+init=epsg:3035"))
plot(tmp2@coords)
#I plot them ignoring the extreme locations (note that there are several
# Spatial outliers owing to various "outlying" European territories.
plot(tmp2@coords, xlim=c(2000000, 7000000), ylim=c(90, 6000000))
title(main="Airbase Stations", sub="All stations",
     cex.main = 1.5,   font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")

#Transform back to a data frame and change the column names to "x"and "y"
tmp3      <- as.data.frame(tmp2)
colnames(tmp3)[8]  <-  "easting"
colnames(tmp3)[9]  <-  "northing"
#Transform to kilometers
tmp3$easting <- tmp3$easting/1000
tmp3$northing <- tmp3$northing/1000

tmp3      <- data.frame(tmp3, lat=lat, lon=lon)

# This data frame contains the station information.
st.info.airbase <- data.frame(lat=lat, lon=lon, easting=tmp3$easting, northing=tmp3$northing,
                    height=height, code=st.code, name=st.name, type=st.type, type_area=st.type_area)

########## Save to .Rdata files
save(st.info.airbase, pm25.airbase, pm25.LE, file="airbase25.Rdata")
#For pm10
#save(st.info.airbase, pm10.airbase, pm10.LE, file="airbase10.Rdata")
```

```
##########Work out Regions

tmp4  <- tmp3
tmp4["region"] <- NA

tmp4$region[which(tmp4$coun == "BE")]     <- "NWE"
tmp4$region[which(tmp4$coun == "LU")]     <- "NWE"
tmp4$region[which(tmp4$coun == "NL")]     <- "NWE"
tmp4$region[which(tmp4$coun == "FR" & tmp4$lat >= 45)]     <- "NWE"

tmp4$region[which(tmp4$coun == "IT")]      <- "SE"
tmp4$region[which(tmp4$coun == "ES")]      <- "SE"
tmp4$region[which(tmp4$coun == "PT")]      <- "SE"
tmp4$region[which(tmp4$coun == "FR" & tmp4$lat < 45)]  <- "SE"

tmp4$region[which(tmp4$coun == "AT")]  <- "CEE"
tmp4$region[which(tmp4$coun == "CH")]     <- "CEE"
tmp4$region[which(tmp4$coun == "CZ")]     <- "CEE"
tmp4$region[which(tmp4$coun == "PL")]      <- "CEE"
tmp4$region[which(tmp4$coun == "DE")]  <- "CEE"

head(tmp4)#data frame with all fields

##########Inspect the data with regions
# Get the number of stations for each region
summary(as.factor(tmp4$region))

#Show only the data without missing values
tmp4 <- tmp4[complete.cases(tmp4),]
head(tmp4)
plot(tmp4$easting, tmp4$northing, xlab="Easting", ylab="Northing")
title(main="Airbase stations", sub="Countries of interest",
      cex.main = 1.5,   font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")
dim(tmp4)

#Define the bounding box
plot(tmp4$easting, tmp4$northing, xlim=c(1000, 6000), ylim=c(1000, 5000))
tmp4      <- tmp4[which(tmp4$easting > 2000),]
tmp4      <- tmp4[which(tmp4$easting < 6000),]
#tmp4    <- tmp4[which(tmp4$y > 1000),]
plot(tmp4$easting, tmp4$northing)
title(main="Airbase stations", sub="Countries of interest", cex.main = 1.5,
      font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")
summary(as.factor(tmp4$type_area))

#Remove rows where type of area is unknown
tmp5 <- tmp4[!tmp4$type_area == "unknown", ]
tmp5 <- tmp5[!tmp5$type_area == "", ]
```

```
#Plot stations Rural in green, suurban in red and urban in black
plot(tmp5$easting, tmp5$northing, xlab="easting", ylab="northing",
    col=ifelse(tmp5$type_area == "rural", 'green', ifelse(tmp5$type_area == "suburban", 'red', 'black')))
title(main="European Stations", sub="Rural, Urban and Suburban stations", cex.main = 1.5,
     font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")


#Name of the table pm2.5
pm2.5 <- tmp5

#Name of the table pm10
#pm10 <- tmp5

########### Save file
#Create a file with measurements for April 5 2009
save(pm2.5, pm10, file="compositions.Rdata")

# Delete all variables and then restore the key data
rm(list=ls())

################################END##################################

# Code to import and process the PM data for group of regions B
# from Arjo Seger's NetCDF file

# Author: Dr Nicholas Hamm
# Modifications M.A.
setwd("C:/GFM MSc/Thesis/Preprocessing/Data")

##########Load the necessary libraries
library(ncdf4) # Version 1.4 (1.3 doesn't seem to work)
library(gstat)
library(rgdal)

#opens the netcdf data file(Run it either for PM2.5 or PM10)
#Let's create first the PM2.5
#ap  <- "C:/GFM MSc/Thesis/Preprocessing/airbase/airbase/aqord__obs__1d__pm25_mass__pm25.nc"
#LE  <- "C:/GFM MSc/Thesis/Preprocessing/airbase/airbase/aqord__LE_eu__1d__pm25_mass__pm25.nc"
#For PM10
LE  <- "C:/GFM MSc/Thesis/Preprocessing/airbase/airbase/aqord__LE_eu__1d__pm10_mass__pm10.nc"
ap <- "C:/GFM MSc/Thesis/Preprocessing/airbase/airbase/aqord__obs__1d__pm10_mass__pm10.nc"

nc1 <- nc_open(ap)
nc2 <- nc_open(LE)

# now you can inspect the contents of nc (by typing "nc")
nc1
nc2
```

```
########## Extract the 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")

##########Create data frames
# PM2.5 data
# This dataframe contains the PM2.5 in situ observations
#pm25.airbase           <- ncvar_get(nc1, "pm25_mass__pm25")
# This dataframe contains the PM2.5 LE simulation outputs
#pm25.LE                <- ncvar_get(nc2, "pm25_mass__pm25")

# PM10 data
# This dataframe contains the PM10 in situ observations
pm10.airbase  <- ncvar_get(nc1, "pm10_mass__pm10")
# This dataframe contains the PM10 LE simulation outputs
pm10.LE                 <- ncvar_get(nc2, "pm10_mass__pm10")

# Close the links to the NetCDF files
nc_close(nc1)
nc_close(nc2)

#Create a data frame with the coordinates for April 5 2009

#tmp0  <- data.frame(lat=lat, lon=lon, height=height, st.code=st.code, type=st.type, type_area=st.type_area,
#             coun=substr(st.code, 1, 2), pm25.obs=pm25.airbase[,826], pm25.LE=pm25.LE[,826])
#For PM10
tmp0  <- data.frame(lat=lat, lon=lon, height=height, st.code=st.code, type=st.type, type_area=st.type_area,
coun=substr(st.code, 1, 2), pm10.obs=pm10.airbase[,826], pm10.LE=pm10.LE[,826])

tmp1                    <- tmp0
coordinates(tmp1) <- ~ lon + lat

##########Coordinates
#Projected coordinates WGS84
proj4string(tmp1) <- CRS("+proj=longlat")
proj4string(tmp1)

# Here I transform from Lat/Lon to ETRS89, see
# http://spatialreference.org/ref/epsg/3035/
tmp2                    <- spTransform(tmp1, CRS("+init=epsg:3035"))
plot(tmp2@coords)
#I plot them ignoring the extreme locations (note that there are several
# Spatial outliers owing to various "outlying" European territories.
plot(tmp2@coords, xlim=c(2000000, 7000000), ylim=c(90, 6000000))
title(main="Airbase Stations", sub="All stations",
    cex.main = 1.5,   font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")

#Transform back to a data frame and change the column names to "x"and "y"
tmp3                          <- as.data.frame(tmp2)
colnames(tmp3)[8]  <-  "easting"
colnames(tmp3)[9]  <-  "northing"
#Transform to kilometers
tmp3$easting <- tmp3$easting/1000
tmp3$northing <- tmp3$northing/1000

tmp3 <- data.frame(tmp3, lat=lat, lon=lon)

# This data frame contains the station information.
st.info.airbase <- data.frame(lat=lat, lon=lon, easting=tmp3$easting, northing=tmp3$northing,
                          height=height, code=st.code, name=st.name, type=st.type, type_area=st.type_area)

##########Save data

# Save to .Rdata files
#save(st.info.airbase, pm25.airbase, pm25.LE, file="airbase25.Rdata")
#For pm10
save(st.info.airbase, pm10.airbase, pm10.LE, file="airbase10.Rdata")

########## Work out regions
tmp4  <- tmp3
tmp4["region"] <- NA

tmp4$region[which(tmp4$coun == "BE")]     <- "BENELUX"
tmp4$region[which(tmp4$coun == "LU")]     <- "BENELUX"
tmp4$region[which(tmp4$coun == "NL")]     <- "BENELUX"


tmp4$region[which(tmp4$coun == "SK")]     <- "SLOCZ"
tmp4$region[which(tmp4$coun == "CZ")]     <- "SLOCZ"

tmp4$region[which(tmp4$coun == "PL")]      <- "GERPOL"
tmp4$region[which(tmp4$coun == "DE")]  <- "GERPOL"

head(tmp4)#data frame with all fields


##########Analysis of the regions

# Get the number of stations for each region
summary(as.factor(tmp4$region))

#Show only the data without missing values
tmp4 <- tmp4[complete.cases(tmp4),]
head(tmp4)
plot(tmp4$easting, tmp4$northing, xlab="Easting", ylab="Northing")
```

```
title(main="Airbase stations", sub="Countries of interest: BENELUX, SLOCZ, GERPOL",
      cex.main = 1.5,   font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")
dim(tmp4)

#Define the bounding box
plot(tmp4$easting, tmp4$northing, xlim=c(3500, 6000), ylim=c(2500, 4000))
tmp4  <- tmp4[which(tmp4$easting > 2000),]
tmp4 <- tmp4[which(tmp4$easting < 6000),]

plot(tmp4$easting, tmp4$northing)
title(main="Airbase stations", sub="Countries of interest", cex.main = 1.5,
      font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")
summary(as.factor(tmp4$type_area))

#Remove rows where type of area is unknown
tmp5 <- tmp4[!tmp4$type_area == "unknown", ]
tmp5 <- tmp5[!tmp5$type_area == "", ]
summary(as.factor(tmp4$region))
summary(as.factor(tmp4$type_area))

#Plot stations Rural in green, suurban in red and urban in black
plot(tmp5$easting, tmp5$northing, xlab="easting", ylab="northing",
     col=ifelse(tmp5$type_area == "rural", 'green', ifelse(tmp5$type_area == "suburban", 'red', 'black')))
title(main="European Stations", sub="Rural, Urban and Suburban stations", cex.main = 1.5,
      font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")


#Name of the table pm2.5
#pm2.5 <- tmp5

#Name of the table pm10
pm10 <- tmp5

#Create a file with measurements for April 5 2009
save(pm2.5, pm10, file="compositions2.Rdata")

# Delete all variables and then restore the key data
rm(list=ls())

################################END##############################
#Code to extract prediction Grid
library(ncdf4) # Version 1.4 (1.3 doesn't seem to work)
library(gstat)
library(rgdal)

##########Open the file
#opens the netcdf data file(grid)
#Grid
CTMg  <- "C:/GFM MSc/Thesis/Preprocessing/LE-eu-y2009-daily/LE-eu-y2009/LE_eu-y2009_conc-sfc_20090405_daily.nc"
```

```
nc <- nc_open(CTMg)

# Now you can inspect the contents of nc (by typing "nc")
nc

########## Retrieve the Information
lat  <- ncvar_get(nc, "lat")
lon <- ncvar_get(nc, "lon")
tpm25 <- ncvar_get(nc, "tpm25")
tpm10 <- ncvar_get(nc, "tpm10")

nc1 <- as.matrix(as.numeric(t(tpm25)))
nc2 <- as.matrix(as.numeric(t(tpm10)))

# The units are in kg*m-3 and it is necessary to tranform them to mg*m-3
#Transform the tpm25 and tpm10
pm25.LE <- nc1*1000000000
pm10.LE <- nc2*1000000000

# SP object from grid data
att1 <- expand.grid(lat=lat, lon=lon)
#create a table with the measurements
j <- cbind(att1, pm25.LE, pm10.LE)
coordinates(j) <- ~lon+lat
proj4string(j)<- CRS("+proj=longlat")

##########Save the data
write.csv(j, sep=";", file="C:\\GFM MSc\\Thesis\\Preprocessing\\Data\\ctm_data.csv")

##########Modify data externally
#The names for the country for each code were assigned in ArcGis

##########Open new data
#Open the csv data with the name of the country
ctm.data <- as.data.frame(read.csv("C:\\GFM MSc\\Thesis\\Preprocessing\\Data\\ArcGis\\ctmdata.csv"))

#Create columns for the regions
# Region A
tmp1  <- ctm.data
tmp1["region"] <- NA

tmp1$region[which(tmp1$CNTRY_NAME == "Austria")]  <- "CEE"
tmp1$region[which(tmp1$CNTRY_NAME == "Switzerland")]  <- "CEE"
tmp1$region[which(tmp1$CNTRY_NAME == "Czech Republic")]          <- "CEE"
tmp1$region[which(tmp1$CNTRY_NAME == "Poland")]          <- "CEE"
tmp1$region[which(tmp1$CNTRY_NAME == "Germany")]  <- "CEE"


tmp1$region[which(tmp1$CNTRY_NAME == "Belgium")]  <- "NWE"
tmp1$region[which(tmp1$CNTRY_NAME == "Luxembourg")]              <- "NWE"
```

```
tmp1$region[which(tmp1$CNTRY_NAME == "Netherlands")] <- "NWE"
tmp1$region[which(tmp1$CNTRY_NAME == "France" & tmp1$lat >= 45)]        <- "NWE"

tmp1$region[which(tmp1$CNTRY_NAME == "Italy")]            <- "SE"
tmp1$region[which(tmp1$CNTRY_NAME == "Spain")]            <- "SE"
tmp1$region[which(tmp1$CNTRY_NAME == "Portugal")]         <- "SE"
tmp1$region[which(tmp1$CNTRY_NAME == "France" & tmp1$lat < 45)]  <- "SE"


#Region B
tmp1["regionB"] <- NA

tmp1$regionB[which(tmp1$CNTRY_NAME == "Belgium")]  <- "BENELUX"
tmp1$regionB[which(tmp1$CNTRY_NAME == "Luxembourg")]  <- "BENELUX"
tmp1$regionB[which(tmp1$CNTRY_NAME == "Netherlands")]            <- "BENELUX"

tmp1$regionB[which(tmp1$CNTRY_NAME == "Germany")]  <- "GERPOL"
tmp1$regionB[which(tmp1$CNTRY_NAME == "Poland")]  <- "GERPOL"

tmp1$regionB[which(tmp1$CNTRY_NAME == "Slovakia")]  <- "SLOCZ"
tmp1$regionB[which(tmp1$CNTRY_NAME == "Czech Republic")]  <- "SLOCZ"

#Assign coordinates
coordinates(tmp1) <- ~lon+lat
proj4string(tmp1) <- CRS("+proj=longlat")
proj4string(tmp1)
tmp2 <- spTransform(tmp1, CRS("+init=epsg:3035"))

ctm.data  <- as.data.frame(tmp2)
colnames(ctm.data)[9]  <-  "easting"
colnames(ctm.data)[10]  <-  "northing"

#Transform to kilometers
ctm.data$easting <- ctm.data$easting/1000
ctm.data$northing <- ctm.data$northing/1000

#coordinates(ctm.data) <- ~easting+northing
#proj4string(ctm.data) <- CRS("+init=epsg:3035 +units=km")

#Subset the data
tmp3 <- subset(ctm.data, !is.na(ctm.data$region) | !is.na(ctm.data$regionB))
tmp3$pm25.LE <- tmp3$pm25_LE
tmp3$pm10.LE <- tmp3$pm10_LE
R1.ctm.data <-  tmp3[c(7, 9:12)]
R2.ctm.data <-  tmp3[8:12]
R1.ctm.data <- R1.ctm.data[complete.cases(R1.ctm.data),]
R2.ctm.data <- R2.ctm.data[complete.cases(R2.ctm.data),]
###Save data
save(tmp3, R1.ctm.data, R2.ctm.data, file="regions_ctm_data.Rdata")
##############################END#####################################
```

```
## Code for prediction with cokriging
# Make two data sets, one for regions A: Northwest Europe, one for Southern and one for Central Europe and one
for regions B: BENELUX, GERPOL and SLOCZ
# The table contains all the countries for each region and both pollutants PM10 and PM2.5

########## Require Libraries
library(gstat)
library(sp)

##########Load the data
setwd("C:/GFM MSc/Thesis/Preprocessing/Data")
load("compositions.Rdata")
#load("compositions2.Rdata")

########## Create validation and prediction data set
#Retrieve the validation data set  (25% of random observation)
df <- merge(pm10, pm2.5, by="st.code")
ss <- sort(sample(1:255, 75))#for region A
ss <- sort(sample(1:255, ))#for region A
df <- df[ss,]#for region A
vd <- df[,c(6:9, 17:18, 23)]
colnames(vd)[7]  <-  "region"
row.names(vd) <- NULL

#Retrieve the prediction data set
pd <- pm2.5[!(pm2.5$lat %in% vd$lat),]

################ Descriptive statistics

#Summary statistics
summary(pm10[,6:7])
summary(pm2.5[,6:7])
summary(vd)
summary(pd[,6:11])

#Histogram
hist(pm10$pm10.obs, xlab="PM10 observations", main="Histogram of PM10")
abline(v=median(pm10$pm10.obs), col="green")
abline(v=mean(pm10$pm10.obs), col="red")
hist(log(pm10$pm10.obs), xlab="log PM10 observations", main="Histogram of log PM10")
abline(v=median(log(pm10$pm10.obs)), col="green")
abline(v=mean(log(pm10$pm10.obs)), col="red")

hist(pm2.5$pm25.obs, xlab="PM2.5 observations", main="Histogram of PM2.5")
abline(v=median(pm2.5$pm25.obs), col="green")
abline(v=mean(pm2.5$pm25.obs), col="red")
hist(log(pm2.5$pm25.obs), xlab=" log PM2.5 observations", main="Histogram of log PM2.5")
abline(v=median(log(pm2.5$pm25.obs)), col="green")
abline(v=mean(log(pm2.5$pm25.obs)), col="red")
```

```
#qqplot PM10
qqnorm(pm10$pm10.obs, main="Q-Q plot of PM10")
qqline(pm10$pm10.obs)
qqnorm(log(pm10$pm10.obs), main="Q-Q plot of log PM10")
qqline(log(pm10$pm10.obs))

#qqplot PM2.5
qqnorm(pm2.5$pm25.obs, main="Q-Q plot of PM2.5")
qqline(pm2.5$pm25.obs)
qqnorm(log(pm2.5$pm25.obs), main="Q-Q plot of log PM2.5")
qqline(log(pm2.5$pm25.obs))


# Check the correlations of the observations and the CTM

cor(log(pm2.5[,6:7]))
cor(log(pm10[,6:7]))


########## Coordinates

#Define the coordinates
coordinates(pd) <- ~easting+northing
proj4string(pd) <- CRS("+init=epsg:3035 +units=km")
coordinates(vd) <- ~easting.x +northing.x
proj4string(vd) <- CRS("+init=epsg:3035 +units=km")
coordinates(pm2.5) <- ~easting+northing
proj4string(pm2.5) <- CRS("+init=epsg:3035 +units=km")
coordinates(pm10) <- ~easting+northing
proj4string(pm10) <- CRS("+init=epsg:3035 +units=km")

plot(pd@coords)
title( main="Locations of the Prediction Data")
plot(vd@coords)
title( main="Locations of the Validation Data")

#Find the values that repeat
#Find point pairs with equal spatial coordinates
#When using kriging, duplicate observations sharing identical spatial locations result in
#singular covariance matrices. This function may help identify and remove spatial duplices
zd <- zerodist2(pm10, pd)
zd
#Remove duplicates
ndpm10 <- pm10[-zd[,1], ]

plot(pm2.5@coords)
title(main="PM2.5 Stations")
plot(pm10@coords)
title(main="PM10 Stations")
```

```
#########Define a model
#Establish regions as factors to use them on as covariates

factor(pd$type_area)
factor(pd$region)

factor(ndpm10$type_area)
factor(ndpm10$region)

#Create the model
mdl2.5 <- lm(log(pm25.obs) ~ log(pm25.LE)+type_area+region, data=pd)
par(mfrow=c(2,2))
plot(mdl2.5)
anova(mdl2.5)
summary(mdl2.5)

samdl2.5 <- lm(log(pm25.obs) ~ log(pm25.LE)+region, data=pd)
par(mfrow=c(2,2))
plot(samdl2.5)
anova(samdl2.5)
summary(samdl2.5)

mdl10 <- lm(log(pm10.obs) ~ log(pm10.LE)+type_area+region, data=ndpm10)
par(mfrow=c(2,2))
plot(mdl10)
anova(mdl10)
summary(mdl10)

samdl10 <- lm(log(pm10.obs) ~ log(pm10.LE)+region, data=ndpm10)
par(mfrow=c(2,2))
plot(samdl10)
anova(samdl10)
summary(samdl10)

########## Empirical Variogram
# Compute variograms for PM10 and PM2.5
pm2.5.ev <- variogram(log(pm25.obs)~ log(pd$pm25.LE)+region, data=pd, cutoff=1000)
#pm2.5.ev <- variogram(log(pm25.obs)~ log(pd$pm25.LE)+region, data=pd, cutoff=700)#For region B
plot(pm2.5.ev)
pm2.5.ev

pm10.ev <- variogram(log(pm10.obs)~ log(ndpm10$pm10.LE)+region, data=ndpm10)
#pm10.ev <- variogram(log(pm10.obs)~ log(ndpm10$pm10.LE)+region, data=ndpm10, cutoff= 1100)#For region
B
plot(pm10.ev)
pm10.ev

########## Variogram Model
pm2.5.mv <- fit.variogram(pm2.5.ev, model=vgm(0.12, "Sph", 300, 0.04))
#pm2.5.mv <- fit.variogram(pm2.5.ev, model=vgm(0.06, "Exp", 150, 0.02))#For region B
```

```
pm10.mv <- fit.variogram(pm10.ev, model=vgm(0.12, "Sph", 400, 0.10))
#pm10.mv <- fit.variogram(pm10.ev, model=vgm(0.16, "Sph", 450, 0.14))#For region B

# Plot the Empirical Variogram and Variogram Model
plot(pm2.5.ev, pm2.5.mv)
plot(pm10.ev, pm10.mv)
str(pm2.5.mv)
str(pm10.mv)

########## Sample variograms and cross-variograms
# Create a gstat object, g, to hold the data for log(PM10)
rm(g)
# We use the prediction sample for pm2.5 (primary) and pm10 (secondary) variable
g <- gstat(NULL, "ln.pm2.5", log(pm25.obs) ~ log(pm25.LE)+region, pd)

# Append the other variables to g
g <- gstat(g, "ln.pm10", log(pm10.obs) ~ log(pm10.LE)+ region, ndpm10)

v <- variogram(g)
# Check the output and plot the result
v
plot(v)
# Define the initial variogram model and append it to g
g <- gstat(g, model=vgm(0.04, "Sph", 600, 0.1), fill.all=TRUE)
#g <- gstat(g, model=vgm(0.08, "Sph", 200, 0.04), fill.all=TRUE)# for region B
# Use the LMC for fitting
g.fit <- fit.lmc(v, g)
g.fit
plot(v, g.fit)

#vgm.map = variogram(g, cutoff = 1000, width = 1000/15, map = TRUE)
#plot(vgm.map, threshold = 15, col.regions = bpy.colors(), xlab = "", ylab = "")

# Append the fitted model to g
g <- g.fit

##########Co-kriging
#Use the grid for prediction
#grid
load("regions_ctm_data.Rdata")

row.names(R1.ctm.data) <- NULL
coordinates(R1.ctm.data) <- ~easting+northing
proj4string(R1.ctm.data) <- CRS("+init=epsg:3035 +units=km")
spplot(R1.ctm.data, zcol=1, edge.col=TRUE)
plot(R1.ctm.data@coords)


row.names(R1.ctm.data) <- NULL
colnames(R2.ctm.data)[1]  <-  "region"
```

```
coordinates(R2.ctm.data) <- ~easting+northing
proj4string(R2.ctm.data) <- CRS("+init=epsg:3035 +units=km")
spplot(R2.ctm.data, zcol=1, edge.col=TRUE)
plot(R2.ctm.data@coords)


tmp1 <- predict.gstat(g, newdata=R1.ctm.data)
#tmp1 <- predict.gstat(g, newdata=R2.ctm.data) #for regions B
spplot(tmp1, zcol="ln.pm2.5.pred", scales=list(draw=TRUE),  col.regions=bpy.colors(20))
tmp2 <- as.data.frame(tmp1)
tmp2$pm2.5.pred <- exp(tmp2$ln.pm2.5.pred)
coordinates(tmp2) <- ~easting+northing
proj4string(tmp2) <- CRS("+init=epsg:3035 +units=km")
spplot(tmp2, zcol="pm2.5.pred", scales=list(draw=TRUE), col.regions=bpy.colors(20))

#Perform accuracy assesment using the sub-sample
aa1  <- predict.gstat(g, vd)
pm2.5.err1 <- log(vd$pm25.obs) - aa1$ln.pm2.5.pred
sum(pm2.5.err1)/length(pm2.5.err1) # Mean Error
sum(pm2.5.err1^2)/length(pm2.5.err1) # RMSE

########## Universal Kriging
pb.mv <- fit.variogram(pm2.5.ev, model=vgm(0.10, "Sph", 500, 0.9))
aa2  <- krige(log(pm25.obs) ~ log(pm25.LE), pd, newdata=vd, model=pb.mv)
pb.err2 <- log(vd$pm25.obs) - aa2$var1.pred
sum(pb.err2)/length(pb.err2) # Mean Error
sum(pb.err2^2)/length(pb.err2) # RMSE

##############################END####################################

#Code for analysis of Compositional Data
#Author: Marisol Amador
#Set the libraries
library(sp)
library(compositions)
library(rgdal)
library(gstat)
library(plyr)
library(ggplot2)
library(MASS)
library(lattice)
library(energy)

##########Load the data
setwd("C:/GFM MSc/Thesis/Preprocessing/Data")

#load("compositions.Rdata")
load("compositions2.Rdata")

#Join the tables by the station code "st.code
```

```
comp <- merge(pm10, pm2.5, by="st.code")
head(comp)

#Check the correlation between PM2.5 and PM10
cor(comp[,c(6,17)])

#Create a table with variables of interest
ncom <- comp[,c(1:12, 17:18)]
head(ncom)

#Compute PM coarse for the observations
ncom$coarse.obs <- ncom$pm10.obs-ncom$pm25.obs
#Compute PM coarse for the model
ncom$coarse.LE <- ncom$pm10.LE-ncom$pm25.LE

########## Get the ratios and coarse fraction
#Compute the ratios of the observations
ncom$ratios.obs <- ncom$pm25.obs/ncom$pm10.obs
#Compute the ratios of the model
ncom$ratios.LE <- ncom$pm25.LE/ncom$pm10.LE

#Coarse fraction ratios
#Compute the ratios of the observations
ncom$cratios.obs <- ncom$coarse.obs/ncom$pm10.obs
#Compute the ratios of the model
ncom$cratios.LE <- ncom$coarse.LE/ncom$pm10.LE

##########Get basic statistics
obs <- ncom[,c("pm25.obs", "coarse.obs", "pm10.obs", "ratios.obs", "cratios.obs")]
LE <- ncom[,c( "pm25.LE", "coarse.LE", "pm10.LE", "ratios.LE", "cratios.LE")]

summary(obs)
summary(LE)

#The coarse fraction seems to be negative and 0 in some cases
#Delete Rows with negative values or equal to 0
ncomp <- ncom[!(ncom$coarse.obs <= 0),]
#show the new statistics
obs <- ncomp[,c("pm25.obs", "coarse.obs", "pm10.obs", "ratios.obs", "cratios.obs")]
LE <- ncomp[,c( "pm25.LE", "coarse.LE", "pm10.LE", "ratios.LE", "cratios.LE")]
summary(obs)
summary(LE)

#Inspect the number of stations in each country, region and type of area
summary(as.factor(ncomp$coun.x))
summary(as.factor(ncomp$region.x))
summary(as.factor(ncomp$type_area.x))

#Analyze the mean and standard deviation of the ratios
#By the type of area
```

```
aggregate(ncomp$ratios.obs, list(type_area= ncomp$type_area.x), mean)
aggregate(ncomp$ratios.obs, list(type_area= ncomp$type_area.x), sd)

#By the type of area and country
count(ncomp, c("type_area.x", "coun.x"))
aggregate(ncomp$ratios.obs, list(type_area= ncomp$type_area.x, country=ncomp$coun.x), mean)
aggregate(ncomp$ratios.obs, list(type_area= ncomp$type_area.x, country=ncomp$coun.x), sd)

#By type of area and region
count(ncomp, c("type_area.x", "region.x"))
aggregate(ncomp$ratios.obs, list(type_area= ncomp$type_area.x, region=ncomp$region.x), mean)
aggregate(ncomp$ratios.obs, list(type_area= ncomp$type_area.x, region=ncomp$region.x), sd)

#Plot the stations of the table
plot(ncomp$easting.x, ncomp$northing.x, xlab="easting", ylab="northing",
     col=ifelse(ncomp$type_area.x == "rural", 'green', ifelse(ncomp$type_area.x == "suburban", 'red', 'black')))
title(main="European Stations", sub="Rural, Urban and Suburban stations", cex.main = 1.5,
      font.main= 2, cex.sub = 0.75, font.sub = 3, col.sub = "blue")

ggplot(ncomp)+geom_line(aes(x=coun.x, y=ratios.obs, colour=type_area.x))+
  facet_wrap(~region.x)+guides(col=guide_legend(ncol=3))

##########Compositional Data Analysis
#Close the compositions and compare with the ratios
oc <- ncomp[,c("pm25.obs", "coarse.obs")]
#This ratios will be the same as the ratios calculated above
coi <- clo(oc, parts=c("pm25.obs", "coarse.obs"), total=1)
#compare the summary statistics of "coi" with the ratios of "obs" (They have to be the same)
#Give scale to the complete data set
y= acomp(oc)
#Transformation of the complete data set
ct <- clr(y)
it <- clr2ilr(ct)
plot(ct)
title(main="Clr Transformed Compostition")
lines(ct)

plot(it)
title(main="Ilr Transformed Compostition")
lines(it)


##########Linear model
#Unified Model
y= acomp(oc)

covariables= ncomp[, c("pm25.LE", "coarse.LE", "type_area.x", "region.x")]
x1= acomp(covariables[,1:2])
x2=factor(covariables$type_area.x)
x3=factor(covariables$region.x)
```

```
#Including all the covariates
(umodel <- lm(ilr(y)~ ilr(x1)*x2*x3))
summary(umodel)
anova(umodel)
par(mfrow=c(2,2))
plot(umodel)

#Including CTM and region
(umodel2 <- lm(ilr(y)~ilr(x1)*x3))
summary(umodel2)
anova(umodel2)
plot(umodel2)

#Including CTM and type of area
(umodel3 <- lm(ilr(y)~ilr(x1)*x2))
summary(umodel3)
anova(umodel3)
plot(umodel3)

#Including region and type of area
(umodel4 <- lm(ilr(y)~x2*x3))
summary(umodel4)
anova(umodel4)
plot(umodel4)

#Including only CTM
(umodel5 <- lm(ilr(y)~ilr(x1)))
summary(umodel5)
anova(umodel5)
plot(umodel5)

##########Analysis by region
#Dependent Variable
Y <- ncomp[,c("pm25.obs", "coarse.obs")]
compY= acomp(Y)
#compY1 <- acomp(compY[ncomp$region.x=="CEE",])
#compY2 <- acomp(compY[ncomp$region.x=="NWE",])
#compY3 <- acomp(compY[ncomp$region.x=="SE",])

compY1 <- acomp(compY[ncomp$region.x=="BENELUX",])
compY2 <- acomp(compY[ncomp$region.x=="GERPOL",])
compY3 <- acomp(compY[ncomp$region.x=="SLOCZ",])

#Independent Variable
X <- ncomp[,c("pm25.LE", "coarse.LE")]
compX=acomp(X)
#compX1 <- acomp(compX[ncomp$region.x=="CEE",])
#compX2 <- acomp(compX[ncomp$region.x=="NWE",])
#compX3 <- acomp(compX[ncomp$region.x=="SE",])
```

```
compX1 <- acomp(compX[ncomp$region.x=="BENELUX",])
compX2 <- acomp(compX[ncomp$region.x=="GERPOL",])
compX3 <- acomp(compX[ncomp$region.x=="SLOCZ",])

#Linear Model for CEE ===== plot, coefficients and Anova test
opar <- par(mar=c(4,4,0,0), oma=c(3,3,0.1,0.1))
pairwisePlot(clr(compX1), clr(compY1))
#mtext(text=c("model", "Central Eastern Europe observations"),side=c(1,2), at=0.5, line=2, outer=TRUE)
mtext(text=c("model", "Belgium, Netherlands and Luxemburg observations"),side=c(1,2), at=0.5, line=2,
outer=TRUE)

(modelR1 <- lm(ilr(compY1)~ilr(compX1)))
summary(modelR1)
anova(modelR1)

#Linear Model for NWE ===== plot, coefficients and Anova test
opar <- par(mar=c(4,4,0,0), oma=c(3,3,0.1,0.1))
pairwisePlot(clr(compX2), clr(compY2))
#mtext(text=c("model", "North-Western Europe observations"),side=c(1,2), at=0.5, line=2, outer=TRUE)
mtext(text=c("model", "Germany and Poland observations"),side=c(1,2), at=0.5, line=2, outer=TRUE)

(modelR2 <- lm(ilr(compY2)~ilr(compX2)))
summary(modelR2)
anova(modelR2)


#Linear Model for Southern Europe area ===== plot, coefficients and Anova test
opar <- par(mar=c(4,4,0,0), oma=c(3,3,0.1,0.1))
pairwisePlot(clr(compX3), clr(compY3))
#mtext(text=c("model", "Southern Europe observations"),side=c(1,2), at=0.5, line=2, outer=TRUE)
mtext(text=c("model", "Slovakia and Czech Republic observations"),side=c(1,2), at=0.5, line=2, outer=TRUE)

(modelR3 <- lm(ilr(compY3)~ilr(compX3)))
summary(modelR3)
anova(modelR3)

##########Descriptive Statistics
##########Analysis by region

###Central Eastern Europe
###Second Analysis for BENELUX

#R1 <- subset(ncomp, region.x=="CEE", select=c(pm25.obs, coarse.obs))
R1 <- subset(ncomp, region.x=="BENELUX", select=c(pm25.obs, coarse.obs))
R1 <- clo(R1, parts=c("pm25.obs", "coarse.obs"), total=1)
#Give scale to the measurements
SR1 <- acomp(R1)

#Transformation of the data
#Centered Log-ratio transformation
```

```
CTSR1 <- clr(SR1)
ITSR1 <- ilr(SR1)
ITSR1 <- clr2ilr(CTSR1)

plot(CTSR1)
title(main="Clr Transformed Region 1 Compostition")
lines(CTSR1)

plot(ITSR1)
title(main="Ilr Transformed Region 1 Compostition")
lines(ITSR1)

#Test for compositional normality
#Testing Marginals
qqnorm(SR1, alpha=0.05, main="Normal Q-Q Plot: Region 1 Composition")
qqnorm(CTSR1, main="Normal Q-Q Plot Clr: Region 1")
qqline(CTSR1)
qqnorm(ITSR1, main="Normal Q-Q Plot Ilr: Region 1")
qqline(ITSR1)

#Mean of the composition
mean(SR1)
#Metric Variance
mvar(SR1)
#Metric Standard Deviation
msd(SR1)
#variation matrix
variation(SR1)

###North-Western Europe
###Second Analysis for GERPOL

#R2 <- subset(ncomp, region.x=="NWE", select=c(pm25.obs, coarse.obs))
R2 <- subset(ncomp, region.x=="GERPOL", select=c(pm25.obs, coarse.obs))
R2 <- clo(R2, parts=c("pm25.obs", "coarse.obs"), total=1)
#Give scale to the measurements
SR2 <- acomp(R2)

#Transformation of the data
#Centered Log-ratio transformation
CTSR2 <- clr(SR2)
ITSR2 <- ilr(SR2)
ITSR2 <- clr2ilr(CTSR2)

plot(CTSR2)
title(main="Clr Transformed Region 2 Compostition")
lines(CTSR2)
plot(ITSR2)
title(main="Ilr Transformed Region 2 Compostition")
lines(ITSR2)
```

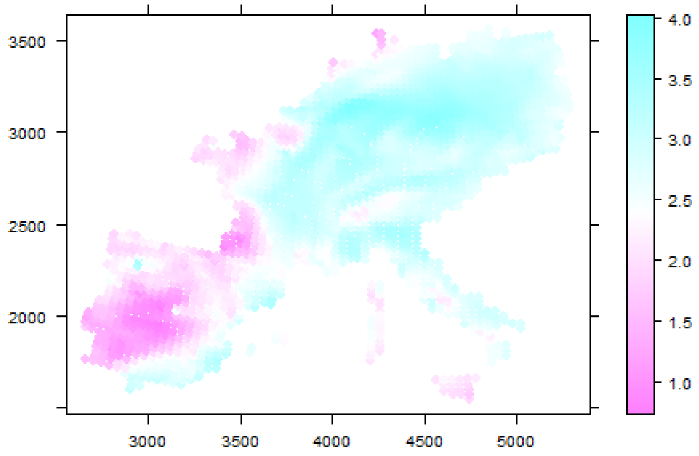
```
#Test for compositional normality
qqnorm(SR2, main="Normal Q-Q Plot: Region 2 Composition", alpha=0.05)
qqnorm(CTSR2, main="Normal Q-Q Plot Clr: Region 2 composition")
qqline(CTSR2)
qqnorm(ITSR2, main="Normal Q-Q Plot Ilr: Region 2 Composition")
qqline(ITSR2)

#Mean of the composition
mean(SR2)
#Metric Variance
mvar(SR2)
#Metric Standard Deviation
msd(SR2)
#variation matrix
variation(SR2)

###Southern Europe
###Second Analysis for SLOCZ

#R3 <- subset(ncomp, region.x=="SE", select=c(pm25.obs, coarse.obs))
R3 <- subset(ncomp, region.x=="SLOCZ", select=c(pm25.obs, coarse.obs))
R3 <- clo(R3, parts=c("pm25.obs", "coarse.obs"), total=1)
#Give scale to the measurements
SR3 <- acomp(R3)

#Transformation of the data
#Centered Log-ratio transformation
CTSR3 <- clr(SR3)
#Isometric log-ratio transformation
ITSR3 <- ilr(SR3)
ITSR3 <- clr2ilr(CTSR3)

plot(CTSR3)
title(main="Clr Transformed Region 3 Compostition")
lines(CTSR3)

plot(ITSR3)
title(main="Ilr Transformed Region 3 Compostition")
lines(ITSR3)

#Test for compositional normality
qqnorm(SR3, main="Normal Q-Q Plot: Region 3 Composition")
qqnorm(CTSR3, main="Normal Q-Q Plot Clr: Region 3 Composition")
qqline(CTSR3)
qqnorm(ITSR3, main="Normal Q-Q Plot Ilr: Region 3 composition")
qqline(ITSR3)

#Mean of the composition
```

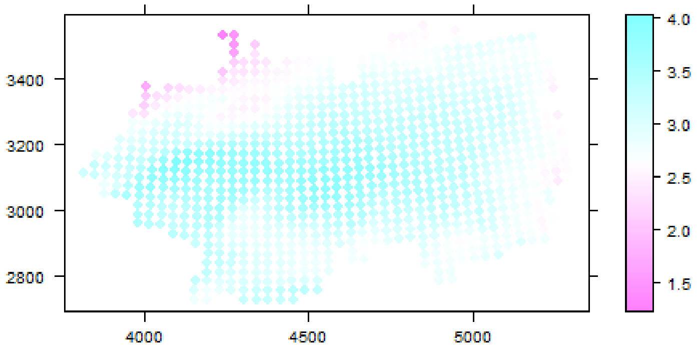
```
mean(SR3)
#Metric Variance
mvar(SR3)
#Metric Standard Deviation
msd(SR3)
#variation matrix
variation(SR3)


##########Variogram
#Try different methods to include the model in the variogram
#Best Model=umodel2

#Data frame
gr <- data.frame(ncomp$pm25.obs, ncomp$coarse.obs, ncomp$easting.x, ncomp$northing.x)
colnames(gr) <- c("pm25", "coarse", "x", "y")
head(gr)

###For Central Eastern Europe
### Second analysis for Belgium, Netherlands and Luxemburg
#Y1 <- gr[ncomp$region.x=="CEE",]
Y1 <- gr[ncomp$region.x=="BENELUX",]
coord1 <- Y1[,c("x", "y")]
Y1 <- clo(Y1[,1:2])
compY1 <- acomp(Y1)

lrv1 <- logratioVariogram(compY1, coord1)
plot(lrv1)
lrvModel1 <- CompLinModCoReg(~nugget()+R1*sph(500), compY1)
vgmModel1 <- vgmFit2lrv(lrv1, lrvModel1, print.level=0)
vgmModel1
plot(lrv1, lrvg=vgram2lrvgram(vgmModel1$vg))

###North Western Europe
###Second Analysis for Germany and Poland
#Y2 <- gr[ncomp$region.x=="NWE",]
Y2 <- gr[ncomp$region.x=="GERPOL",]
coord2 <- Y2[,c("x", "y")]
Y2 <- clo(Y2[,1:2])
compY2 <- acomp(Y2)

lrv2 <- logratioVariogram(compY2, coord2)
plot(lrv2)
lrvModel2 <- CompLinModCoReg(~nugget()+R1*gauss(100), compY1)
vgmModel2 <- vgmFit2lrv(lrv2, lrvModel2, print.level=0)
vgmModel2
plot(lrv2, lrvg=vgram2lrvgram(vgmModel2$vg))


###Southern Europe
```

```
###Second Analysis for Slovakia and Czech Republic
#Y3 <- gr[ncomp$region.x=="SE",]
Y3 <- gr[ncomp$region.x=="SLOCZ",]
coord3 <- Y3[,c("x", "y")]
Y3 <- clo(Y3[,1:2])
compY3 <- acomp(Y3)

lrv3 <- logratioVariogram(compY3, coord3)
plot(lrv3)
lrvModel3 <- CompLinModCoReg(~nugget()+R1*sph(200), compY1)
vgmModel3 <- vgmFit2lrv(lrv3, lrvModel3, print.level=0)
vgmModel3
plot(lrv3, lrvg=vgram2lrvgram(vgmModel3$vg))

############################################Use the CTM grid

load("regions_ctm_data.Rdata")
plot(R1.ctm.data$easting, R1.ctm.data$northing)
plot(R2.ctm.data$easting, R2.ctm.data$northing)

#Grids for group of regions A
R1G <- R1.ctm.data[which(R1.ctm.data$region=="CEE"),]
R1G <- R1G[,2:3]
R2G <- R1.ctm.data[which(R1.ctm.data$region=="NWE"),]
R2G <- R2G[,2:3]
R3G <- R1.ctm.data[which(R1.ctm.data$region=="SE"),]
R3G <- R3G[,2:3]


#Grids for group of regions B
R1G <- R2.ctm.data[which(R2.ctm.data$regionB=="BENELUX"),]
R1G <- R1G[,2:3]
R2G <- R2.ctm.data[which(R2.ctm.data$regionB=="GERPOL"),]
R2G <- R2G[,2:3]
R3G <- R2.ctm.data[which(R2.ctm.data$regionB=="SLOCZ"),]
R3G <- R3G[,2:3]

###################################################Prediction
CK1 <- compOKriging(compY1, coord1, R1G, vg=vgmModel1$vg)
summary(CK1)
str(CK1)
a1 <- as.data.frame(CK1$Z)
a2 <- as.data.frame(CK1$X)
aa = cbind(a1, a2)
head(aa)

CK2 <- compOKriging(compY2, coord2, R2G, vg=vgmModel2$vg)
summary(CK2)
str(CK2)
b1 <- as.data.frame(CK2$Z)
```

```
b2 <- as.data.frame(CK2$X)
bb = cbind(b1, b2)
head(bb)

CK3 <- compOKriging(compY3, coord3, R3G, vg=vgmModel3$vg)
summary(CK3)
str(CK3)
c1 <- as.data.frame(CK3$Z)
c2 <- as.data.frame(CK3$X)
cc = cbind(c1, c2)
head(cc)

m <- rbind(aa, bb, cc)
#n <- cbind(m, R1.ctm.data)
n <- cbind(m, R2.ctm.data)

coordinates(m) <- ~easting+northing
proj4string(m) <- CRS("+init=epsg:3035 +units=km")
spplot(m, zcol="pm25", scales=list(draw=TRUE),  col.regions=bpy.colors(20))
###############################END####################################
```

## Appendix 2: Prediction Grid for group of regions A and B


3500
3000
2500
2000
3000
3500
4000
4500
5000
4.0
3.5
3.0
2.5
2.0
1.5
1.0


3400
3200
3000
2800
4000
4500
5000
4.0
3.5
3.0
2.5
2.0
1.5