NON-STATIONARY LINEAR MIXED MODELLING OF AIR QUALITY

FEVEN SOLOMON DESTA

February, 2012

SUPERVISORS:
Dr. N. A. S. Hamm
Prof. Dr. Ir. A. Stein

NON-STATIONARY LINEAR MIXED MODELLING OF AIR QUALITY

FEVEN SOLOMON DESTA
Enschede, The Netherlands, [February, 2012]

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

SUPERVISORS:
Dr. N. A. S. Hamm
Prof.Dr. Ir. A. Stein

THESIS ASSESSMENT BOARD:
Chair: Prof.Dr. Ir. A. Stein
External Examiner: Dr.Ir. G.B.M. Heuvelink

ABSTRACT

The issue of air quality has become a major concern, since the quality of air has deteriorated from time to time. Air quality is essential for human health and quality of life that it needs to be assessed at every location. Among the different pollutant sources PM_{10} was modelled for this study based on the fact that, it takes a large portion of pollutant load in the air and it presents a health risk which is of increasing concern. Air quality monitoring stations are limited in space that there is a need for interpolation of information at every location in space. Thus, geostatistical methods are considered to be efficient for modelling of air quality and predicting at unsampled sites. For this study, Linear mixed model (LMM), a model-based geostatistics approach was applied to model the non-stationarity in variance and spatial correlation of PM_{10}.

PM_{10} showed different distribution over different temporal aggregations. Moreover, the spatial distribution of PM_{10} differs from region to region and there are areas on which the concentration exceeds the standard limit set by EU.

Most studies assume that pollution processes are stationary but in real application environmental processes are non-stationary (in the mean and in variance). This research explored how the LMM can be extended to model the heteroskedasticity (non-constant variance) and spatial correlation of air pollution process. For this study, a simple extension of LMM model with parametric non-stationary variance model resulted in an improved model for PM_{10} .

For LMM specifications four models were compared, the first model (Model A) a simple LMM with ordinary least square (OLS) estimation (which do not take in to account the heteroskedasticity and spatial correlations), for the second one (Model B) the LMM was extended to model the spatial correlation but not the heteroskedasticity, for the third one (Model C) the LMM was extended to model the heteroskedasticity but not the spatial correlations and finally the fourth LMM (Model D) took in to account both the spatial correlation and the heteroskedasticity. Model D has a lower AIC value and higher log likelihood value than the other models; therefore it was found that the non-stationary model is an improved model over the stationary model, this is further supported by likelihood ratio test result. In general, the step wise specification of the LMM clearly indicated the improvement of the model with spatial correlation structure and heteroskedasticity modelling.

Prediction was done at 326 prediction sites, and accuracy of the models was evaluated using RMSE and ME, Model D which accounts both heteroskedasticity and spatial correlation has a higher accuracy than the other models. Moreover, the accuracy was checked for different level of grouping structure, the prediction at level two (multi-level specification) gave a better result than the population level and level one specification.

The extended LMM was applied for summer, winter and yearly average data. The spatial structures of the variogram models for these temporal aggregations show difference in the models parameters that the choice of temporal aggregation should be taken in to account for modelling of air quality.

The assumption for LMM was checked and the study is supported by the explanatory analysis that the approach is scientifically plausible. In general, this work showed the prospect of LMM application for air quality modelling.

Key words: Air quality, Linear Mixed Model, Non-stationarity, Temporal aggregations, Geostatistics

ACKNOWLEDGEMENTS

Thank you GOD for everything that you have done for me.

I owe my deepest gratitude to my supervisor Dr. Nicholas Hamm, his guidance and encouragement from the initial to the final of the thesis work enabled me to develop an understanding of the subject. I would like to appreciate his swift reply for the e-mails and his office was open any time for consultations. Above all, his critical view and criticism helped for the successful accomplishment of this thesis work. Thank you Dr. Hamm.

It is an honour for me to thank my second supervisor Prof. Alfred Stein. I appreciate your critical view on the results, it helped me to think widely and see results from different directions. Thank you Prof. Stein.
I am indebted to all my instructors in ITC.

I would like to thank my sponsor WORLD BANK.

I am grateful to my family; my dad Solomon Desta and my mom Segged Kifle, I am always thankful for everything you did to me. My thanks also go to my brothers and sisters. Heni, Thank you so much.

It is a pleasure to thank my friends who supported me in any respect during the completion of the thesis work.

TABLE OF CONTENTS

Abstract i
Acknowledgements ii
List of figures v
List of tables vi
List of acronomy vi
1. Introduction 1
1.1. Motivation 1
1.2. Problem statement 3
1.3. Research objectives 3
1.3.1. Overall research objectives 3
1.3.2. Specific objectives 3
1.4. Research questions? 4
1.5. Innovation aimed at 4
1.6. Thesis structure 4
2. Related work 5
2.1. PM_{10} over Europe 5
2.2. Classical geostatistics and air quality modelling 5
2.3. Non-stationarity in the variance for environmental processes 5
2.4. Type (background) area, elevation and air quality 6
2.5. Temporal aggregations and environmental processes 7
2.6. AOT and PM_{10} 7
3. Study area and data description 9
3.1. Study area 9
3.2. Data description 9
4. Methodology 11
4.1. Data pre-processing 11
4.2. Data exploration 11
4.3. Data projection 12
4.4. Temporal aggregations 13
4.4.1. Daily pattern analysis 13
4.5. Linear Mixed-Effects Modelling 13
4.6. The hypothesis of stationarity 14
4.7. GLS and LME functions for LMM 16
4.8. LMM specification using lme function 16
4.9. Fitting a linear mixed model 16
a) LMM without heteroskedasticity and spatial correlation modelling (Model A) 17
b) Correlation Structures specifications for LMM (Model B) 19
c) Modelling Heteroskedasticity (Model C) 19
d) Modelling both heteroskedasticity and spatial auto correlation (Model D) 20
4.10. LMM model specification using gls function 20
4.11. Examining a Fitted Model 20
4.12. Prediction at unsampled location 21
4.13. Software implementation 22
5. Results 23
5.1. Data exploration results 23
5.2. Temporal aggregations results 25
5.3. LMM results 31

a) LMM without heteroskedasticity and spatial correlation (Model A) ... 33
b) Modeling spatial correlation (Model B) ... 34
c) Modeling heteroskedasticity (Model C) ... 35
d) Modeling both heteroskedasticity and spatial correlation (Model D) ... 40
5.4. LMM model specification using GLS ... 41
5.5. LMM for summer and winter temporal aggregations ... 41
5.6. Prediction at unsampled location ... 43
6. Discussion ... 46
6.1. Temporal aggregations analysis ... 46
6.2. The LMM implementation ... 46
6.3. Prediction assesment ... 48
7. Conclusions and recommendations ... 50
7.1. Conclusions ... 50
7.2. Recommendations ... 51
List of apendices ... 53
List of references ... 61

LIST OF FIGURES

Figure 1: Location map of air quality monitoring stations 9
Figure 2: General methodology 12
Figure 3: Methodology for LMM 15
Figure 4: Histogram plot for year 2007, 2008 and 2009 in situ PM_{10} 23
Figure 5 : QQ plot for three years data 24
Figure 6: The PM_{10} (ppm) concentration over different temporal aggregations 25
Figure 7 : PM_{10} (ppm) concentration with high variability over different temporal aggregations 26
Figure 8: PM_{10} over different temporal aggregations having a horizontal line which indicates the limit value for PM concentration in air (which is 40 ppm for yearly average data and 50 ppm for daily average data) 26
Figure 9: Summer PM10 concentration trend over 5 consecutive days 27
Figure 10: Winter PM10 concentration trend over 5 consecutive days 27
Figure 11: PM10 (ppm) pattern for three years data 27
Figure 12: PM_{10} concentration variation a)Winter b) Summer (it shows values above and below the standard limit) 28
Figure 13: PM_{10} concentration variation over temporal aggregations 29
Figure 14: Variogram models for different temporal aggregations 30
Figure 15: Variogram model for different direction 32
Figure 16: Variograms for four different directions (directional variogram) 32
Figure 17: Concentration of in situ measurements and their distribution, filled circles with grey shades proportional to the measured in situ values 33
Figure 18: A plot for OLS residuals versus their spatial coordinates (Black dots represent negative residuals and grey dots positive residuals 34
Figure 19: Variogram of OLS residuals 35
Figure 20: Boxplots of residuals per group (country) for LMM fit (with no heteroskedasticity modelling) 36
Figure 21: Residual plots corresponding to LMM (type areas, as grouping factor) 36
Figure 22: Scatter plots of standardized residuals versus fitted values for LMM fit by type areas 37
Figure 23: Scatter plots of standardized residuals versus fitted values for LMM fit by country 37
Figure 24: Observed versus fitted values plot for model C 38
Figure 25: Normal plot of residuals for the Model C fit per type area 39
Figure 26: Normal plot of residuals for the Model C fit per country 39
Figure 27: Normal plot of residuals for the Model C fit 40
Figure 28: A variogram of Model D LMM 41
Figure 29: Variogram model for Summer season (Model D) 42
Figure 30: Variogram model for winter season (Model D) 43

LIST OF TABLES

Table 1: Research questions....4
Table 2: Summary of datasets used....10
Table 3: Summary of background areas (year 2008 in situ measurements)....10
Table 4: Summary of software's implementation for this study....22
Table 5: Summary of year 2007, 2008 and 2009 PM_{10} data....24
Table 6: In situ PM_{10} (ppm) concentration summary per background area for the temporal aggregations....25
Table 7: p_value for PM_{10} distribution trend check over type areas....31
Table 8: p-value for coefficients significance check....31
Table 9: Summary of LMM output for multi and single level groping specification of random effects....33
Table 10: Likelihood ratio test for single level versus multilevel (type area in country) grouping specification....34
Table 11: Summary of model output for Model A and Model B....35
Table 12: Likelihood ratio test for Model A versus Model B....35
Table 13: Summary of model output for heterokedastic (Model C) and homoskedastic (Model A) models....38
Table 14: Likelihood ratio test for heterokedastic versus homoskedastic model....38
Table 15: Summary of Model output for Model A and Model D....40
Table 16: Summary of model output for Model A, B, C and D....40
Table 17: Summary of LMM model outputs using `gls` function....41
Table 18: Multi-level and single level grouping structure for summer data....42
Table 19: Likelihood ratio test for single level versus multilevel grouping specification for summer data....42
Table 20: Summary of the LMM model outputs for summer data (using lme function)....42
Table 21: Summary of the model outputs for winter data....43
Table 22: RMSE summary of summer validation sites prediction using summer LMM developed by lme function....43
Table 23: ME summary of summer validation sites prediction using summer LMM developed by lme function....44
Table 24: Summer validation sites prediction using yearly average data LMM developed by lme function....44
Table 25: Winter validation sites prediction using winter LMM developed by gls function....44
Table 26: Summer validation sites prediction using summer LMM developed by gls function....44
Table 27: Winter and summer validation sites prediction using LMM (of gls function) developed for yearly average data....44
Table 28: Winter validation sites prediction using summer LMM developed by gls function....44

LIST OF ACRONYMS

AIC Akaike Information Criteria

AOT Aerosol Optical Thickness

CK Cokriging

CTM Chemical Transport Model

ELM Empirical Line Method

ELPI Electrical Low Pressure Impactor

EU European Union

GLS Generalized least square

IID Independent and identically normally distributed

LMM Linear Mixed Model

MBG Model-based geostatistics

ML Maximum Likelihood

ME Mean Error

UK Universal Kriging

OK Ordinary Kriging

OLS Ordinary Least Square

PM Particulate Matter

REML Residual Maximum Likelihood

RK Regression Kriging

RMSE Root Mean Square Error

SDO Saharan Dust Outbreak

SKlm Simple Kriging with Varying Local Means

SW South West

WHO World Health Organization

ELPI Electrical Low Pressure Impactor

1. INTRODUCTION

1.1. Motivation

Air pollution is recognized as one of the leading environmental problem, even in countries with relatively low concentrations of air pollutants. Air pollution affects the environment; as well as human health and quality of life. The World Health Organization (WHO) estimates that 1.5 billion people living in urban areas all over the world breathe dangerous levels of air pollution (Clean AIR Systems, 2007). WHO also says that air pollution ranks within the top 10 causes of worldwide death and disability. Moreover, aerosols affect the environment by modifying the radiative budget of the Earth (Koelemeijer et al., 2006).

There are natural and anthropogenic sources of air pollution, but generally natural sources are not as much of a problem as are human-generated pollutants or anthropogenic sources (Monks et al., 2009). The most common pollutants having anthropogenic sources include; particulate matter (PM_{10}), NO, CO and CO_2 (Huang et al., 2011).

Among the different pollutant sources, PM_{10} was modelled in this project. PM_{10} is a major constituent of air pollution that threatens both our health and our environment (Phalen, 2003). Smaller particles to be inhaled into the deepest parts of the lung are less than 10 μm diameter, and known as PM_{10}. In addition, toxic or harmful elements such as S, As, Ni and Mn are enriched mainly in fine particles. Major sources for PM_{10}, in both urban and rural areas, include: industries, motor vehicles and dust from construction, landfills and agriculture. Furthermore, particulate matter also forms when gases emitted from vehicles and industry undergoes chemical reactions in the atmosphere (Great Basin Unified Air Pollution Control District, 2007).

Understanding the spatial distribution of air pollution and having spatial predictions at unsampled locations is crucial for proper control of air pollution. Consequently, comprehensive understanding of spatial distribution and modelling is achieved by fusing different information sources using geostastical models (Kanevski, 2010).

Model-based geostatistics (MBG) means the application of explicit parametric stochastic models to geostatistical problems. MGB is explicit specification of a Gaussian process model, it gives specification of the distribution for the residuals and has explicit link to the linear model. A major advantage of MBG approaches is that they provide a flexible statistical platform for handling and representing different sources of uncertainty, providing plausible and robust information on the spatial distribution of phenomena (Diggle et al., 1998). Linear mixed model (LMM) is one of model-based approach. It is a linear model of both fixed and random effects. Classical geostatistics allows us to relax the assumption of stationarity in the mean, however, the intrinsic hypothesis remains that they do not allow as non-stationarity in variance (Lark., 2009). LMM framework with residual maximum likelihood (REML) estimation can relax the assumption of stationarity in the variance with relatively simple parametric variance models (Lark., 2009). In general, LMM allows modelling trend, spatial correlation and heteroskedasticity (Hamm, in review).

In order to improve the insight in PM_{10} distributions over Europe, integration of different variables namely; chemical transport model (CTM) output, elevation and ground-based measurements (PM_{10}); using

LMM was carried out. In situ field measurements are a direct method for data capturing and they provide better air quality information (Cinzia Mazzetti & Todini, 2002). However, it is unable to provide complete coverage of an area of interest because of the local character of the measurements.

Chemical Transport Model (CTM) LOTOS-EUROS is a 3D chemistry aimed to simulate air quality in the lower troposphere. Its output provides air quality information basing on knowledge of chemical and physical processes (Schaap et al., 2008). However, it has limitations for example, models require detailed information on relative humidity, wind speed, temperature, precipitation rates, pollution sources distributions, height of the source, (Beelen et al., 2009). Furthermore, models tends to under estimate pollutants concentration; for example LOTOS-EUROS model tends to underestimate $PM_{2.5}$ (Denby et al., 2008). Though, it was indicated that integrating data from different sources with in situ measurements was found to be successful for increasing prediction accuracy of air pollution maps (Singh et al., 2011), so that for this research AOT , model output and in situ field measurement will be integrated in order to produce air pollution model for Europe.

Elevation is the other covariate used for modelling of air quality, elevation was considered because various studies showed topographical influence on pollutants dispersion. For example, Beelen, et al., (2009) indicated that air pollution concentration related to elevation. Carvalho et al., (2006) analyze how mesoscale circulations induced by topography and/or land use control pollutants dispersion in a coastal region, it is also indicated that topography represented as the main driving force mechanism on air pollutants injection in higher tropospheric levels. Kim & Stockwell, (2008) indicated that complex terrain was shown to have an important influence on the vertical transport of air pollutants on the regional scale.

Moreover, Aerosol optical thickness (AOT) was considered for this study, compared to ground measurements. Satellite imagery, owing to their wide spatial coverage and reliable repeated measurements, provide another important tool to monitor aerosols and their transport patterns (Emili et al., 2010). Furthermore, PM10 represents point observations, they do not capture the pollution over wide areas and satellite data can be used in areas where ground measurements are not available. One important aerosol parameter retrieved from satellite sensors is AOT.

The transport pattern and dispersion of air pollutants in the air are influenced by, global and regional weather patterns and the different weather pattern over temporal aggregations makes it interesting to look PM_{10} distribution over temporal aggregations. Meteorological conditions play an important role in the formation, emission and deposition as well as spatial distribution of PM (Gomiscek et al., 2004a). For example, high anthropogenic emissions in combination with frequently occurring stagnant atmospheric conditions in the Po valley (in Northern Italy) cause very high PM concentrations in winter (Pernigotti et al., 2012). Meteorological conditions vary from region to region, moreover, local topographical conditions affect the way that pollutants are transported and dispersed (EPA., 2011) that it causes variation of PM distribution over regions.

In general, to apply proper air pollution regulation policy and mitigation, there is a need to understand the distribution of PM over geographic regions and temporal aggregations.

1.2. Problem statement

Most studies assume that pollution process (for example PM_{10}) are stationary, however, it is widely recognized that in real applications spatial processes are rarely stationary and isotropic. Air pollutants concentration often varies in response to metrological conditions, geographical differences and other factors. For this study, in order to have a better understanding of the distribution pattern of PM_{10} concentration, the non-stationarity over the mean and the variance were modelled using LMM (which provide more flexibility to model non-stationarity). Moreover, for proper understanding of PM_{10} concentration over different temporal aggregations and to comprehend the temporal aggregation signal on the correlation and cross correlation of the dependent and explanatory variables, analysis and modelling of PM_{10} over different temporal aggregation were conducted. So that, it will be possible to distinguish the geographic differences and temporal aggregations signal in the correlation and cross-correlation of elevation, PM_{10} and model output; and apply proper air pollution controlling mechanism.

1.3. Research objectives

1.3.1. Overall research objectives

To explore and model PM_{10} over different temporal aggregations and extend LMM to account for heteroskedasticity (non-constant variance) and spatial correlation of PM_{10} pollution process using different covariates.

1.3.2. Specific objectives

The specific objectives are:

- to explore the spatial structure of the correlation and cross-correlation of in situ data (PM_{10}), (CTM) model output and elevation over regions and over temporal aggregations (daily, monthly, seasonal and yearly);
- to extend LMM to account the spatial correlation and non-stationarity in variance of PM_{10} distribution ;
- to predict concentrations of pollutants at validation sites.

1.4. Research questions?

Objective Number	Research Questions?
1	What is the spatial distribution of in situ PM_{10} over regions?
	What is the spatial structure of the correlation of CTM, PM_{10} and elevation over temporal aggregations (daily, monthly, seasonal and yearly)?
2	How do we specify the fixed and random effects in LMM?
	How can we model the correlation structure in LMM?
	How do we model the non-stationarity in the variance using the LMM?
	Does the non-stationary model offer an improvement over stationary model?
3	Does the non-stationary model offer better prediction accuracy than stationary model?

Table 1: Research questions

1.5. Innovation aimed at

LMM which is one of the model based approach was applied since it provides flexibility to address non-stationarity in the variance of PM_{10} air pollution process; this is a new application for air quality modelling and gave an improved model.

1.6. Thesis structure

This thesis has 7 chapters. Chapter 1 informs the basis of the study in which the motivation, problem and objectives and research questions of this study are addressed. Chapter 2 provides some related works. Chapter 3 describes the data and study area. Chapter 4 provides methodologies adopted for this study. Chapter 5, 6 and 7 are results, discussion and conclusion respectively.

2. RELATED WORK

2.1. PM_{10} over Europe

In Europe, particulate matter is the most significant air pollutant that causes loss of human health. The EU wide standard limit value for PM_{10} concentration in the air is 40ppm for yearly average value and 50ppm for daily average value (Koelemeijer, et al., 2006).

2.2. Classical geostatistics and air quality modelling

Meul & Van Meirvenne., (2003) applied four geostatistical interpolation methods namely; ordinary kriging (OK) , universal kriging (UK), simple kriging with varying local means (SKlm) and ordinary cokriging (OCK); to compare their ability to account different types of non-stationarity over the mean for the topsoil silt content, they found out that the global trend was best accounted for by OCK and the local non-stationarity in the mean by UK ; so that they combined the results of the two prediction methods, and found out more precise overall estimation of silt content for the top soil than any single method used. Meul & Van Meirvenne., (2003) model the non-stationarity in the mean however, they did not account the non-stationarity in the variance.

Using the classical geostatistical methods, Beelen, et al., (2009) produce a map of air pollution at a fine spatial scale across the European Union and they used altitude and topography as predictor variables. They found out that universal kriging performed better than either regression models or ordinary kriging. This is consistent with the presence of spatial correlation in the concentrations even after specifying regression models, and systematic trends related to climate, geography (altitude especially).This study shows, the importance of elevation for air quality mapping.

Mwenda., (2011) applied regression kriging (RK), and cokriging (CK) to integrate in situ measurements (PM10), models (PM2.5) and remotely sensed data (AOT) and predict PM10 daily annual mean concentration over parts of Europe for the year 2003. He found out that RK gave better results as compared to CK. Moreover, he compared ordinary kriging (OK) and universal kriging (UK), and showed that both RK and UK gave similar results of RMSE (0.096) and correlation (0.72). However, his approach is based classical geostatistics that it does not give flexibility to handle non-stationarity in the variance.

2.3. Non-stationarity in the variance for environmental processes

Using a hierarchical Bayesian approach, modelling and prediction of ozone concentrations over different geopolitical boundaries across the USA was done by Fuentes.,(2002).The developed model captured the lack of stationarity of the air pollution process; and by using a Bayesian approach for spatial prediction; they successfully accounted for the uncertainty of the covariance parameters in interpolation.

Spatial processes in soil science, environmental sciences, oceanography, and many other disciplines are generally non-stationary (Fuentes, 2003). Fuentes., (2003) develop a method (which consider a hierarchical Bayesian approach to model and take into account the spatial structure of data when estimating the parameter) to test the lack of stationarity of a time series environmental processes. The method also used to examine the character of the non-stationarity and the potential anisotropy.

Lark, (2009) implement a simple extension of the random effects variance model in a LMM for the slope of soil surface, and extension of the random effects variance model in LMM allows non-stationarity in the variance to be modelled, hence tests on the log-likelihood ratio gave evidence in favour of the non-stationary model, and the results of prediction at validation sites revealed that it characterized the uncertainty of the predictions better than does a stationary correspondent. Lark, (2009) concluded that, a relatively slight relaxation of the stationarity assumptions can result in an improved spatial modelling and prediction of a quite complex environmental variable. Hamm et al., (in review) showed LMM implementation of the empirical line method (ELM), based on the linear relationship between at sensor radiance and DN to at surface reflectance. They model the heteroskedasticity and spatial correlation the two variables using LMM and obtained an improved model.

Various studies indicated that non-stationary models for the random effect fitted the data better than stationary models; and the difference is statistically significant. Moreover, it is pointed out that non-stationary models pronounce the error variance of predictions at the validation sites better than stationary models. For example, Haskard & Lark., (2009) carried out modelling of soil potassium by using linear mixed model, and allowing both the variance and autocorrelation of the property of interest to adjust locally in response to a set of covariates, relaxing the stationarity assumption gave a better account of the uncertainty of predictions at validation sites than did a simpler stationary LMM, and indicated that it is important to model non-stationarity in the variance.

2.4. Type (background) area, elevation and air quality

Gomiscek, et al., (2004b) indicated that an annual average mass concentration for PM in Austria, at the urban sites is higher than at the rural site. It is also pointed out that the number concentrations at the urban sites are in the upper European level and show a distinct seasonal cycle; however at the rural site no seasonal influence was seen. Moreover, at urban sites higher values were observed during winter time and at the rural site higher values were found during the summer period. This study shows the difference in PM concentration amount and characteristics for different surrounding areas (rural and urban).

Sanchez et al., (2007) studied the hourly variability in height of Saharan dust outbreak (SDO) in central Spain. They indicated that PM_{10} spread took 2 days in the upper plateau and 3 to 4 in the lower plateau, in agreement with the geographical location of the monitoring stations. They also pointed out that the greater impact of SDO was linked to the lower altitudes.

Held et al., (2008) used an electrical low pressure impactor (ELPI) to study atmospheric aerosol particle mass concentrations, and size distributions over a diameter range of 7 nm – 10 μm at urban, rural and high-alpine locations along an alpine altitude transect across southern Germany. They showed that long-term measurements at the rural site (Hohenpeißenberg, Germany) revealed distinct seasonal patterns with the highest number concentrations in summer and the highest mass concentrations in spring and fall. In addition, relatively clean air (PM10 < ¼ 5 mg m3) was generally advected from the Alps (SW), whereas urban air from Munich (NE) clearly contributed to elevated particle mass loadings (PM10 > ¼ 10 mg m^3). This study indicated that contribution of particle mass in air differs for different background areas (e.g urban and rural), moreover it also indicated that the Alps (elevated areas), has relatively clean air. Their findings suggest that improvement could be gained by modelling the variability in variance (heteroskedasticity) of air pollution process by considering elevation and background areas as covariates.

2.5. Temporal aggregations and environmental processes

Van Bussel et al., (2011) analysed the sensitivity of different modelling approaches (to model leaf area index development and associated radiation interception and biomass productivity), to the temporal resolution (temporal aggregations) of weather input data. The acquired results for different climatic regions in Europe showed that simulated biomass differs between model simulations using actual or aggregated temperature and/or radiation data. This study showed the differences between model results while using different temporal aggregation of input data for modelling of biomass productivity, and van Bussel et al., (2011) concluded from the implication of their result that the choice of a specific approach to model a certain process depend on the available temporal resolution of input data. This study indicates the importance of temporal aggregations for modelling of environmental processes.

2.6. AOT and PM_{10}

Correlations between satellite-derived AOT and PM surface concentration measurements were studied by many researchers. Comparison of temporal and spatial variations of AOT and particulate matter over Europe was done using ordinary least square method by Koelemeijer et al., (2006). The correlation of the AOT and PM were investigated under different meteorological conditions. Correlation coefficients were calculated between the one-year time-series of AOT and PM, and it is indicated that the spatial correlation between fitted and observed yearly average PM2.5 levels is 0.82, with a RMS-error of 2.8 μg/m3. Moreover, Koelemeijer et al., (2006) indicated that modelling yearly average PM2.5 distribution using both model output and measured AOT as explanatory variables showed RMS-errors decrease by about 25% compared to fitting with only one explanatory variable. However, the spatial correlation of the covariate was not considered in their study.

Using regression analysis, Gupta et al., (2006) derived empirical relationship between 24 h PM2.5 mass concentration and MODIS AOT over global cities and conclude that the satellite derived AOT is an excellent tool for air quality studies over large spatial area. Gupta et al., (2006) noted that satellite data are a remarkable asset for studying PM air quality over large spatial extent that is not possible from ground measurements alone.

Emili et.al. (2010) investigated the capability of spaceborne remote sensing data to predict ground concentrations of PM10 over the European Alpine region; using satellite derived Aerosol Optical Depth (AOD) from the geostationary Spinning Enhanced Visible and InfraRed Imager (SEVIRI) and the polar-orbiting MODerate resolution Imaging Spectroradiometer (MODIS) by adopting linear regression model. However, their study did not consider the spatial correlation of the covariate.

Emili et al., (2011) applied a linear model (including aerosol optical depth (AOD) and meteorological boundary layer height (BLH)) for mapping of PM10 over the alpine regions in 2008–2009. It is indicated that the validation of the satellite maps shows higher accuracy in flat areas than in alpine valleys and elevated sites. Moreover, the inverse distance interpolation of in-situ measurements is able to produce more accurate PM10 maps than satellite maps. Moreover, they indicated that satellite data has limited benefit in the study area due to good spatial coverage of the ground networks and the difficulties inherited to the satellite PM retrieval over rugged topography .They pointed out that AOT retrieval from satellite is not good over rugged topography, since the acquired accuracy represents a serious limitation to the applicability of satellites for ground PM mapping. Moreover, it is concluded that satellite data are of higher interest for regions with a sparser distribution of measurement sites (e.g., distance > 100 km between sites). This study investigates the efficiency of AOT for PM mapping over rugged topography, and the study implies modelling of PM needs to consider the difference over physical regions (elevation ranges).

Various studies address the correlation of AOT, CTM, elevation and PM_{10}; and model air quality using different geostatistical methods. However, this research deviates from the previous works in such a way that it applies model-based approach to explore and model the heteroskedasticity (non-constant variance). Many researchers have addressed the non-stationarity over the mean when they define the mean function, however the heteroskedasticity (non-constant variance) for air quality modelling have not been considered. Researchers indicated that spatial process has different forms of non-stationarity (for example, heteroskedasticity and geometric anisotropy). Heteroskedasticity can be well explained and modelled by simple extension of the variance model. For this study, different covariates (e.g. type area and country) were assessed for modelling the heteroskedasticity in optimal way.

3. STUDY AREA AND DATA DESCRIPTION

3.1. Study area

The study area is most of Europe, comprising 16 countries namely Germany, Netherlands, France, Austria, Belgium, Switzerland, Check Republic, Italy, United Kingdom, Spain, Poland, Ireland, Hungary, Portugal, Slovenia and Slovakia.

The availability of different data sources for modelling of air pollution was the reason for selecting the study area. The area has a wide network of monitoring stations and the region occupies many industries which makes the area suitable for study. Furthermore, the LOTOS-EURO model provides air pollution concentration maps for the whole Europe.

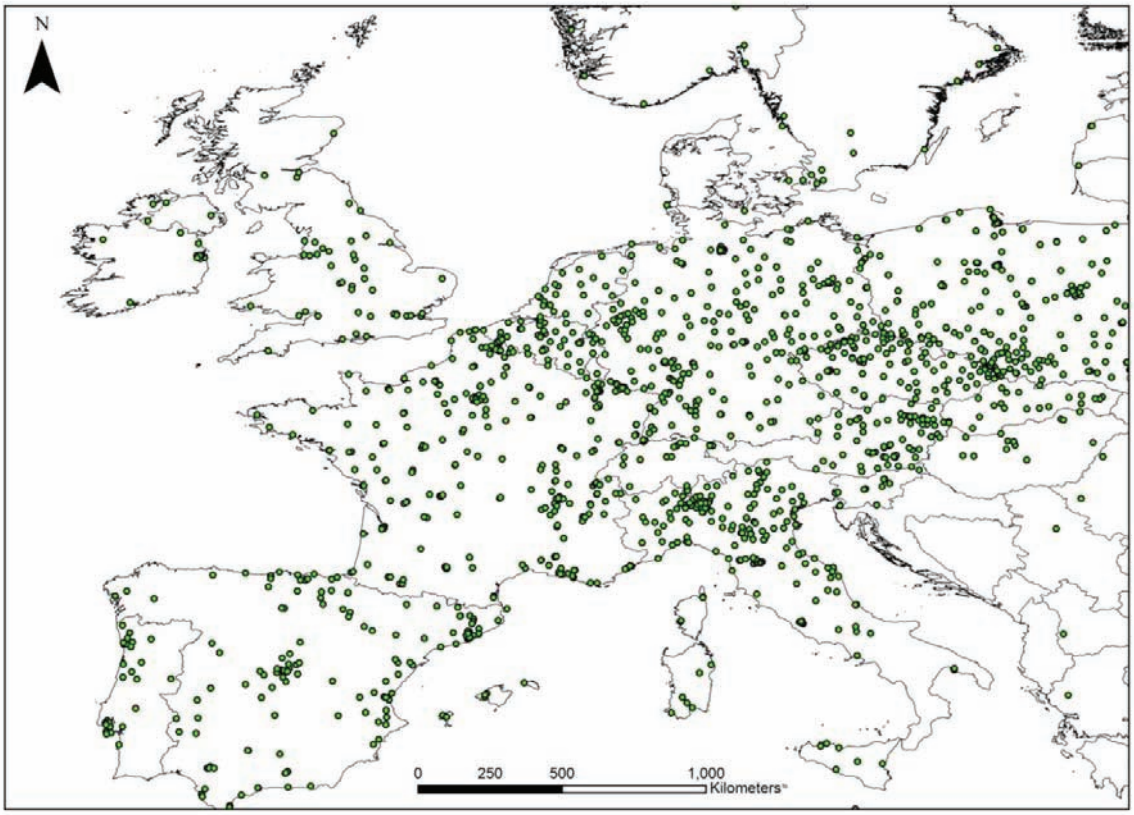


Figure 1: Location map of air quality monitoring stations

3.2. Data description

The data sets used for this study were in situ field measurements, model output, elevation and type areas. Moreover, AOT is the other dataset which was considered for this study. The acquired data has a daily reading of PM_{10} and CTM data over three years (2007, 2008, and 2009). Table 2 and Table 3 gives summary of the datasets used for this study, Figure 1 shows the distribution of monitoring stations over the study area.

Data Type	Year		
	2007	2008	2009
Model output	Full coverage	Full coverage	Full coverage
In situ observations	1177	1346	1396
Elevation	Full coverage	Full coverage	Full coverage
AOT	Full coverage	Full coverage	Full coverage

Table 2: Summary of datasets used

Type area is the background area at which the air pollution monitoring stations is found. The data used for this study has urban, rural, suburban and unknown type areas.

Type (background) area	No. of readings
Rural	285
Suburban	354
Unknown	5
Urban	703

Table 3: Summary of background areas (year 2008 in situ measurements)

The yearly average data for year 2008 data set contains readings from different background areas Table 3 . The monitoring stations with reading for yearly average data has an elevation range of -4m to 2258m, from this we infer that there is large elevation range variation. The minimum value for in situ field measurement is 7.57ppm and a maximum value of 66.41ppm. Model output has a minimum value of 6.84ppm and a maximum value of 29.27 ppm. Every reading has associated station elevation and type area information, Moreover, the country from where the reading was taken is also specified.

Model output data for PM_{10} is available over the year 2007, 2008 and 2009. This three years data has daily PM_{10} readings at any location of the study area. This dataset is available as raster format and as point locations over any in situ stations location. The model output data which is available in raster format has a resolution of 0.50 by 0.250. For model output values at any in situ stations location, there are 4299 stations, however out of 4299 stations ; 90 stations have missing values and 84 stations has duplicated locations. This makes the total number of model output data at any in situ stations locations 4125. The number of stations with reading is the same for all years moreover; there are 4125 stations readings for each day of three years.

The in situ dataset consists of daily readings for three years namely 2007, 2008 and 2009. In situ field measurements show a day to day variation of readings. Due to the day to day variation of stations readings; the number of readings for different temporal aggregations varies, for this study stations with more than 75% data coverage per temporal aggregation were considered.

The AOT data was downloaded from NASA website, by specifying the area over which the data was needed, first the yearly average data (for year 2008) was downloaded and exported to ArcGIS, and then a spatial join based on spatial location was done for AOT and in situ data that, the AOT information is available at any in situ stations location.

4. METHODOLOGY

This chapter presents the methodology followed to achieve the research objectives and the software implementations. The first few sections discuss the method followed for data-pre-processing, data exploration and data projection. Section 4.4 discusses the method how the temporal aggregations are organized and how the analysis was conducted over each temporal aggregation. Sections 4.5 up to 4.10 provide information on the method adopted to extend the LMM, section 4.11 is about how the fitted models were evaluated. Section 4.12 provides information on how prediction was conducted at validation sites and how the accuracy of the prediction was evaluated. The last section is about the software implementation.

4.1. Data pre-processing

The acquired datasets are in netcdf format. The netCDF files provide all the information needed to interpret the data, since the required information about how many dimensions, the length of each dimension, the units of the quantity, etc. are all available. This format combines1D, 2D and 3D data, so that the first step of exploration the datasets was to get information on the metadata, and this was done using Panoply software. Moreover, manipulating data from netCDF file needs proper understanding of the data, and for that appropriate R code was applied to extract subset of the data which is needed for this study.

The model output dataset consists of station code, station name, station location, station height, time and mass PM_{10} data. These data are combined in the package having array and matrix class. For opening the netcdf format dataset on R software, a package called ncdf4 was used and the data is imported in to R. The different variables in the datasets were extracted and assigned a name. The class, structure and dimension of the variables were checked in order to have understanding on the nature of the data.
After extracting the desired variables, a new dataset was created by combining the longitude, latitude, station height, station type, type area and mass particulate matter variables together. During the process rows with missing values and duplicate coordinates were deleted.

In situ field measurement dataset has; station code, station name, station location, station height, time, air base station type of the area and mass PM_{10} variables information. Same procedure was followed to unpack the in situ dataset. After successfully unpacked the datasets in to R and extracted the required variables, descriptive and quantitative data exploration was done on the datasets.

4.2. Data exploration

Descriptive and quantitative data analyses were used to understand the nature of data on hand. Combinations of summary statistics, histogram and bubble plots were employed during data exploration.

4.3. Data projection

Annon et al. (2001); recommended that ETRS LAEA 1989 is the best projection to be used for statistical mapping in the Europe. The in situ and model output data were projected to European conventional Terrestrial Reference System Lambert Azimuth Equal Area 1989 (ETRS LAEA 1989). First, coordinates reference system (CRS) datum WGS84 was assigned to the dataset, then it is transformed to CRS ("+init=epsg: 3035")) using rgdal library.

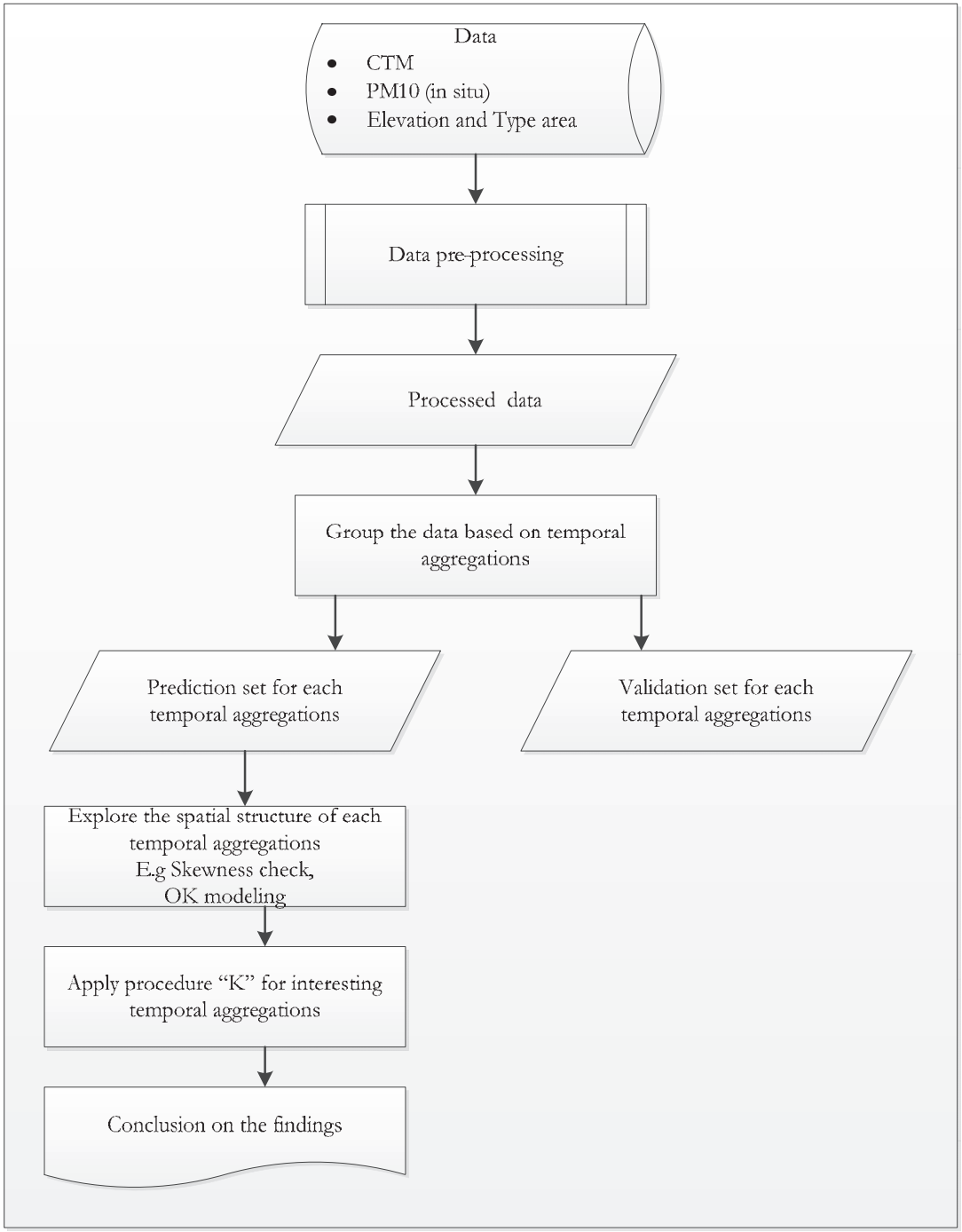


Figure 2: General methodology

4.4. Temporal aggregations

The acquired data is a daily reading of PM_{10} and CTM over three years. The daily base data are organized in different temporal aggregations; thus the mean value at every station locations was computed for each month, for each season and for each year that the data is available on monthly, seasonal and yearly basis. In this way, different temporal aggregations data were generated for year 2008, that all the temporal aggregations data used for this study fit in year 2008.

The data organized seasonal basis based on the fact that; there are four seasons in Europe; summer (June, July and August), autumn (September, October and November), winter (December, January and February) and spring (March, April and May).

For exploration purpose, a graph was prepared to explore the variability of PM_{10} concentration for various temporal aggregations over different locations. To understand the pattern of PM_{10} concentration for different temporal aggregations, a variogram models using ordinary kriging were developed for the above listed temporal aggregations and restricted maximum likelihood (REML) was used for the parameter estimation. Ordinary kriging (OK) is a spatial estimation method, which is most commonly used type of kriging. OK assumes constant but unknown mean and used to interpolate values of a random field at unobserved location from observation at nearby locations (Denby, et al., 2008).

Moreover, to understand the PM_{10} distribution over spatial extent, maps were produced for each temporal aggregation.

4.4.1. Daily pattern analysis

In order to have insight on the daily pattern of particulate matter concentration, plots were produced for 5 consecutive days of summer and 5 consecutive days of winter.

4.5. Linear Mixed-Effects Modelling

LMM was developed by considering in situ field measurements as dependent (response) variable and CTM and elevation data as independent (explanatory) variable of fixed effects, the assumption is based on the linear relationship between dependent and explanatory variables. As it is indicated by different researchers (Hamm et al., in review; Lark., 2009; Pinheiro. & Bates., 2000) , LMM can be used to account for non-stationarity and the spatial autocorrelation of spatially varying random process.

The linear mixed model is written as:

$$Y(s) = X(s)\beta + \eta(s) + \varepsilon \quad (1)$$

Where:

X is $n \times k$ matrix with 1s in the first column and the predictor in the other columns.β is fixed effect coefficients (contains intercept and slope), η is a vector of n spatially correlated random effects describing the spatially correlated environmental variations, ε is a vector of independent random errors, s is location, k is number of covariates and n is number of observation.

The random terms (spatially correlated random variable and independent random errors) are independent of each other, and are assumed to be jointly normally distributed:

$$\begin{bmatrix}\eta\\ \varepsilon\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\nu\sigma^2\boldsymbol{\Lambda R\Lambda} & 0\\ 0 & (1-\nu)\sigma^2\boldsymbol{\Lambda I\Lambda}\end{bmatrix}\right) \quad (2)$$

Where:
σ^2 is the overall variance of ε ,v gives the fraction of the total variance σ^2 attributed to the correlated variance T_s^2, T_n^2 is the variance of ε . $(1-\mathrm{v})$ is the fraction attributed to uncorrelated variance T_n^2 , Λ is diagonal $\{(\lambda_1 \dots \dots \dots \lambda_n)\}$ and the λ s adjust the variance for each measurement to account for heteroskedasticity, R refers$\mathrm{n} \times \mathrm{n}$ matrix describing the correlation between any two data points and I is the $\mathrm{n} \times \mathrm{n}$ identity matrix .

Between any two points, the correlation is modeled as a function of their geographic separation, therefore

$$\mathrm{R}_{\mathrm{ij}} = \mathrm{f}(\mathrm{h}_{\mathrm{ij}}, \emptyset) \quad (3)$$

Where:f is the correlation function, h_{ij} is the distance between s_{i} and s_{j} and $\emptyset$ is the range.

4.6. The hypothesis of stationarity

Most geostatistical analysis presumes some form of stationarity of the variable under study, however different types of stationarity exist and often spatial data show different form of non-stationarity (Meul & van Meirvenne, 2003).

When all the moments of random field distribution are invariant under translation, this is termed strict stationarity (Webster & Oliver, 2008); however these conditions cannot be verified and probably are not met . Under second order (weak) stationarity it is assumed that constant covariance function that depends on separation not on absolute location. This form of stationarity is often considered in geostatistical analysis. Second order (weak) stationarity can be applied when assuming stationarity of the first two moments of the increments of the random fields, termed intrinsic stationarity. Mean may vary across a region and variance continues to increase as the region is expanded, this is the case for intrinsic stationarity which assumes stationarity of the increment.

The above mentioned designations of stationarity lead to different definitions of nonstationarity. Nonstationarity conditions are present for example:

- when the mean value changes with location;
- when the covariance (semi variance) does not depend on separation distance, but changes with direction or location (geometric anisotropy);
- when there is heteroskedasticity (non-constant variance) of the random process over a region.

For this study, LMM which models the within group heteroskedasticity and spatial correlation was developed. Under second order stationarity, the spatially correlated variance can be modeled using a variance function. Moreover, the heteroskedasticity can be modeled using the variance function model (Lark., 2009; Pinheiro. & Bates., 2000).
Heteroskedasticity is one form of nonstationarity because the variance is not constant across a region.

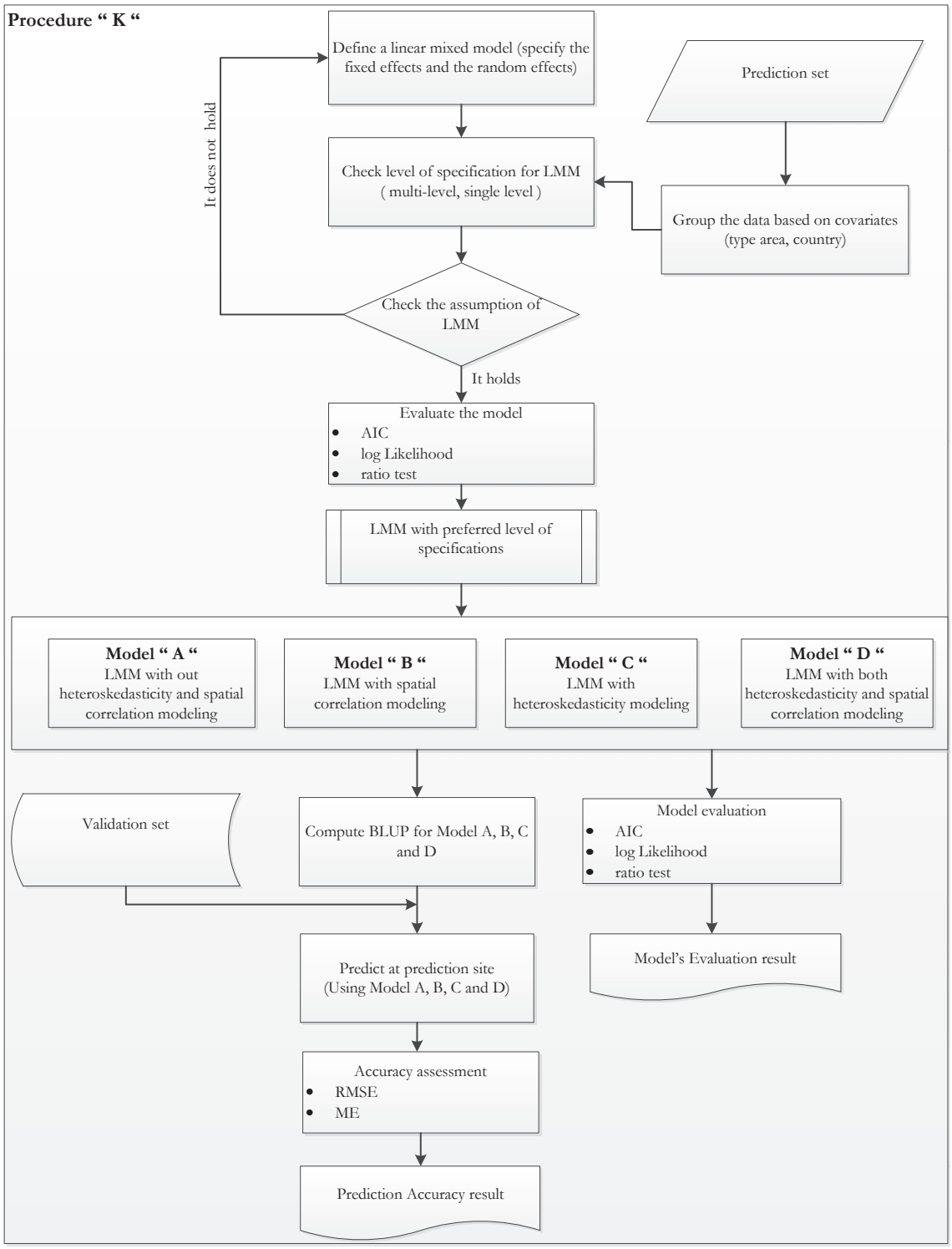


Figure 3: Methodology for LMM

4.7. GLS and LME functions for LMM

Two methods namely LME (linear mixed effect) and GLS (generalized least square) functions were used to develop LMM. Both functions were executed in the nlme package. Model specification for variance function model and correlation model is the same for the two functions. Both functions can use the same classes for example varIdent (a function generally used to allow different variances according to the levels of a classification factor) or corExp (to specify exponential correlation structure) for variance and correlation models respectively. The difference between gls and lme is, the latter needs specifications of random terms as a list of objects per grouping level (Pinheiro. & Bates., 2000). However, for gls we do not specify the random terms in the model specification. Models from the two functions can be compared trivially, using anova test.

The first step followed for the specification of the LMM was, defining a non-constant mean function by parameterizing the mean function in terms of regressor variable (model output). Then estimate of the variance parameters was carried out using the likelihood function for a subset of data (prediction set). Once, the estimates of fixed effects coefficient and prediction of random effects is acquired; LMM and it is associated BLUP were computed.

Model evaluation was carried out by; computing likelihood ratio test, comparing the log-likelihoods and AIC values.

4.8. LMM specification using lme function

The LMM model was developed based on the following assumptions:

- Normality for the within-group errors

The quantities used to check this assumption are the within-group residuals, which are the difference between the observed response and the within-group fitted value.
The normality assumption for the within-group errors was assessed with the normal probability plot of the residuals.

- Assessing Assumptions on the Random Effects

Pinero & Bates., (2000) stated that the estimated BLUPs of the random effects are the primary quantities for assessing the distributional assumptions about the random effects.

For this study, the random effect method was used to extract the estimated BLUPs of the random effects from `lme` objects.

Normal probability plots of the estimated random effects were examined at each level of grouping when assessing the adequacy of a multilevel model fit.

4.9. Fitting a linear mixed model

A linear mixed model of air quality data was fitted using OLS, for which the residuals error are assumed to be independent and identically normally distributed (iid). For this model the non-stationarity and the spatial correlation were not considered and the model was named, Model A. Again the same model was fitted using generalized least square (GLS), in this case the spatial correlation was considered but the non-stationarity in the variance was not considered, Model B. Moreover, the linear mixed model was fitted by

considering the heteroskedasticity but not the spatial correlation, Model C. Finally, the model was extended by modelling both heteroskedasticity and the spatial correlation, Model D.

a) LMM without heteroskedasticity and spatial correlation modelling (Model A)

The model developing process involved elevation and CTM (Chemical Transport Model) for fixed effect specification and country and type area for grouping structure specification. The approach for model developing is step by step that the improvement of the model was investigated with the specification of each covariate. For example, adding the elevation for fixed effect specification improved the model.
An `lme` function was used to fit linear mixed-effects model for PM10 data and restricted maximum likelihood was used for the parameter estimation. Despite the different optional arguments, the general form of the `lme` function is:

```
lme(fixed ,data ,random)
```

The first one specifies the fixed effect and the last one specifies the random effects and the grouping structure in the model. Use of LMM needs a grouped data, Pinheiro. & Bates., (2000) indicated the option to work on grouped-Data object or it is possible to work on the data frame and specify the grouping in the model formula. Mixed-effects models flexibly represent the covariance structure induced by the grouping of the data. So there is a need for specifying the grouping structure.

For this study the general formulation of LMM used for fixed effects and random effects were

`fixed= insitu ~ le_model+ elevation` and
`random = ~ le_model + elevation | coun/type_area` respectively.

(insitu represents the in situ PM10 field measurements , le_model represents the model output data , coun represents country and type_area represents the background areas (for example , urban and rural)) . The model output data and elevation were used both in the fixed effects and random effects specifications. This is because in the LMM every fixed effect has an associated random effect (Pinheiro. & Bates., 2000). Type area and country were used for specification of grouping structure of the data, so that they were used as a grouping factor that divides the observations into the distinct groups of observations. Restricted Maximum Likelihood (REML) was used for parameter estimates since it provide a conservative estimates than Maximum Likelihood (ML).

Elevation was used for fixed effect specification because various studies showed the topographical influence on pollutants dispersion. For example, Carvalho et al., (2006) analyze how mesoscale circulations induced by topography and/or land use control pollutants dispersion in a coastal region, it is also indicated that topography represented as the main driving force mechanism on air pollutants injection in higher tropospheric levels. Kim & Stockwell., (2008) indicted that complex terrain was shown to have an important influence on the vertical transport of air pollutants.

The specification of fixed effect with CTM and elevation gives a better fit to the data than the model which only use CTM for fixed effect specification. For fitted objects with different fixed effects the AIC comparisons are not meaningful so the adjusted R^2 value was checked.

In order to come up on the above indicated general formulation of LMM the below mentioned conditions were assessed.

1. Is there a pattern for PM10 distribution over type areas (background areas)?
   To check if different surrounding areas have different PM10 distribution pattern, the type of the surrounding areas is included in the fixed effect and LMM was fitted to the data, and the p-value was checked.
2. Is there a pattern for PM10 distribution over countries?
   To check if different country have different PM10 distribution pattern, graphical display was prepared, and the distributions were compared.
3. Is the multilevel grouping or single level grouping more substantial for the PM10 data?
   Covariates are used for specification of multilevel and single level model and the significance of using multilevel grouping over single level grouping was checked using the model summary from the multi-level LMM and single level LMM output specifically AIC and log likelihood values.

In general, the step wise approach was adopted to develop the LMM.

Equation form expression of the multilevel model for PM10 concentration y_{ij} in the j^{th} type area with in i^{th} country is expressed, for $i = 1, \dots 16 \; and \;\; j = 1,2,3$.

$$y_{ij} = X_{ij}\,\beta + Z_{i.j}b_i \; + Z_{ij}b_{ij} + \varepsilon_{ij} \tag{4}$$

$$i = \; 1, \dots\dots, M \qquad j = 1, \dots\dots k$$

$$b_i \sim N\,(0, \psi_1) \qquad b_{ij} \sim N\,(0, \psi_2) \qquad \varepsilon_i \sim N\,(0, \sigma^2 I)$$

Where b_i is level - 1 (country level) random effects (q_1-vectors), b_{ij} is level - 2 (type area in country level) random effects (q_2 -vectors), $\mathrm{Z_{i,j}}$ is level-1 (country level) regressor matrix $(\mathrm{n_{ij}} \times \mathrm{q1})$, $\mathrm{Z_{ij}}$ is level - 2 regressor matrix $(\mathrm{n_{ij}} \times \mathrm{q2})$, ψ_1 is covariance matrix, ψ_2 is covariance matrix , $\varepsilon_{\mathrm{ij}}$ is within group error, β is fixed effects, M is the number of first-level groups (which is 16 since there are 16 countries) and K is the number of second-level groups (which is 3since there are 3 type areas).

For this study, the fixed effect regressor matrices are CTM and elevation covariates and the fixed effects; β_0 is the intercept, and β_1 and β_2 are the slope for CTM and elevation. $Z_{i,j}$ is regressor matrices for level one specifications which are the model output values, station heights, country and the spatial covariates namely easting and northing (x and y). $Z_{ij}$is regressor matrices for level two specifications which are the model output values, station heights, type area in country and the spatial covariates namely easting and northing (x and y). $\mathrm{b_i}$and $\mathrm{b_{ij}}$are the random effects or parameters to be estimated for level one and level two respectively. $\mathrm{b_{ij}}$ refers to random effects nested within the $\mathrm{b_i}$ random effects. q and p are random and fixed effect matrices .

b) Correlation Structures specifications for LMM (Model B)

Mixed-effects models are used to analyse grouped data, since they flexibly model the within-group correlation often present in this type of data (Pinheiro. & Bates., 2000).

A plot of `ols` residuals versus their spatial coordinates was produced to ensure the need for modelling spatial dependence. Correlation structures are used to model dependence among observations (spatial dependence) in general; it is used to model dependence among the within-group errors. Correlation structures are specified in `lme` through the correlation argument. Different standard classes of correlation structures are available in the `nlme` library. Among these, the most commonly used exponential spatial correlation (`corExp`) and spherical spatial correlation (corSpher) were used for this study.

The correlation structure was specified accordingly

```
correlation = corExp (Value=c(80,0.3),form = ~ x + y |
coun/type_area, nugget= TRUE)
```

A two-dimensional position vector with coordinates x (longitude) and y (latitude) was specified with form, these are the spatial covariates. Moreover, the grouping factor which is a nested type (type area in country) was specified. Exponential spatial correlation structure was used and the spatial correlation structure was computed based on the Euclidean distance between x and y.

There are different class of correlation structure specifications in `nlme`, for this study exponential correlation structure was used because the `ols` variogram shows exponential nature.

c) Modelling Heteroskedasticity (Model C)

To account for heteroskedasticity, the LMM was expanded to incorporate the parameters described in Equation 2. Variance functions was used to model the variance structure of the within group errors using covariates. The `lme` function allows the modelling of heteroskedasticity of the within-group error through a weights argument. Weights may be modelled as a function of covariates or assigned in an ad hoc fashion (Hamm, et al., in review). There are different available standard classes of variance function structures. Among the different standard classes, the `varIdent` variance function structure allows different variances for each level of a factor and was used to fit the heterokedastic model for the PM_{10} data.

Apascaritei et al., (2009) pointed out the difference in pollutant concentration between rural and urban areas. The different PM_{10} distribution pattern over different type areas was assessed and type area was used to specify weight for modelling the heteroskedasticity. REML was used for parameter estimation since it gives unbiased estimate than the ML estimates. The specification of the joint distribution of random terms (Equation 2) is the key for estimating the variance parameters.

The variance function model was specified accordingly

```
weights = varIdent (form= ~ 1 | type_area)
```

The LMM which models the heteroskedasticity has a general form

```
wreml.model<- update(ols.model, weights=varIdent (form= ~ 1 |
type_area))
```

The need for a heterokedastic model for the PM_{10} data was formally assessed with `anova` test.

d) Modelling both heteroskedasticity and spatial auto correlation (Model D)

The LMM model was extended to model both the spatial autocorrelation and heteroskedasticity, and it has a general form

```
wsreml.model<-update (ols.model, weights=varIdent (form= ~ 1
┤|type_area), correlation = corExp (value=c (80, 0.3), form = ~ x +
y | coun/type_area, nugget= TRUE))
```

The specified model was applied on the PM_{10} data and model evaluation was conducted to assess the significance of modelling heteroskedasticity and spatial correlation.

4.10. LMM model specification using gls function

The `gls` function is the other function used to fit the extended linear mixed model, using restricted maximum likelihood for variance parameter estimation. LMM fit using `gls` can be viewed as an `lme` function without the argument random (Pinheiro. & Bates., 2000).

The correlation structure and variance model specification is the same as `lme` function, so that the same procedure as `lme` function was followed to extend the LMM using `gls` function. The same four conditions namely LMM without heteroskedasticity and spatial correlation modelling (model A), modelling spatial correlation (model B), modelling heteroskedasticity (model C), and modelling both spatial correlation and heteroskedasticity (model D) were conducted for `gls` function.

4.11. Examining a Fitted Model

To evaluate the improvement offered by differently specified LMM models (e.g., modelling heteroskedasticity), the following model evaluation techniques were conducted;

- assessment of the distribution of the residuals using diagnostic plots;
- computing likelihood ratio test and ;
- comparing the log-likelihoods and Akaike Information Criterion (AIC) values.

In order to address whether the coefficients of the model are different from zero, hypothesis testing was done. Lark, (2009) stated that, the log likelihoods and AIC for different models can be compared if the fixed effect do not change.
The AIC is used to assess the improvement by more complex model and it is computed by:

$$AIC = 2(n_{par}) - 2logLik \quad (5)$$

Where:
n_{par} is the number of parameters and $logLik$ is the log- likelihood value.

The log-Likelihood ratio test was used to compare the models , since nested models can be compared by likelihood ratio test (Lark., 2009).

4.12. Prediction at unsampled location

For the spatial application of LMM, the prediction at a given location would be the sum of $X\beta$ and the interpolated random effect (Lark, 2006). Where the distance between the prediction location and the nearest measurements of y is larger than the variogram range the predicted value is $\mathrm{X}\widehat{\beta}$ (Hamm, et al., in review).

The LMM's of summer temporal aggregation data were used to do prediction at sampled locations. The validation data set consisted of 326 sites, and prediction at validation points was done using the BLUPs of the LMM developed using `lme` function. Estimated BLUPs of the individual coefficients are obtained and, because of the multiple grouping levels, a level argument is used to specify the desired grouping level. The coefficient estimates were assessed at each level and the predict method was used to obtain predictions at the validation points.

In order to achieve this new data frame was created with the required information for example; station location, elevation and model output values. Then predict function was specified;

`predict (model, new data frame, level)` in order to get prediction values at validation sites.

For multi-level grouping structure specifications, the parameter estimation is for each level of grouping, so that the prediction can be conducted at each level of specification. Specifying the level from which the estimated parameter will be plugged-in in to the BLUP gives prediction values at the specified level. For example, if level is specified to 1 it means use the variance parameter estimate for country level. In order to get the prediction values at the second level (type area in country), the level was specified by 2. The prediction values were acquired at the three levels of specifications namely at population level (the entire prediction set as one group), at level one and at level two.

Finally, accuracy assessment which is the process of assessing the result was carried out. Root Mean Square Error (RMSE) Equation 6 and Mean Error (ME) Equation 7, were the two component of accuracy assessment used for this study. Finally, result comparison was carried out for the stationary and non-stationary models.

$$RMSE = \left[\frac{1}{N}\sum_{i=1}^{N}(y^*(s_i) - y(s_i))^2\right]^{0.5} \quad (6)$$

and

$$ME = \frac{1}{N}\sum_{i=1}^{N} y^*(s_i) - y(s_i) \quad (7)$$

Where:
$y^*(s_i)$ is the estimated value at location s_i, $y(s_i)$ is the observed value at location s_i, and N is the number of prediction points

Moreover, prediction was done at summer and winter seasons' validation sites using the yearly average model, and the accuracy was compared with summer model prediction at summer validation sites and winter model prediction at winter validation sites.

4.13. Software implementation

Data exploration and analysis was implemented on the following software's:

Software		Purpose of Implementation for this study
ArcGIS		To visualize the data with country boundary and to produce location map of stations
		To display the AOT data and make a spatial join with CTM and PM_{10} data
		To display the netCDF raster file
		To project the datasets to the same coordinate system
Panoply		To see the meta data information for netCDF format and to understand the nature of the data
R software	Packages	
	nlme	For developing LMM model
	ncdf4	For unpacking the netCDF file
	maptools	Used under data processing step
	geoR	To convert an object to the class geodata and to check duplicated coordinates
	rgdal	For projection transformation
	lattice	Used under data processing step
	proj4	For defining projection of different data layer
	raster	To read the raster data in to R
	ctv	Used under data processing step
	maps	Used under data processing step
	reshape	Used under data unpacking process (to change data class)
	Stringer	Used under data unpacking process
	gstat	To do the OK variogram analysis and change an object to sp class
Beam software		To visualize the raster map and understand the nature of netCDF format data

Table 4: Summary of software's implementation for this study

5. RESULTS

5.1. Data exploration results

Using quantitative and descriptive statistics PM_{10} data of in situ field measurements over three years were explored and it was observed that the annual average data were skewed. So that the data was log transformed Figure 4.

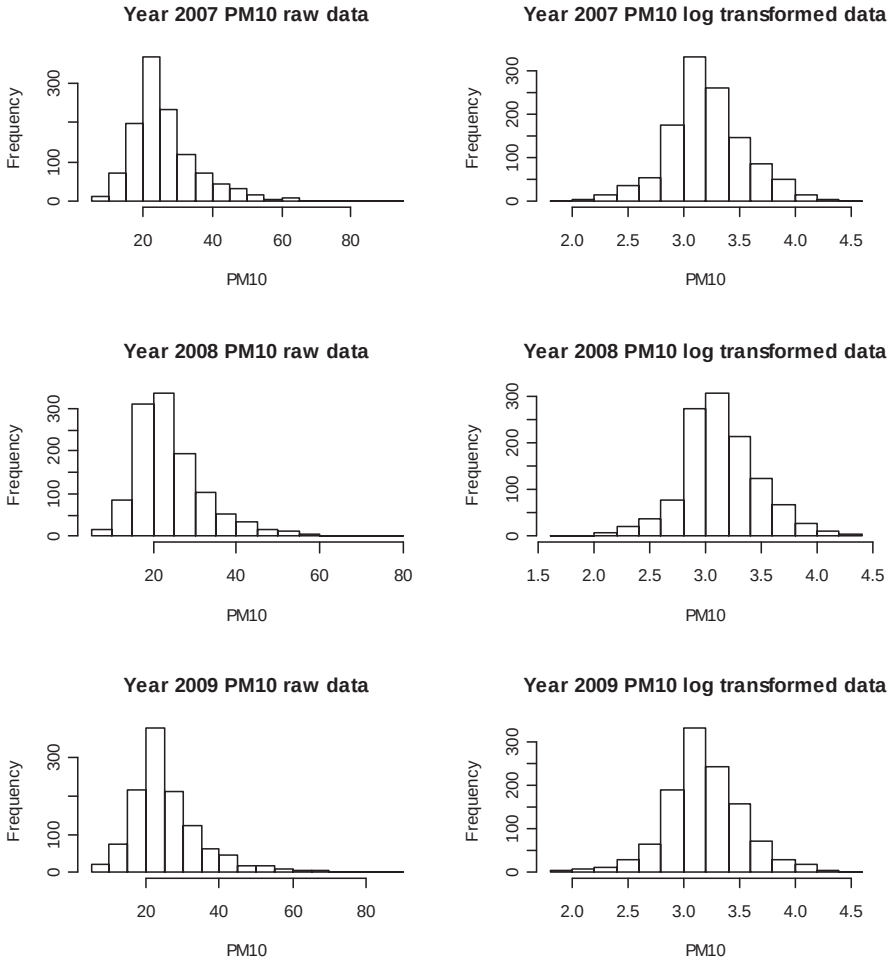


Figure 4: Histogram plot for year 2007, 2008 and 2009 in situ PM_{10}

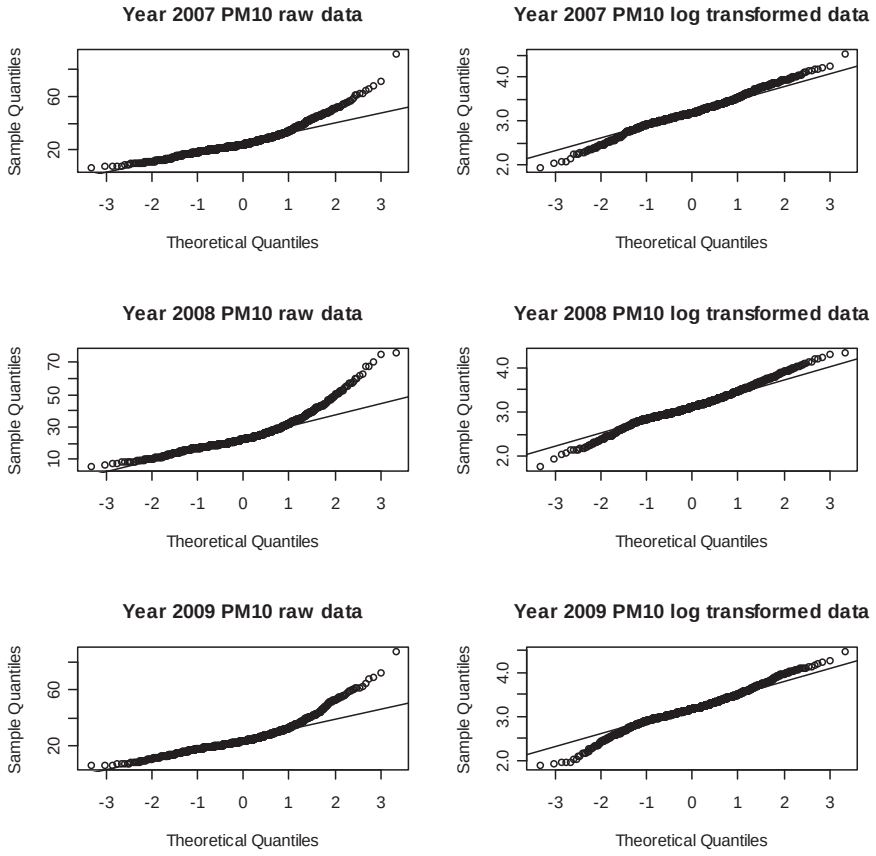


Figure 5 : QQ plot for three years data

Year	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2007	6.98	20.26	24.12	26.22	30.09	90.82
2008	7.57	18.73	22.61	24.44	28.05	75.26
2009	6.59	19.87	23.78	25.7	29.7	87.05

Table 5: Summary of year 2007, 2008 and 2009 PM_{10} data

Referring Figure 4 and

Figure 5, the three years data are log normally distributed. The data exploration results for each temporal aggregation are presented on the Appendix list.

5.2. Temporal aggregations results

Table 6 shows summary of PM_{10} insitu measurement per background area over different temporal aggregations. The daily reading has a minimum value of 2ppm for urban area and maximum value of 249 ppm.

Temporal Aggregations	Urban			Suburban			Rural		
	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
Year 2008	11.69	26.46	66.4	10.75	23.58	51.95	7.57	19.03	48.28
Summer	8.62	20.73	52.83	0.85	19.85	49.55	5.09	16.91	48.02
Winter	10.8	48.28	150.5	7.12	32.5	72.58	4.12	23.87	84.87
Day 211	2	30.9	249.9	5.79	32.71	138	0.5	20.82	144
January	8.42	33.95	176.25	4.27	32.62	218.76	2.67	21.18	68.17

Table 6: In situ PM_{10} (ppm) concentration summary per background area for the temporal aggregations

As it is inferred from Figure 6 and Figure 7, the concentration of PM_{10} in the air varies significantly over different temporal aggregations.

Data aggregation means we are averaging over a given time span and this smoothen the data and the result might be significantly different and gives different inference Figure 8.
The graphs show PM_{10} concentration variation at every station location over different temporal aggregations. The x–axis represents station locations, every number on the x-axis has an associated easting and northing coordinates. For all temporal aggregations, the same sequential order of stations was maintained that it is possible to see the PM_{10} variation over temporal aggregations. Especially, it is easy to infer PM_{10} concentration variation over temporal aggregations at every station location.

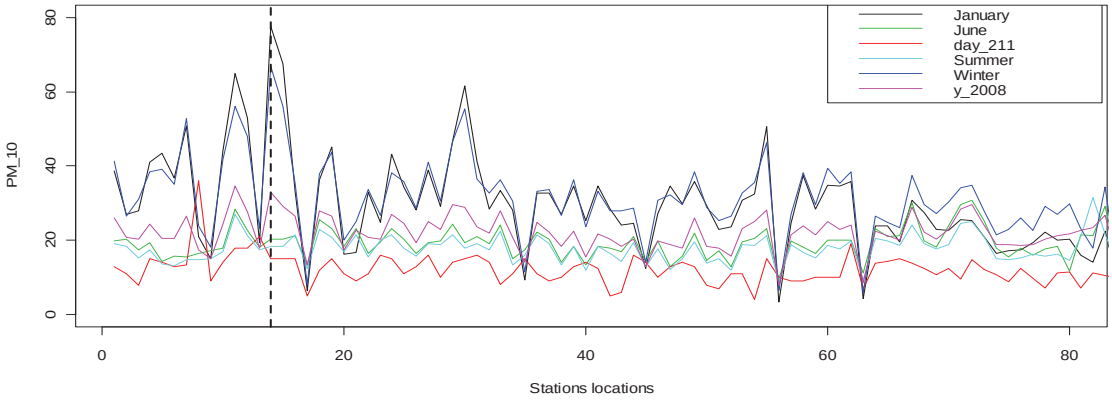


Figure 6: The PM_{10} (ppm) concentration over different temporal aggregations

For example, the vertical broken line indicates a location at 15.43 longitude and 47.04 latitude, the variability at this point is significant , even if the variability exists at all locations.

At some locations the variability is significant that there is a variation of more than 150 ppm, over different temporal aggregations. This is clearly observed at the broken vertical line, which represent a location at certain longitude and latitude.

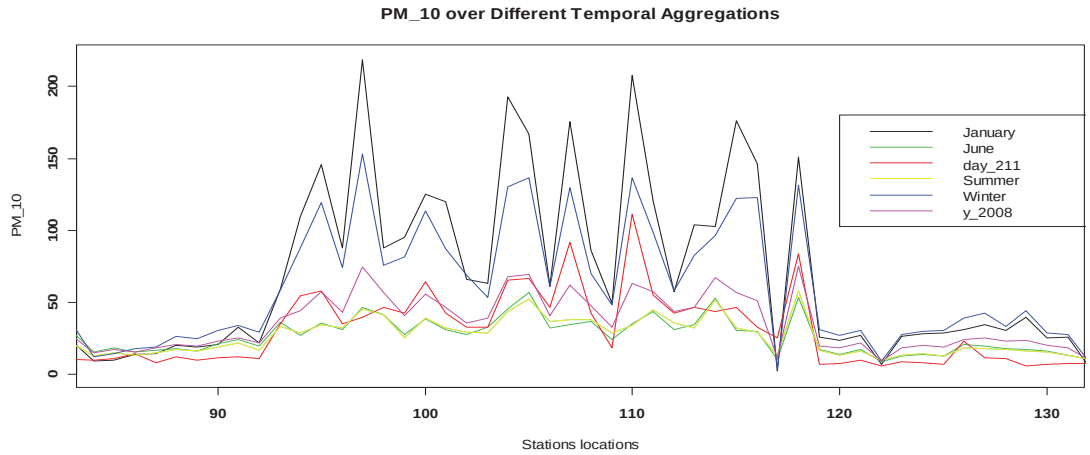


Figure 7 : PM_{10} (ppm) concentration with high variability over different temporal aggregations

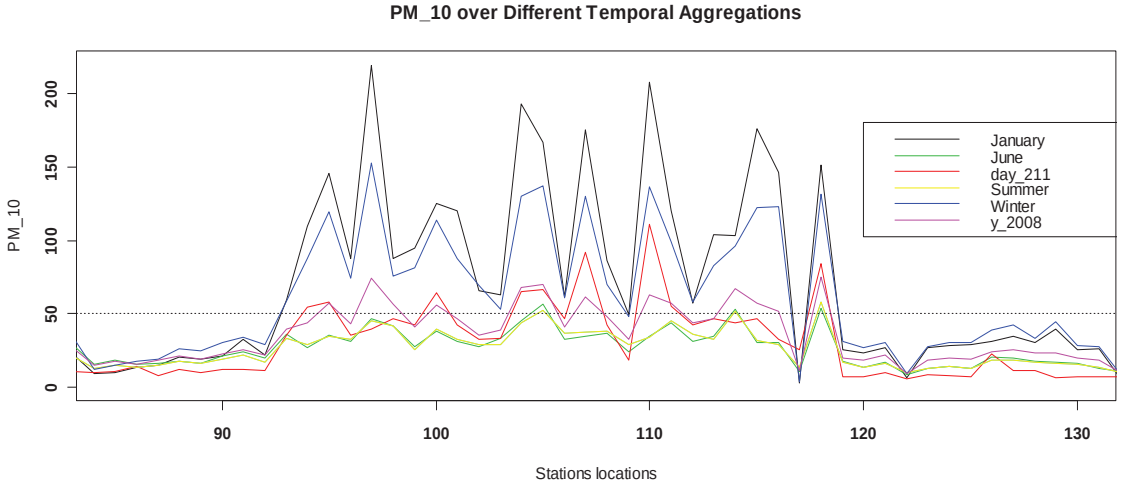


Figure 8: PM_{10} over different temporal aggregations having a horizontal line which indicates the limit value for PM concentration in air (which is 40 ppm for yearly average data and 50 ppm for daily average data)

This high temporal variability should be considered especially when the values exceed the threshold values set by Environmental Protection Authorities. As it is inferred from Figure 8.the value exceeds for some temporal aggregations and become lower for other aggregations.

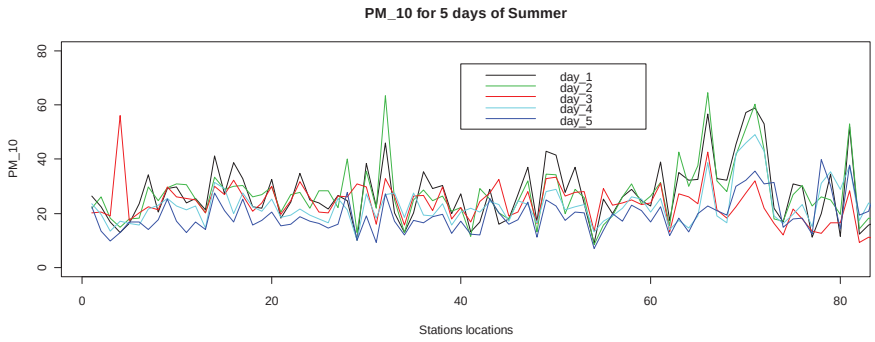


Figure 9: Summer PM10 concentration trend over 5 consecutive days

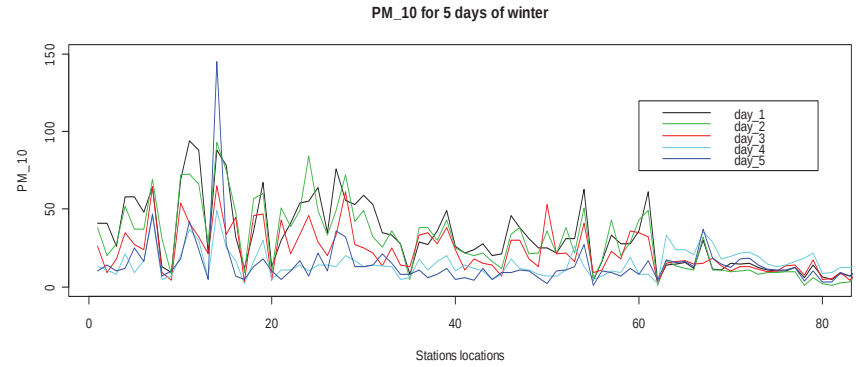


Figure 10: Winter PM10 concentration trend over 5 consecutive days

Five consecutive days were taken from summer and 5 days from winter and a graph is produced in order to see the day to day variation of PM10 concentration over winter and summer Figure 9 and Figure 10, it is inferred that there is a day to day variation in both cases however, the concentration value variation is much higher for winter than for summer.

The three years data was combined and there were 1174 stations with readings. Furthermore, the pattern over the three years was examined. The number of stations is more than 1000; however the first 100 stations were presented on the graph Figure 11 in order to show the trend clearly.

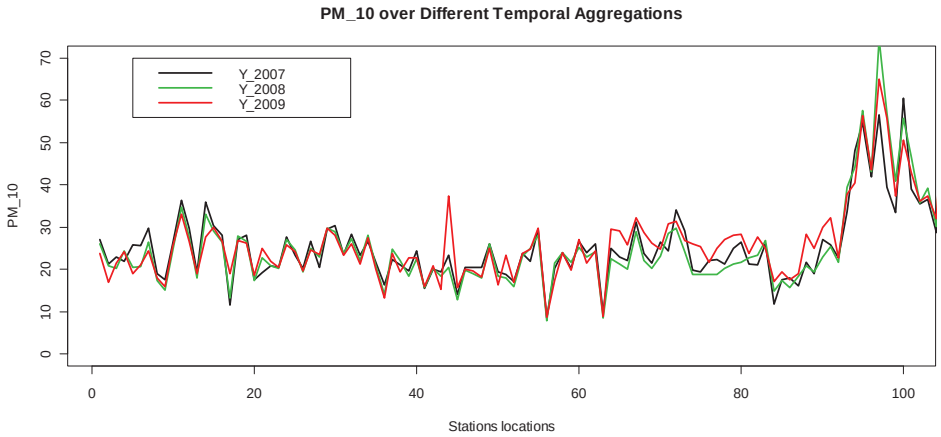


Figure 11: PM10 (ppm) pattern for three years data

Yearly average values over the three years do not show different pattern Figure 11, so that rather than considering the time serious analysis, it is reasonable to consider and analyse the data based on temporal aggregations.

As it is inferred from Figure 8 PM_{10} levels at many stations exceeded EU air quality daily mean standards, 50 ppm. The yearly average values over many station locations also show exceedances of EU air quality limit 40 ppm for yearly average data.

Moreover, Figure 13 shows the variation of PM_{10} concentration over different temporal aggregations. The legend for each temporal aggregation map represents the same value that, maps can be easily compared.
The concentration of PM_{10} over most parts of the study area becomes higher during, winter, January and for daily data (for the selected specific day).The concentration over most areas becomes relatively lower during summer, June and year 2008.

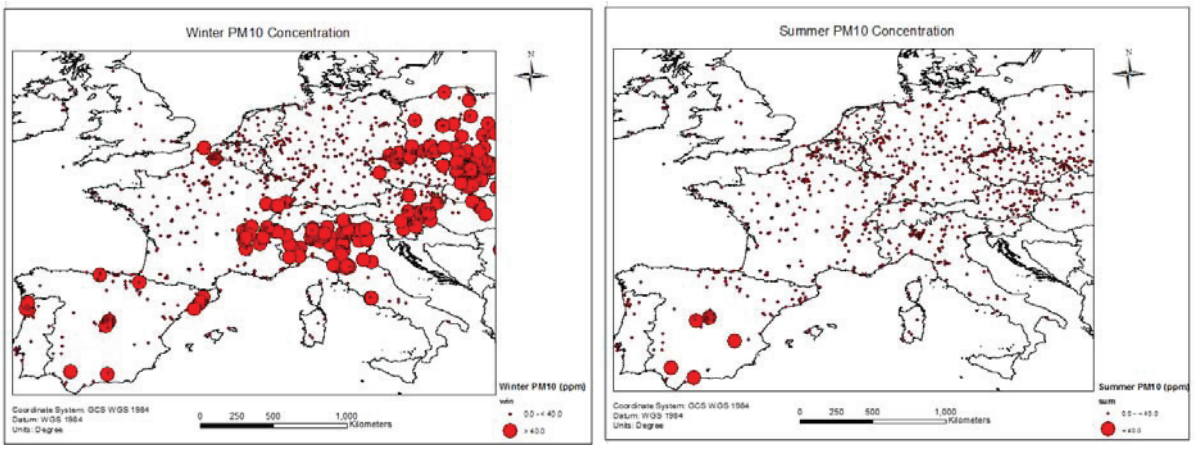


Figure 12: PM_{10} concentration variation a)Winter b) Summer (it shows values above and below the standard limit)

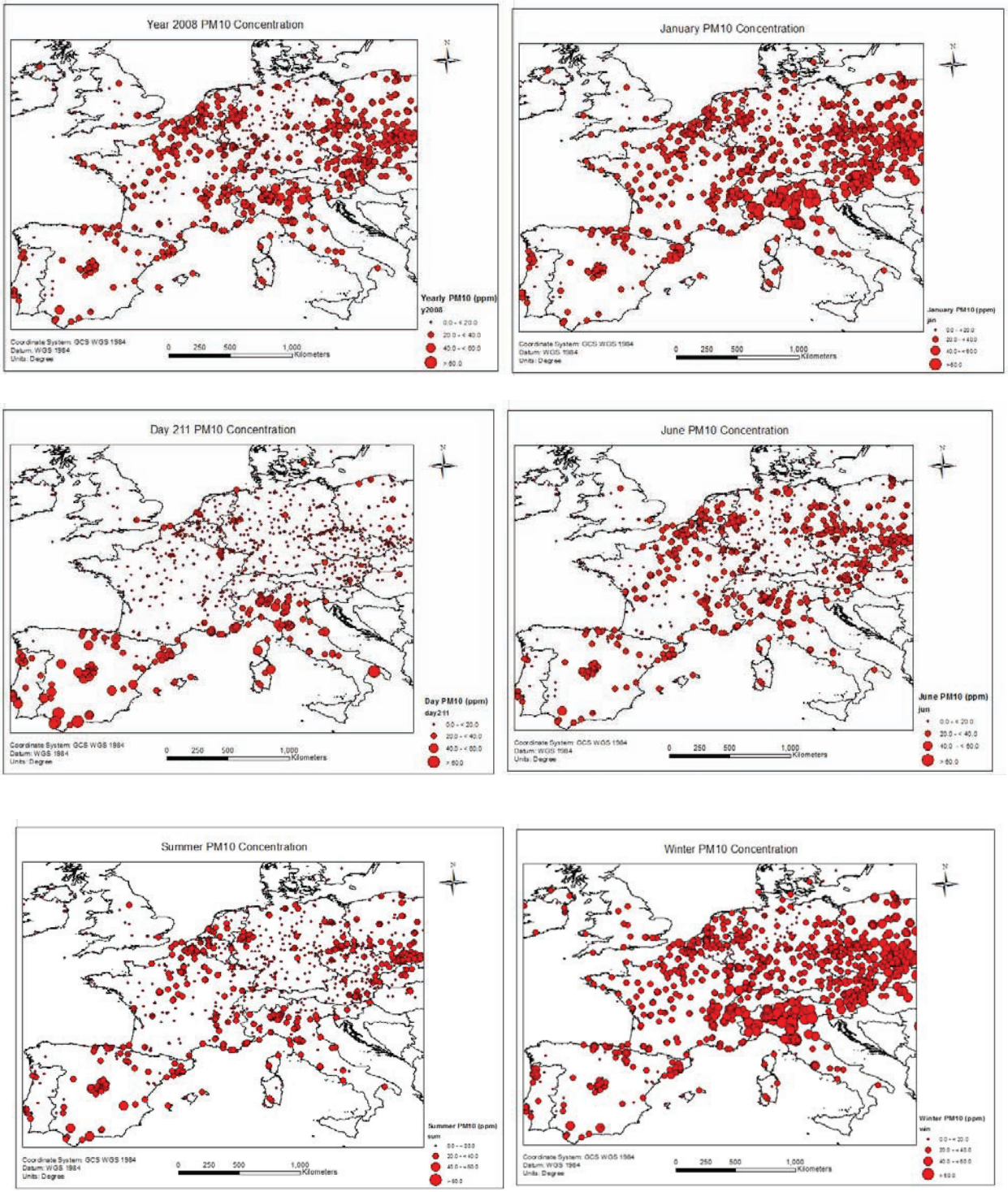


Figure 13: PM_{10} concentration variation over temporal aggregations

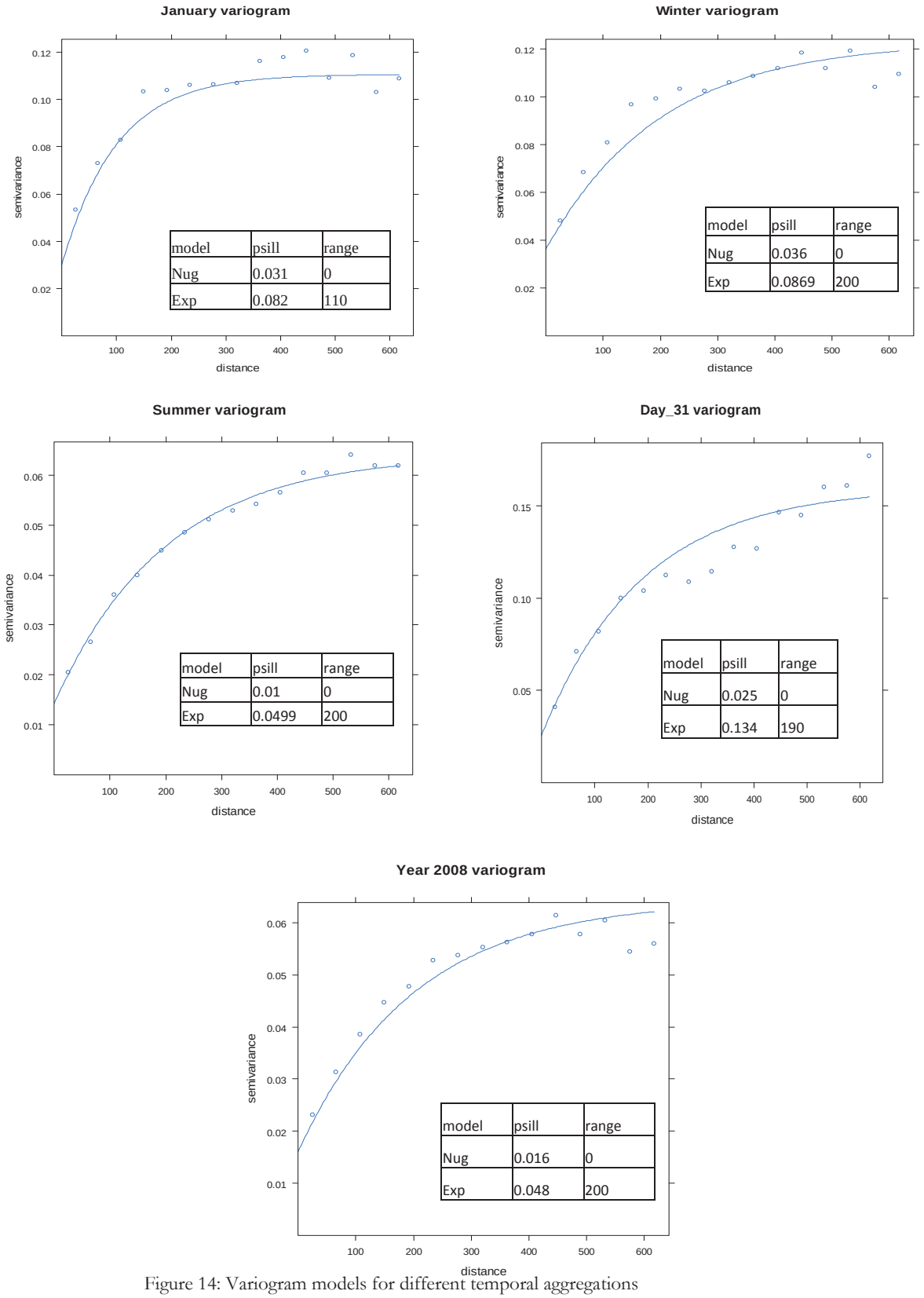


Figure 14: Variogram models for different temporal aggregations

The general spatial structures vary over different temporal aggregation, the variograms model Figure 14 differs over the sill and range.

5.3. LMM results

To assess if different surrounding areas have different PM_{10} distribution pattern, the type of the surrounding areas is included in the fixed effect and LMM was fitted to the data, from the resulting summary output, it is inferred that small p-values associate with all type areas except unknown surrounding areas (Pinheiro. & Bates., 2000). The resulting small p-values (Table 7 except for unknown areas) show different PM_{10} patterns over different surrounding areas.

	P-value
Intercept	0.0152
suburban	0.0000
unknown	0.7934
urban	0.0000
CTM	0.0000
suburban: CTM	0.0000
unknown: CTM	0.8240
urban: CTM	0.0000

Table 7: p_value for PM_{10} distribution trend check over type areas

The number of stations data with unknown surrounding area is 5 (out of the total 1348 monitoring stations). So the stations with unknown surrounding areas are excluded and lower p-values were acquired for the other type areas.

The acquired correlation value between in situ field measurement and AOT is very low (4%) that, AOT variable is excluded from the model since it did not give improvement to the model. Moreover, the significance of AOT and other terms (CTM and elevation) was assessed; and the acquired p-value for AOT is large (0.22) Table 8 that, it is excluded from the model . This is further supported by low value of adjusted R^2 value for linear model of in situ and AOT data.

Coefficients	p-value
Intercept	0
CTM	0
elevation	0
AOT	0.22

Table 8: p-value for coefficients significance check

Isotropy – means that spatial dependence of residuals are the same in any directions , if this is not the case you add more covariate and model the anisotropy.

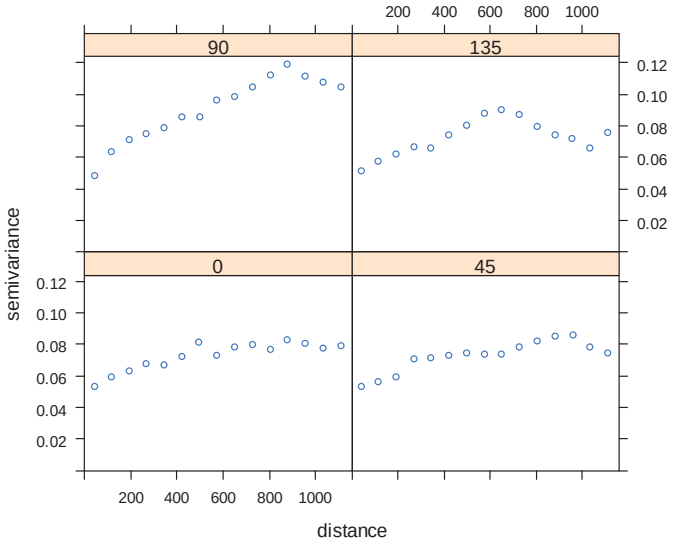


Figure 15: Variogram model for different direction

The spatial pattern over different directions varies especially for 90^0; the spatial structure of the variogram varies. If we have horizontal points of a variogram then we can assume there is no spatial dependence.

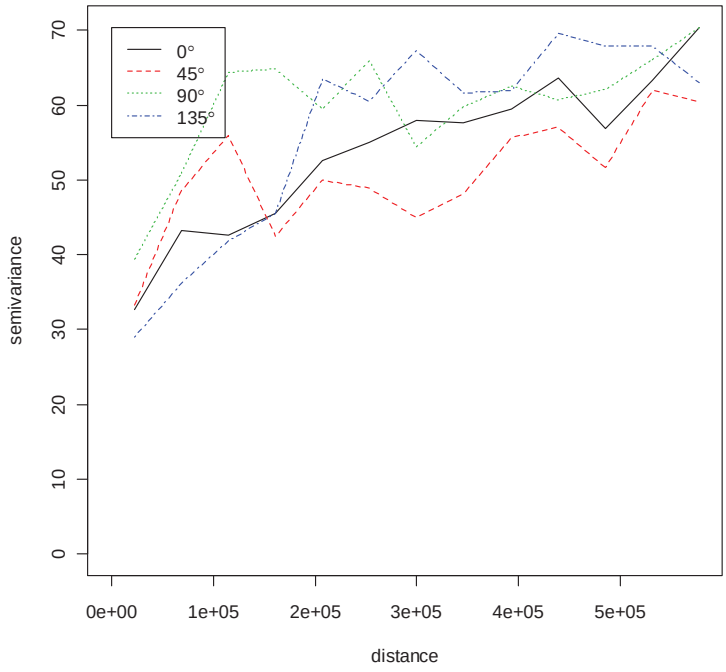


Figure 16: Variograms for four different directions (directional variogram)

Figure 15 and Figure 16 show how the trend varies for each location.

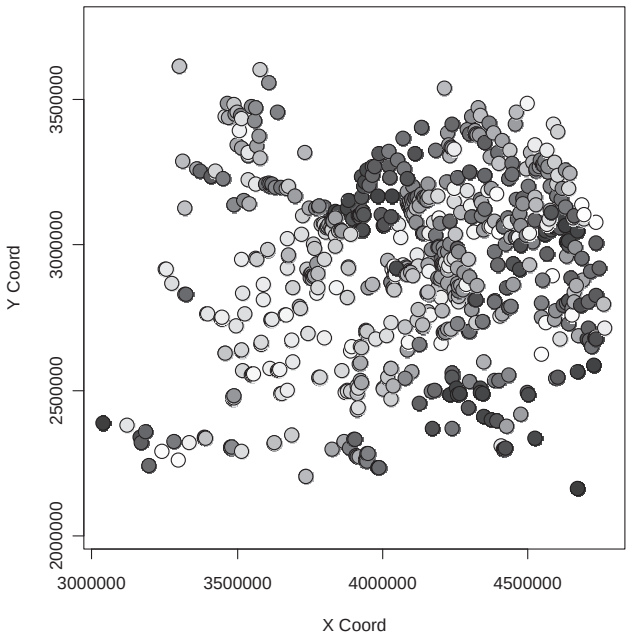


Figure 17: Concentration of in situ measurements and their distribution, filled circles with grey shades proportional to the measured in situ values

As it is inferred from scatterplot Figure 17 higher values of in situ measurements are found to northern, north-eastern and south-eastern parts of Europe, this suggests that a trend surface term in the model might be appropriate. The exploratory analysis suggested a model with a non-constant mean.

We use a linear trend surface to describe a spatially varying mean, our exploratory analysis suggested a model with a non-constant mean, non-stationarity in the mean can be modelled using covariates, and for this study elevation and CTM were used as a covariate to explain the spatially varying mean.

a) LMM without heteroskedasticity and spatial correlation (Model A)

For the specification of the LMM model using lme function, multi-level and single level specifications were compared and the result is presented on Table 9.

Grouping level	AIC	logLik	No. of par.
Multilevel (type area in country)	-177.9	104.95	16
Single level (country)	-36.76	28.38	10

Table 9: Summary of LMM output for multi and single level groping specification of random effects

Moreover, likelihood ratio test and p-value were calculated for single level grouping specification versus multi-level specifications. The null hypothesis is the single level grouping specification gives a better fit for the data than multi-level specifications.

$$H_0 = no\ difference$$

$$H_a = two\ level\ specifications\ gives\ a\ better\ fit$$

Model	L.Ratio	p-value
Multi-level vs single level	153.14	<0.0001

Table 10: Likelihood ratio test for single level versus multilevel (type area in country) grouping specification

The very high value of the likelihood ratio test statistic and low p-value confirms that the significance of the multilevel specification in the model. In general, from Table 9 and Table 10 it is observed that for the PM_{10} data multilevel specification describe the data more than single level random effect specification. So for this study, two levels of nesting was used (type area in country). Then the estimated random effects at both grouping levels were assessed.

b) Modeling spatial correlation (Model B)

For the PM_{10} data used for this study, the need for modelling the spatial correlation was assessed. Figure 18 show residuals obtained by gls (ols) plotted versus their spatial coordinates. Black dots represent negative residuals and grey dots positive residuals. The size of the dots is proportional to the value of residuals, the graph show spatial pattern (e.g positive residuals and negative residuals show some clustering), which indicates spatial correlations or dependence between observations.

Figure 18: A plot for OLS residuals versus their spatial coordinates (Black dots represent negative residuals and grey dots positive residuals.

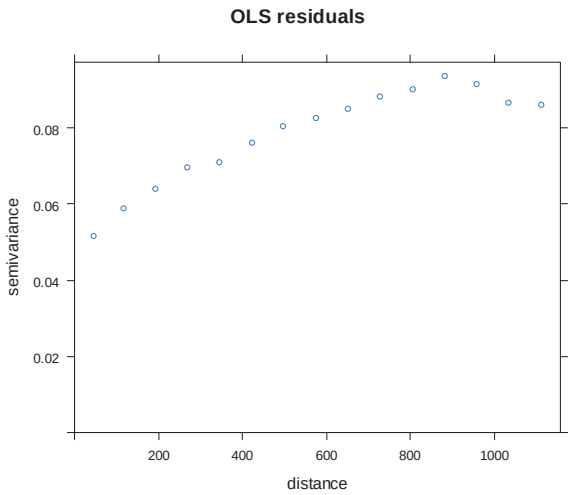


Figure 19: Variogram of OLS residuals

Both the variogram and the bubble plots show the spatial dependence in the residuals. This is one way of assessing the occurrence of spatial dependence between observations.
LMM was extended to model the spatial correlation, as it is inferred from Table 11 there is a significant increase in the log-restricted-likelihood, as evidenced by the large value for the likelihood ratio test, indicating adding the correlation structure improved the model or it gives a better fit to the data.

Model	AIC	logLik	No. of par.
Model A	-177.9	104.95	16
Model B	-391.2	213.6	18

Table 11: Summary of model output for Model A and Model B

Moreover, likelihood ratio test and p-value were assessed for correlation structure specifications in LMM, the result shows rejection of the null hypothesis (LMM without the spatial correlation specifications) in favour of the alternative hypothesis (model with spatial correlation specifications). This also supports the idea of spatial structure specifications in the LMM.

Model	L.Ratio	p-value
Model A vs Model B	217.3	<0.0001

Table 12: Likelihood ratio test for Model A versus Model B

c) Modeling heteroskedasticity (Model C)

The within group residuals, which are the difference between the observed response and the within-group fitted value are used to check, whether there is a non-constant variance (heteroskedasticity) with in group. As it is inferred from

Figure 20 the residuals are centered at zero, but the variability changes with group.

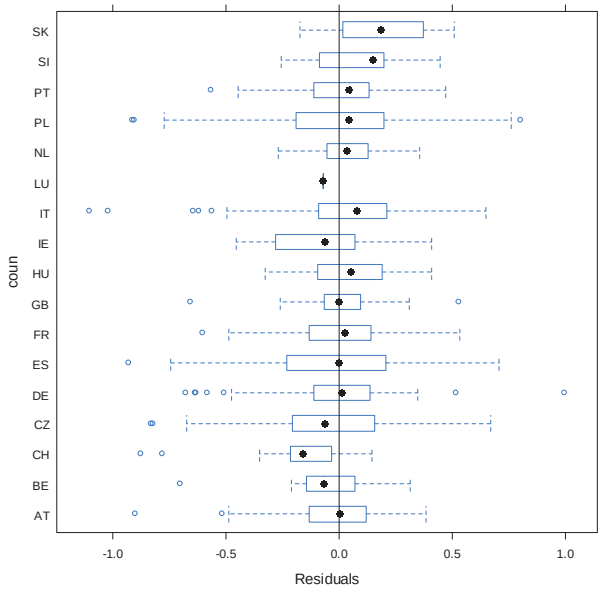


Figure 20: Boxplots of residuals per group (country) for LMM fit (with no heteroskedasticity modelling)

There is an outlying observation and large residuals for few countries. A pattern suggested by the individual boxplots is that there is more variability with some countries than the others. Plot for the standardized residuals versus fitted values was used to look at the pattern by type areas.

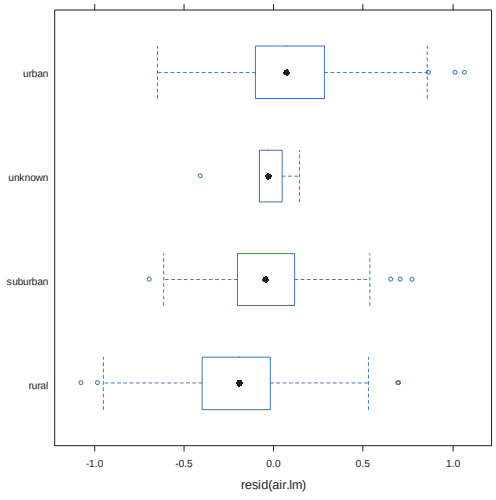


Figure 21: Residual plots corresponding to LMM (type areas, as grouping factor)

The type of the surrounding areas is included in the fixed effect and LMM was fitted to the data, from the resulting summary output, it is inferred that small p-values are associated with all type areas except unknown surrounding areas. The resulting small p-values show different PM_{10} distribution patterns over different surrounding areas.

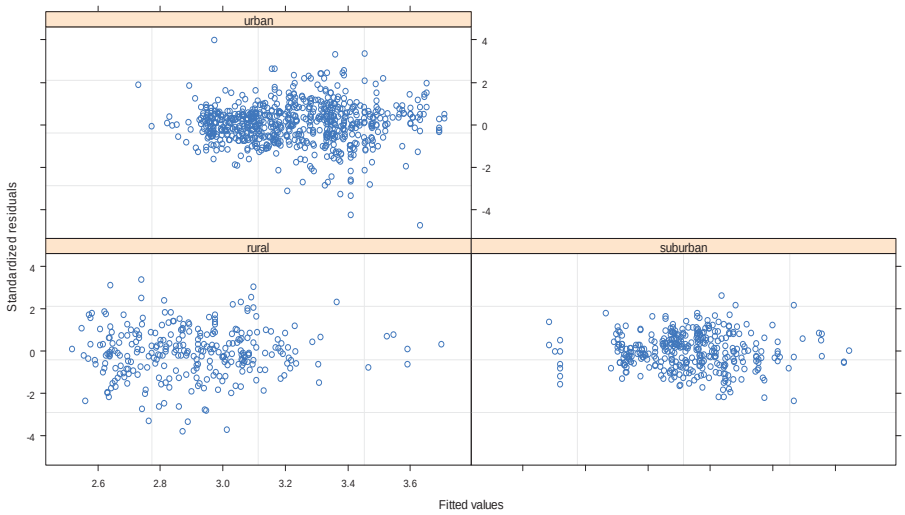


Figure 22: Scatter plots of standardized residuals versus fitted values for LMM fit by type areas

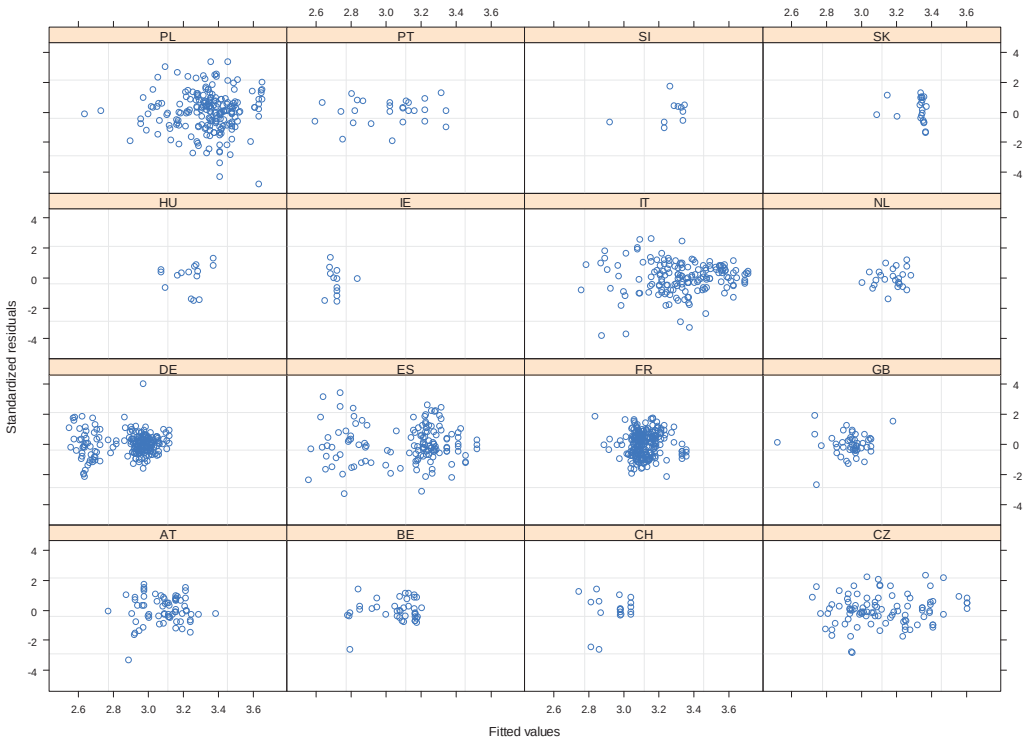


Figure 23: Scatter plots of standardized residuals versus fitted values for LMM fit by country

PM_{10} data allows different variances per country and per type areas for within-group errors. So it seems plausible to model the non-constant variance per type areas. In order to extend the LMM for modelling heteroskedasticity, a variance function structure was used to allow different variances for each level of a factor (type area).

Referring the above results there is non-constant variance across type (background) areas. So that the LMM was extended to model the heteroskedasticity, as it is inferred from Table 13 the heterokedastic

model has a lower AIC value than the homoskedastic model and it has a higher log likelihood value than the homoskedastic model. Furthermore, the low p-value of the likelihood ratio test and higher likelihood ratio test values suggests that modelling the heteroskedasticity for linear mixed-effects model provides a much better description of the data than the homoskedastic case.

Model	AIC	logLik	No. of par.
Model A	-177.9	104.95	16
Model C	-189.39	112.7	18

Table 13: Summary of model output for heterokedastic (Model C) and homoskedastic (Model A) models

Model	L.Ratio	p-value
Heteroskedastic vs homoskedastic model	15.49	<0.0004

Table 14: Likelihood ratio test for heterokedastic versus homoskedastic model

The small value of the standardized residuals was better seen by looking at a plot of the observed responses versus the within-group fitted values. The LMM fitted values are in close agreement with the observed insitu measurement, except for few extreme observations.

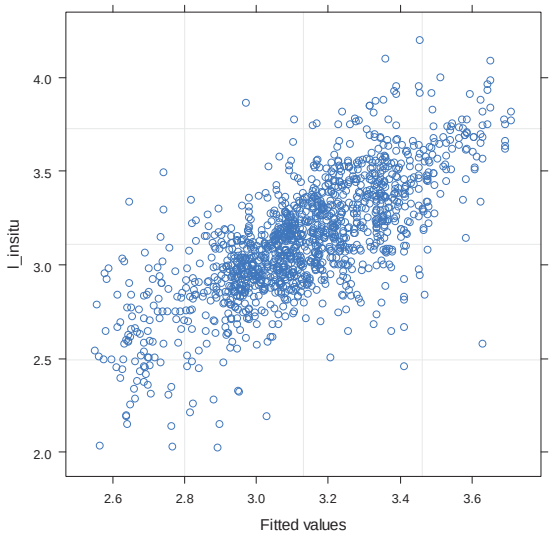


Figure 24: Observed versus fitted values plot for model C

The assumption of normality for the within-group errors was assessed with the normal probability plot of the residuals. Analysis of histograms and normal plots of the standardized residuals showed that residuals are normally distributed. Moreover, normal probability plots of the estimated random effects must be examined at each level of grouping when assessing the adequacy of a multilevel model fit.

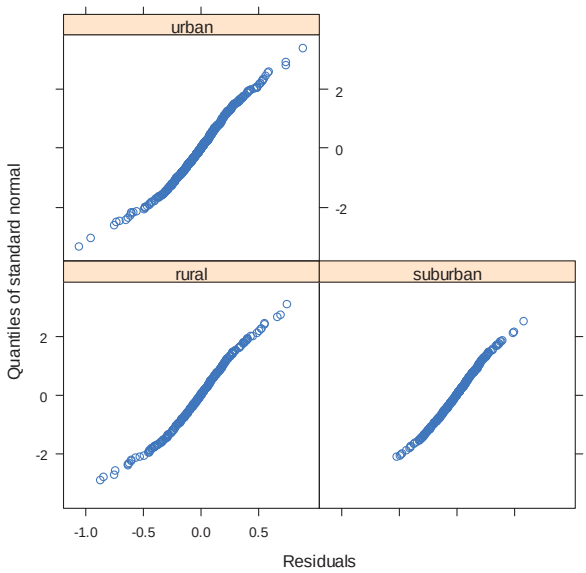


Figure 25: Normal plot of residuals for the Model C fit per type area

As it inferred from

Figure 25 and Figure 26 normality assumption seems plausible.

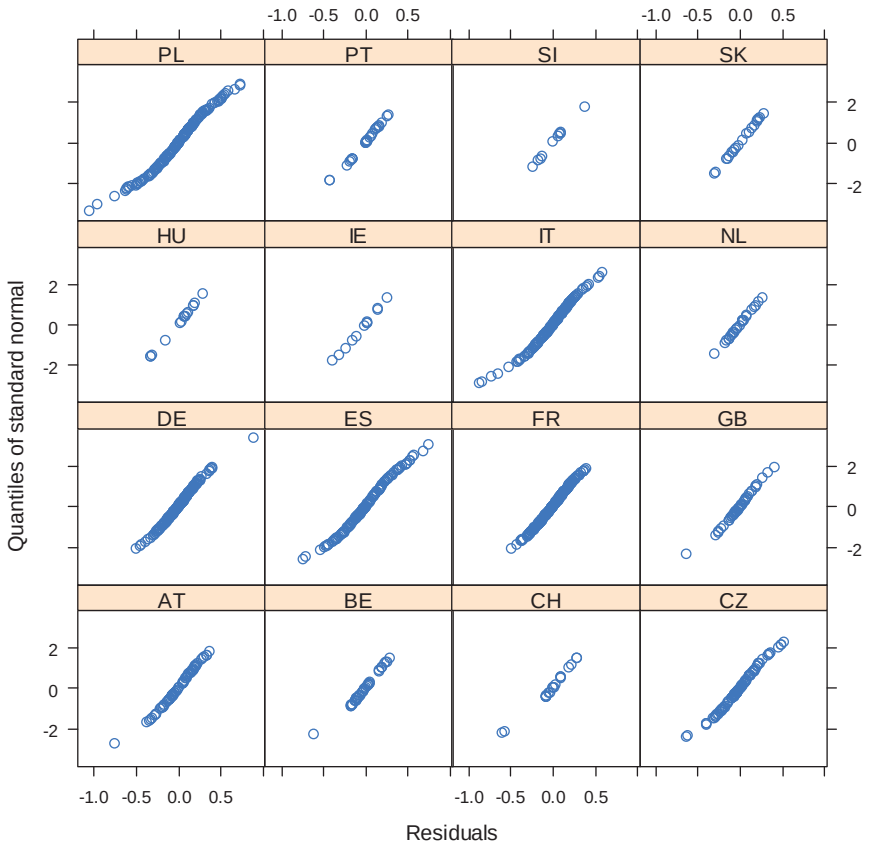


Figure 26: Normal plot of residuals for the Model C fit per country

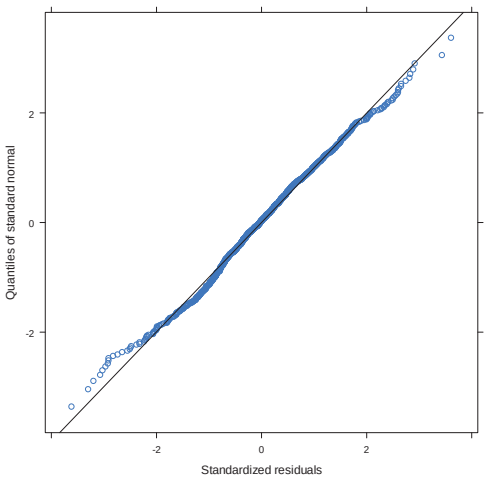


Figure 27: Normal plot of residuals for the Model C fit

d) Modeling both heteroskedasticity and spatial correlation (Model D)

The LMM as further extended to model both the heteroskedasticity and the spatial correlation. From the above result we have seen that modelling heteroskedasticity and modelling spatial correlation improves the LMM. Here both terms were modelled and the model was compared with the other three LMM specifications. As it is inferred from Table 15 the higher log-likelihood value, lower AIC value, higher likelihood ratio test, the LMM which considers both the heteroskedasticity and spatial correlation structure is an improved model over the other.

Model	AIC	logLik	No. of par.
Model A	-177.9	104.95	16
Model D	-410.04	225.02	20

Table 15: Summary of Model output for Model A and Model D

Table 16 summarizes the output of the Models; as it is inferred from the table, model D which takes in to account both the heteroskedasticity and the spatial correlation has a lower AIC value and high log Likelihood value than the other three models. This indicates the need for considering Heteroskedasticity and spatial correlation for air quality modelling.

Model	AIC	logLik	No. of par.
Model A	-177.9	104.95	16
Model B	-391.20	213.60	18
Model C	-189.39	112.7	18
Model D	-410.04	225.02	20

Table 16: Summary of model output for Model A, B, C and D

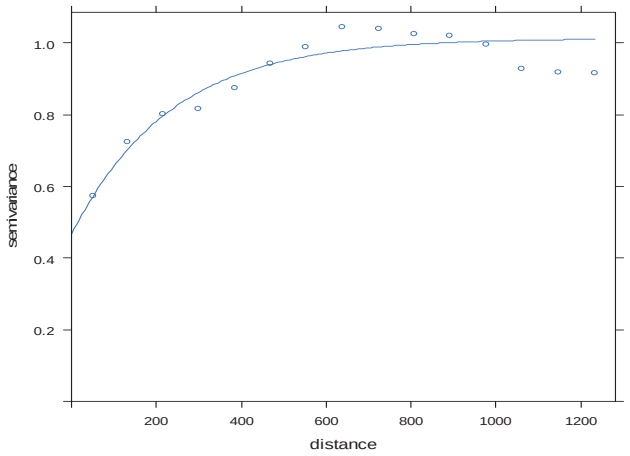


Figure 28: A variogram of Model D LMM

The LMM is fitted directly to the data, rather than fitting a model to the sample variogram. Hence the exact match between the sample variogram and modelled variogram should not be expected (Hamm, et al., in review).

5.4. LMM model specification using GLS

With the same variance function model and correlation structure specifications a LMM model was developed using gls function. The output of the four models was presented on Table 17.

gls model	AIC	logLik	No. of par.
Model A	-130.6	69.3	4
Model B	-184.4	98.2	6
Model C	-152.2	82.1	6
Model D	-211.5	113.8	8

Table 17: Summary of LMM model outputs using `gls` function

Even though, `gls` and `lme` functions could weakly compared using AIC, the model output values for LMM models developed using `lme` function and `gls` function, the modelled developed using `lme` functions have lower AIC values and a higher log-likelihood values so in general for the data set used for this study, the LMM model fit using `lme` function gives a better fit. Pinheiro. & Bates., (2000) indicated that for comparison of `gls` and `lme` there are other factor to consider. For example they indicate that if the data has a multi-level grouping then `lme` is recommended. We cannot do likelihood ratio test since they are not a nested models.

5.5. LMM for summer and winter temporal aggregations

The same procedure was adopted and LMM was developed for winter and summer temporal aggregations.

Grouping level	AIC	logLik
Multilevel (type_area in country)	83.6	-25.8
Single level (country)	121.5	-50.7

Table 18: Multi-level and single level grouping structure for summer data

Model	L.Ratio	p-value
Mulitlevel vs single level	49.84	<0.0001

Table 19: Likelihood ratio test for single level versus multilevel grouping specification for summer data

Model	AIC	logLik
Model A	83.57	-25.79
Model B	69.16	-17.6
Model C	77.4	-20.7
Model D	60.8	-11.4

Table 20: Summary of the LMM model outputs for summer data (using lme function)

As it is observed for yearly average data the non-stationary model (Model D) has a lower AIC and higher log- likelihood value than stationary model.
For visualization purpose, a variogram model was developed by computing sample semivariogram estimates corresponding to the standardized residuals of Model D for summer Figure 29 and winter data

Figure 30.

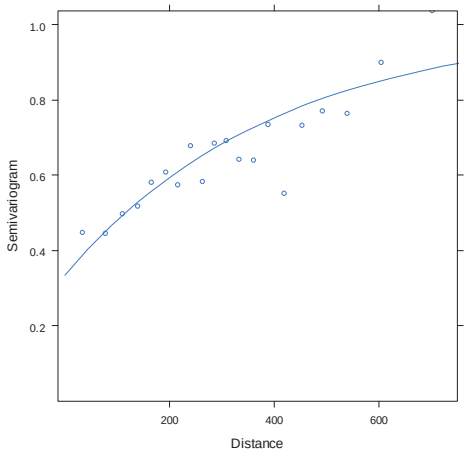


Figure 29: Variogram model for Summer season (Model D)

Model	AIC	logLik	No. of par.
Model A	430.3	-211.15	4
Model B	329.52	-158.76	6
Model C	385.2	-186.6	6
Model D	305.6	-144.8	8

Table 21: Summary of the model outputs for winter data

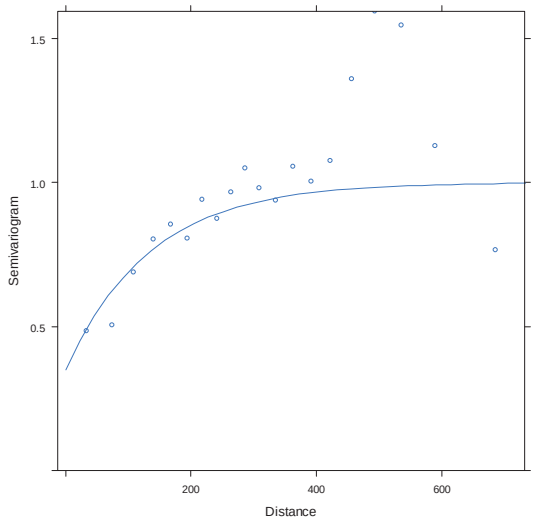


Figure 30: Variogram model for winter season (Model D)

5.6. Prediction at unsampled location

The LMM's of summer temporal aggregation data were used to do prediction at sampled locations. Moreover, RMSE (which tells how far on average the observed values are from the true values) and ME (which tells us whether a set of measurement underestimate or overestimate the true value) were computed using observed values and predicted values at validation points. The acquired values are presented in Table 22 and Table 23.

Model	RMSE		
	Level_0	Level_1	Level_2
Model A	0.3014	0.2729	0.2672
Model C	0.2805	0.2533	0.2476
Model B	0.2720	0.2431	0.2374
Model D	0.2510	0.2235	0.2180

Table 22: RMSE summary of summer validation sites prediction using summer LMM developed by lme function

This RMSE values are for log transformed values so when the values are back transformed the difference becomes relatively larger.

Model	ME		
	Level_0	Level_1	Level_2
Model A	-0.0494	-0.0410	-0.0255
Model C	-0.0479	-0.0415	-0.0239
Model B	-0.0423	-0.0402	-0.0237
Model D	-0.0402	-0.0379	-0.0229

Table 23: ME summary of summer validation sites prediction using summer LMM developed by lme function

Temporal Aggregation	Model	RMSE			ME		
		Level_0	Level_1	Level_2	Level_1	Level_2	Level_3
Summer	Model B	0.2679	0.2464	0.2448	0.0356	0.0313	0.0834
	Model D	0.2581	0.2435	0.2416	0.0436	0.0415	0.0873

Table 24: Summer validation sites prediction using yearly average data LMM developed by lme function

Temporal Aggregation	Model	RMSE	ME
Winter	Model B	0.4380	-0.1599
	Model D	0.4111	-0.1349

Table 25: Winter validation sites prediction using winter LMM developed by gls function

Temporal Aggregation	Model	RMSE	ME
Summer	Model B	0.2602	-0.0514
	Model D	0.2521	-0.0459

Table 26: Summer validation sites prediction using summer LMM developed by gls function

Temporal Aggregation	Model	RMSE	ME
Summer	Model B	0.2582	0.0614
	Model D	0.2541	0.0422
Winter	Model B	0.5169	-0.3676
	Model D	0.5021	-0.3485

Table 27: Winter and summer validation sites prediction using LMM (of gls function) developed for yearly average data

Temporal Aggregation	Model	RMSE	ME
Winter	Model B	0.6266	-0.4998
	Model D	0.6130	-0.4952

Table 28: Winter validation sites prediction using summer LMM developed by gls function

Referring the RMSE and ME values for summer and winter model predictions Table 25 and Table 26 the summer season has a lower value of RMSE and a mean value which is closer to zero than winter season

prediction. Moreover, predictions were done at winter and summer validation sites using the yearly average model. The yearly average model gave a better accuracy at summer validation sites prediction than winter validation sites predictions Table 27. Moreover, the winter model was used to predict at summer validation sites and the summer model was used to predict at the winter validation sites, again the accuracy is lower than the predictions values from summer model at summer validation sites and winter model at winter prediction sites. In general, for all cases, it is observed that model D has a better accuracy than the other models.

PM_{10} shows different trend over different temporal aggregations Figure 13. Moreover, the spatial distribution of PM_{10} concentration varies from region to region.

The step wise development of LMM was supported by the explanatory analysis result, and the model which took in to account the heteroskedasticity and the spatial correlation has a lower AIC value and a higher log likelihood value. The Likelihood ratio tests also support this. Moreover, the non-stationary model has higher prediction accuracy than stationary model.

6. DISCUSSION

The aim of this thesis was to investigate the distribution of PM_{10} over different temporal aggregations and develop a LMM for modelling of air quality over most of Europe.

6.1. Temporal aggregations analysis

Referring Figure 5 - 12, PM_{10} distribution over different temporal aggregations varies moreover; the concentration and variability difference between the different temporal aggregations is substantial Figure 13. Following the explanatory analysis, the seasonal temporal aggregations (winter and summer) are found to be interesting to look at because their spatial structure differs (variogram model Figure 14) in addition these seasons represent two weather conditions and weather has influence on concentration and variability of PM_{10} in air (Gomiscek, et al., 2004a). The winter variogram has a higher nugget and semi variance value than the summer variogram. Summer variogram seems stable on the other hand the winter variogram looks unstable, one possible reason for this might be, the high variability of PM_{10} during winter period (minimum value of 4 ppm and maximum value of 150.5 ppm). Summer season has a minimum PM_{10} concentration value of 0.86 ppm and a maximum value of 52.83 ppm.

The correlation of in situ field measurements and CTM data for winter and summer is 35% and 38% respectively, from this it can be inferred that the summer season correlation is higher than winter season correlation; this is due to the fact that correlation depends on local meteorological conditions (Koelemeijer, et al., 2006). Normally, owing to wet deposition, the concentration of PM10 is expected to be lower during winter than summer time. However, this is reversed for the data used for this study, and Koelemeijer, et al., (2006) explain such kind of cases as outcome of lower boundary mixing layer height which results on higher concentration of PM_{10} near the surface during rainy seasons.

Most studies indicated that pollutant load in the air decreases with elevation (Beelen, et al., 2009; Held et al., 2008). This situation is also observed for the dataset used for this research. The correlation of in situ field measurements with elevation is -19% and -33% for summer and winter seasons respectively, this might be due to comparatively stable (over most of the study area) and lower concentration of PM_{10} during summer time and high variability and large value of PM_{10} concentration for winter period.

From Figure 8 it is inferred that the daily PM_{10} levels at many monitoring station areas exceeded EU air quality daily mean standards (50ppm). At some location the maximum value reached up to 200 ppm. The yearly average values over some station locations also shows exceedances of EU air quality limit 40 ppm for yearly average data. Therefore, it is recommended for policy makers or concerned authorities to consider the exceedances and apply appropriate measures.

6.2. The LMM implementation

Even though, various studies (Emili et al., 2011; Gupta et al., 2006; Koelemeijer, et al., 2006) showed the correlation between AOT and in situ measurements, the acquired AOT data has low correlation (4%) with in situ data that , adding AOT as explanatory variable did not improve the model. So that AOT variable was excluded from the model. Koelemeijer, et al., (2006) divided AOT by the boundary layer height and made a correction for the growth of aerosols with relative humidity that they acquired a better correlation value between AOT and in situ PM_{10}. However, this need extra work that it is not considered in this study.

From the explanatory data analysis, clustering of negative and positive residuals was observed and this implies the spatial dependence between observations. For this condition, the typical OLS approach (which assumes residuals are independent and normally distributed) do not give plausible result (Pinheiro & Bates, 2000). Moreover, residuals were found to be heterokedastic and it is pointed out by different researchers (Hamm, et al., in review; Lark., 2009) that, if heteroskedasticity is present it should be modelled, if not it leads to biased estimates of parameters. Thus, to get plausible result these two conditions (heteroskedasticity and the spatial correlations) were accounted for the modelling process. LMM model has three advantages. First, it helped to model the heteroskedasticity observed over surrounding areas. Second, the approach can be used for modelling of spatial correlation between observations and third it was used to model the spatially varying mean.

This research has shown how the extension of LMM by variance model and the correlation structure model specification resulted in an improved model. The approach yielded a precise parameter estimates (a better fit to the data) comparing to the typical (OLS) parameter estimation method.

Understanding the structure of the spatial data is the prior for developing a plausible LMM, since LMM requires grouping factor that divides the observations into distinct groups of observations. Single level and multi-level grouping structure specifications were explored, for the PM_{10} data the multi-level specification gives a better fit to the data than a single level specification. This might be due to the structure of PM_{10} data is best explained by grouping using country and the variability is again well explained by variability with in type or background area in a country. This is the case because we have different background areas in a country, and there is different pattern of PM_{10} distribution for each background areas. The variability of PM_{10} distribution among countries is due to measurement techniques difference, socio economic difference, the adopted different environmental policies and some other reasons. Moreover, due to the kind and distribution of emission sources there is variability of PM_{10} distribution among different background areas (Held, et al., 2008), so grouping by type area within country gave a better description of the data. In general, two-level model was adopted because there are two levels of random variation of PM_{10} data, there is variation over country and there is variation over background areas.

The use of within group variance, leads to an increase in the estimated between group variability, heterokedastic model reduces the estimated between group variability so that it resulted on better fit to the data. The high value of the likelihood ratio test statistic confirms the significance of modelling heteroskedasticity in the model.

Spatially correlated variance can be modelled using the variogram or covariance function, Smyth., (2002) provides a flexibility of models, that model σ_i^2 as a function of any covariate. In this research, the non-stationarity in the variance was modelled through grouping the data by country and type (background) areas and use type area for weight specification to model the heteroskedasticity.

The non-constant variance over the three type (background) areas in countries (rural, suburban and urban) might be due to difference on kind of emission sources and the distribution of pollution sources. Since, air pollution has different pollutant sources; industry, agriculture; residential etc. that the spatial distribution of the pollutants is not uniform because the distribution of the sources is not even. Even though wind has capability to harmonize the pollutant concentration to local area, it does not mean that it will harmonize the overall concentration.
After assessing the PM distributional variability over type (background) areas, type area was used for weight specification or heteroskedasticity modelling. Weight can be assigned in ad hoc fashion or it can be estimated using a model, for this study the weight for different type areas were estimated using the

variance function model. Consequently, modelling the heteroskedasticity resulted on improved model this implies there is variability of PM_{10} distribution between type areas within country, so this should be taken in to account for modelling of air quality.

The variograms Figure 28, 29 and 30 show different spatial structures for different temporal aggregations, moreover the model parameter values differ among temporal aggregations models.

In general, LMM which model both the heteroskedasticity and the spatial correlation was found to be an improved model than the stationary model. The possible reason for this is PM_{10} has a different distribution pattern over different grouping specification, that it has different variability trend between countries and type areas. Moreover, there is a spatial correlation between observations that modelling the spatial structure yielded a better fit to the data.

A LMM was developed both for winter and summer temporal aggregations, and prediction was done. The winter period prediction has a lower accuracy than the summer model. This might be due to the high fluctuation and variability of PM_{10} concentration in air during winter time.

6.3. Prediction assesment

Even though, the estimates at the prediction sites from the non-stationary and stationary models are very close to each other they are not the same, moreover, referring the RMSE values the precision (measure of uncertainty) varies between models and among the grouping levels of the data.

The RMSE of the prediction at the grouping levels of the models differs, (level 0 population level, level 1 country level and level 2 type areas in country level have different RMSE values). The prediction for level 2 parameter estimates gives more accurate prediction than the other two levels. The possible explanation for this is, the parameter estimates at level two which are estimated for the within group variability better fit the data than the parameter which are estimated for the entire dataset.

The acquired BLUPs for the stationary and non-stationary models differs, in general the non-stationary model has a higher accuracy than the stationary model Table 20, 22 and 25. For model D, the RMSE and ME values for summer model prediction is 0.252 and -0.046 respectively subsequently, winter model prediction has RMSE value of 0.411and ME value of -0.135 for model D. So that the summer prediction has a higher accuracy than winter prediction, one possible reason for this is the winter data shows high variability in concentration than summer data. Moreover, predictions were done at winter and summer validation sites using the yearly average model. Subsequently, the yearly average model gave a better accuracy at summer validation sites prediction than winter validation sites predictions Table 27 again this is possibly explained by high variability of PM_{10} concentration during winter than during summer or for yearly average data.

For this study, stationary auto correlation was assumed in the non-stationary model, so that the geometric anisotropy was not modelled that the stationarity on the correlation was not relaxed, however it differs to some extent from the auto correlation under the full assumptions of stationarity in the variance. Lark..,(2009) suggests that the accuracy of the non-stationary model will improve if the geometric anisotropy is modeled.

In general, predicting at validation sites of a certain temporal aggregation has a better accuracy when the model parameter is estimated for that particular temporal aggregation. So that considering temporal aggregations for air quality modeling is crucial.

The specification of the variance models give a better result for nonstationary model than stationary model, however when we plugged in the estimated parameter in to BLUP the prediction accuracy difference it not significant however, for some points it shows that the prediction of PM_{10} value has different values and give different inferences when it is compared to threshold values (standard limits).

The results presented in this thesis have shown that non-stationary model yielded more accurate prediction than stationary model. This is due to the fact that spatial processes are spatially correlated and has a non-constant variance across a region. The developed LMM is consistent with explanatory data analysis that we can say that the LMM is scientifically reasonable.

7. CONCLUSIONS AND RECOMMENDATIONS

7.1. Conclusions

To conclude, the objective of the study was to explore and model PM_{10} over different temporal aggregations and extend LMM to account for heteroskedasticity (non-constant variance) and spatial correlation of PM_{10} pollution process using different covariates. Moreover, it was aimed to do prediction at the validation sites using LMM. To achieve the research objectives seven research questions were formulated Table 1. Referring, the results obtained and the discussion made the below mentioned conclusions were drawn.

What is the spatial distribution of in situ PM_{10} over regions?
The spatial distribution of PM_{10} differs from region to region. For most parts of Europe; the concentration of PM_{10} in the air is lower than the standard limit. However, there are areas on which the concentration exceeds the limit.

What is the spatial structure of the correlation of CTM, PM_{10} and elevation over temporal aggregations (daily, monthly, seasonal and yearly)?
PM_{10} shows different distribution trend over the temporal aggregations. During summer and June the concentration gets lower whereas, during winter the concentration highly increases over most part of Europe and exceeds the EU standard limits (which are 40ppm for yearly average data and 50ppm for daily average data). In general, areas show different pollutant concentration trend over different temporal aggregations that the choice of temporal aggregation should be taken in to account for modelling of air quality.

The extended LMM was applied for summer, winter and yearly average data. The spatial structures of the variogram models of these temporal aggregations (for standardized residuals) show difference in the model parameters. Moreover, predicting at validation sites of a certain temporal aggregation has a better accuracy when the model parameter is estimated for that particular temporal aggregation that considering temporal aggregations for air quality modeling is crucial.

How do we specify the fixed and random effects in LMM?
The correlation of PM_{10} and covariates (namely AOT, CTM and elevation) was assessed and these terms except AOT (since the acquired correlation value between PM_{10} and AOT is very low, it was excluded from the model) were specified for fixed effect definition.

A grouping structure specification was carried out for random effect definitions using country and type (background) areas as grouping variables. Moreover, multi and single level grouping structures of the data were compared using AIC, log Likelihood and likelihood ration test values of the model outputs. It was found that the model with multi-level grouping structure gave a better description of the data structure than the single level specification.

How can we model the correlation structure in LMM?
The need for modelling the spatial correlation was assessed and the correlation structure in the LMM was specified using the spatial covariates (longitude and latitude) with in nested type grouping factor. Moreover, the initial model parameters (range and nugget) were specified in correlation structure definition. Referring the model evaluation criteria's used for this study LMM which modelled the spatial correlation (Model B) is an improved model over Model A.

How do we model the non-stationarity in the variance using the LMM?
A variance function was used to model the non-constant variance structure of PM10; the function allowed modelling the heteroskedasticity of the within-group error through a weights argument. The weight was modelled as a function of covariate (type area for this study after assessing the different PM_{10} distribution pattern over different type areas). LMM which modelled the heteroskedasticity (Model C) is an improved model over Model A.

Does the non-stationary model offer an improvement over stationary model?
In general, a simple extension of LMM model with parametric non-stationary variance model and correlation function model resulted on improved model for PM_{10} spatial dataset. For this study, the mean and the variance are non-stationary, but the autocorrelation was assumed to be stationary (since the model is based on intrinsic stationarity). The explanatory analysis of the data showed that the assumption made for the LMM was plausible.

The non-stationary LMM, which modelled both the heteroskedasticity over type (background) areas and the spatial correlation, has AIC and log likelihood value of -410.04 and 225.02 respectively and the stationary model has AIC value of -177.9 and log likelihood value of 104.95, this shows that non-stationary model is an improved model over the stationary model. This is further supported by likelihood ratio test result, 223.2 with p-value <0.0001. These result was acquired for yearly average data however, the improvement of non-stationary model over stationary model is also observed for summer and winter temporal aggregations.

Does the non-stationary model offer better prediction accuracy than stationary model?
Referring the validation of the BLUPs at a set of test sites the prediction from the non-stationary summer model was more accurate (has lower RMSE 0.21) than the prediction done from the stationary summer model, 0.27 (the RMSE is for log transformed value). The acquired accuracy difference between the two models become more significant when the value of the PM_{10} value is close to the standard limit, so in this case the stationary model underestimates the result and gives a wrong impression for PM_{10} status in the air. The acquired prediction accuracy of the models further supports the conclusion from a log-likelihood ratio test that the non-stationary model was better than stationary model. In general for summer, winter and yearly average data the non-stationary model has a better accuracy than stationary model.

The results presented on this paper have shown that non-stationary model yielded a better fit to the data and better prediction accuracy than stationary model. This is due to the fact that spatial processes are spatially correlated and has a non-constant variance across a region. The developed LMM is consistent with explanatory data analysis that we can say that the LMM is scientifically reasonable.

7.2. Recommendations

For air quality modelling, there is a need to take in to account the temporal aggregations signal on the pattern of pollutant distribution.

As it is indicated by Koelemeijer, et al., (2006) AOT and in situ correlation was increased by correcting AOT using meteorological parameters for example Relative humidity and boundary layer height, I recommend exploring this further and use the AOT term for fixed effect specification of LMM.

There are different forms of non-stationarity, for this study the non-stationarity in the mean and non-stationarity in the variance were modelled, however further work is needed to model the geometrical

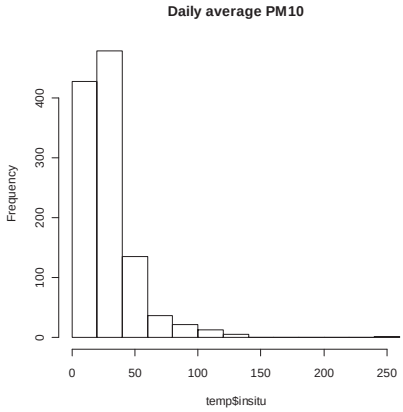
anisotropy or there is a need to relax the assumption of stationarity in the autocorrelation, because Lark, (2009) pointed out that such non-stationarity will have effects on precision of the predicted values.
For modelling anisotropy it will be optimal to use more covariate (for example pollutant source locations, wind direction) which could explain the directional variations.

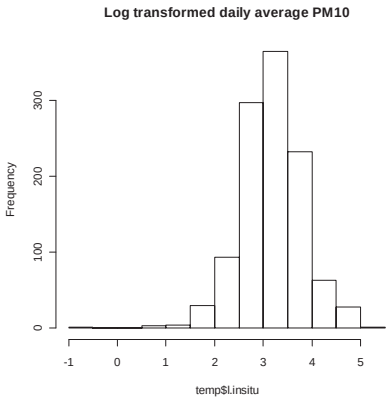
Typically to improve the predictive ability of the model, the below listed points are recommended

- Modeling the geometric anisotropy
- Use consistent data (e.g use same measurement techniques over areas)
- use almost equal number of data from different background areas so that the data will be balanced with respect to the number of observations
- Use more data for a training set
- Add other covariates (like distance from the cost, precipitation) since (Gomiscek, et al., 2004a) and (Gomiscek et al., 2004b) indicates the potential of in-situ field measurements to characterize the local environment strongly depends on the meteorological condition, the pollutant type and topographic features of the area.

LIST OF APPENDICES

Appendix 1- Explanatory data analysis results of temporal aggregations



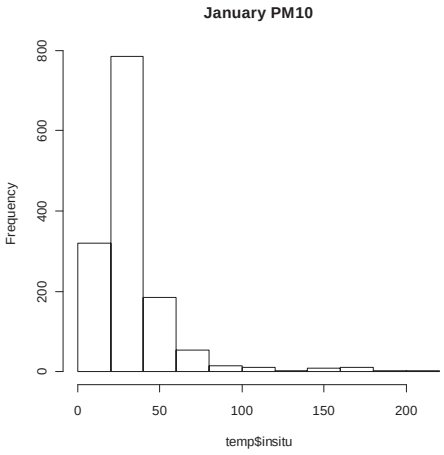


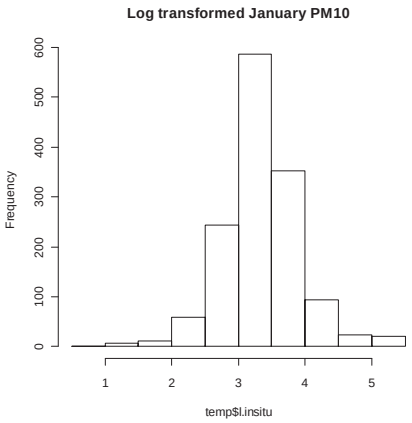
Summary of daily insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.5	16.42	24.67	29.51	35.25	242.9

Summary of daily log transformed insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.69	2.79	3.2	3.19	3.56	5.49



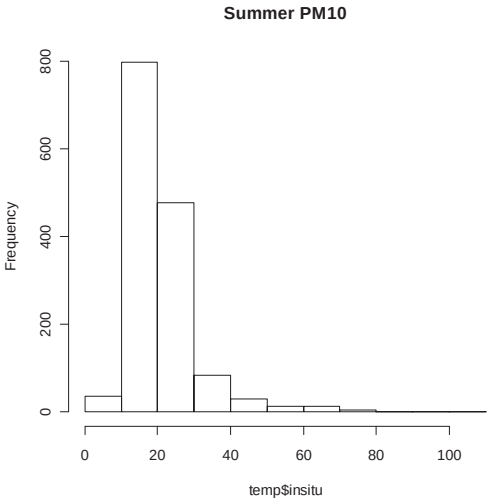


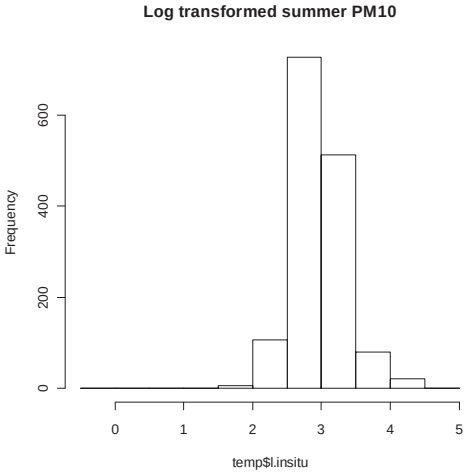
Summary of January insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.68	20.72	26.9	33.27	37.81	218.8

Summary of January log transformed insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.985	3.031	3.29	3.33	3.63	5.39



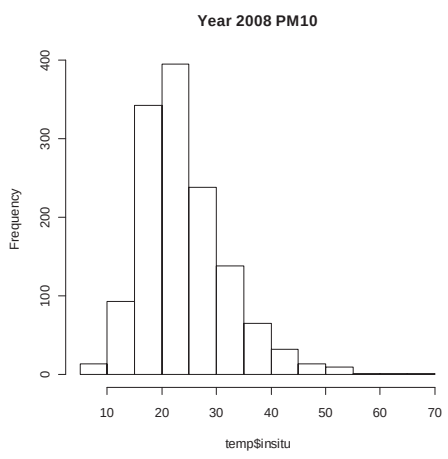


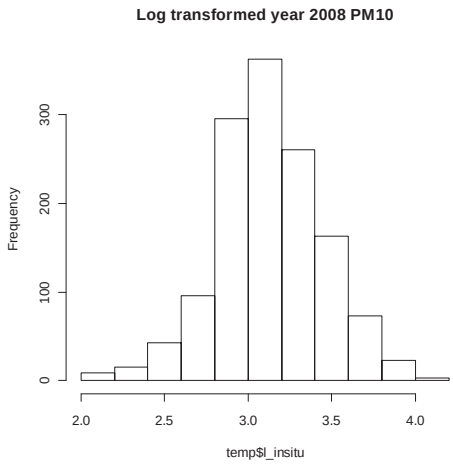
Summary of summer insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.86	15.48	19.03	21.08	23.98	107.6

Summary of summer log transformed insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.15	2.74	2.95	2.97	3.18	4.68



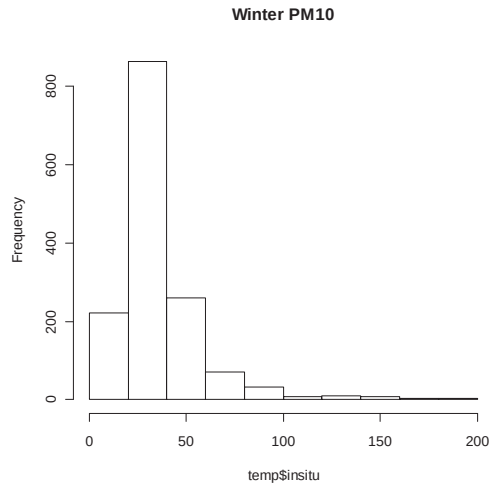


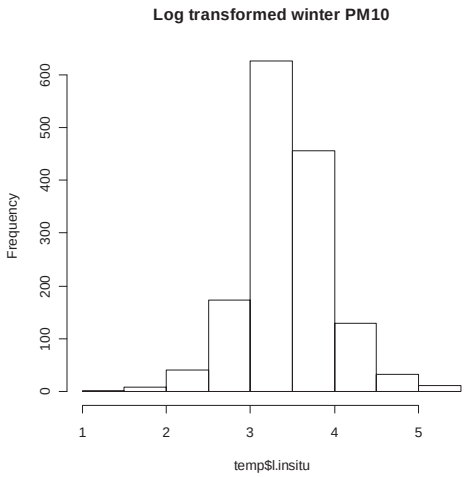
Summary of year 2008 insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7.57	18.82	22.78	24.12	28.06	66.41

Summary of year 2008 log transformed insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.02	2.94	3.13	3.13	3.33	4.19





Summary of winter insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.75	23.94	30.6	36.03	41.2	192

Summary of winter log transformed insitu data

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.32	3.18	3.42	3.45	3.72	5.26

Appendix -2 LMM output

LMM for yearly average data (using lme function)

Model A l.insitu~l.le.model+height

Data Yearly average

AIC	BIC	logLik
-177.9	-94.69	104.95

Random effects:

Formula: ~l_le_model + height | coun

Structure: General positive-definite,Log-Cholesky parametrization

	StdDev
Intercept	0.811
l.le.model	0.372
elevation	0.0003

Random effects:

Formula: ~l_le_model + height |type area in coun

Structure: General positive-definite,Log-Cholesky parametrization

	StdDev
Intercept	0.368
l.le.model	0.119
elevation	0.0003

Fixed effects: l_insitu~l~l_le_model+height

	Value	p-value
Intercept	1.93563	0
l.le.model	0.4882	0
elevation	-0.0003	0.0072

LMM for yearly average data (using lme function)

Model B l.insitu~l.le.model+height

Data Yearly average

AIC	BIC	logLik
-391.2	-297.59	213.6

Random effects:

Formula: ~l_le_model + height | coun

Structure: General positive-definite,Log-Cholesky parametrization

	StdDev
Intercept	0.155
l.le.model	0.119
elevation	0.0002

Random effects:

Formula: ~l_le_model + height |type area in coun

Structure: General positive-definite,Log-Cholesky parametrization

	StdDev
Intercept	0.419
l.le.model	0.103
elevation	0.0003

Correlation Structure: Exponential spatial correlation

Formula: ~x+y|coun/type_area

Parameter estimate(s)

range	nugget
208.19	0.4

Fixed effects: l_insitu~l~l_le_model+height

	p-value
Intercept	0
l.le.model	0
elevation	0

LMM for yearly average data (using lme function)

Model C l.insitu~l.le.model+height

Data Yearly average

AIC	BIC	logLik
-189.39	-95.78	112.69

Random effects:

Formula: ~l_le_model + height | coun
Structure: General positive-definite,Log-Cholesky parametrization

	StdDev
Intercept	0.8
l.le.model	0.371
elevation	0.00027

Random effects:

Formula: ~l_le_model + height |type area in coun
Structure: General positive-definite,Log-Cholesky parametrization

	StdDev
Intercept	0.383
l.le.model	0.126
elevation	0.0003

Fixed effects: l_insitu~l~l_le_model+height

	p-value
Intercept	0
l.le.model	0
elevation	0.0091

LMM for yearly average data (using lme function)

Model D l.insitu~l.le.model+height

Data Yearly average

AIC	BIC	logLik
-410.04	-306.04	225.02

Random effects:

Formula: ~l_le_model + height | coun

Structure: General positive-definite,Log-Cholesky parametrization

	StdDev
Intercept	0.162
l.le.model	0.124
elevation	0.0002

Random effects:

Formula: ~l_le_model + height |type area in coun

Structure: General positive-definite,Log-Cholesky parametrization

	StdDev
Intercept	0.416
l.le.model	0.148
elevation	0.0002

Correlation Structure: Exponential spatial correlation

Formula: ~x+y|coun/type_area

Parameter estimate(s)

range	nugget
204.08	0.4

Fixed effects: l_insitu~l~l_le_model+height

	p-value
Intercept	0
l.le.model	0
elevation	0

LIST OF REFERENCES

Apascaritei, M., Popescu, F., & Ionel, H. I. (2009). *Air pollution level in urban region of bucharest and in rural region.* Athens: World Scientific and Engineering Acad and Soc.

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

Carvalho, A. C., Carvalho, A., Gelpi, I., Barreiro, M., Borrego, C., Miranda, A. I., & Pérez-Muñuzuri, V. (2006). Influence of topography and land use on pollutants dispersion in the Atlantic coast of Iberian Peninsula. *Atmospheric Environment, 40*(21), 3969-3982.

Cinzia Mazzetti, & Todini, E. (2002). Development and application of the block Kriging technique to rain-gauge data. [University of Bologna].

Clean AIR Systems. (2007). Emissions control systems for today's air quality standards. Retrieved May 26 / 2011, from http://www.cleanairsys.com

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

Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model Based Geostatistics. New york: Springer.

Diggle., P. J., & Ribeiro., P. J. (2007). Model-based Geostatistics. Springer Verlag, New York.

Emili, E., Popp, C., Petitta, M., Riffler, M., Wunderle, S., & Zebisch, M. (2010). PM10 remote sensing from geostationary SEVIRI and polar-orbiting MODIS sensors over the complex terrain of the European Alpine region. *Remote Sensing of Environment, 114*(11), 2485-2499.

Emili, E., Popp, C., Wunderle, S., Zebisch, M., & Petitta, M. (2011). Mapping particulate matter in alpine regions with satellite and ground-based measurements: An exploratory study for data assimilation. *Atmospheric Environment, 45*(26), 4344-4353.

EPA. (2011). United States Environmental Protection Agency. Retrieved 23 September 2011, from http://www.epa.gov/gateway/learn/

Fuentes, M. (2003). A formal test for nonstationarity of spatial stochastic processes. *96*, 30 – 54.

Fuentes., M. (2002). Interpolation of nonstationary air pollution processes: a spatial spectral approach, SAGE. *2*, 281–298.

Gomiscek, B., Frank, A., Puxbaum, H., Stopper, S., Preining, O., & Hauck, H. (2004a). Case study analysis of PM burden at an urban and a rural site during the AUPHEP project. *Atmospheric Environment, 38*(24), 3935-3948.

Gomiscek, B., Hauck, H., Stopper, S., & Preining, O. (2004b). Spatial and temporal variations of PM1, PM2.5, PM10 and particle number concentration during the AUPHEP—project. *Atmospheric Environment, 38*(24), 3917-3934.

Great Basin Unified Air Pollution Control District. (2007). Particulate Matter Air Pollution. Retrieved 05/29/2011, from http://www.gbuapcd.org/index.htm

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

Hamm, N. A. S., Atkinson, P. M., & Milton, E. J. (in review). A linear mixed model (LMM) implementation of the empirical line method (ELM). Remote Sensing of Environment.

Haskard, K. A., & Lark, R. M. (2009). Modelling non-stationary variance of soil properties by tempering an empirical spectrum. *Geoderma, 153*(1-2), 18-28.

Held, A., Zerrath, A., McKeon, U., Fehrenbach, T., Niessner, R., Plass-Dülmer, C., Kaminski, U., Berresheim, H., & Pöschl, U. (2008). Aerosol size distributions measured in urban, rural and high-alpine air with an electrical low pressure impactor (ELPI). *Atmospheric Environment, 42*(36), 8502-8512.

Huang, C., et al. (2011). Emission inventory of anthropogenic air pollutants and VOC species in the Yangtze River Delta region, China. *Atmos. Chem. Phys., 11*(9), 4105 - 4120.

Jan van de Kassteele. (2006). Statistical Air Quality Mapping.

Kanevski, M. (Ed.). (2010). *Advanced mapping of environmental data: geostatistics, machine learning, and Bayesian maximum entropy.* US: Iste.

Kim, D., & Stockwell, W. R. (2008). An online coupled meteorological and air quality modeling study of the effect of complex terrain on the regional transport and transformation of air pollutants over the Western United States. *Atmospheric Environment, 42*(17), 4006-4021.

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

Lark., R. M. (2009). Kriging a soil variable with a simple nonstationary variance model;Journal of Agricultural, Biological, and Environmental Statistics*., 14*(3), 301-321.

Meul, M., & van Meirvenne, M. (2003). Kriging soil texture under different types of nonstationarity. *Geoderma, 112*(3-4), 217-233.

Monks, P. S., et al. (2009). Atmospheric composition change - global and regional air quality. *Atmospheric Environment, 43*(33), 5268-5350.

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

Pernigotti, D., Georgieva, E., Thunis, P., Cuvelier, C., & Meij, A. (2012). The Impact of Meteorology on Air Quality Simulations over the Po Valley in Northern Italy Air Pollution Modeling and its Application XXI. Springer Netherlands. In D. G. G. Steyn & S. T. Trini Castelli (Eds.), (Vol. 4, pp. 485-490).

Phalen, R. F. (2003). Harmful effects of particulate air pollution, The particulate air pollution controversy. Springer US., 1-13.

Pinheiro, J. C., & Bates, D. M. (2000). Linear Mixed-Effects Models: Basic Concepts and Examples Mixed-Effects Models in Sand S-PLUS. Springer New York (pp. 3-56).

Pinheiro., J. C., & Bates., D. M. (2000). Mixed-Effects Models in Sand S-PLUS.e-book (Second edition ed.).Springer Verlag New York, LLC

Sanchez, M. L., Garcıa, M. A., Perez, I. A., & de Torre, B. (2007). Ground laser remote sensing measurements of a Saharan dust outbreak in Central Spain. Influence on PM10 concentrations in the lower and upper Spanish plateaus. *Chemosphere, 67*(2), 229-239.

Schaap, M., Timmermans, R. M. A., Roemer, M., Boersen, G. A. C., & Builtjes, P. J. H. (2008). The LOTOS–EUROS model: description, validation and latest developments, Int. J. Environment and Pollution,. *32*(2).

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.

Smyth, G. K. (2002). An efficient algorth for reml in heteroskedastic regression. Journal of Computational and Graphical Statistics. *11*(4), 836-847.

van Bussel, L. G. J., Muller, C., van Keulen, H., Ewert, F., & Leffelaar, P. A. (2011). The effect of temporal aggregation of weather input data on crop growth models' results. *Agricultural and Forest Meteorology, 151*(5), 607-619.

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