Spatio-temporal analysis and interpolation of PM_{10} and PM_1 in Eindhoven

Module 13: Advanced Geostatistics

Mini Project: Report

Ipsit Dash s6019552

Faculty ITC, University of Twente

Enschede, Netherlands

July 2015

Table of Contents

List of Tables

1. Introduction

1.1 Background and rationale of study

The advancement of industrialization and urbanization to provide a better life for humans has today, severely jeopardized the aim for which it was intended. It has led to severe deteoriation in the quality of environment by means of pollution and toxicity making humans and other life forms highly vulnerable to health risks (Iii et al., 2002). Thus, this calls for proper monitoring of the causality of these long term life-threatening factors and provide an alternate or a sustainable solution that would re-instate the balance in the environment without compromising the urbanization needs.

Particulate matter (PM) are a major contributing factors for the urban air pollution and cardiovascular diseases such as heart attacks in humans.(Murad, 2012) Based on the diameter of these "semi-solid semi-fluid" substances, these are categorized as fine (≤ 2.5µm) like PM1, ultrafine particles (UFP) and coarse particles like PM2.5 and PM10 (W.H.O, 2003). Coarse particles are chemically composed of earth crust materials and elements from roads and industries and fine particles are generally secondarily formed aerosols, combustion particles and recondensed organic and metal vapours. As the size is relatively larger for coarse materials, they exhibit high levels of variability in temporal scale as compared to finer particles which generally have more constancy in variation.

This study tries to predict levels of concentration of particulate matters specifically PM1 and PM10 levels in around the city of Eindhoven, Netherlands from 1st October 2014 to 14th October 2014. Municipality of Eindhoven is location in south of the Netherlands and is one of the highly urbanized city of the Netherlands. This could be considered as a primal step in quantifying the air quality levels in the city and would be of considerable importance while deciding statergies to reduce air pollution. Furthermore, as a step towards the smart city governance involving participatory approach of citizens it could be utilized as an informative tool to aware the citizens about the local levels of pollution.

Owing to spatial and temporal variability of the pollutants, prediction and forecasting of the concentration levels by conventional modelling methods of air quality (Collett & Oduyemi, 1997) doesn't provide satisfactory. Geostatistics could be a better solution in producing optimized and robust models that take into account both spatial and temporal dependancies of the pollutants (Jerrett et al., 2009). Furthermore, utilization of other location specific information in tandem with geostatistical analysis could aid to the improvisation of accuracy of prediction and forecasting (Desta, 2012 cited in Enkhtur, 2013).

Research by Knox et al., 2013 have suggested that utilization of dense network of cost effective and highly reliable sensors could be effective in providing data about ground level pollutants at a greater temporal frequency. AiREAS (AiREAS, 2014)have established such a network to obtain the levels of particulate matter in around Eindhoven. They operate with high temporal frequency (~10 minutes) and data from this network has been utilized for analysis.

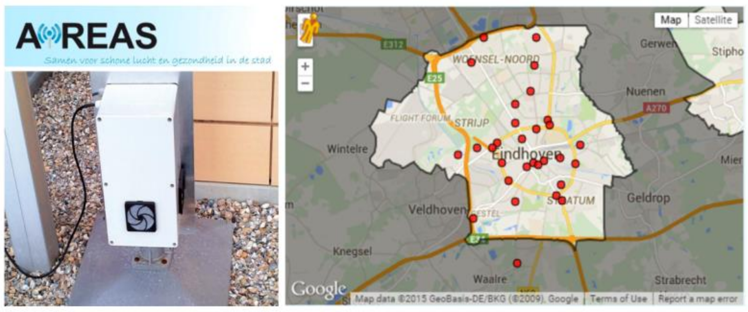


Figure 1Location of AirBoxes (left) around Eindhoven. (www.aireas.com)

1.2 Research objectives

This study is based on investigating the potential of spatio-temporal kriging approaches for prediction and forecasting the levels of PM10 and PM1 concentration for 2 weeks (1^{st} Oct-14^{th} Oct, 2014) around the municipality of Eindhoven, the Netherlands. The data utilized for the space time geostatistical analysis is a hourly averaged values of original PM10 and PM1 values obtained from the AirBase dataportal (EEA, 2014). Methods and usability of incorporating other geo-ancillary information like land use, traffic density to improve the prior knowledge of dataset and also validate the prediction were also incorporated. Finally cross validation was performed on 3 stations randomly to comment on the accuracy of the prediction.

2. Methodology

Statistical procedures outlined in Sterk & Stein, 1997 give a detailed study about geostatistical analysis to handle spatio-temporal data. Various interpolational techniques reviewed and analysed in Gräler, Gerharz, & Pebesma, 2012 provided relevant ways to model PM10 values at a continental level. These were taken as the base materials along with guided description of "spacetime" library (E. Pebesma, 2012) and "gstat" library (E. J. Pebesma, 2004) in R working environment (R Team, 2014) to model concerntration levels of PM1 and PM10 for the city of Eindhoven. A detailed flow chart of methodology is presented below and the explanation following to it.

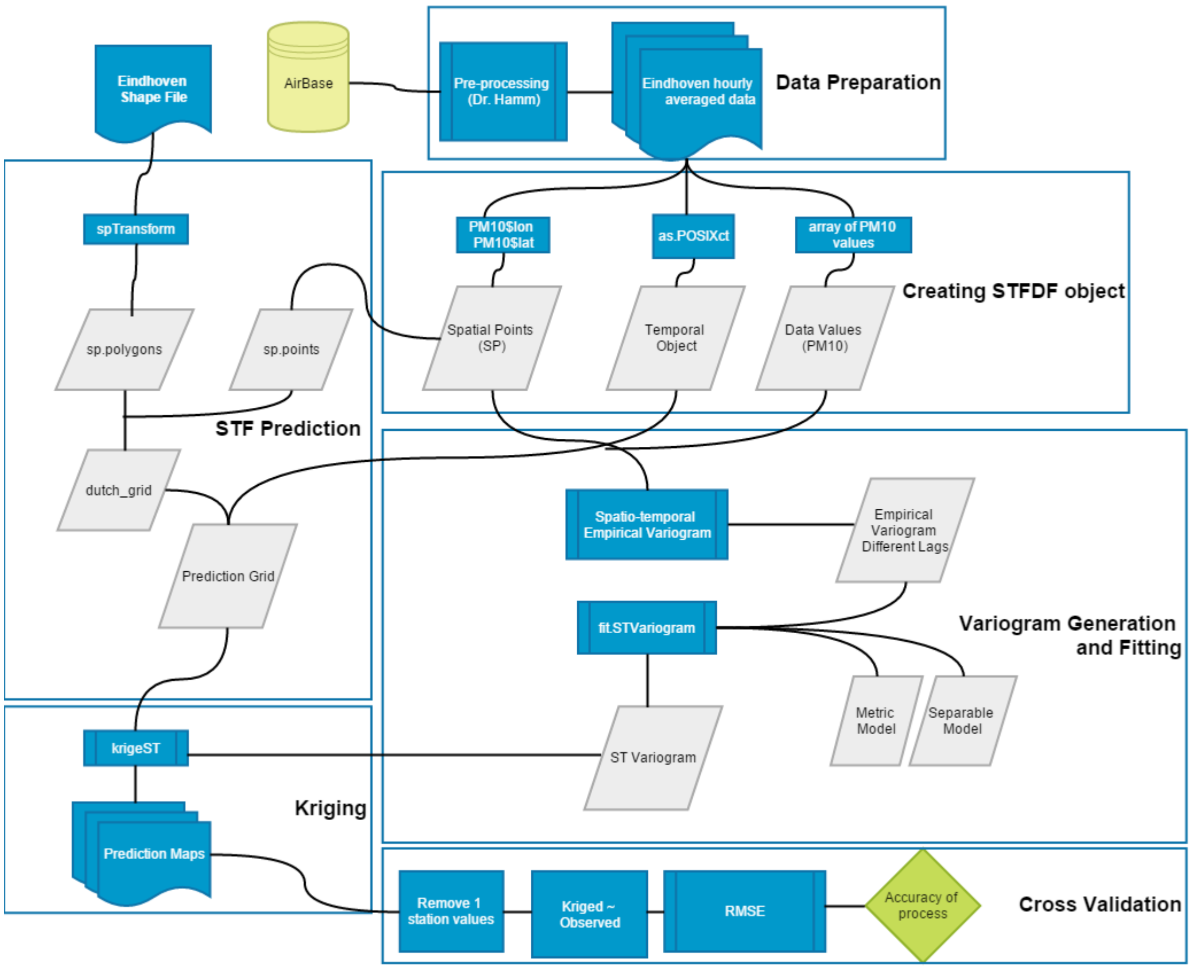


Figure 2Flow chart of methodology adopted

2.1 Data preparation

The data utilized for the analysis is the hourly averaged PM10 and PM1 concerntration levels at 32 stations in and around Eindhoven, the Netherlands from the AirBase database. It was previously pre-processed to cater specifically for the study. The temporal stamp was hourly and spanned for 14 days therby 336 observations. A sample representation of data is depicted below that represents a temporal view of a station at location (5.431, 51.413) and a spatial view of PM1 observation for all locations for 2 consecutive timestamps (1st Oct 2014- 1 am and 2am). From the temporal view it is evident that there is relatively a high amount of variability for PM10 whereas it is low for PM1 and PM2.5.

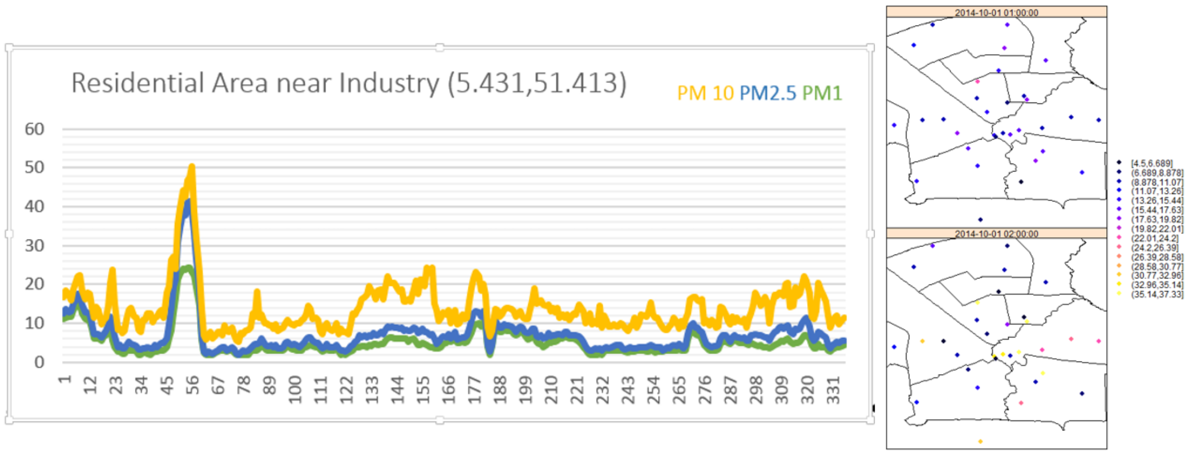
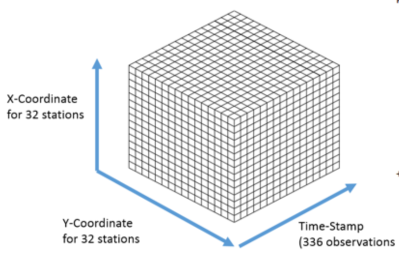


Figure 3 Left: Temporal observation for station located at 5.431, 51.413 and Right: PM1 observations at all locations for 2 consecutive time stamps at 1st Oct 2014 – 1am and 2 am

2.1.1 Creating STFDF object

For geostatistical analysis of the data as per norms of spacetime and gstat library data was converted to a STFDF object, i.e. a 3-dimensional matrix containing spatial points (coordinates), timestamps (temporal locations) and values. For this study, 2 STFDF objects were created each for PM10 and PM1 respectively. They were defined as:

- Spatial points : Locations of stations (Latitude and Longitude) - 32 x 2 matrix
- Timestamps: Starting from 2014-10-01 : 00:00 to 2014-10-14 : 23:59 – 336 x1 matrix
- Value: corresponding PM1 or PM10 values for each station and each timestamp
  i.e.list of (32 x 336 = 10752 values)




```
> str(st.all.obs)
Formal class 'STFDF' [package "spacetime"] with 4 slots
  ..@ data   :'data.frame': 10752 obs. of  2 variables:
  .. ..$ value: num [1:10752] 9.67 12.33 13 11 9.83 ...
  .. ..$ ID   : Factor w/ 10752 levels "ID 1","ID 10",..: 1 1865 2976 4087 5198 6309 7420 8531 9642 2 ...
  ..@ sp     :Formal class 'SpatialPoints' [package "sp"] with 3 slots
  .. .. ..@ coords     : num [1:32, 1:2] 5.51 5.47 5.48 5.46 5.48 ...
  .. .. .. ..- attr(*, "dimnames")=List of 2
  .. .. .. .. ..$ : NULL
  .. .. .. .. ..$ : chr [1:2] "x" "y"
  .. .. ..@ bbox       : num [1:2, 1:2] 5.42 51.39 5.52 51.49
  .. .. .. ..- attr(*, "dimnames")=List of 2
  .. .. .. .. ..$ : chr [1:2] "x" "y"
  .. .. .. .. ..$ : chr [1:2] "min" "max"
  .. .. ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
  .. .. .. .. ..@ projargs: chr "+init=epsg:3035 +proj=laea +lat_0=52 +lon_0=10 +x_0=4321000 +y_0=3210000 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m
+no_defs"
  ..@ time   :An 'xts' object on 2014-10-01 01:00:00/2014-10-15 containing:
  Data: int [1:336, 1] 1 2 3 4 5 6 7 8 9 10 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr "timeIndex"
  Indexed by objects of class: [POSIXct,POSIXt] TZ:
  xts Attributes:
 NULL
  ..@ endTime: POSIXct[1:336], format: "2014-10-01 02:00:00" "2014-10-01 03:00:00" "2014-10-01 04:00:00" "2014-10-01 05:00:00" ...
```

Figure 4 Sample representation of STFDF object and its parameters

2.1.2 Creating STF prediction grid

For prediction and forecasting of values (PM1 or PM10) a grid was created which contained all the possible locations where prediction was to be made and timestamps on which forecasting was to be done. Then as per requirements STF object containers was made to store the kriged values. These had 2 parameters which were defined as follows:

- Spatial grid : Utilizing the information from bounding box of study area a grid was created with distance of 0.015 spatial units
- Temporal grid: Since the forecasting was to be done for all the timestamps, this STF grid had the same temporal length as that of the STFDF grid, i.e. 336 timestamps in a loop of 14 (one grid for 1 day).

The reduction in spatial grid and looping of temporal grid was done to ease off the computational efficiency of the program.

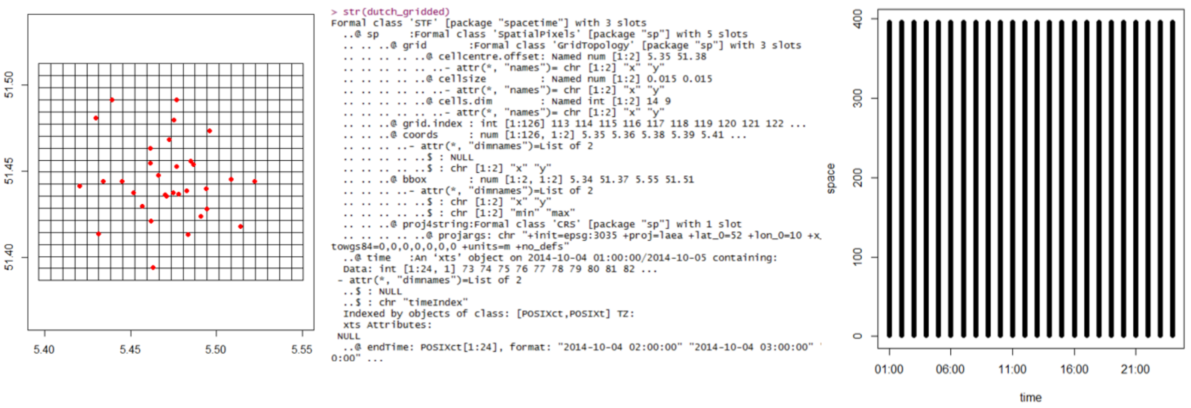


Figure 5 STF object creation and parameters. Left (Spatial grid) and Right (Temporal grid instance for 1st Oct 2014)

2.2 Variogram generation and fitting

The preliminary step for performing the kriging was to compute the spatio-temporal empirical variogram. To generate this function variogramST was utilized, which takes in as parameter the STFDF object (spatial locations, values and timestamps) and calculates the sample variogram. Although another option of time lag of 0:48 was tried to see if there is any temporal patterns emerging out of the variogram.

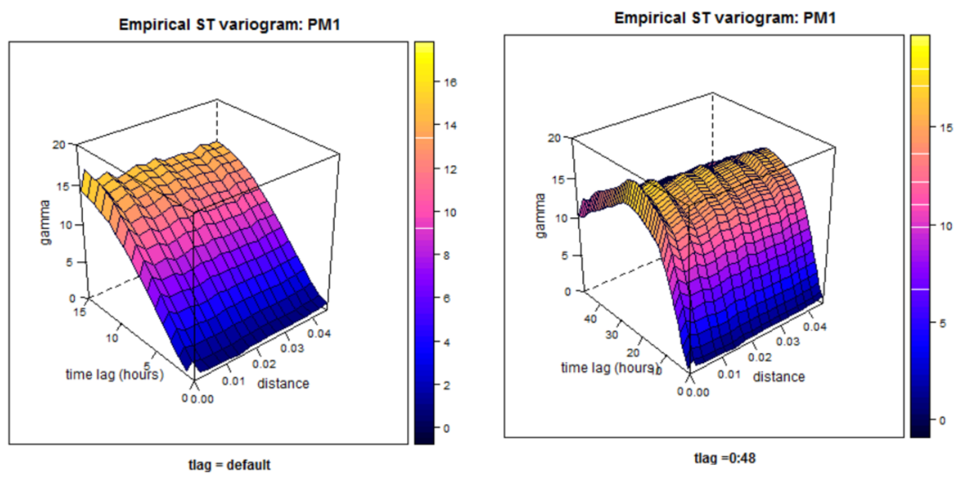


Figure 6 Empirical variograms for PM1

It is seen from the empirical variogram for timelag 0:48 hours that there is a peak around 25 hours but within a day the trends of increase or decrease of PM1 levels were not discernible. For PM10, the empirical variogram was very fluctuating with values, and even fitting the spatial variogram was poor.

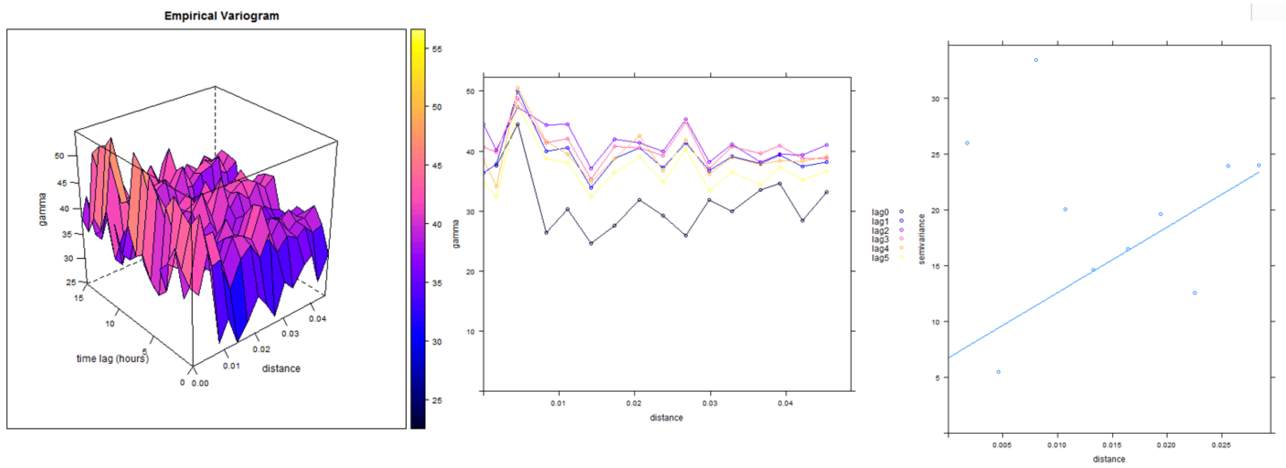


Figure 7Left: Empirical variogram for PM10; Centre: Sample variogram for each time lag, Right: Spatial variogram for a time instance

The empirical variogram was fitted with two models – Metric Model and Separable covariance function model.

2.2.1 Metric model: As an extension of spatial kriging, time as a 3rd dimension was used and comparing the timescales and spatial scales, an anisotropic parameter "stAni" which is the equivalent timescales in spatial scales was given to fit the variogram. However estimating the value of stAni was not possible in the library of gstat and spacetime. Owing to limitation of time, by visual observation of the default empirical variogram, it was noted that 0.045 spatial units equalizes to 15 hours of temporal units, so a rough estimate of "stAni = 330" was utilized.

2.2.2 Separable covariance model: Here, separate variogram models for spatial and temporal values are given to fit the empirical variogram. The separable covariance function takes into account both spatial and temporal distances while fitting the model to the entire dataset. It is fairly easy to compute separate temporal and spatial variograms and no other parameter needs to be estimated. Another advantage of using this model, is due to availability of complete dataset for all temporal locations which would not have been possible if incomplete dataset was available.

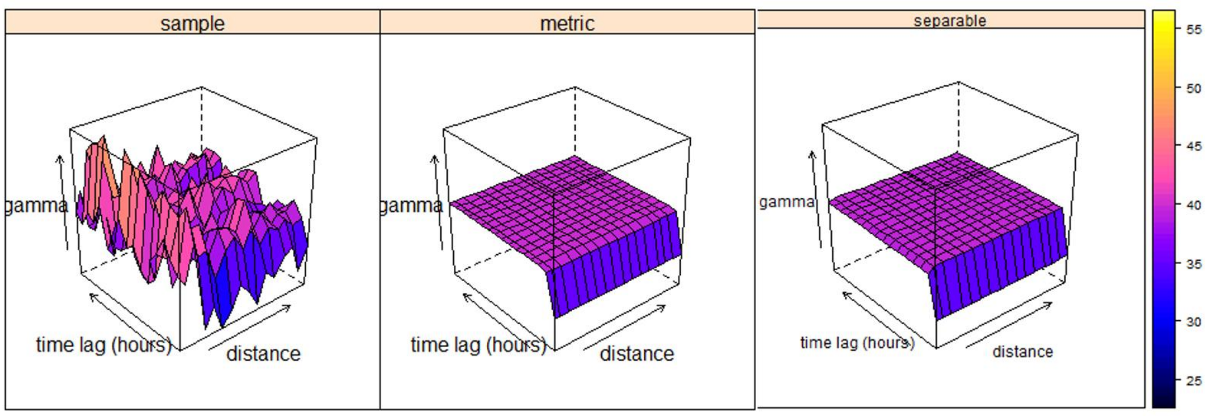


Figure 8 Sample variogram for PM10 and fitting the models

Since the fitting models for PM10 values were evidently poor as they didn't show the variation of the values and produced constant values, the analysis for PM10 was not continued. A follow up to this study may be modelling properly the variogram for PM10 to properly model them. Thus the analysis was followed up for PM1 values only.

The empirical variogram and the modelled variogram for PM1 are depicted below.

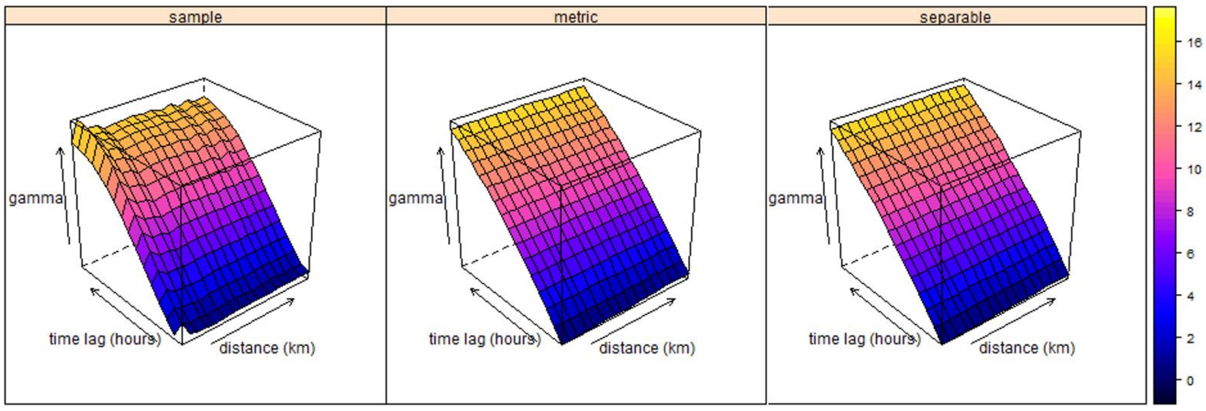


Figure 9 Sample variograms for PM1 and model fitting

The comparison of the models was done with "optim parameter" provided in the "spacetime" library which evaluates the relative fit of the models to the empirical variogram. It is basically the goodness of fit value for model and sample variogram calculated by the root mean square difference of surfaces. The parameters for the models and the optim values are as depicted below:

```
> attr(metricFit,"optim.output")$value
[1] 0.4381815
> attr(separableFit,"optim.output")$value
[1] 0.4395659
> extractPar(metricModel)
  sill  range nugget   anis
   0.8  150.0    0.2  100.0
> extractPar(metricFit)
     sill     range    nugget      anis
 29.96213 210.38933   0.00000  10.35735
> extractPar(separableModel)
 range.s nugget.s  range.t nugget.t     sill
   100.0      0.7     22.0      0.9     50.0
> extractPar(separableFit)
  range.s  nugget.s   range.t  nugget.t      sill
100.00626   0.00000  20.34535   0.00000  29.98683
```

Figure 10 Parameters for PM1 variogram fitting

Since the parameters for separable model were approximatedly similar to the fitted model, and the optim value being almost similar for both the models, it was chosen over metric mode. Furthermore, as the "stAni" parameter was not estimated properly, metric model was not favoured.

Thus, as the final step was computing the prediction maps or kriged maps based on PM1 values and for the entire timestamp, separable covariance fit model was utilized in the "krigeST" function to compute the maps. These parts is discussed under results in the following section.

3. Results and discussions

3.1. Spatio-temporal kriging

The kriged values for corresponding timestamps were done for PM1 values and the results were stored in the previously designed STF grid. It can be seen that the levels of PM1 tend to increase as the time increases, however low values are evident in the south centre regions of the city while the values are comparatively higher for zones in the north-west provinces By considering the separable covariance model with cumulative values till previous days, the kriged values tend to get smoothed as the timestamp increases. It can be seen from the range of values for day 1 (left) and day 10 (right) For Day 1 the kriged value lies from 8 to 13 whilst for day 10 they lie from 10.1663 to 10.1664.

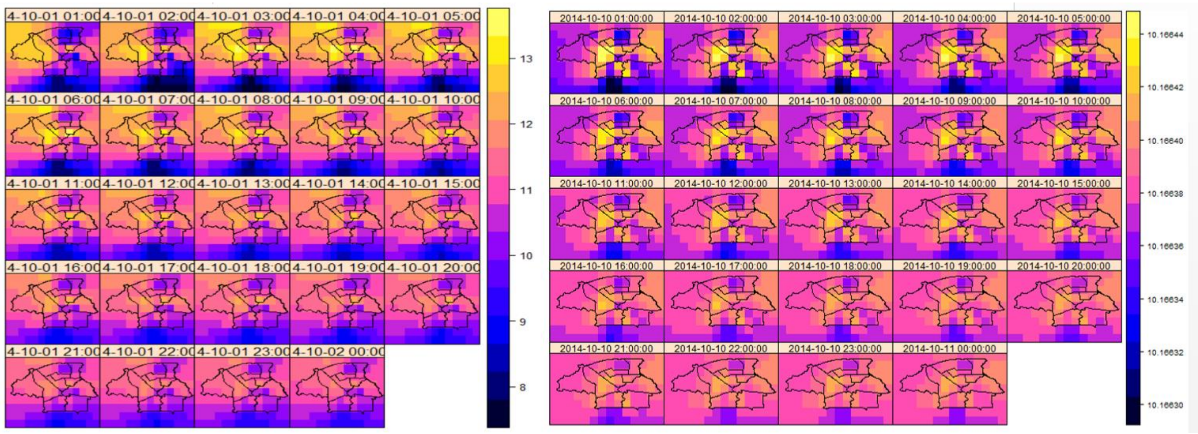


Figure 11Hourly prediction maps for Left: 2014-10-01: (Day 1), Right:2014-10-10: (Day-10)

Generally, peak hours of traffic calls in for higher levels of PM1 values, and time between 8 am to 10 am is the time when mostly office hours exist. This is also the time period when most citizens including children and senior citzens along with other office goers by biking are vulnerable to the effects of PM1. Furthermore, the Netherlands being a bike-loving country, this fact is justified. Therefore, prediction maps for 8am to 10 am for weekdays are depicted below.

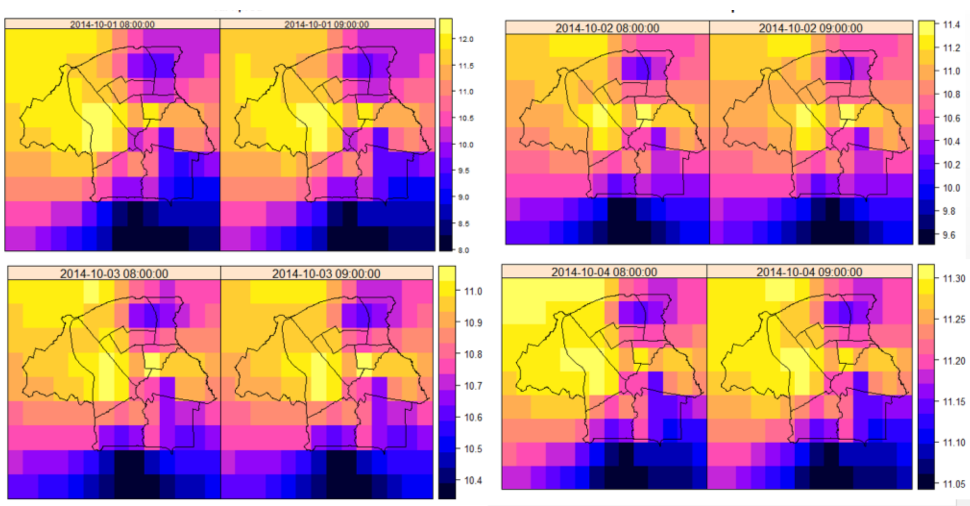


Figure 12 Peak hour predictions for Day 1, 2, 3 and 4

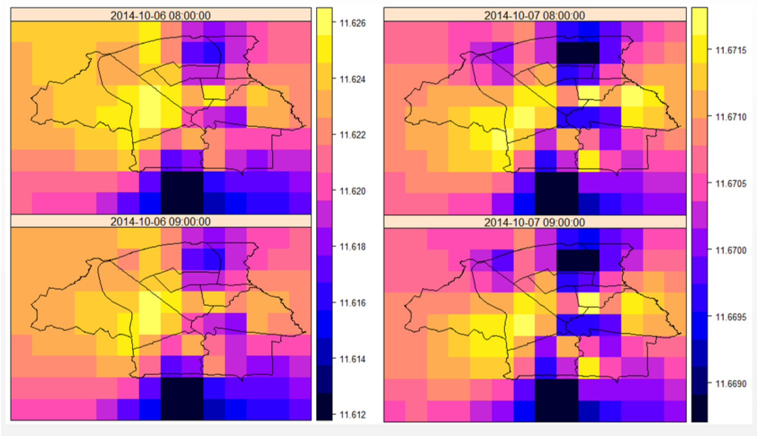


Figure 13Peak hour prediction for Day 6 and Day 7

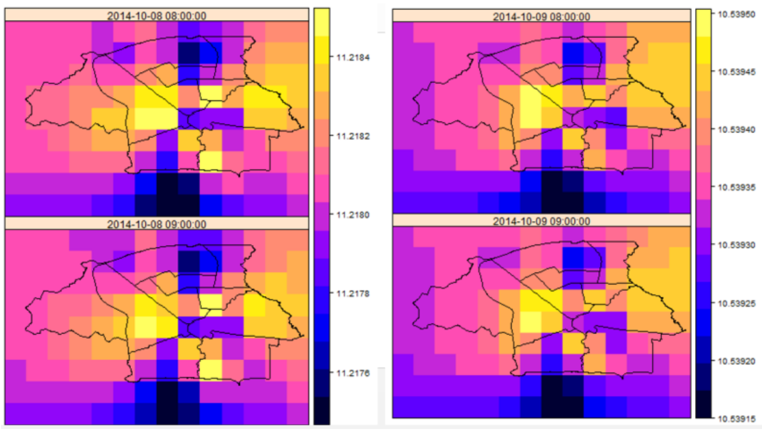


Figure 14 Peak hour prediction for Day 8 and Day 9

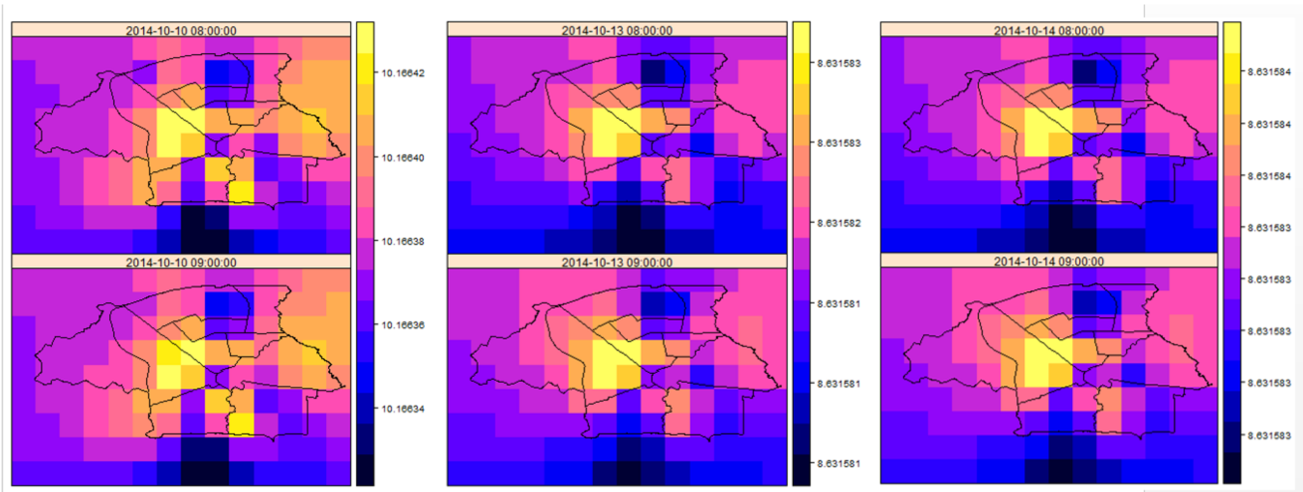


Figure 15 Peak hour prediction for Day 10, 13 and 14

An observation that could be pointed out is that there is a shift of high value of PM1 from North West to north east for consecutive days. However the region of south centre still remains to be lowest value of PM1 and the region in the North West near the centre tends to be the region with highest concerntration of PM1 for all days. This region corresponds to Boschdijk, a major traffic hub of the city and the sample predicted value for a date and the region is depicted below for better visualization.

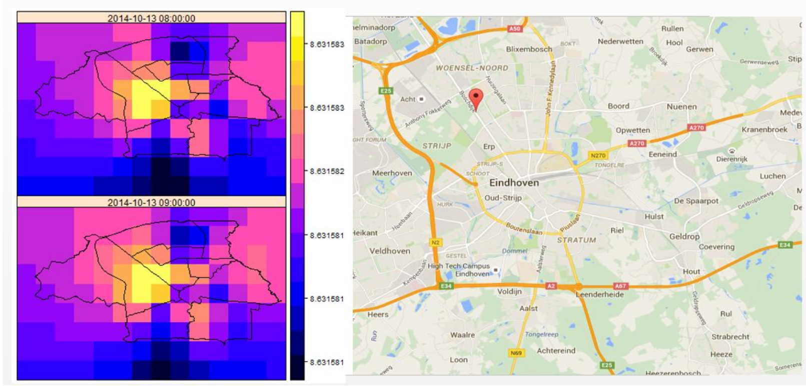


Figure 16 Boschdijk region (OSM) and corresponding predicted values at those region (left)

For cross validation, one of the stations is chosen at random and the values pertaining to it are removed and the process is repeated to obtain the kriged values. Then those values are compared to give comment on the accuracy.

By using the random function utility in R, 3 station numbers are removed one by one and then the predicted values and observed values to be compared. Alternatively, by using spatiotemporal library in R, cross validation could be easily adopted by the incorporated function present. However due to time limitations, this step was not able to be incorporated in the report and remains as a follow up approach for the future.

4 Conclusions and future works

In this study, analysis of average hourly data for PM1 and PM10 was done for a time span of 14 days for in and around Eindhoven. Utilizing various functionalities of spacetime library and gstat, spatio temporal kriging was performed on PM1 data to obtain predictions and forecasting. Fitting of empirical variogram was done using separable covariance model.

However the analysis was not done for PM10 data due to very poor modelling of the variograms and very high variability of values in the empirical variogram. A suitable solution to be explored could be taking the log values and then proceeding with the analysis. This could be incorporated in the future as a follow up. Furthermore estimation of "stAni" could not be done properly hindering the utility of metric model in PM1 data. Utilizing maximum likelihood estimation function, it could be possible to estimate the value and utilize the metric model. This could help in better commenting on the fitness for use of models more accurately.

The grid size for prediction was taken relatively smaller so as to incorporate efficiency in running the script. Making the grid more finer would lead to a better accuracy. Utility of "spatiotemporal" library could also be incorporated as it is devised specifically to model the spatio temporal variability of air pollutants. Cross validation of predicted values could not be carried out due to limitations of time. However the methodology is aptly described to carry it out and it will be a follow up to the study.

Finally, utilization of other ancillary information like land use, altitude could have been implemented as a tool for improving levels of prediction. Due to limitation of time and scope this couldn't be carried out in this study but may be included in future.

References

AiREAS. (2014). AiREAS. Retrieved July 19, 2015, from http://www.aireas.com/

Collett, R. S., & Oduyemi, K. (1997). Air quality modelling: a technical review of mathematical approaches. *Meteorological Applications*, *4*(03), 235–246. doi:10.1017/S1350482797000455

Desta, F. S. (2012). *Non-Stationary Linear Mixed Modelling of Air Quality*. University of Twente Faculty of Geoinformation and Earth Observation (ITC). Retrieved from http://www.itc.nl/library/papers_2012/msc/gfm/desta.pdf

EEA. (2014). AirBase database. Retrieved July 19, 2015, from http://acm.eionet.europa.eu/databases/airbase

Enkhtur, B. (2013). *Geostatistical Modelling and Mapping of Air Pollution*. University of Twente Faculty of Geoinformation and Earth Observation (ITC). Retrieved from http://www.itc.nl/library/papers_2013/msc/gfm/enkhtur.pdf

Gräler, B., Gerharz, L., & Pebesma, E. (2012). Spatio-temporal analysis and interpolation of PM10 measurements in Europe. *ETC/ACM Technical Paper 2011/10*, (November 2011), 37. Retrieved from http://acm.eionet.europa.eu/reports/ETCACM_TP_2011_10_spatio-temp_AQinterpolation

Iii, C. A. P., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., & Thurston, G. D. (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *The Journal of the American Medical Association*, *287*(9), 1132–1141. doi:10.1001/jama.287.9.1132

Jerrett, M., Finkelstein, M. M., Brook, J. R., Arain, M. A., Kanaroglou, P., Stieb, D. M., … Sears, M. R. (2009). A cohort study of traffic-related air pollution and mortality in Toronto, Ontario, Canada. *Environmental Health Perspectives*, *117*(5), 772–777. doi:10.1289/ehp.11533

Knox, A., Mykhaylova, N., Evans, G. J., Lee, C. J., Karney, B., & Brook, J. R. (2013). The expanding scope of air pollution monitoring can facilitate sustainable development. *Science of the Total Environment*, *448*, 189–196. doi:10.1016/j.scitotenv.2012.07.096

Murad, M. H. (2012). Main Air Pollutants and Myocardial Infarction- A Systematic Review and Meta-analysis. *The Journal of the American Medical Association*, *307*(7), 713–721. doi:10.1001/jama.2012.126

Pebesma, E. (2012). spacetime: Spatio-Temporal Data in R. *Journal of Statistical Software*, *51*(7), 1–30. Retrieved from http://www.jstatsoft.org/v51/i07/paper

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers and Geosciences*, *30*(7), 683–691. doi:10.1016/j.cageo.2004.03.012

Sterk, G., & Stein, a. (1997). Mapping Wind-Blown Mass Transport by Modeling Variability in Space and Time. *Soil Science Society of America Journal*, *61*(1), 232. doi:10.2136/sssaj1997.03615995006100010032x

Team, R. C. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://wwwr-project.org/

W.H.O. (2003). Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide. *Report on a WHO Working Group Bonn, Germany 13–15 January 2003*, (January), 98. Retrieved from http://www.euro.who.int/__data/assets/pdf_file/0005/112199/E79097.pdf