A geostatistical method for the analysis and prediction of air quality time series: application to the Aburrá Valley region

Master's Thesis for the Study Program
"Environmental Planning and Engineering Ecology"
at the Technische Universität München (TUM)

Summer Term 2016

Student: Juan Baca Cabrera
Supervisors: (1) Prof. Dr. Karl Auerswald (TUM)
(2) Prof. Dr. Uwe Schlink (UFZ)

A geostatistical method for the analysis and prediction of air quality time series: application to the Aburrá Valley region

Abstract

A geostatistical method was developed to analyze air pollution time series in the Aburrá Valley (Colombia) at different time scales (diurnal, weekly and yearly) and use this information for estimation of missing values or prediction purposes. The method was based on the calculation of omnidirectional semivariograms, by using time as coordinates in a geographical space, thus obtaining the air pollution variability associated to the different pollution cycles. The resulting semivariograms were valid until small lag distances. The kriging technique was afterwards applied for the estimation of missing data (interpolation) or the prediction of future events (extrapolation). The selected method was able to accurately capture the diurnal, weekly and monthly variability of PM_{10}, $PM_{2.5}$ and NO_2 in the Aburrá Valley. Satisfactory results were obtained by using the method for the prediction of $PM_{2.5}$ during days at which the Colombian Air Quality Norm was exceeded (Index of Agreement = 0.85; R^2 = 0.55)

Key words

Air pollution cycles, time series analysis, omnidirectional semivariogram, kriging interpolation

TABLE OF CONTENTS

FIGURES

TABLES

Abbreviations

AMVA	Área Metropolitan del Valle de Aburrá
EEA	European Environmental Agency
EPA	US Environmental Protection Agency
GAM	Generalized Additive Model
NO	Nitrogen oxide
NO_2	Nitrogen dioxide
NO_x	Mono-nitrogen oxides NO and NO_2
PM_{10}	Particulate matter with a maximum diameter of 10 μm
$PM_{2.5}$	Particulate matter with a maximum diameter of 2.5 μm
SO_x	Sulfur oxides
WHO	World Health Organization

1 Introduction

Air pollution is one of the biggest environmental concerns in urban areas. The adverse effects of air pollutants such as NO_x, particulate matter, CO, O_3, or SO_2 to the population (especially vulnerable groups like children, pregnant women, and elderly people) have been widely investigated and tested, especially regarding cardio-respiratory diseases (Hoek et al., 2013; Raaschou-Nielsen et al., 2013; WHO, 2006).

In the developed world air quality levels have improved throughout the last decades, what is mainly associated with the reduction of emissions due to better control technologies (EEA, 2015; EPA, 2015). However, there are still urban areas worldwide where air pollution levels have not yet been sufficiently controlled and this represents an enormous hazard for their population. This is mainly the case in fast growing cities in the developing world, where emissions and population grow at a fast rate (Cohen et al., 2005). It has been estimated that about 88% of all premature deaths associated to air pollution occur in low income countries (WHO, 2014).

Besides that, non-anthropogenic variables also have a significant influence on the pollution levels in cities. Urban areas located in terrains with a complex topography (e.g. valley bottoms, mountain slopes or mountain basins) often experience high pollution episodes, because of a limited removal of pollutants. This is caused by topographic barriers which reduce the effect of wind dispersion, thus resulting in higher pollution levels in inter-mountain regions when compared with what would be expected in flat locations (Rendón et al., 2014; Steyn et al., 2013). The boundary layer depth is also highly influenced by the effects of a complex topography. Extremely high pollution concentrations can often be observed in mountain regions during episodes of atmospheric stability (Anttila et al., 2015). This makes the monitoring, analysis, control and management of air pollution in these areas a complicated environmental issue for scientist, authorities and technicians.

Considering this, the analysis of air quality data and critical pollution episodes of cities in mountain regions is a very relevant topic in the atmospheric research. This is especially important for big cities in developing countries, which also face other problems like a rapid population growth, a partially uncontrolled urban development

and the pressure of multiple emission sources (high traffic, presence of industrial facilities, use of fuels with low quality, etc.). In the Andean Region in South America there are several big cities (over 1 million inhabitants) which are in this situation, both facing anthropogenic pressures that affect the air quality and are located in terrains with high topographical complexity. One of these cities is Medellín, in the Aburrá Valley Region in Colombia, and it will be the focus of the present study as long data series are available.

A special characteristic of air pollution is the temporal variability of pollutants. Studies in very different locations like Northern China, Augsburg, Massachusetts or Tenerife have shown that air pollution concentrations can highly vary (differences of 100% and more between different periods) when comparing different seasons, times of the day or days of the week (Baldasano et al., 2014; Gu et al., 2013; Ji et al., 2014; Padró-Martínez et al., 2012). The strong temporal variability of air pollutants represents a challenge for air quality management, because average values cannot characterize exposure levels of the population. Therefore a deep knowledge about the temporal patterns of air pollution in cities is needed for improved air quality and health management.

Recently, for the Aburrá Valley a complex temporal pattern for rainfall has been observed. There is a strong variability at monthly, daily and yearly rainfalls over the study region and there are evidences of phase locking among the different time scales. While for the period between May-September most of the precipitation is observed during the night (between 22:00-04:00), during the rest of the months the highest rainfall events occur in the afternoon hours between 13:00-17:00 (Figure 1). Effects of El Niño Southern Oscillation (ENSO), mesoscale convective systems and the diurnal cycle of insolation are probably combined to create this phenomenon (Poveda et al., 2015). It is possible that such patterns could also be present in data of air pollutants of Medellín, either caused by the rain pattern via outwash or caused by the same factors that also create the rain pattern. Therefore it is important to define an adequate methodological framework to obtain the temporal distribution of air pollutants at different time scales (daily, monthly, yearly cycles).

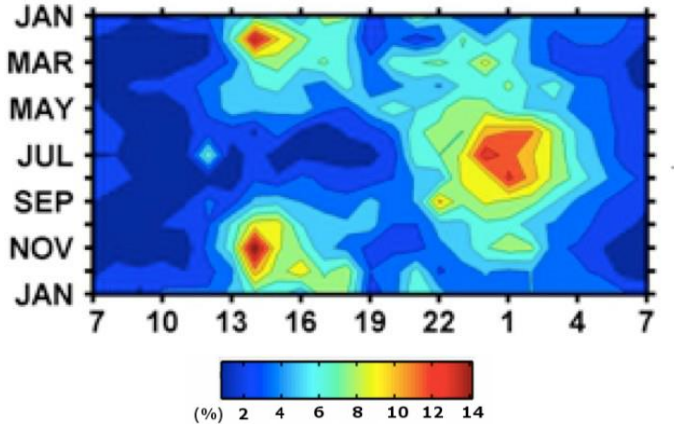


Figure 1: Diurnal and annual rainfall cycle in Aburrá Valley. Rainfall percentage for hourly time steps, during different months (Poveda et al., 2015)

Since 2003 the city of Medellín has been measuring the most important air pollutants at different locations. This information can be used to identify the temporal trends that describe air pollution levels in the city, considering the high temporal variability of these values. For this purpose multi-annual data series have to be analyzed. Time series have the special property that "[the] correlation between adjacent points in time is best explained in terms of a dependence of the current value on past values" (Shumway & Stoffer, 2011, p. 2). This means that temporally dependent measurements are not independent from each other (what is usually an assumption in classical statistics) and therefore statistical methods taking into account autocorrelation have to be applied for a successful analysis.

Statistical methods like distributed lag models, autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), Discrete Fourier Transform, Spectral Analysis, State-Space Models or Principal Components Analysis (PCA) have been widely used for time series analysis in different scientific disciplines (Shumway & Stoffer, 2011). Most of these methods make use of the temporal autocorrelation of the data for the analysis. Geostatistical analysis is also commonly applied for autocorrelated data, however mostly from a spatial perspective. That means geostatistical approaches usually account for the spatial autocorrelation of measurements. On the contrary, in the present investigation a geostatistical approach will be used for the time series analysis of air quality data (PM_{10}, $PM_{2.5}$ and NO_2) of the city of Medellín and the surrounding Aburrá Valley. The applied analysis utilizes two

time dimensions (day and hour of the day or day and month of the year) in place of the two spatial dimensions latitude and longitude. In other words, diurnal-annual plots (similar to Figure 1) or diurnal-weekly plots are analyzed like a spatial map. The main goal of the investigation is to detect the principal temporal patterns of air pollutants of the city, as well as to find the variability cycles affecting the pollution levels. This can be further used for prediction purposes or data reconstruction

Consequently, the objectives of the master project are the following:

i) Determination of temporal cycles of air pollution (PM_{10}, $PM_{2.5}$ and NO_2) and meteorological data for the city of Medellín and the Aburrá Valley, at daily, weekly and monthly scales.
ii) Development of a method for analysis of air pollution data series by using a geostatistical approach, which allows the estimation of missing values and the prediction of future events
iii) Application and evaluation of the performance of the method for selected periods of the air pollution datasets in the Aburrá Valley

2 Literature overview

2.1 Air pollution in Medellín and the Aburrá Valley

Medellín is one of the cities with the highest air pollution levels in Colombia (Ministerio de Ambiente, Vivienda y Desarrollo Territorial, 2010). Because of this, research institutions have been concerned with this topic during the last years. The city has an air quality monitoring network working since 2003, which measures major air pollutants in Medellín on an hourly basis. Based on the information obtained by this network and on the different analyses performed by universities and other institutions, several research studies have been published about air quality and air pollution in Medellín and its surroundings. Some of the most important studies are presented below.

Air pollution in Medellín follows clear temporal patterns, which principally depend on emission sources and meteorological conditions in the valley surrounding the city. Bedoya and Martínez (2009) found a clear daily pattern for the principle air pollutants of the city (PM_{10}, NO_x, SO_x), which can be associated with the traffic peaks throughout

the day. A yearly cycle of the air pollution was also observed in this study. The detected concentration oscillations within a period of 6 months might result from differences in the emission sources or from meteorological phases influencing the pollutants' concentration levels.

Zapata et al. (2015) identified a daily pattern for the pollutants PM_{10} and $PM_{2.5}$, with the highest peak in the morning hours and a second large peak during the evening. The effect of ENSO was also investigated to understand the yearly differences in the concentrations of particulate matter. Yearly variations in the concentrations of PM_{10}/$PM_{2.5}$ in Medellín are not significantly associated with the by El Niño phenomenon. However, a very clear annual cycle was observed for the period 2007-2013 (bimodal cycle), which is probably linked to the effect of complex interactions between meteorology and emissions.

Different kinds of exposure and vulnerability analyses have also been performed for the city of Medellín. Concentrations of air pollutants above thresholds recommended by health organizations have been observed in zones of Medellín affected by multiple emission sources, like the Medellín city center. These pollution levels increase the risk to suffer from respiratory or cardiac diseases, especially when considering vulnerable population strata who are continuously exposed to these hazards (Gaviria et al., 2012; Orduz et al., 2013). Research is still needed to determine the temporal factors affecting the exposure and vulnerability levels regarding air pollution.

Finally, Londoño et al. (2015) used a geostatistical approach for the spatial characterization of PM_{10} in Medellín. Monthly values were analyzed for a 5 year period (2003-2007) in 9 stations along the Aburrá Valley. Different variogram models were applied to simulate the spatial distribution of the pollutants in the study area and then utilized for a kriging interpolation. J-bessel and Hole effect were the best possible models obtained. This study demonstrates that a geostatistical approach can be helpful for the analysis of air quality data in Medellín The temporal autocorrelation of the data was however not considered in the analysis. Further efforts are needed to obtain detailed temporal patterns of air pollutants at different time scales (especially daily and yearly cycles and their interactions).

2.2 Geostatistics and time series analysis

Geostatistics and its tools –variography and kriging- have not been commonly used for time series analysis so far. However, Iaco et al. (2013) argue that linear Geostatistics are clearly linked to time series analysis, especially in the statistical approach where the Box–Jenkins methodology (Box & Jenkins, 1976), based on the autocorrelation function (ACF), is applied. Even though the variogram has not been generally used in time series analysis, the literature shows that it can be successfully applied for the analysis of stationary (Gevers, 1985; Ma, 2004) and non-stationarity (Cressie, 1988) time related data. This is the case because variograms can properly describe stochastic processes.

A geostatistical approach for time auto correlated data has several advantages over traditional time series analysis methods. A detailed literature review about this topic by Iaco et al. (2013) suggested the following advantages: 1) the variogram is able to describe a wider class of stochastic processes than the ACF (e.g. second-order stationary stochastic processes); 2) unlike the estimation of covariance, it is not necessary to know the expected value of the stochastic process to calculate the variogram. This assures the unbiasedness of its estimator; 3) geostatistics is useful for the identification of trends and periodicities of the data, because of the capacity of the variogram to capture the details of its structure; 4) the variogram can assess the different scales of variation which are typical for time series; and 5) geostatistics is able to handle time series with missing data better than other methods of time series analysis.

During the last years geostatistics has been applied more frequently for time series analysis in various scientific disciplines. The multiple advantages of variography have contributed to its propagation in scientific research and its application under different methodological approaches.

In the field of hydrology variograms have been used to estimate the variation of the uncertainty of streamflow rating curves over time. For that, discharge estimations at multiple time stages were analyzed through variography (Jalbert et al., 2011). The elements that describe a variogram (nugget, sill and range) were used to represent the small scale variation, the variance of the random variable and the long term variation of

the uncertainty. This approach was considered robust, due to the capacity of variograms to capture trends in the data and different temporal correlation structures.

Variography has also been applied in hydrological research to detect and attribute changes in data series, in the context of a changing climate. Change detection and attribution is a complex task for non-linear systems, such as hydrometeorological systems. The methodology used by Chiverton et al. (2015) was based on the calculation of empirical variograms and its application to moving windows in a river flow time series. This allowed the identification of changes and its adequate attribution (e.g. changes caused by meteorological forcing). Different relevant values like seasonality, measurement error and sub-daily variability were obtained by using this method.

Another scientific field where the use of variograms plays an important role is the reconstruction and interpretation of very long time series. Enzi et al. (2014) reconstructed time series of extraordinary snowfall episodes in Italy over 300 years, by applying a geostatistical approach. They used variography to determine the temporal autocorrelation of snow recurrence into the time domain. By choosing a hole-effect model for the empirical semivariogram, the non-stationarity of the data was reflected. This method showed the internal structure of the time data very accurately. Meanwhile, the challenge of analyzing misaligned irregular time series has also been resolved through the application of variograms. In a research study of Greenland ice core data (Doan et al., 2015) empirical variograms were used to integrate data at different temporal resolutions and observe the variability of the time series. In result of this procedure consistent ice core data series were obtained.

Furthermore, variogram analysis has been used to handle complex data at multi spatiotemporal scales. Computer simulations of chemical catalytic reaction face the problem that different time scales must be coupled into a consistent model. The research done by Gur et al. (2016) shows the role of variography in addressing this issue. A wavelet-based model was developed, which allows the temporal up- and downscaling of data. Empirical variograms were used to determine data sets with convergent statistics, observe the autocorrelation of data series, estimate the temporal variation of the data and define the length of cycles for prediction purposes. As a result of this study, a robust model for simulation of multi temporal scale catalytic reactions was obtained.

Finally, in the specific field of atmospheric research, Iaco et al. (2013) proposed a geostatistical method for the estimation of missing values and the prediction of future pollution events for NO_2 time series. Hourly NO_2 values measured over one month were used to calculate empirical variograms that represent the temporal variability of the data at different time scales (hourly, daily, weekly variation). The kriging technique was then applied, using the selected hole effect variogram model, for estimation of missing values (data interpolation, imputation) or for prediction purposes (extrapolation) of the NO_2 hourly values. This research showed the flexibility of kriging for reconstruction of air quality time series and its accuracy for predicting air pollution events.

This brief literature review highlights the multiple application fields of geostatistical techniques for time series analysis and its flexibility for addressing research topics like: calculation of time series uncertainty, estimation of missing values and prediction of future events, integration of data at multi temporal scales, reconstruction of very long time data sets, detection and attribution of changes. In consequence, a geostatistical approach was considered as reliable for the analysis of air quality time series and was applied for the air pollution data of the Aburrá Valley.

3 Materials and methods

3.1 Area of study

The city of Medellín is located alongside the Aburrá Valley, which is a narrow valley in the Colombian Andean mountains. Together with another eight municipalities they form the Metropolitan Area of the Aburrá Valley (AMVA). This metropolitan region extends over 60 km and has a variable width between 10 and 20 km. Its coordinates are 6.0° - 6.5° latitude north and 75.2-75.7° longitude west. The AMVA has a total population of 3.594.198 and Medellín is its most populated city with a population of 2.486.723 (DANE, 2007).

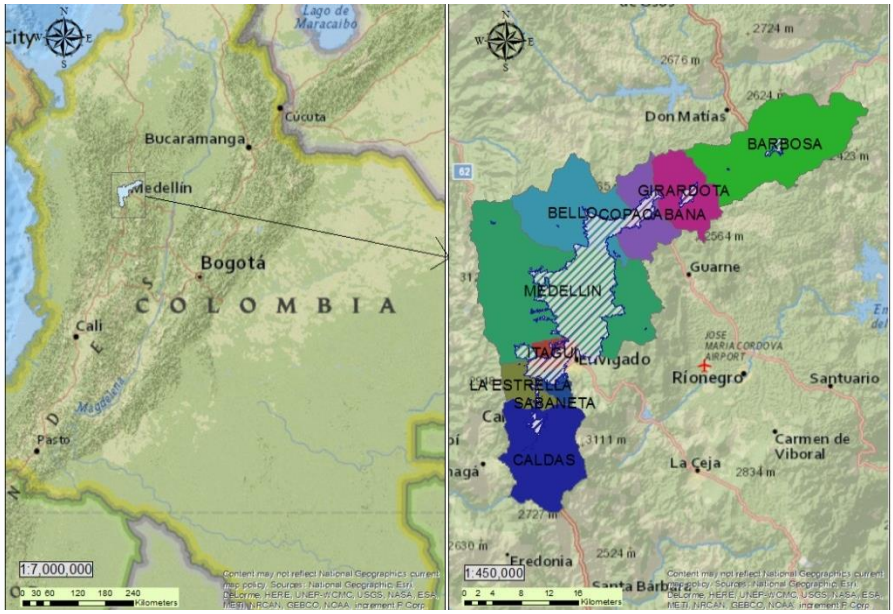


Figure 2: Location of the Metropolitan Area of Aburrá Valley (AMVA) and the city of Medellín. The hatched area on the map represents the urban areas of AMVA (mainly located in Medellín municipality)

The Aburrá Valley has a medium height of around 1500 m, with variations from 1400 m in Barbosa (North) until 1800 m in Caldas (South). The valley is surrounded by mountains with a maximum height of more than 3000 m and plateaus between 2000-2600 m. The valley is dominated by the basin of the Medellín river, which crosses the city of Medellín from south to north. The orographic conditions of the region favor the development of thermal inversion layers in the valley during the early morning hours, especially in Medellín downtown (Laverde, 1988). Under such conditions air pollution episodes can occur, due to a low dispersion of pollutants produced with increasing stability.

A distribution of precipitation with two peaks can be observed in the research area: two rainy seasons between April-May and October-November and two drier seasons the rest of the months. The bimodal regime shapes the monthly differences in the rainfall amount, with minimum monthly values around 50mm/month and maximum values over 200mm/month. The mean temperature is around 22°C, with low seasonal variability (less than 1.5°C between maximum and minimum monthly temperatures). Temperature variability can be observed because of height differences alongside the Aburrá Valley and due to the day-night regime.

The AMVA is a highly dynamic metropolitan region with a continuously growing population. Between 1993 and 2005 the region had a population increase of 25%, while for the period 2005-2012 an increase of 10% was estimated (Gobernación de Antioquia, 2013). Most of this increase corresponds to the city of Medellín, where over 65% of the total population is concentrated. Furthermore, the AMVA is the second most important industrial region of Colombia, with most of the industrial facilities concentrated in the south of the valley (Betancur et al., 2001). Finally, the vehicular fleet of the region has been growing at very fast rates lately. Between 2000 and 2011 the number of vehicles increased from 300,000 to 800,000. Vehicles and industries are the principle sources of air pollutants in the Aburrá Valley, what can be associated to the socioeconomic development of the region during the last years (AMVA, 2012).

As a consequence of its industrial characteristics and its complex orography, the urban areas of the Aburrá Valley are constantly faced with air quality problems. The principle pollution problems are caused by PM_{10} and $PM_{2.5}$, mostly in the municipality of Medellín and other municipalities located in the south. These two pollutants have exceeded the values considered as healthy by air quality guidelines during the last years and are therefore an issue of concern for the environmental authorities (UNAL, 2015; UPB, 2013).

3.2 Data description

The Air Quality Network of The Metropolitan Area of the Aburrá Valley (REDAIRE[1]) has been in charge of air quality monitoring in Medellín and the AMVA since the year 2003. It consists of 22 fixed and 1 mobile measuring stations (Figure 3), combining automatic and semiautomatic sampling techniques. The stations are distributed alongside the Aburrá Valley, with 14 of those stations being equipped with automatic samplers for the monitoring of air pollutants (PM_{10}, $PM_{2.5}$, NO, NO_2, O_3 and SO_2) and meteorological variables (temperature, relative humidity, solar radiation, wind speed, wind direction, precipitation and atmospheric pressure). The automatic monitoring stations provide information for all variables of interest at hourly intervals, what allows a detailed analysis of air pollution patterns at different time scales.

[1] Additional information about the air quality network and its measurement procedures are available on: http://www.metropol.gov.co/CalidadAire/Paginas/redaire.aspx

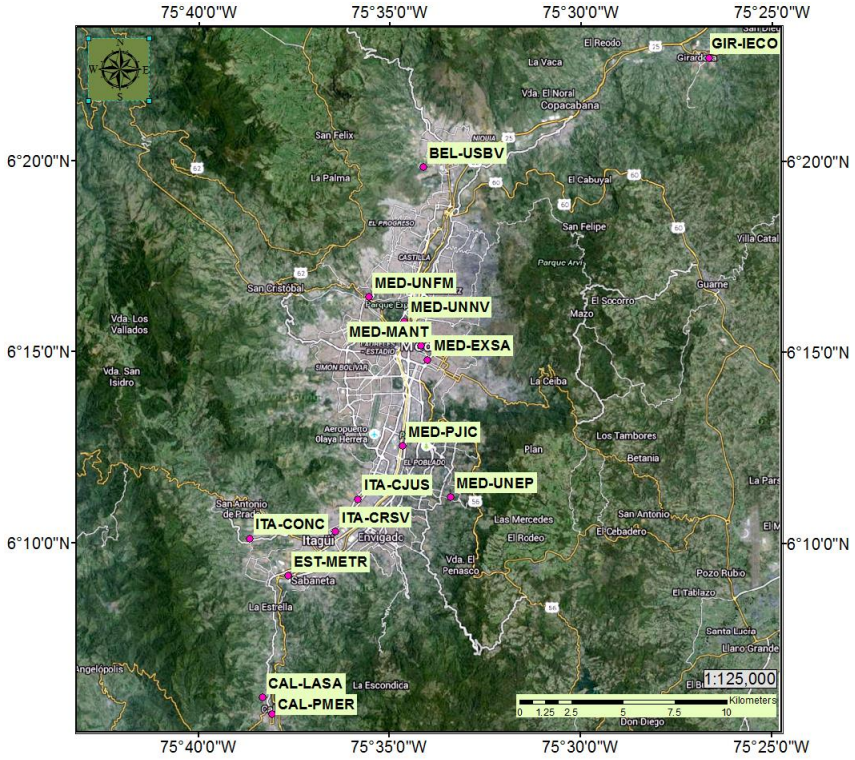


Figure 3: Location of the automatic monitoring stations of REDAIRE alongside the Aburrá Valley

All data used for this master's thesis have been sampled and validated by REDAIRE, following strict quality assurance and quality control procedures. From the entire datasets, following criteria were used for the selection of the data included in the study:

- Only data sampled by automatic stations was included
- PM_{10} and $PM_{2.5}$ were selected due to their major contribution to air pollution problems in the Aburrá Valley. NO_2 has also been selected because it is both a primary and a secondary air pollutant and therefore its temporal variability is of special interest. All other pollutants were excluded
- The automatic stations have been recording data for all air pollutants since 2012 (before that only PM_{10} was registered). For that reason data recorded from 2012 onwards was used.

Consequently, hourly values of PM_{10}, $PM_{2.5,}$ NO_2 (expressed in units of mass concentration of pollutants in $\mu g \cdot m^{-3}$) and meteorological variables measured by the automatic stations of REDAIRE for the period October 2012-September 2015 were used for this investigation. Only stations with complete air quality data series for the

entire period were considered, to avoid a temporal bias in the analysis. Following this condition, samples collected by the station EST-METR were removed from the analysis.

3.3 Methodology

3.3.1 Exploratory data analysis

Exploratory data analysis was performed as a first approach for the determination of temporal trends of air pollutants and meteorological variables. Daily average values were examined over the entire time period to detect the most general patterns associated with short and long term variability. A Generalized Additive Model (GAM) was used for the calculation of smooth trends for the entire study period, based on the daily values. For this purpose the mgcv package (Wood, 2011) was used, which automatically finds the most appropriate smoothing fit for the trends by using natural splines.

Diurnal, day of the week and monthly variation plots were generated for the area of interest. Additionally, contour plots were created for a comparison of hourly profiles of pollutants and meteorology during different months of the year. The interactions between the two different time scales could also be visualized through this method.

The results of this analysis allowed for initial conclusions about the temporal cycles dominating the air pollution behavior in the Aburrá Valley and the city of Medellín, as well as the principal factors influencing its variation. The complete exploratory data analysis was performed with the statistical software R (R Development Core Team, 2015). The R packages 'openair' (Carslaw & Ropkins, 2012) and 'lattice' (Sarkar, 2008) were used for the generation of special plots.

3.3.2 Variogram analysis

Based on the results of the previous section, a variographic analysis for the air pollutants time series (PM_{10}, $PM_{2.5}$ and NO_2) was performed. The basic element of variography is the empirical semivariogram, which according to Cressie (1993) is defined as

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)}(Z(s_i) - Z(s_j))^2 \ , \qquad (1)$$

where N(h) denotes the set of pairs of observations i and j separated by distance h, |N(h)| is the number of distinct elements of N(h) and Z is the value observed at a point s.

Since the variogram is usually used for spatial analysis, all terms are related with distances between points in the geographical space. However, for this research the distance h represents the **time** separation between 2 measurements of the data series, thus allowing an analysis of the variability of pollutants in the temporal (not spatial) dimension.

For a geostatistical analysis of such temporal data the empirical semivariogram was used to estimate the theoretical semivariogram, which is valid for all possible time distances h. The theoretical variogram is usually described by three parameters:

nugget: y-intercept of semivariogram, represents the small scale variation or the measurement error of the data,

sill: value at which the variogram levels off and when the lag distance tends to infinity,

range: lag distance at which the sill is reached; autocorrelation is presumably 0 after this point.

There are many different theoretical semivariogram models proposed in the literature (Journel & Huijbregts, 1978). For the analysis of air pollutants in the Aburrá Valley two types of models were tested: spherical (equation 2) and Gaussian model (equation 3). The chosen models are some of the most typically used models in geostatistics and can be applied in multiple research questions due to their flexibility. The formulas for the selected models are presented below:

Spherical model (||h|| being the Euclidean distance between two points in $\mathbb{R}^1$, $\mathbb{R}^2$, or $\mathbb{R}^3$) with nugget c_o, sill (c_o+c_s) and range a_s:

$$\gamma_h = \begin{cases} 0, & h = 0, \\ c_0 + c_s\{(3/2)(\|\boldsymbol{h}\|/a_s) - (1/2)(\|\boldsymbol{h}\|/a_s)^3\}, & 0 < \|h\| \le a_s, \\ c_0 + c_s, & \|h\| \ge a_s, \end{cases} \qquad (2)$$

Gaussian model (||h|| being the Euclidean distance between two points in $\mathbb{R}^1$, $\mathbb{R}^2$ and $\mathbb{R}^3$) with nugget c_o, sill (c_o+c_g) and range a_g:

$$\gamma_h = \begin{cases} 0, & h = 0, \\ c_0 + c_g\left\{1 - \exp(-(\|\boldsymbol{h}\|/a_g)^2\right\}, & h \neq 0, \end{cases} \quad (3)$$

Multiple methods (Cressie, 1993) can be used to obtain a best model fit, when selecting the theoretical semivariogram and its parameters (nugget, sill and range). For this research study a combination of automatic and "fit by eye" methods were used. The automatic fit was based on the weighted least squares method, which minimizes the sum of squared residuals with different weights, depending on the number of data pairs and the lag distances.

The eyefit method was performed for improving the automatic fit outputs. This was done when the parameters of the theoretical semivariogram from the automatic fit highly differed from the expected results, based on the knowledge of the data. For this method an automatic fit was first performed and its results were taken as start values to manually fit the most accurate model parameters. The R package gstat (Pebesma, 2004) was used for the estimation of the empirical and theoretical semivariograms. RMSE values were calculated to test the accuracy of the model outputs.

3.3.3 Kriging of PM_{10}, $PM_{2.5}$ and NO_2 temporal data

Kriging is an interpolation method commonly used in geostatistics. It relies on the knowledge of the autocorrelation of observed data to make inferences on unobserved values of a random process (Journel & Huijbregts, 1978; Matheron, 1963). The kriging method uses the theoretical semivariogram to predict a value Z_0 based on the weighted observations Z at the sample points i by following Eq. (4)

$$\hat{Z}_0 = \sum_{i=1}^{n} \lambda_i \times Z_i, \quad (4)$$

where λ_i are unknown real coefficients (weights). In the case of Ordinary Kriging the weights are obtained by solving the following kriging system:

$$\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} & -1 \\ \gamma_{21} & \cdots & \gamma_{2n} & -1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} & -1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{n0} \\ 1 \end{bmatrix}, \quad (5)$$

where γ_{ij}=0.5Var(Z_i - Z_j), γ_{i0}=0.5Var(Z_i – Z_0) and μ is a Lagrange multiplier. In Ordinary Kriging the best predictor $\hat{Z}_0$ is obtained by minimizing the mean square prediction error. In the case of time series i and j represent the different time points in the temporal dimension.

Semivariogram analysis and kriging were mainly used to analyze and interpolate the interaction of the diurnal and seasonal variation of air quality parameters. This analysis was based on the long-term mean averages per month and hour to derive the general pattern. The cyclic nature of diurnal and seasonal variation, which has no predefined starting point, had to be considered in the analysis to avoid artifacts at arbitrarily defined starting points like the end of the previous day or the end of the previous year. To this end, the data were recycled prior to analysis. After analysis and kriging, the recycled parts were deleted again to arrive at a kriged representation of the diurnal and seasonal variation without distortions and artifacts at the margins of the day and of the year. This is illustrated in Figure 4. It has to be noted that the semivariance has to become zero at a lag of 12 months in the seasonal domain and at a lag of 24 hours in the diurnal domain due to the use and recycling of long-term averages for each month and hour.

For the joint analysis of the seasonal and the diurnal variation, the months were used as y coordinates and the hours of the day as x coordinates (Figure 4). This leads to two problems. (i) In contrast to geographical coordinates, both coordinates have different units. (ii) It is rather unlikely that the autocorrelation length during a day is identical to the autocorrelation length during a year. To account for these problems, anisotropic semivariograms were calculated that either strictly followed the x coordinate or the y coordinate. For kriging, omnidirectional semivariograms were necessary. These were obtained by scaling the x coordinate and the y coordinate in a way that both directional semivariograms became identical at least up to the maximum lag that was needed during kriging. It turned out that both directional semivariograms became near identical,

if the diurnal coordinate was hourly scaled while the seasonal coordinate was monthly scaled (see Results). The omnidirectional semivariogram should not be used for other purposes than kriging and may only be interpreted at short lags (maximum distance of 4). Its interpretation becomes especially invalid at a lag of 12, for which the semivariance in the seasonal domain must be zero while a large semivariance can be expected in the diurnal domain because this lag would include the difference between midday and midnight.

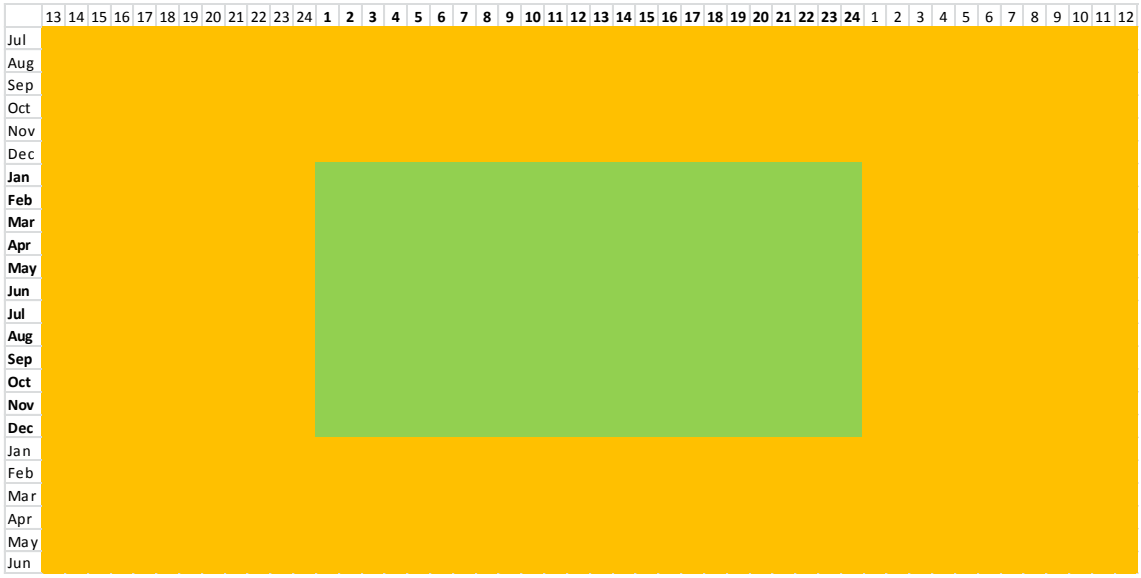


Figure 4: Recycling of data to avoid artifacts at the start and end of the diurnal cycle and of the seasonal cycle. The yellow area displays the recycled data taken from the green area. The data of both, the green and the yellow area were then used for semivariogram analysis and kriging. The x coordinates represent the hours of day and the y coordinate the months of a year.

The same analysis and the developed procedure for the construction of omnidirectional semivariograms were repeated for kriging at smaller time intervals, i.e. for the joint analysis of the weekly and diurnal pollution cycles. In this case the x coordinates corresponded to the hour of the day and the y coordinates to the days of the week (Monday-Sunday). This analysis was performed to obtain the pollution variability within a week cycle, what is needed for prediction or estimation of missing values at day-distances.

Additionally to the study of the interactions between, diurnal, weekly and yearly variation of air pollutants, kriging was used as a tool for 1) estimation of missing values (interpolation); and 2) prediction of air pollution concentrations in the Aburrá Valley (extrapolation). Since the calculated omnidirectional semivariograms were only valid at small lag distances in x (hours) and y (months or days of the week) directions, the prediction or estimation of missing values was performed for short time windows (i.e. adjacent months or days of the week). The calculation of the target values through the developed method is shown schematically in Figure 5.

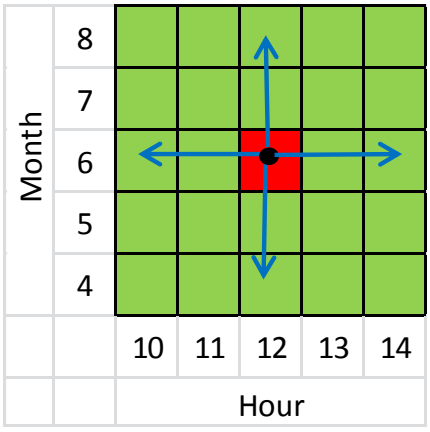


Figure 5: Scheme of the values used for the kriging procedure. The green cells represent the observed values, the red cell the missing data and the arrows the maximum distance considered for kriging

For the calculation of every missing value, kriging was performed by only using data at very short distances (in this representation a maximum distance of 2). In the example, the missing value at hour 6/month 12 is calculated based on the data of months 4-5 and 7-8 and the hours 10-11 and 13-14. Notice how the kriging method makes use of both the diurnal and the yearly pollution cycles. This type of procedure is considered adequate for the modelling of air quality data of the Aburrá Valley, due to similar behavior of the diurnal pollution cycle at different months of the year and days of the week (see Results 4.1.3 and 4.3.2). In study regions with high variability of the diurnal cycle this method would not be effective. In the case of the prediction of values an extrapolation was actually performed. Only data observed prior to the target values was used for the kriging procedure. Similarly to the estimation of missing data, short time distances were used for the calculation.

Estimation of missing data and prediction was performed for all months of year 2014. This year was chosen because of data availability for all months and presence of pollution events. Hourly average values for every month were used as input data. The results were afterwards compared with the observed values to test the accuracy of the model. In addition, specific days in February-March 2014 and 2015 (periods where exceedances of the Colombian Air Quality Norm were observed) were also simulated. Model evaluation was performed based on the following criteria: R^2, RMSE and Index of Agreement. Additionally, the distribution of the residuals was tested for normality. The kriging interpolation and the statistical analysis were performed using the R package gstat

Before the semivariograms of air pollutants were calculated and kriging interpolation was performed, the pollution datasets were transformed to obtain normally distributed data, which is required for an optimal applicability of geostatistical methods. PM_{10}, PM_{25} and NO_2 values corresponding to hour, day of the week and month averages were checked for normality based on histograms and the Lilliefors-Test. The Box-Cox method (Box & Cox, 1964) was used to select the best possible value for the transformation. A fourth root power transformation was applied for all air pollutants, with the resulting values showing normal distribution (see Appendix 7.3 for histograms and normality tests). To avoid ambiguity, the fourth root transformed data will be called rPM_{10}, $rPM_{2.5}$ and rNO_2 in the Results section. After kriging the results were back transformed to obtain the definite concentrations of air pollutants.

4 Results and discussion

4.1 Air pollutants and meteorology explorative analysis

4.1.1 Seasonal variations

A yearly periodicity in the air pollution data was found for the Aburrá Valley. During the study period the pollutants PM_{10}, $PM_{2.5}$ and NO_2 continuously showed their highest daily averages during February-March. After these pollution maximums the concentrations decreased until the yearly minima around June-July and then started to increase again to complete the yearly pollution cycle (Figure 6). This yearly periodicity

was more pronounced and stable for PM_{10} and $PM_{2.5}$ than for NO_2. For PM_{10} and $PM_{2.5}$ the described periodicity practically did not show any alterations during the three years of analysis. Meanwhile, the pollution cycle of the pollutant NO_2 changed over the years. In 2013 the differences between seasonal maximum and minimum values were small and the expected trend was difficult to identify. On the contrary, during the years 2014 and 2015 NO_2 concentrations followed the same pollution cycle as PM_{10} and $PM_{2.5}$ and the yearly peaks were considerable higher than the minimum daily averages (around 4-times higher both in 2014 and 2015).

During the pollution peaks in February-March the Colombian Norm of Air Quality for $PM_{2.5}$ (50 [$\mu g \cdot m^{-3}$] daily average) was exceeded more than 10 times in 2014 and 2015, while the PM_{10} Norm (100 [$\mu g \cdot m^{-3}$] daily average) was only exceeded in rare occasions (3 times during the entire study period). In the case of NO_2 the seasonal peaks did not represent a danger for the population, since the observed levels were considerably lower than the suggested threshold (150 [$\mu g \cdot m^{-3}$] daily average).

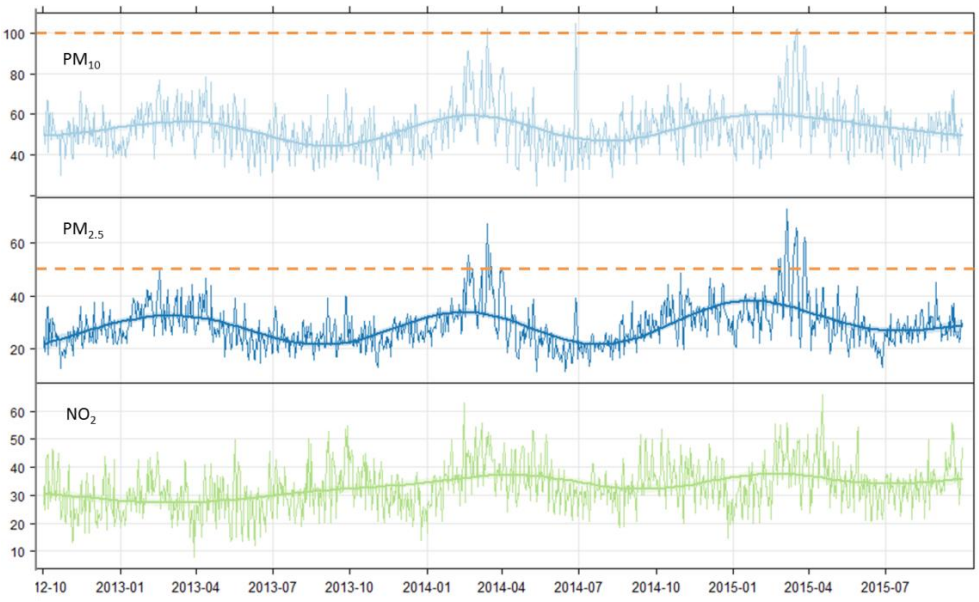


Figure 6: Seasonal distribution of PM_{10}, $PM_{2.5}$ and NO_2 in the Aburrá Valley. Daily averages [$\mu g \cdot m^{-3}$] and smooth curves (Oct. 2012 - Sep. 2015). The dashed lines show the Colombian Air Quality Norm for PM_{10} and $PM_{2.5}$

Additionally to the pollution cycles presented in Figure 6, the corresponding statistics to the seasonal trends are summarized in Table 1. Twice as much of the variability of $PM_{2.5}$ as of the two other pollutants was explained though the yearly periodicity. The

concentrations of PM_{10} and NO_2 were also influenced by this seasonal effect, but to a lesser degree than $PM_{2.5}$

Table 1: Output statistics of the Generalized Additive Model (GAM) for PM_{10}, $PM_{2.5}$ and NO_2. Calculations based on daily averages [$\mu g \cdot m^{-3}$].

Pollutant	Estimated degrees of freedom	Adjusted R-squared	Deviance explained [%]
PM_{10}	8.78	0.15	15.9
$PM_{2.5}$	8.89	0.31	31.5
NO_2	8.64	0.15	15.3

The yearly periodicity of the meteorological parameters solar radiation, wind speed, temperature and precipitation was not homogenous among them and similarities in comparison with the air pollution periodicity were difficult to identify. Solar radiation was the meteorological variable which showed the most similar periodical behavior compared with the air pollution patterns. Inversely to the pollutants' concentrations, the radiation peaks occurred during the months of June-July and the minimum values around January-February. For wind speed and temperature a consistent seasonal trend over the 3 years of analysis was difficult to identify, because of higher short term variability than long term periodicity. Finally, rainfall peaked around May and October-November. However, the yearly periodicity of rain was not similar to the one of air pollutants in the Aburrá Valley.

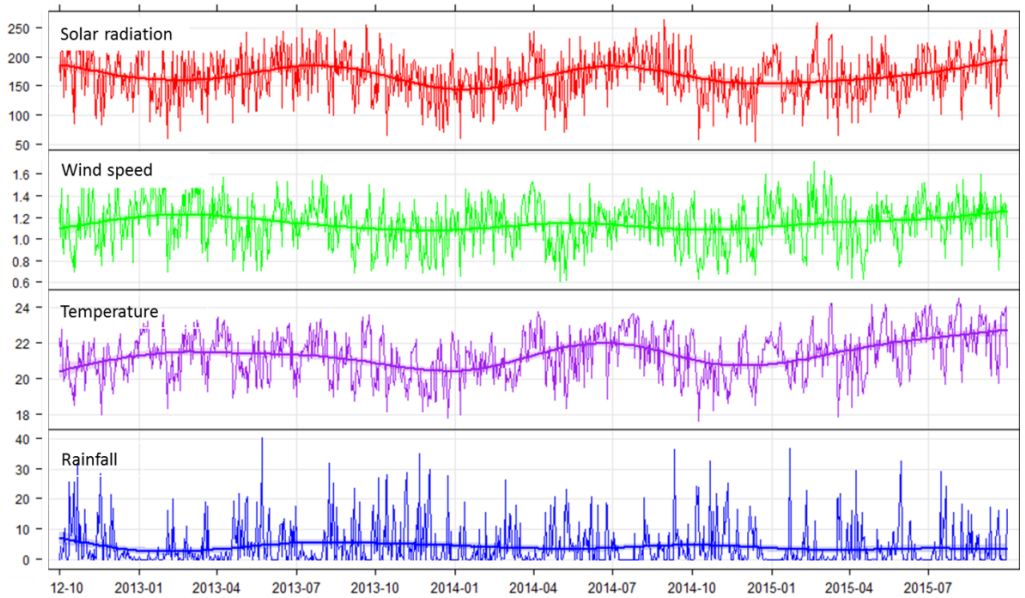


Figure 7: Seasonal distribution of meteorological variables in the Aburrá Valley. Daily averages and smooth curves (Oct. 2012 - Sep. 2015): solar radiation [$W \cdot m^{-2}$], wind speed [$m \cdot s^{-1}$], air temperature [°C] and rainfall amount [$mm \cdot day^{-1}$]

According to the orography of the study region, the main winds came from the North-East direction and crossed to the South alongside the Aburrá Valley (Figure 8). Under favorable atmospheric conditions this kind of winds helped to disperse accumulated air pollutants in the valley. The wind roses varied little among the different months of the year. In all cases, the predominant winds came from North-East direction, while winds coming from the South, South-West and West directions had the lowest frequency. Differences in wind speed were also low, with monthly average values ranging between 1.14-1.22 [$m \cdot s^{-1}$]. The temporal trends observed for air pollutants could not be found for wind speed/direction at a monthly scale.

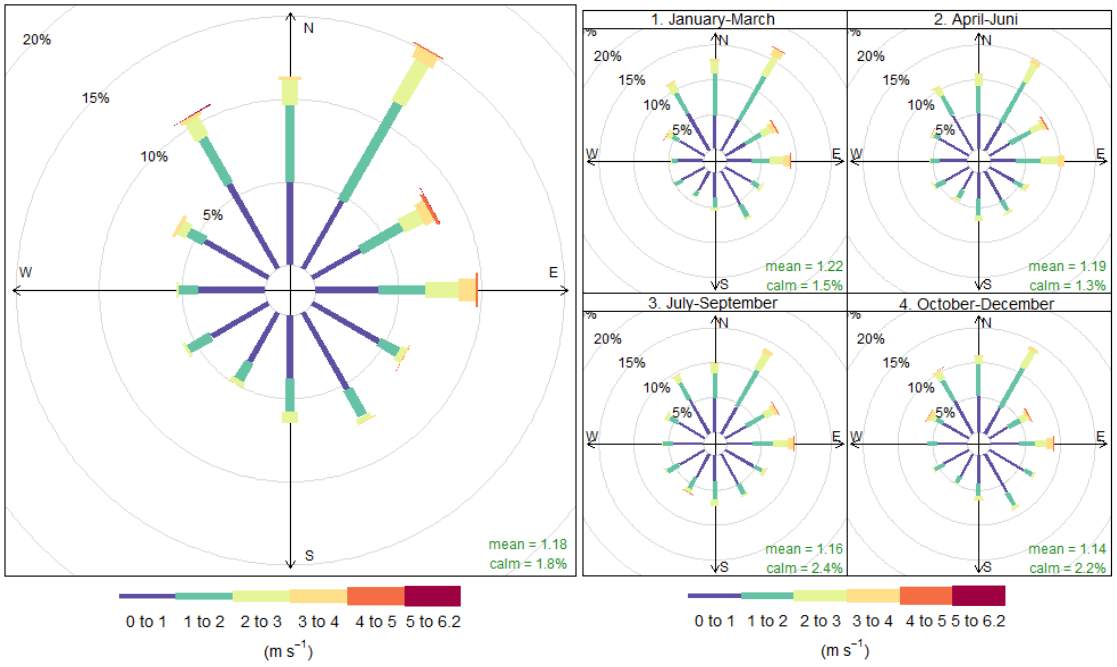


Figure 8: Wind roses for the Aburrá Valley. The graph on the left presents all data for the study period, while in the graph on the right the data has been divided into the different months of the year

4.1.2 Hourly, weekly and monthly distribution

All pollutants exhibited a maximum peak value around 07:00-10:00 and a second, less pronounced peak in the evening hours (18:00-21:00) (Figure 9). This bimodal behavior was associated with the morning and evening traffic flows in the Aburrá Valley, as

observed by Zapata et al. (2015). During the afternoon pollutants were generally reduced, which can be partly explained by less stable atmospheric conditions (solar radiation and wind speed maximum values). Minimum values during the night matched the minimum in traffic flow at late hours.

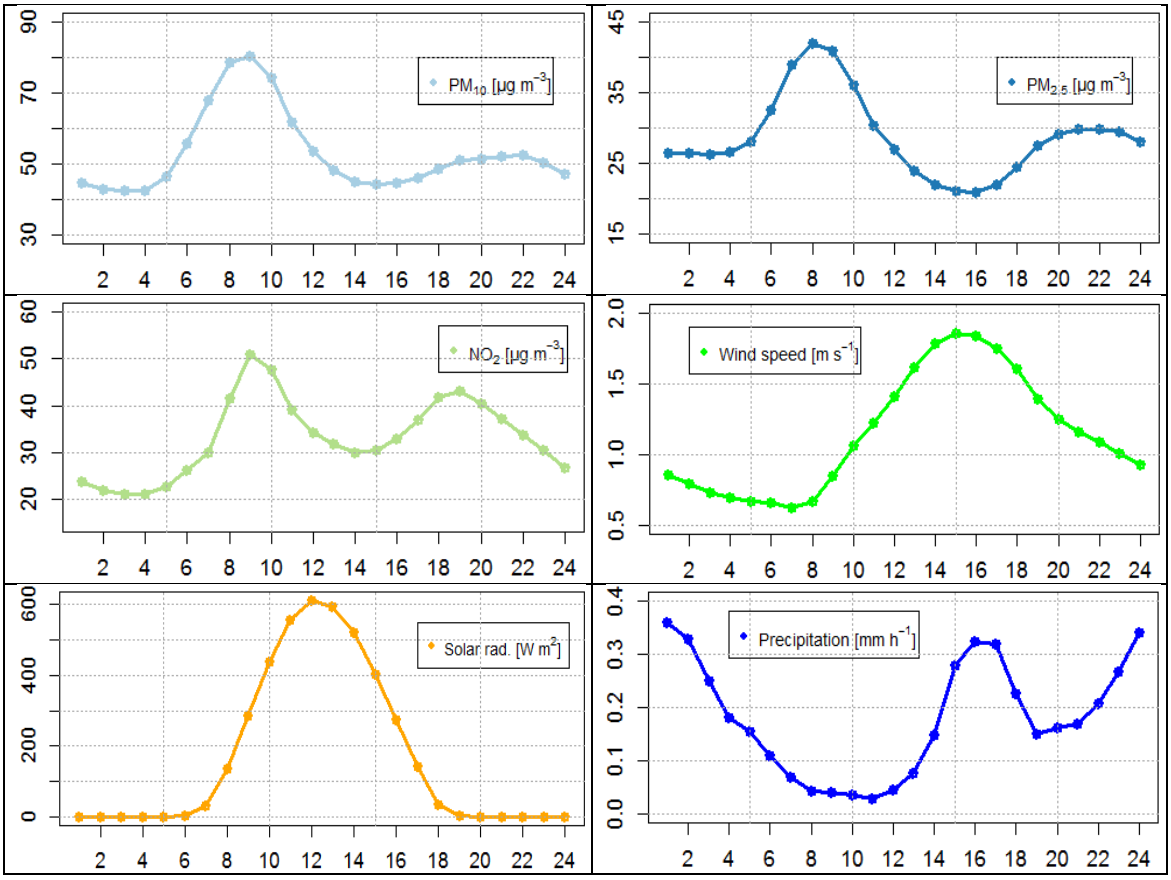


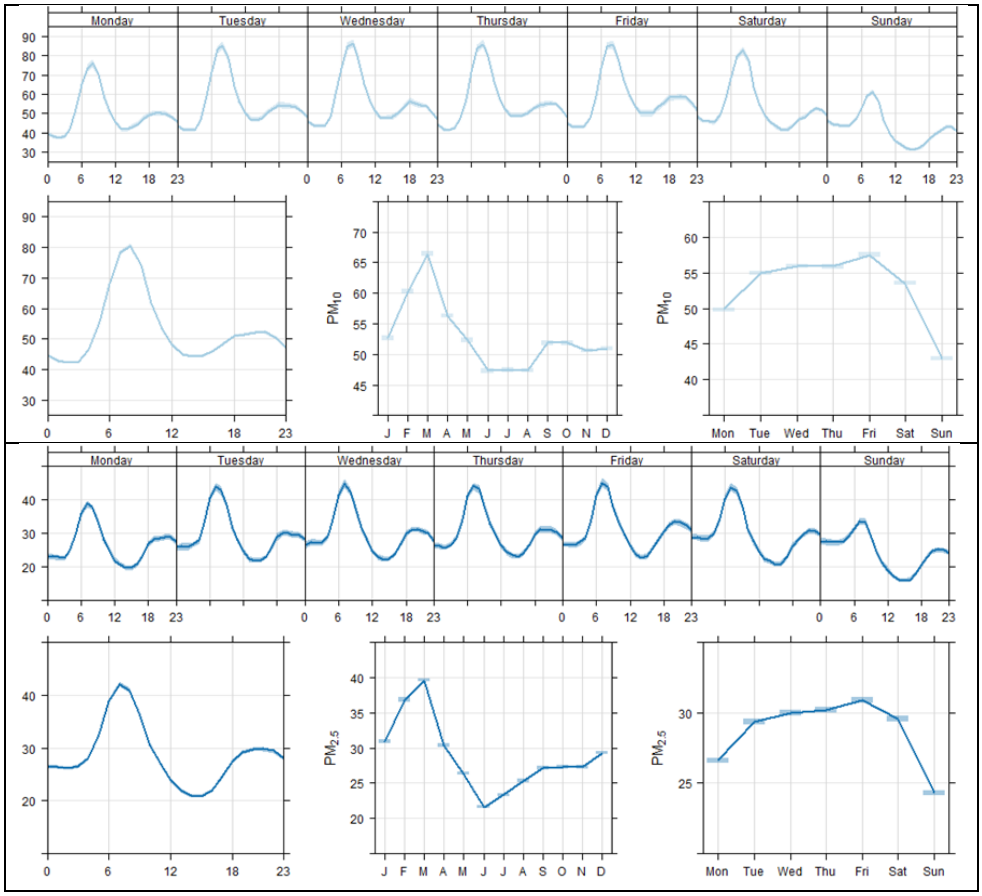
Figure 9: Hourly distribution of air pollutants and meteorological variables in the Aburrá Valley

Wind speed and temperature showed the expected daily cycle, with maximum hourly values after midday and daily minima just before sunrise (06:00). Because of the latitude of the study region near the equator this daily profile presented low variability around the year. On the contrary, the rainfall profile showed a pronounced pattern. Rain events occurred mainly during the afternoon hours (15:00-17:00) and around midnight (23:00-02:00). This behavior is typical for precipitation in the Aburrá Valley region.

The comparison between hourly, weekly and monthly distributions of PM_{10}, $PM_{2.5}$, NO_2 (Figure 10) allowed a better understanding of the interactions between the different time scales. The daily profiles of all pollutants showed a bimodal regime with a morning and an evening peak. This profile was very similar for all weekdays (Monday-Friday) and

even Saturdays. Only Sundays had considerable lower pollutants' concentrations associated to lower anthropogenic activities (vehicular traffic, industrial production). This indicates a strong dependency of the hourly distribution of pollutants on the emission sources and their temporal variability.

The monthly distribution of air pollutants showed maximum mean values around March and minimum monthly values in June-July. The differences between minimum and maximum monthly values were large, reaching 33% for NO_2 (28.2 - 37.4 $\mu g \cdot m^{-3}$), 40% for PM_{10} (47.3 - 66.4 $\mu g \cdot m^{-3}$) and 85% for $PM_{2.5}$. These differences can hardly be explained by variations in the emission sources alone and could be related to changes in the atmospheric stability. This effect has been observed in similar Colombian regions (González-Duque et al., 2015), however further research regarding this topic is needed for the Aburrá Valley.



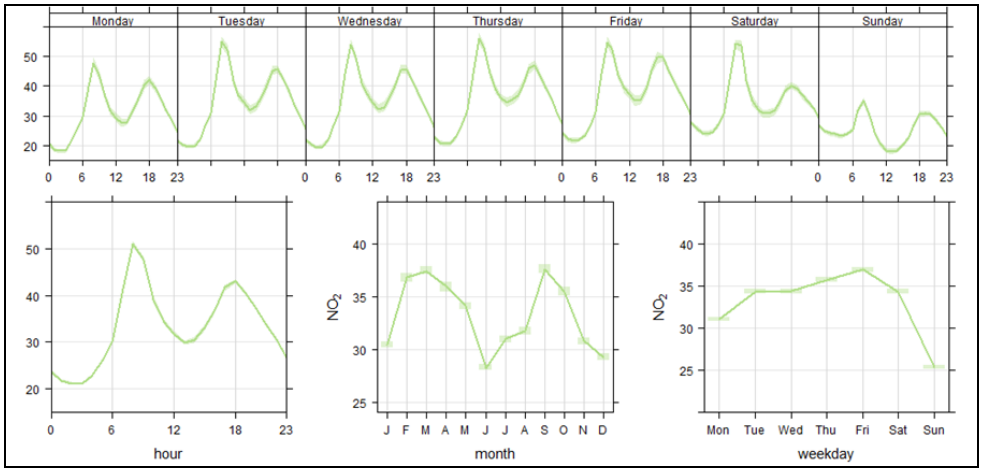


Figure 10: Daily, weekly and monthly time variation PM_{10}, $PM_{2.5}$ and NO_2. Mean values [$\mu g \cdot m^{-3}$]

4.1.3 Yearly and diurnal cycles

For the observance of air pollution and meteorological cycles their yearly and diurnal distributions were combined. The maximum hourly peak for air pollutants values was observed between 08:00-10:00, with no exception for all the months of the year. The diurnal cycle also showed a secondary peak during the evening, however it was much more pronounced for NO_2 in comparison with PM_{10} and $PM_{2.5}$. The daily cycle did not present important variations over the year, March being the month with the maximum pollution concentrations for all pollutants. During this month the pollution levels were constantly high at all hours.

The lowest air pollution levels occurred in June-July, at afternoon hours (14:00-16:00). This could be associated with two combined effects: favorable atmospheric conditions during afternoon hours (reduction of hourly values) and lower traffic emissions during the holiday season (decrease in the monthly averages). The lowest monthly average concentrations occurred during these months. The differences between total minimum and maximum pollution values were very high (+169% PM_{10}, +294% $PM_{2.5}$ and +233% NO_2).

Wind speed and solar radiation showed a consistent diurnal cycle during the different months. This was expected because of the proximity of the study region to the earth's equator. Maximum values occurred in June-July, between 14:00-16:00 for wind speed and in July-August between 12:00-13:00 for solar radiation. Minimum values for both

variables were observed in June-July, just before sunrise. On the contrary, there were large variations in the daily profile of precipitation over the year. Between May-October there was presence of rainfall events at late night hours and the rest of the months the events occurred mainly in the afternoon. This shift in the diurnal precipitation profile agrees with the observations by Poveda et al. (2015). Between 07:00-13:00 very low rainfall amounts were recorded.

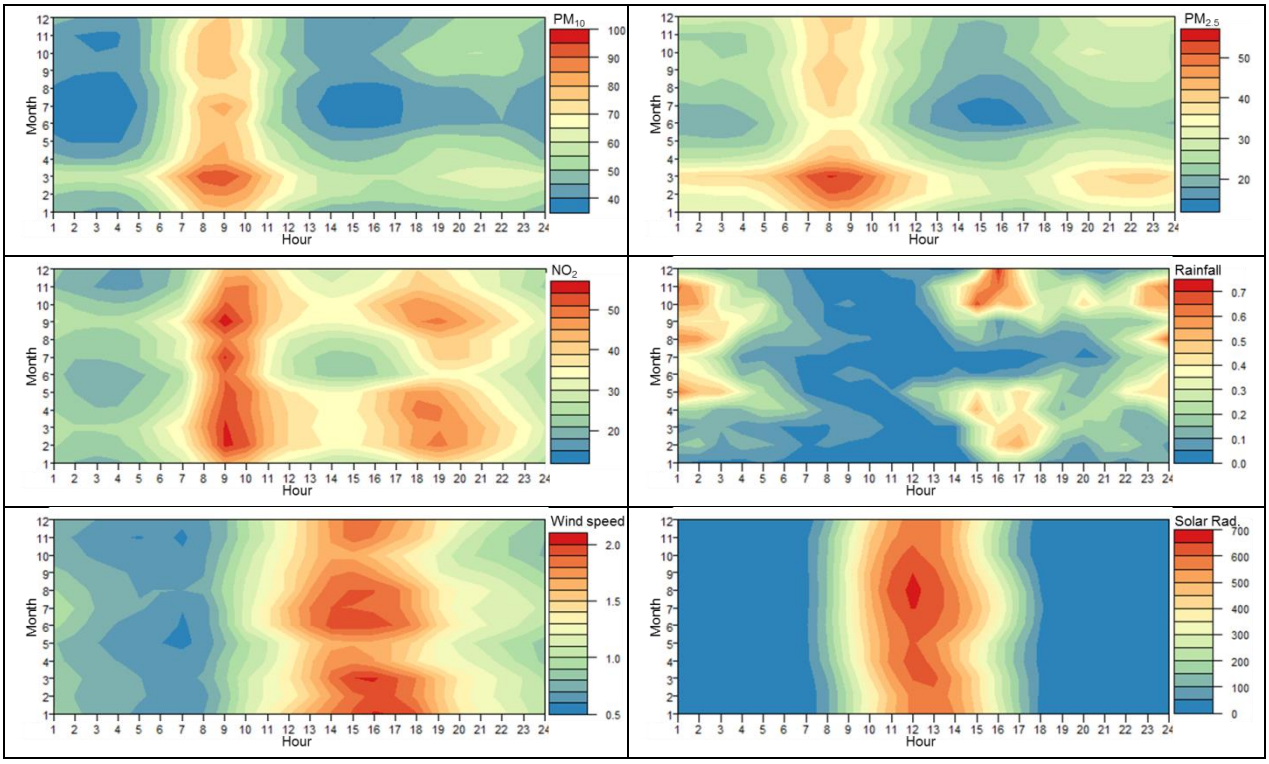


Figure 11: Yearly and diurnal cycle of air pollutants and meteorology in the Aburrá Valley. PM_{10} [µg·m^{-3}], $PM_{2.5}$ [µg·m^{-3}], NO_2 [µg·m^{-3}], precipitation [mm·h^{-1}], wind speed [m·s^{-1}] and temperature [°C].

Throughout this section hourly, weekly and monthly distribution profiles for air pollutants and meteorology in the Aburrá Valley were presented. However, a clear relationship between these variables could not be identified. A variogram analysis is therefore needed for an effective description and prediction of PM_{10}, $PM_{2.5}$ and NO_2.

4.2 Variogram analysis

4.2.1 Diurnal and yearly cycles

Semivariograms of the long trend pollution variation (hourly and monthly averages) for the pollutants PM_{10}, $PM_{2.5}$ and NO_2 were calculated. The results for $PM_{2.5}$ and NO_2 are

presented in detail in this section. Results for PM_{10} were very similar to those of $PM_{2.5}$ and are therefore not shown here, but can be found in Appendix 7.1.

Empirical semivariograms

The empirical semivariogram of $rPM_{2.5}$ (Figure 12a) when using time during the day as x coordinate and month as y coordinate showed two maximum semivariances at lag distances of around 7 and 21 respectively, where the unit of the lag may be hours or months or a combination of both. It is rather unlikely that the diurnal pattern is the same as the seasonal pattern. However, this approach was necessary to interpolate and extract the variation of the diurnal pattern over seasons (months) by kriging. The prerequisite for doing so is that the diurnal semivariogram is sufficiently similar to the seasonal semivariogram for short lags that play a role during kriging. The comparison of Figure 12b and 12c showed that this was the case up to a lag of 4 (either months or hours). Thus, the overall semivariogram can be used up to a lag of 4 (either months or hours) for simultaneously kriging the variation of the diurnal pattern over months.

For the interpretation of the diurnal and the seasonal variation at lags longer than 4, the semivariograms on directional bands 90° (strictly hourly values) and 0° (strictly monthly values) have to be used (Figure 12b and c). The maximum semivariance (total sill) was similar for the monthly and the hourly semivariogram, indicating that the maximum variation during the year was about the same as the maximum variation during the day. Furthermore, both semivariograms showed a bimodal behavior indicating that the day or the year was separated in four phases, two of which were characterized by high values and two with low values. This bimodal behavior appeared to be more pronounced during the year than during the day although the differences were small.

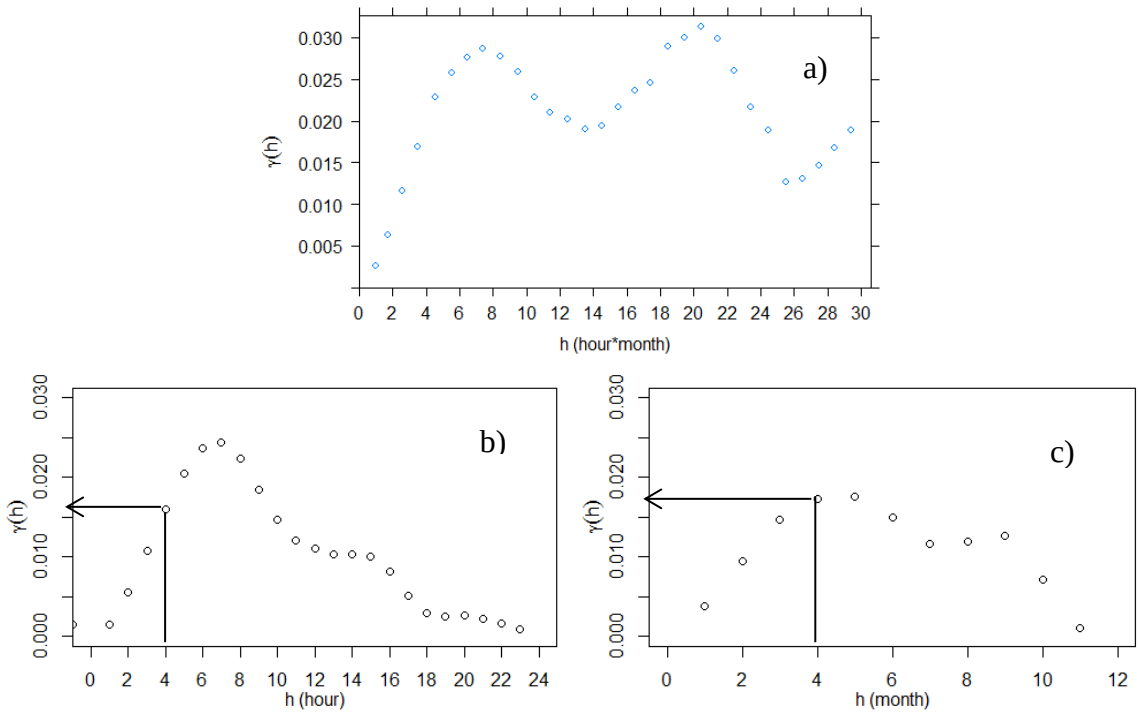


Figure 12: Semivariogram of the diurnal and yearly $rPM_{2.5}$ cycle: a) distance = hour *month; b) distance = hour; c) distance = month; γ-value $[(\mu g \cdot m^{-3})^{0.5}]$

Similarly to $rPM_{2.5}$, the hourly and monthly semivariograms for rNO_2 which combined hourly and monthly average values could be compared up to a lag of 4 (Figure 13 b-c). The semivariance at month 4 equals the semivariance at hour 2, meaning that up to this distance the semivariance of the diurnal cycle is twice higher as the semivariance of the yearly cycle. This ratio was considered for the calculation of the semivariogram and allowed kriging to be applied up to lag of 4. The combined semivariogram showed three peaks at lag distances of around 7, 12 and 21. The additional peak in comparison to the rPM2.5 semivariogram is associated to the presence of a second pronounced daily peak value in the NO_2 data (during the early evening hours) in comparison with only one very pronounced peak by $PM_{2.5}$ (in the morning). The higher variability of the NO_2 diurnal cycle (concentration differences between morning peak and night valley of over 150%, see Figure 9) was also reflected in the presence of continuous semivariance values > 0.03 between lag distances of 6 and 22.

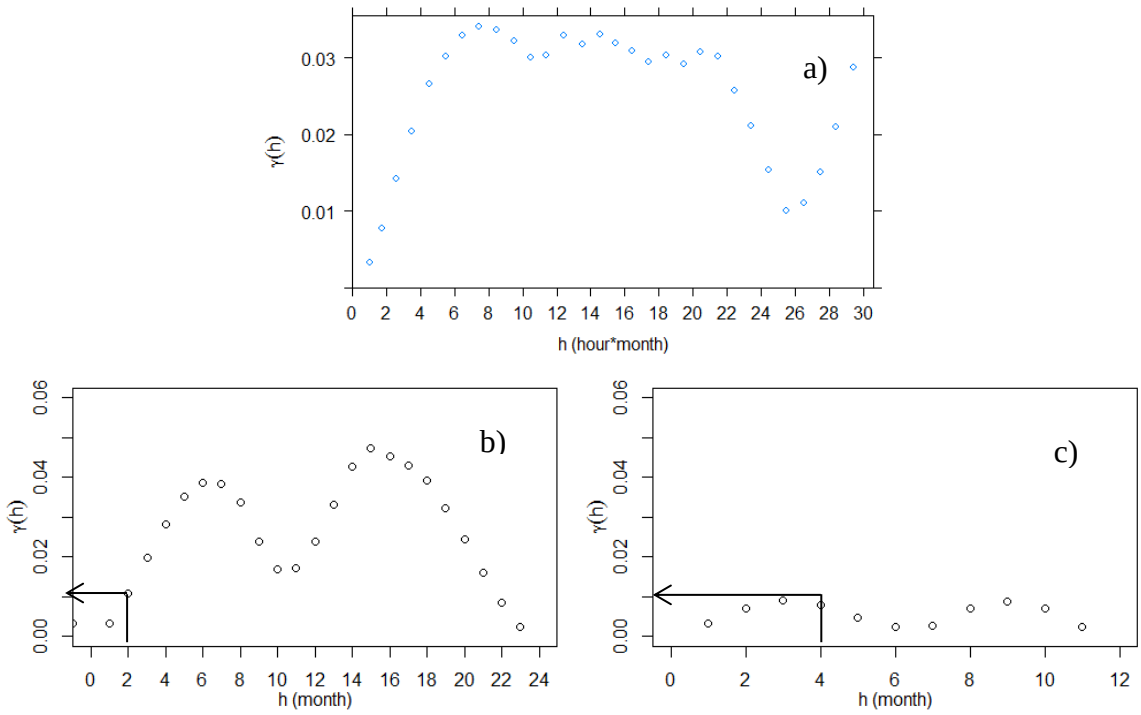


Figure 13: Semivariogram of the diurnal and yearly rNO_2 cycle: a) distance = hour*month; b) distance = hour; c) distance = month; γ-value $[(\mu g \cdot m^{-3})^{0.5}]$

The semivariogram corresponding to the yearly cycle of rNO_2 showed analogous patterns to the one observed for $rPM_{2.5}$, with two maximum semivariance values at lag distances of 4 and 8 months, respectively. This indicates that a general yearly cycle influenced the pollution levels in the Aburrá valley, yet up to a different degree for $PM_{2.5}$ and NO_2. The yearly pollution cycle had a stronger effect on $PM_{2.5}$ than on NO_2.

Two semivariance peaks associated with the difference between pollution concentrations at morning peaks and night valleys were found at lag distances of 6 and 16 hours in the diurnal semivariogram. The diurnal cycle had a bigger effect over the concentrations of NO_2 than the yearly cycle. The semivariogram corresponding to the hourly variation of this pollutant showed maximum γ-values of 0.05, while the maximum semivariance for monthly values lied around 0.010 (at month 4 and 9). At a lag distance of 2 hours the semivariance of the diurnal cycle equaled the maximum of the yearly-cycle semivariogram, showing the maximum distance up to which an omnidirectional semivariogram is valid.

Theoretical semivariograms

Omnidirectional semivariograms were calculated using the method described in section 3.3.3. In Figure 12(b, c) and Figure 13(b, c) the first peak of the monthly semivariogram for both pollutants is observed at a distance of 4 month. This was the maximum distance taken for the month-coordinate. The equivalent semivariance for the hour-coordinates lied for $rPM_{2.5}$ at a distance of 4 hours and for rNO_2 at a distance of 2 hours. Considering this, the theoretical semivariogram for $rPM_{2.5}$ was calculated until a maximum distance of 4 hours and 4 months, without any further data manipulation. On the contrary, by the rNO_2 dataset the hours were multiplied by 2 and then the semivariogram was calculated until a distance of 4 hours and 4 months, thus obtaining a valid semivariogram that simultaneously included the diurnal and yearly variability of air pollutants. A lag distance of 0.5 was chosen for the data pairs. A maximum cutoff value of 6 was selected.

Figure 14 presents the theoretical semivariograms obtained for $rPM_{2.5}$ and rNO_2. In both cases the fit was accurate and matched the γ-values obtained from the empirical semivariograms with precision. The nugget had a value of 0 for $rPM_{2.5}$ and rNO_2, meaning that the averaging of values (hourly + monthly averages) eliminated the measurements error of the air quality samples. The γ-values of the range lied by expected values of 6.0 (Table 2). Finally, a much higher sill than the nugget was associated with the significant effect of the diurnal and yearly cycles over the pollution concentration and its variability in the Aburrá Valley.

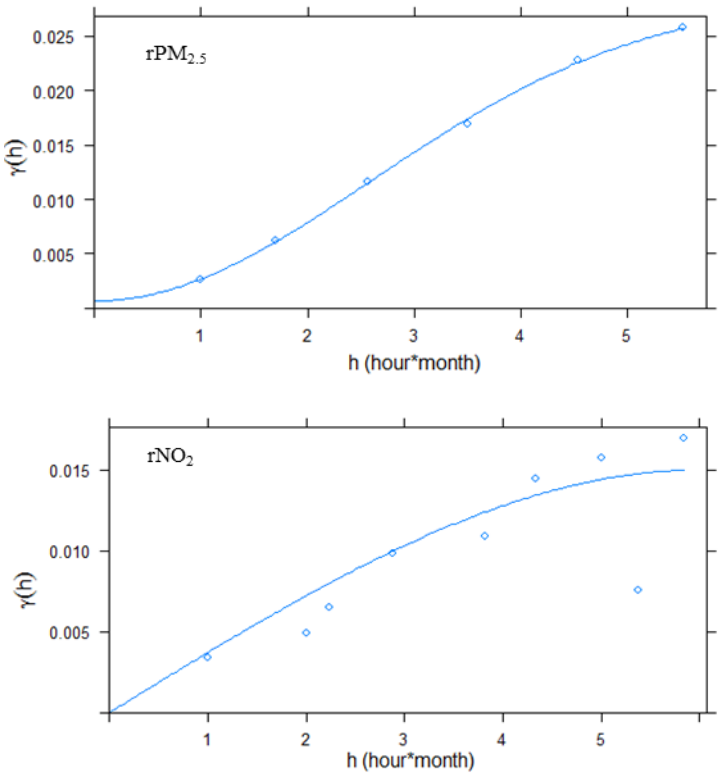


Figure 14: Theoretical semivariogram models for $rPM_{2.5}$ and rNO_2 diurnal and yearly cycles. γ-value $[(\mu g \cdot m^{-3})^{0.5}]$ and distance [hour*month]

Table 2: Parameters for the theoretical semivariograms of $rPM_{2.5}$ and rNO_2 diurnal/yearly cycles.

	$rPM_{2.5}$	**rNO_2**
Semivariogram model	Gaussian (automatic)	Spherical (automatic + eyefit)
Nugget $[(\mu g \cdot m^{-3})^2]$	0.0007	0.0
Sill $[(\mu g \cdot m^{-3})^2]$	0.028	0.015
Range [(hour*month]	3.66	6
RMSE $[(\mu g \cdot m^{-3})^2]$	0.0002	0.0028

The calculated semivariogram models properly described the variability of air pollutants related to diurnal and yearly cycles, thus allowing its application for kriging procedures. Additionally, in the next section the results of semivariogram models based on a higher time resolution will be presented

4.2.2 Diurnal and weekly cycles

Complementary to the semivariograms calculated for the diurnal and yearly pollution cycles, a variographic analysis was performed based on the diurnal and weekly cycles of air pollutants. For this purpose, the average values used for the calculation of the

semivariogram (coordinates x and y) corresponded to the hour of the day (diurnal cycle) and the day of the week, from Monday to Sunday (weekly cycle). The results of this analysis allowed a better understanding of the short term variability of the main pollutants in the Aburrá Valley.

Empirical semivariograms

The empirical semivariogram of the diurnal and weekly cycle of $rPM_{2.5}$ (Figure 15a) showed two maximum semivariances at lag distances of around 7 and 18 respectively. The first peak corresponded to the maximum variability associated to the diurnal cycle, while the second peak is produced by the combination between the diurnal cycle and the weekly cycle and their maximum and minimum values (e. g. differences between a Wednesday morning and a Sunday afternoon). The weekly cycle differs highly from the diurnal cycle, as it was expected. However, until a lag distance of 2 the semivariance of the weekly cycle equaled the semivariance of the diurnal cycle, meaning that until this lag the variability of $rPM_{2.5}$ was as much influenced by the hourly variability as by the differences between days of the week. The overall semivariogram could therefore be used up to a lag of 2 for simultaneous kriging of variation of the diurnal and weekly cycle.

The semivariograms on directional bands 90° (strictly hourly values) and 0° (strictly weekday values) showed a higher effect of the hourly cycle than the weekly cycle over $rPM_{2.5}$ variability. The maximum semivariance (total sill) at 7 hours was around 4 times bigger than the corresponding maximum of the weekly cycle, thus indicating the importance of the diurnal peaks over pollution events in the area of study. Furthermore, the semivariogram of weekdays peaked at distances of 2 and 4 days, which was related to rPM2.5 minimum values on Sundays and high differences compared to the rest of the days with the exception of Mondays which was influenced by the pollution decrease on Sundays and presented the second lowest $rPM_{2.5}$ concentrations of all weekdays.

.

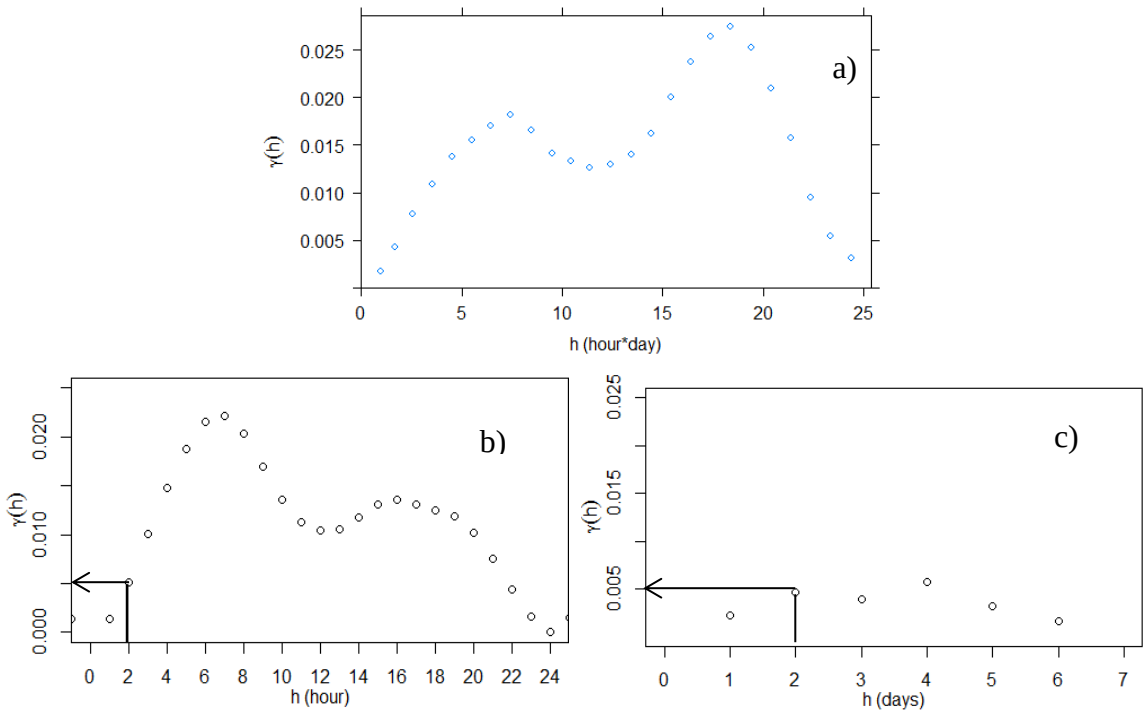
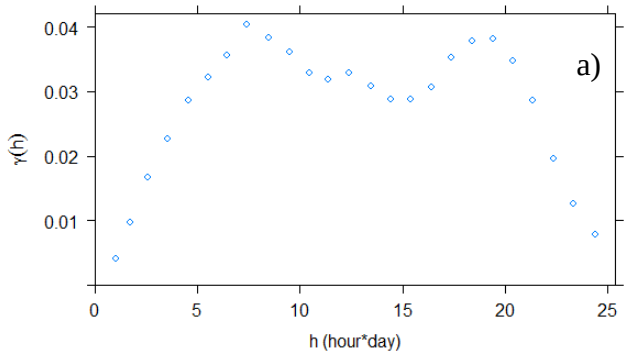


Figure 15: Semivariogram of the diurnal and weekly $rPM_{2.5}$ cycle: a) distance = hour *day; b) distance = hour; c) distance = day; γ-value $[(\mu g \cdot m^{-3})^{0.5}]$

Similarly to $rPM_{2.5}$, the combined semivariogram of the weekly and diurnal cycles of rNO_2 presented two peaks at lag distances around 7 and 19 (Figure 16). The first peak was produced by the maximum diurnal semivariance at 7-hours distance and the second peak combined the effect of the diurnal and the weekly cycle. The second semivariance peak was less pronounced by rNO_2 compared to $rPM_{2.5}$, what could be explained by a general lower effect of the weekly cycle over the total variability of rNO_2.



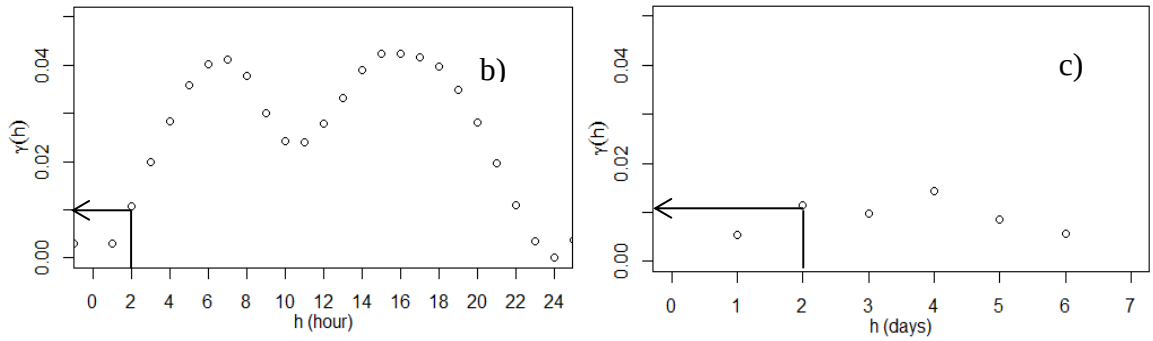


Figure 16: Semivariogram of the diurnal and weekly rNO_2 cycle: a) distance = hour *day; b) distance = hour; c) distance = day; γ-value $[(\mu g \cdot m^{-3})^{0.5}]$

The strictly 0° rNO_2 pollution semivariogram showed a nearly identical pattern to the corresponding $rPM_{2.5}$ semivariogram, with 2 semivariance peaks at lag distances of 2 and 4 days. A common weekly cycle could be observed for the air pollutants in the Aburrá Valley, where the minimum pollution concentrations occurred in Sundays and Mondays, while the pollution maximum values were observed at Fridays/Saturdays. This pattern was constant over the study period and did not change greatly throughout the different phases of the year (the air pollutants showed the same pattern during pollution peaks in March and yearly minima in June).

The weekly cycle for $rPM_{2.5}$ and rNO_2 could be successfully captured by the 0° semivariogram with similar results for both pollutants, showing that a general pattern based on the differences between Sundays (and to a lesser degree Mondays) and the rest of the days dominated the pollution concentrations in the area of study. The weekly cycle, in combination with the hourly pollution distribution, captured the variability of air pollutants' concentrations at short term intervals (days within a same week, at different hours).

Theoretical semivariograms

Omnidirectional semivariograms of the combined diurnal and weekly pollution cycle were calculated using the same method as in section 4.1.2. The directional semivariograms of the diurnal and weekly cycles, both for $rPM_{2.5}$ and rNO_2, showed that at a lag distance of 2 the semivariance of the diurnal cycle equaled the semivariance of the weekly cycle. Considering this, the theoretical semivariograms for $rPM_{2.5}$ and rNO_2 were calculated until a maximum distance of 2 hours and 2 days. Besides the

definition of the maximum possible distance no further data manipulation was performed. A lag distance of 0.5 was chosen for the data pairs.

The fitted omnidirectional semivariograms matched the empirical $rPM_{2.5}$ and rNO_2 with high precision (very low RMSE values of 0.0003 and 0.001, respectively, Table 3). The theoretical semivariograms of both pollutants showed very similar patterns (Figure 17), thus indicating that the combination of the diurnal and weekly pollution cycles until the selected lag distance affected the pollution concentrations of $rPM_{2.5}$ and rNO_2 in a similar way. The nugget for the theoretical semivariograms was near 0 in both cases and the range lied at values around 3, what was expected considering that the maximum selected distance was 2, both in x (hour) and y (day of the week) direction.

Even though the theoretical semivariograms of $rPM_{2.5}$ and rNO_2 presented almost identical patterns, the total sill of rNO_2 was twice as high as $rPM_{2.5}$. The combined weekly/diurnal cycle influenced the pollution concentrations under a common schema, however to a different degree depending on the air pollutant. rNO_2 presented a higher variability under the influence of this common cycle. NO_2 showed during the 3-year study period higher differences than $PM_{2.5}$ between Sundays and rest of the days (Figure 10) and also the diurnal cycle of this pollutant was less stable; this was accurately reflected though the omnidirectional semivariogram.

Table 3: Parameters for the theoretical semivariograms of $PM_{2.5}$ and NO_2 diurnal/weekly cycles

	$PM_{2.5}$	**NO_2**
Semivariogram model	Gaussian (automatic)	Gaussian (automatic)
Nugget $[(\mu g \cdot m^{-3})^{0.5}]$	0.0004	0.0014
Sill $[(\mu g \cdot m^{-3})^{0.5}]$	0.016	0.031
Range [(hour*month]	3.257	3.111
RMSE $[(\mu g \cdot m^{-3})^{0.5}]$	0.00031	0.0010

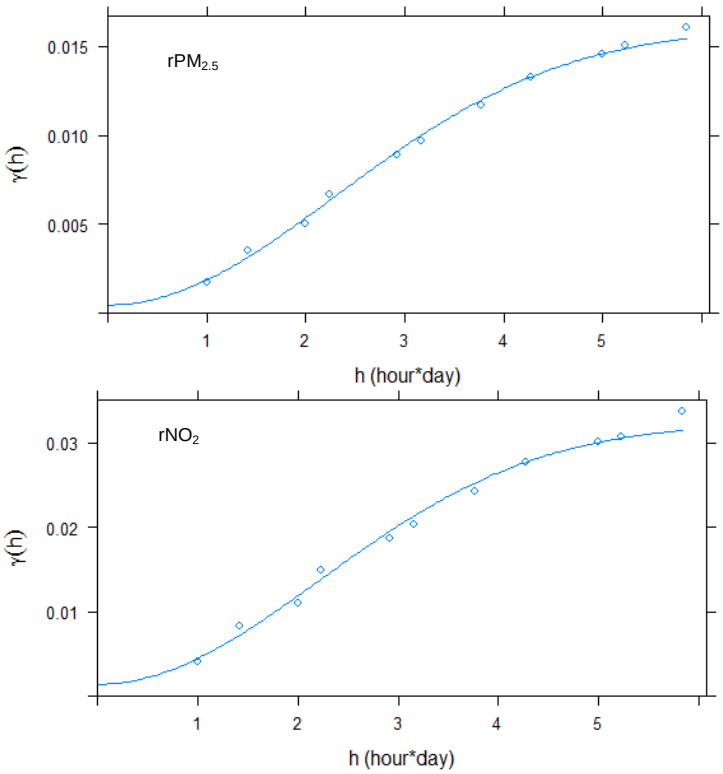


Figure 17: Theoretical semivariogram models for $rPM_{2.5}$ and rNO_2 diurnal and weekly cycles. γ-value $[(\mu g \cdot m^{-3})^{0.5}]$ and distance [hour*day]

Throughout this section the temporal cycles that have an effect over the pollution levels in the study area (diurnal, weekly, yearly cycles) were explored by using a variographic analysis. Its results allowed the calculation of theoretical semivariograms, which can be used for the study of interactions of air quality patterns at different time scales, estimation of missing values or prediction purposes. The next section of this investigation presents the results of this procedure.

4.3 Kriging for time series analysis, missing data estimation and prediction

4.3.1 Kriging modelling at diurnal/yearly scale

The theoretical semivariograms of the diurnal and yearly pollution cycles were used for imputation and prediction of PM_{10} and NO_2 values by using the kriging method. Before kriging was applied for specific time periods, the general validity of the model assumptions was tested by cross validation. Its results confirmed that the selected method was valid for the time interpolation of hourly and monthly averages of $rPM_{2.5}$ and rNO_2. In both cases, the model residuals presented a normal distribution (Figure 18

and Figure 19) and the coefficients of determination between observed and predicted values were near 1 (R^2 = 0.965 for the $rPM_{2.5}$ model and R^2 = 0.9332 for the rNO_2 model)

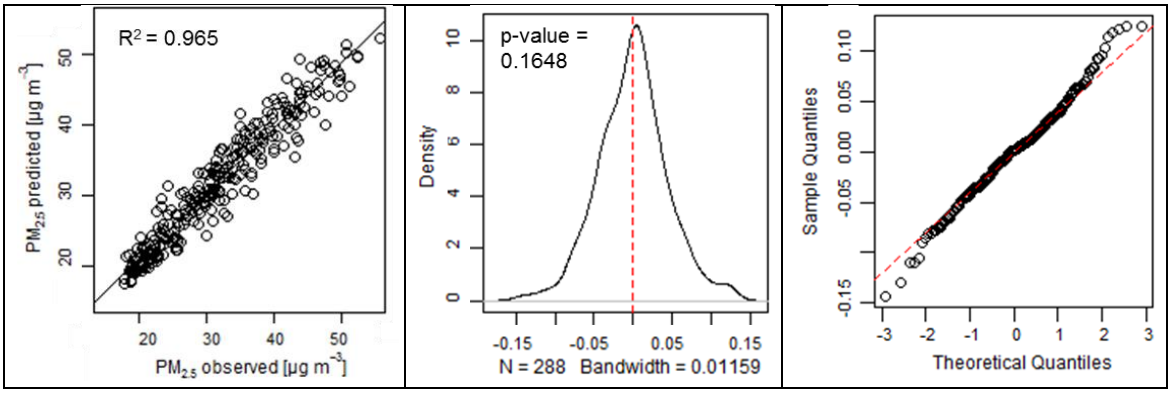


Figure 18: Results of the cross validation for the $rPM_{2.5}$ kriging model diurnal and yearly cycle: a) correlation between observed and predicted values; b) Density plot of model residuals and p-value of Lilliefors-Test; c) Q-q plot of model residuals

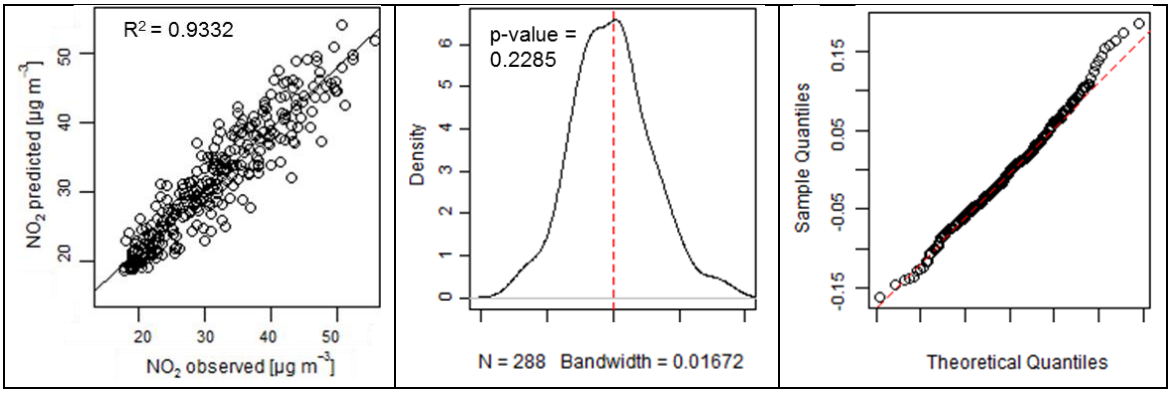


Figure 19: Results of the cross validation for the rNO_2 kriging model diurnal and yearly cycle: a) correlation between observed and predicted values; b) Density plot of model residuals and p-value of Lilliefors-Test; c) Q-q plot of model residuals. Cross validation performed with data corresponding to year 2014

Furthermore the kriging models were applied to reconstruct the diurnal and yearly cycles of air pollution in the Aburrá Valley in a similar way than what was presented in section 4.1.3 (air pollution contour plots). The advantage of kriging over contour plots is that the cycles can be reconstructed even if the air pollution datasets are incomplete and the interpolation between values is based on a robust model (the semivariogram). The diurnal and yearly distribution of $PM_{2.5}$ and NO_2 for year 2014 modeled through a geostatistical approach showed that despite the rather similar diurnal pattern of both pollutants and the rather similar seasonal pattern of the pollutants, both behaved slightly

differently when combining the time domains (Figure 20). The diurnal pattern was rather weak for $PM_{2.5}$ in March, while NO_2 did not show this deviation. The reason for this phenomenon was related to the fact that during March $PM_{2.5}$ accumulated in the Aburrá Valley, thus producing average concentrations of > 30 [$\mu g \cdot m^{-3}$] during all hours of the day. In the case of NO_2 the night levels remained however low because this pollutant rapidly decreases in the absence of emission sources.

The sunset peak was considerably more pronounced for NO_2 than for $PM_{2.5}$. NO_2, in general, showed a stronger diurnal variability than $PM_{2.5}$ with lower values at night hours. Both pollutants have an afternoon depression, which is most pronounced in July, associated with the yearly maximum values of solar radiation and wind speed and therefore with less stable atmospheric conditions.

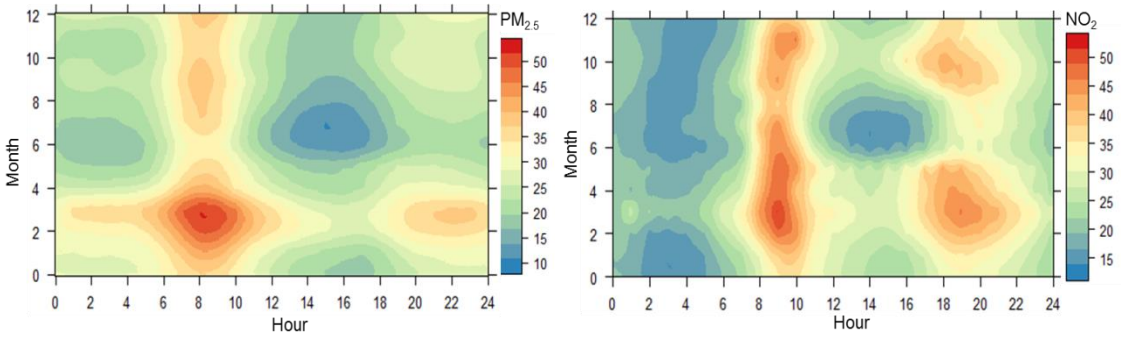


Figure 20: Kriging interpolation of the yearly and diurnal cycle of $PM_{2.5}$ [$\mu g \cdot m^{-3}$] and NO_2 [$\mu g \cdot m^{-3}$]. 2014 hourly and monthly averages were used for the interpolation

Estimation of missing data

Based on the general validity of the kriging interpolation procedure for $PM_{2.5}$ and NO_2 pollution cycles, the quality of the developed model as a modelling tool for air pollution was tested. Two scenarios were analyzed: estimation of missing values and pollution prediction.

It is a common phenomenon in air quality monitoring that because of technical problems a monitoring station stops working during prolonged time intervals or its records are not considered as valid. In this case it is crucial to reconstruct the datasets by calculating the missing data. The accuracy of the kriging technique for this purpose was tested for the available data in the Aburrá Valley.

Coefficients of determination (R^2) between observed and simulated values for all months, with the exception of September and October for NO_2, reached values >0.9 (Table 4), indicating that the estimated values for each month reproduced the diurnal pollution cycle with accuracy. The months with the best simulation results for $PM_{2.5}$ were March and July, while in September and October the least accurate results were obtained. Good estimation results for March are of extreme importance for the overall quality of the model, considering that the highest pollution concentrations in the Aburrá are constantly observed during this month. On the contrary, the estimated values in September-October were too low in comparison with the observed values (monthly estimated averages >30% lower than observed values). This reflects that the model was unable to estimate the rapid increase of the pollution concentrations after the yearly minimum values in June-August. However, since the coefficients of determination remained high, the estimation of missing values for September-October could be improved by using a correction factor based on the difference between simulated and observed values.

Meanwhile, the model could simulate missing NO_2 values with high accuracy in April-May, but showed difficulties with the estimation of missing values in March. A correction for this month will therefore be needed for future simulation, because of the existing pollution peaks during March. The high coefficient of determination achieved (R^2 = 0.93) indicates that the model was able to capture the daily pollution profile accurately, however with a bias towards lower pollution concentrations than the real values. This could be improved through the calculation of a correction factor for this month.

Table 4: Evaluation of the kriging model for estimation of missing $PM_{2.5}$ and NO_2 data. Results for average hourly values Jan-Dec 2014.

		Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
$PM_{2.5}$	Mean observed value [µg·m^{-3}]	29.7	38.2	40.5	28.0	24.1	19.7	20.0	23.3	27.8	32.1	32.9	34.0
	Mean simulated value [µg·m^{-3}]	21.0	31.7	35.8	32.2	32.9	26.2	19.9	18.7	18.7	20.8	24.4	28.4
	Coeff. determ. R^2	0.96	0.95	0.98	0.98	0.91	0.95	0.97	0.96	0.93	0.90	0.97	0.96
	RMSE [µg·m^{-3}]	8.8	6.7	5.0	4.2	8.9	6.7	1.4	4.8	9.3	11.5	8.7	5.9
	Index of Agreement	0.58	0.69	0.87	0.87	0.63	0.76	0.99	0.85	0.63	0.55	0.66	0.82

NO_2	Mean observed value [µg·m^{-3}]	33.0	39.5	42.8	37.9	38.7	29.6	30.9	30.9	36.7	37.1	34.9	32.9
	Mean simulated value [µg·m^{-3}]	25.4	30.4	28.6	35.1	37.4	32.8	30.6	28.6	29.1	30.6	28.7	30.9
	Coeff. determ. R^2	0.90	0.97	0.93	0.93	0.93	0.93	0.93	0.97	0.65	0.62	0.90	0.98
	RMSE [µg·m^{-3}]	8.1	9.3	14.4	3.7	3.0	4.1	2.4	2.6	9.2	8.8	7.2	2.3
	Index of Agreement	0.75	0.81	0.62	0.95	0.98	0.94	0.98	0.96	0.74	0.78	0.85	0.98

The observed and simulated values for the hourly $PM_{2.5}$ and NO_2 averages in March are presented in Figure 21. The simulated diurnal cycle of both pollutants followed with high accuracy the observed diurnal cycle, thus showing that the selected geostatistical model was reliable and stable for the study area. The model could simulate $PM_{2.5}$ almost perfectly (maximum differences around 7 [µg·m^{-3}] between observed and predicted value) and could therefore be applied for estimation of missing data in the Aburrá Valley. For an accurate prediction of NO_2 a correction factor must be included in future calculations to avoid its bias towards the simulation of lower values than the reality.

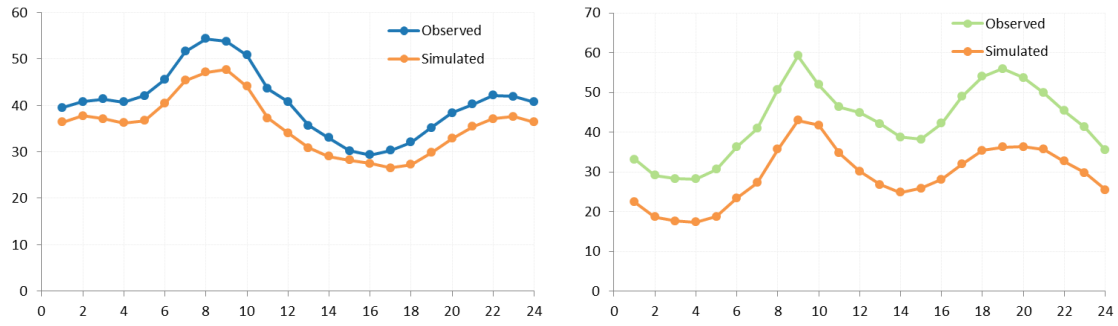


Figure 21: Observed and estimated missing data March 2014. $PM_{2.5}$ [µg·m^{-3}] and NO_2 [µg·m^{-3}] hourly averages

Prediction of air pollution

Besides the estimation of missing data for long data series, the accuracy of kriging for the prediction of NO_2 and $PM_{2.5}$ diurnal averages for all months of the year was tested. Prediction of monthly averages of air pollutants would be very useful for the environmental authorities to determine in advance if actions for the reduction of air pollutants' emissions are needed for a specific month.

The results of the evaluation criteria for the prediction of $PM_{2.5}$ and NO_2 are summarized in Table 5. For $PM_{2.5}$ the coefficient of determination showed values for all

months > 0.8. Together with high values of the Index of Agreement (> 0.8 in all cases with the exception of February and April) this indicates that the kriging interpolation was able to correctly predict both the variability of $PM_{2.5}$ concentration during the diurnal pollution cycle, as well as the differences between months. March was predicted with very high accuracy; this is important for the overall accuracy of the model, considering that March is the month with the highest monthly pollution averages in the Aburrá Valley.

The accuracy by the prediction of NO_2 was a little lower compared to $PM_{2.5}$. The month with the best results was April, while September showed the least accurate predictions. September is a difficult month for the simulation of NO_2, due to its characteristics as a transitional period between pollution cycles. While in 2013 September presented the highest monthly NO_2 average concentrations of the year, in 2013 and 2015 it showed lower values than March, April or May. Furthermore, the diurnal cycle of pollution was well captured by the kriging prediction, what is reflected by values for the Index of Agreement > 0.75, with the exceptions of September and January

Table 5: Evaluation of the kriging model for prediction of $PM_{2.5}$ and NO_2. Results for average hourly values Jan-Dec 2014.

		Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
$PM_{2.5}$	Obs. Value [$\mu g \cdot m^{-3}$]	29.7	38.2	40.5	28.0	24.1	19.7	20.0	23.3	27.8	32.1	32.9	34.0
	Sim. value [$\mu g \cdot m^{-3}$]	25.4	29.2	39.5	38.3	21.8	20.7	16.3	17.8	22.4	26.9	30.3	29.8
	Coeff. determ.R^2	0.82	0.94	0.85	0.90	0.82	0.91	0.95	0.88	0.90	0.92	0.89	0.89
	RMSE [$\mu g \cdot m^{-3}$]	4.8	9.2	4.2	10.5	3.3	2.2	4.3	6.0	5.8	5.5	3.5	5.0
	Index of Agreement	0.77	0.56	0.87	0.59	0.88	0.94	0.87	0.80	0.77	0.79	0.90	0.85
NO_2	Obs. Value [$\mu g \cdot m^{-3}$]	33.0	39.5	42.8	37.9	38.7	29.6	30.9	30.9	36.7	37.1	34.9	32.9
	Sim. value [$\mu g \cdot m^{-3}$]	23.4	30.0	34.5	38.4	34.0	34.8	26.8	27.9	27.2	31.9	33.4	31.7
	Coeff. determ.R^2	0.95	0.93	0.97	0.87	0.88	0.73	0.90	0.81	0.80	0.78	0.83	0.94
	RMSE [$\mu g \cdot m^{-3}$]	9.7	10.0	8.5	3.3	6.0	7.1	5.0	4.9	10.3	6.9	4.4	2.5
	Index of Agreement	0.72	0.75	0.83	0.96	0.90	0.85	0.92	0.89	0.69	0.86	0.94	0.98

The prediction of the diurnal cycle of pollutants during March is detailed in Figure 22. $PM_{2.5}$ values, especially the morning peaks, were predicted with high accuracy. The difference between maximum observed and maximum predicted values was < 5 [$\mu g \cdot m^{-3}$], thus indicating the liability of the selected model to predict the average diurnal cycle of $PM_{2.5}$ during the most polluted month of the year. The model was also able to correctly predict the average pollution decrease at afternoon hours and the second, less

pronounced pollution peak after 20:00. Meanwhile, the prediction of the diurnal pollution cycle of NO_2 in March captured the typical variability throughout the day, however with a bias to predict lower pollution values (around 10 [$\mu g \cdot m^{-3}$]) than the observed concentrations. As the differences between observed and predicted were constant for the diurnal cycle of March, the prediction accuracy of NO_2 could by improved by the calculation of a correction factor.

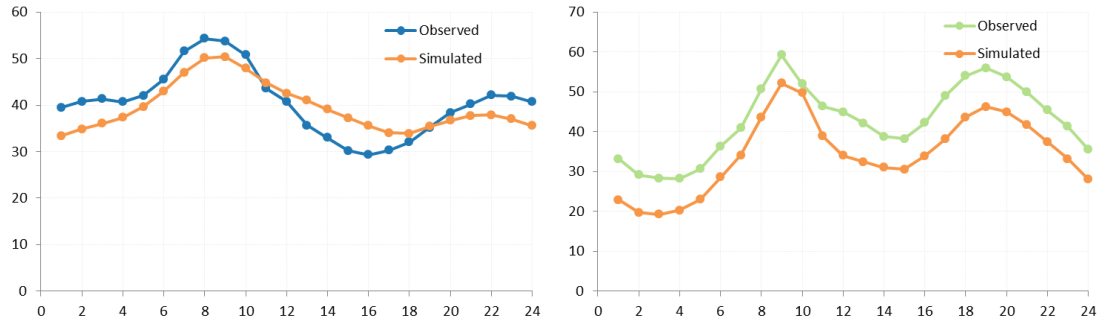


Figure 22: Observed and predicted values March 2014. $PM_{2.5}$ [$\mu g \cdot m^{-3}$] and NO_2 [$\mu g \cdot m^{-3}$] hourly averages

4.3.2 Kriging modeling at diurnal/weekly scale

After the analysis of the yearly and diurnal pollution cycle, kriging was also applied to detect the interactions between the diurnal and the weekly cycles of $PM_{2.5}$ and NO_2. Therefore the general validity of the model assumptions had to be tested by cross validation. The results of the cross validation showed a normal distribution of the residuals, both for $rPM_{2.5}$ and rNO_2 (Figure 23 and Figure 24). The coefficients of determination between observed and predicted values when using all available data for the interpolation were near 1 (R^2 = 0.973 for the $rPM_{2.5}$ model and R^2 = 0.9554 for the rNO_2 model)

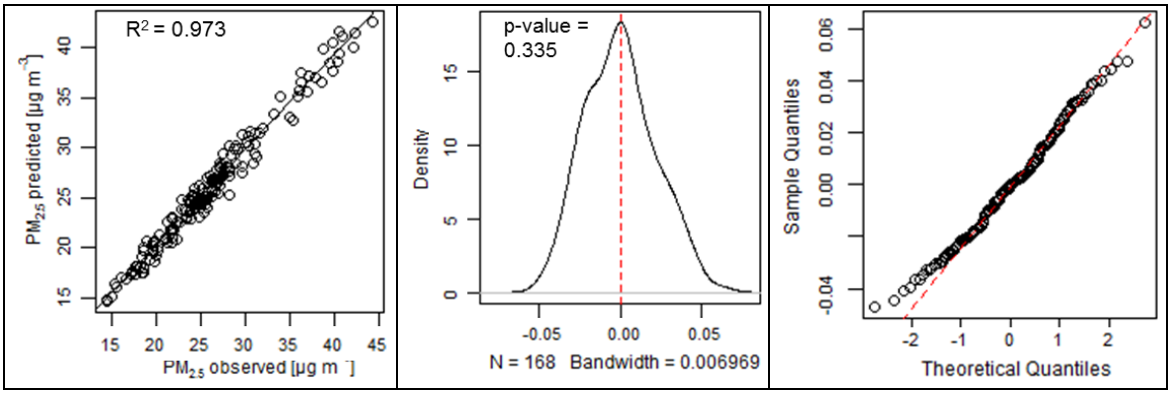


Figure 23: Results of the cross validation for the $rPM_{2.5}$ kriging model diurnal and weekly cycle: a) correlation between observed and predicted value; b) Density plot of model residuals and p-value of Lilliefors-Test; c) Q-q plot of model residuals

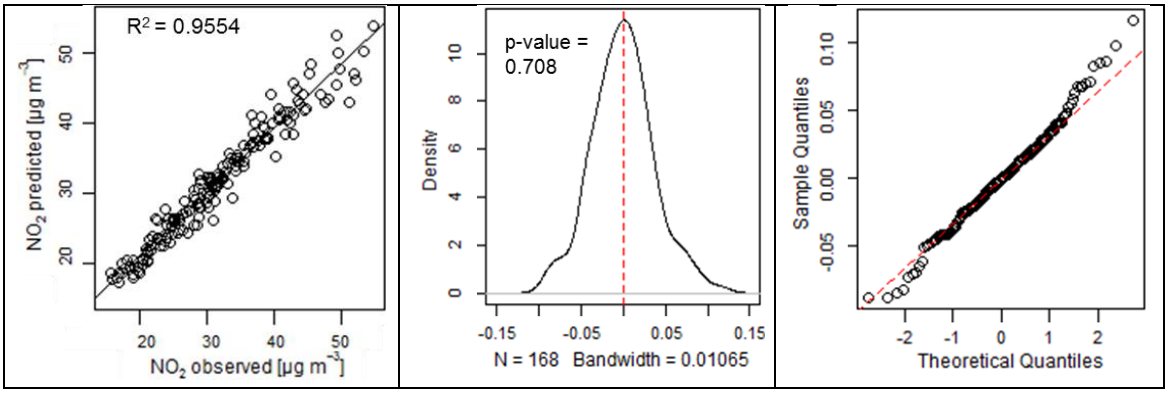


Figure 24: Results of the cross validation for the rNO_2 kriging model diurnal and yearly cycle: a) correlation between observed and predicted value; b) Density plot of model residuals and p-value of Lilliefors-Test; c) Q-q plot of model residuals. Cross validation performed with data corresponding to year 2014

The interactions between the diurnal and weekly air pollution cycles of $PM_{2.5}$ and NO_2 for year 2014 were reconstructed by applying kriging interpolation based on the corresponding semivariogram models. The diurnal cycle of $PM_{2.5}$ and NO_2 behaved quite similar during the morning peak hours, but in the early evening hours the NO_2 peak was considerably higher than the $PM_{2.5}$ peak. During afternoon hours the presence of rain events is common in the Aburrá Valley. Rain has a higher wash out effect over $PM_{2.5}$ and therefore the afternoon pollution valley for this pollutant is more pronounced than for NO_2. Meanwhile at late night NO_2 presented very low concentrations, because of it strong dependency on emissions sources, which at this time of the day are usually at their lowest point. $PM_{2.5}$ on the contrary accumulates for longer time periods in the

Aburrá Valley and therefore the pollution produced by the second traffic peak at the evening remained almost constant until the early morning hours.

The weekly cycle did not have a considerable effect over the diurnal cycle. From Monday to Sunday the maximum daily concentrations were observed between 07:00-09:00, with no exception. The evening $PM_{2.5}$ and NO_2 peaks also showed constant patterns throughout the different weekdays. The weekly cycle showed a constant increase in the average pollution concentrations from Monday to Friday, were absolute maximum $PM_{2.5}$ and NO_2 concentration were observed. On Saturdays there was a low decrease in the concentrations and finally on Sundays the weekly minima were reached, starting again with the weekly cycle. The continuous increase between Monday-Friday was associated to a gradual accumulation of pollutants due to continuous emissions during working days. Due to the reduction of emission levels on weekends the concentrations of air pollutants decreased drastically from Friday until Sunday.

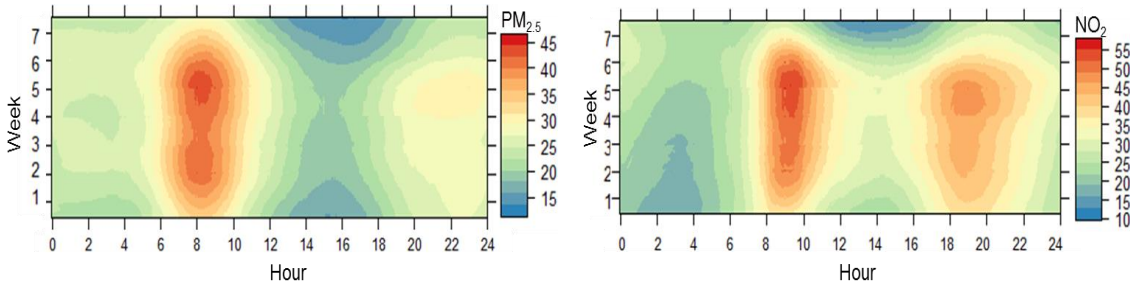


Figure 25: Kriging interpolation of the yearly and diurnal cycle of $PM_{2.5}$ [$\mu g \cdot m^{-3}$] and NO_2 [$\mu g \cdot m^{-3}$]. 2014 hourly and day of the week averages were used for the interpolation

Estimation of missing values

The diurnal cycle of air pollutants in the Aburrá Valley showed a constant average behavior of its peaks and minimum hourly values independently form the day of the week being analyzed. Considering the stability of this pollution pattern, $PM_{2.5}$ and NO_2 concentrations for days with high pollution levels (February-March 2014 and 2015) were simulated by the application of kriging, under the method detailed in section 3.3.3. The target values corresponded in all cases to Fridays, which is usually the day of the week with the highest pollution levels (Figure 11 and Figure 25).

In the first simulation, the target values in February-March were considered as missing data and calculated by kriging interpolation. It should be noted that in 6 out of the 8 selected days exceedances of the Colombian Air Quality for $PM_{2.5}$ were observed. $PM_{2.5}$ was accurately simulated for all selected days with the exception of March 13th 2014, for which the peak concentrations at morning hours were not captured by the model (Figure 26). Especially important for the assessment of the model were the simulations for March 7th 2014 and Feb. 27th 2015 which showed the highest $PM_{2.5}$ concentrations for the study period. In both cases the development of the morning peak was accurately reproduced by the model and reached values only 10-15 [$\mu g \cdot m^{-3}$] below the recorded concentrations.

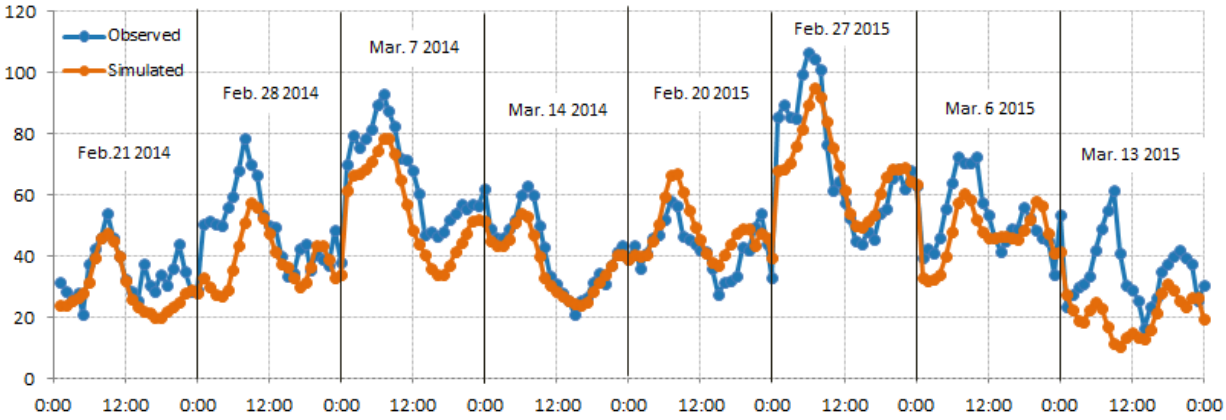


Figure 26: Observed and simulated values by the estimation of $PM_{2.5}$ missing data. Hourly values [$\mu g \cdot m^{-3}$] for every day of analysis

The model was also capable of capturing the diurnal variability of the $PM_{2.5}$ concentrations in the Aburra Valley, what was confirmed by a high Index of Agreement of 0.92 (Table 6). The diurnal cycle of $PM_{2.5}$ was almost constant during the 8 days of analysis. The maximum daily values varied significantly between days, yet the morning and the evening peaks were observed every day approx. at the same time. Because of the stable diurnal behavior of $PM_{2.5}$ the kriging interpolation was able to deliver good results based on the information of the adjacent days.

Estimation of missing NO_2 data delivered less accurate results than for $PM_{2.5}$, thus showing lower values for the coefficient of determination R^2 and the Index of Agreement (Table 6). Though the RMSE was relatively low (11.1 [$\mu g \cdot m^{-3}$]), what indicates that the average concentrations over an entire day were correctly captured, the

model was not entirely consistent, overestimating the observed concentrations during some days and in other occasions being unable to capture the maximum daily peaks (therefore the lower R^2 value compared to $PM_{2.5}$). The difficulties of the model for the estimation of missing NO_2 data were associated with a less stable diurnal pattern of this pollutant. By the majority of simulated days the diurnal cycle showed the expected maximum peak during the morning and a secondary, less pronounced peak at evening hours. However, during days like February 28th 2014 and February 20th 2015 the concentrations reached their hourly peaks during the evening hours (at not on early morning, as it was expected). These changes in the diurnal cycles generated difficulties to obtain better results by applying kriging interpolation. When the expected diurnal pattern was observed (e.g. March 14th 2014 and February 20th 2015) the entire diurnal cycle was accurately reproduced by the model

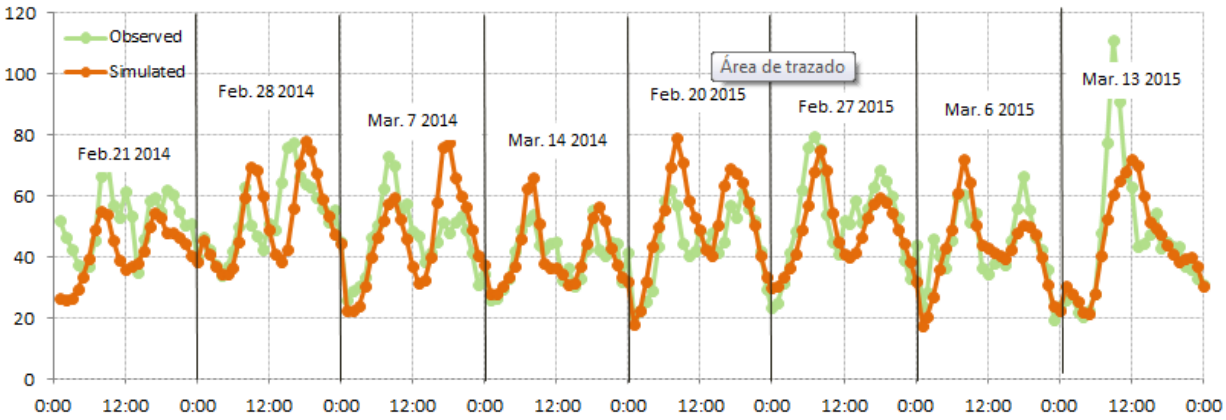


Figure 27: Observed and simulated values by the estimation of NO_2 missing data. Hourly values [$\mu g \cdot m^{-3}$] for every day of analysis

Table 6: Evaluation of the kriging model for estimation of $PM_{2.5}$ and NO_2 missing data. Results for 8 selected days in February-March 2014-2015

pollutant	Coefficient of determination R^2	RMSE [$\mu g \cdot m^{-3}$]	Index of Agreement
PM_{25}	0.77	10.1	0.92
NO_2	0.47	11.1	0.82

Prediction of air pollution

The developed model was additionally tested for prediction of $PM_{2.5}$ and NO_2 concentrations. Kriging was therefore used as an extrapolation method, i.e. the target

values were calculated based on the previous days only. The model was able to predict with good accuracy the $PM_{2.5}$ concentrations for the days of analysis (R^2 = 0.55 and RMSE = 13.57 [$\mu g \cdot m^{-3}$], Table 7). Especially accurate predictions were observed for following days: February 21th 2014, March 7th 2014 and February 27th 2015, which was the day with the highest overall air pollution concentrations (Figure 28). On the contrary, the model showed difficulties to represent the diurnal variability of $PM_{2.5}$ for February 20th 2015 and March 13th 2015. These differences in the model accuracy depend on how stable the diurnal pollution pattern was in the days preceding the prediction. The high Index of Agreement for the simulated values (0.85) shows that the model was in most of the cases capable of both predicting the hourly variations of $PM_{2.5}$, as well as the peak values observed during the days of analysis.

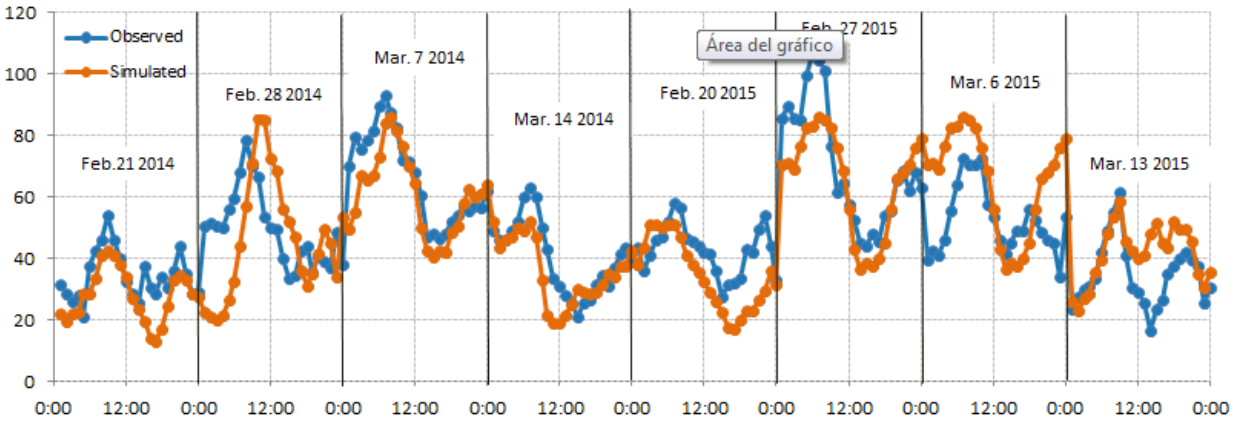


Figure 28: Observed and simulated values by $PM_{2.5}$ prediction. Hourly values [$\mu g \cdot m^{-3}$] for every day of analysis

The applied model showed a lower performance for the prediction of NO_2 hourly values compared to $PM_{2.5}$. A low performance was observed particularly by the prediction of the evening values. This was related with considerable variations in the diurnal NO_2 cycle and the resultant differences between morning and evening peak values. As a result of this, the model over- or underestimated the NO_2 concentrations when predicting the diurnal cycle (relatively low R^2 = 0.27). Nevertheless, in most of the cases the differences between observed values lied within ranges that can be considered as acceptable (± 15 [$\mu g \cdot m^{-3}$]; RMSE = 15.24 [$\mu g \cdot m^{-3}$]).

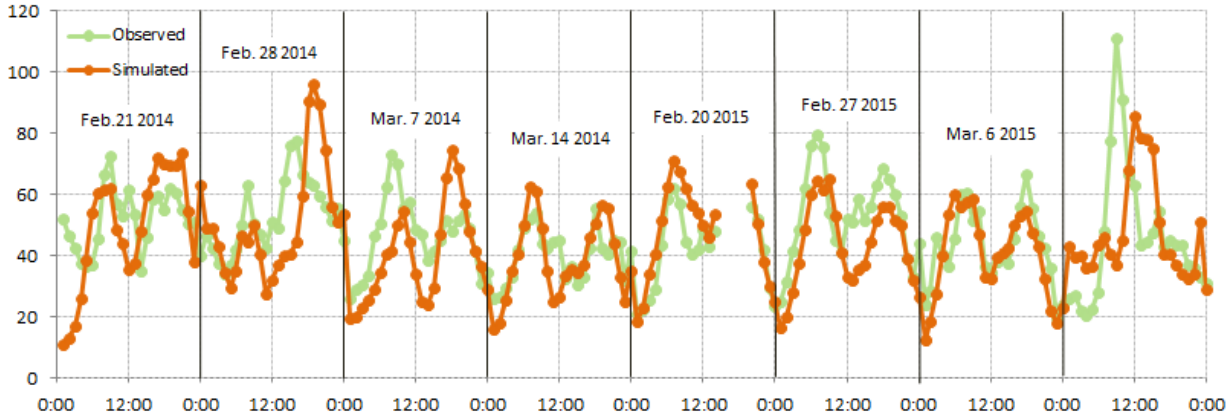


Figure 29: Observed and simulated values by NO_2 prediction. Hourly values [$\mu g \cdot m^{-3}$] for every day of analysis

Table 7: Evaluation of the kriging model for prediction of $PM_{2.5}$ and NO_2. Results for 8 selected days in February-March 2014-2015

pollutant	Coefficient of determination R^2	RMSE [$\mu g \cdot m^{-3}$]	Index of Agreement
PM_{25}	0.55	13.57	0.85
NO_2	0.27	15.24	0.70

5 Conclusions: summary and outlook

Throughout this study the temporal variability of air pollution and meteorological data in the Aburrá Valley was studied, as well as how these patterns can be used for estimation of missing data and prediction purposes. The principle conclusions of the master project are presented below:

Pollution and meteorology cycles in the Aburrá Valley

Daily, weekly and yearly air pollution cycles were detected for the area of study. The diurnal cycle showed two pollution peaks during the day: a daily maximum in the early morning and a second, less pronounced peak in the evening. The peaks of the diurnal cycle were probably associated with the traffic flow throughout the day.

The weekly cycle of air pollutants showed similar values between Monday-Saturday and a very pronounced decrease on Sundays associated to the reduction of the emission sources. Because of a gradual accumulation of the air pollution levels inside the Aburrá

Valley, Friday was in average the day with the highest concentrations in the study period. The observed weekly cycle was constant during the different months of the year, thus showing no evident interaction with the yearly cycle.

A yearly periodicity of the air pollution concentrations showed monthly peaks on February-March and yearly minima in June-July. The monthly variability depends on differences of the monthly average emissions (less traffic flow in June-July because of the holiday season) but probably also on the accumulation of pollutants in February-March because of higher atmospheric stability; further research is needed regarding this hypothesis. $PM_{2.5}$ was the pollutant that was influenced the most by this seasonal effect when compared with PM_{10} and NO_2.

All the meteorological variables with the exception of rainfall presented the expected behavior for regions near the equator. The diurnal cycle was stable throughout the year, with daily maximum values for temperature, solar radiation and wind speed around midday. The monthly variations of meteorological variables were low and not affected by an interaction with the diurnal cycle. A clear association between meteorological and air pollution could not be observed.

Geostatistical method for air pollution time series analysis

A variogram analysis was used for the integration of the diurnal and- weekly (alternatively, diurnal and- yearly) cycles of air pollutants. The different time scales were considered as coordinates in a geographical space and analyzed based on the theoretical principles of geostatistics, i.e. calculation of semivariograms and kriging interpolation. It was observed on the calculated semivariograms that at short lag distances, the overall variability of air pollution is equally dependent on the short time scale (diurnal cycle) as on the long time scale (weekly/yearly cycle).

Consequently, omnidirectional semivariograms were used to represent the air pollution variability related to the different temporal cycles in the Aburrá Valley, as well as the interactions between the different time scales. Spherical and Gaussian models were chosen for the calculation of the omnidirectional semivariograms, achieving a very accurate model fit between empirical and theoretical semivariograms.

Kriging was tested as a technique for time series reconstruction, estimation of missing data or prediction of air pollutants in the Aburrá Valley. Since the omnidirectional semivariograms were only valid at small lag distances, kriging was performed within a short delimited neighborhood (maximum distances of 4 hours /days /months). Model criticism showed validity of the model assumptions. Normal distribution was obtained for the model residuals and the R^2 values by cross validation were > 0.95 in all cases)

Very accurate results were obtained for estimation of missing data or prediction of typical diurnal cycles for different months of the year (interaction between diurnal and yearly pollution cycles). $PM_{2.5}$ showed the best results both for prediction and estimation of missing data; however differences in the accuracy of the model were identified among the months. The least accurate results were obtained for September and October, which are considered as transitional months between yearly maximum and minimum values. On the contrary, March, the month with the highest pollution concentrations in the Aburrá Valley, was simulated with high accuracy, especially for $PM_{2.5}$ ($R^2 > 0.9$).

Kriging simulations for specific days in February-March showed good accuracy for $PM_{2.5}$ prediction or estimation of missing data. The developed method was able to capture the diurnal variability, as well as the very high pollution peaks that can be observed during these months (Index of Agreement > 0.85). Lower stability of the diurnal cycle of NO_2 affected the modelling accuracy for this pollutant. The diurnal cycle of NO_2 is usually more unstable because it does not only depend on emission sources and meteorology (like PM_{10} and $PM_{2.5}$), but also on the balance of the ozone cycle.

The developed kriging method for air pollution data in the Aburrá Valley was reliable for the reconstruction of data series, analysis of interactions between time scales and estimation of missing data and prediction purposes. It analyzed and combined the variability of the different pollution cycles by the calculation of the semivariograms. As this method only uses a small neighborhood of values for the calculation of the target data, it strongly relies on the stability of the diurnal pollution cycle to obtain accurate simulations. Therefore this model should not be applied for study regions where high variability of the diurnal cycle over the year is expected. However, the modelling

procedure presented in this master project could be replicated for other big cities in the Andean region, where the diurnal pollution cycle does not change considerably throughout the year (because of their location near the equator) and is moderately stable due to strong dependency of air pollutants on anthropogenic emission sources at a local scale.

Improvement of the accuracy of the model could be possible by integrating the effect of meteorological variables into the kriging regression (as covariates). Besides that, independent semivariograms could be calculated after grouping the data depending on characteristics like: urban or rural monitoring point; Sundays and rest of the weekdays; rainy days or dry days; major emission sources, etc. This will depend on the major focus of the analysis and the availability of input data. The modeling procedure presented throughout this study could nonetheless be retained in view of the robust theoretical background of the selected approach.

A special advantage of the developed geostatistical method is that it allows integrating two different time dimension (i.e. hours/months or hours/days) for the prediction and estimation of missing data. This is not possible by the application of commonly used methods for time series analysis. This characteristic was of great importance for the analysis of air quality data in the Aburrá Valley, which showed differentiated cycles at diurnal, weekly and yearly scales

In conclusion, the established method was able to identify the pollution cycles in the Aburrá Valley and used this information for statistical modelling purposes. Its precision was especially high for $PM_{2.5}$, the pollutant which most often exceeded the Colombian Air Quality Norm during the period of study and is therefore of major concern for environmental authorities. The developed modelling procedure is flexible and could be applied in other cities, as a good alternative to the complex numerical models commonly used for the prediction of air pollution in the Andean region. Further research will be needed to identify the ideal parameters and grouping factors to obtain the most accurate predictions.

6 References

AMVA, Área Metropolitana del Valle de Aburrá. (2012). *Inventario de Emisiones Atmosféricas del Valle de Aburrá, año base 2011*.

Anttila, P., Stefanovska, A., Nestorovska-Krsteska, A., Grozdanovski, L., Atanasov, I., Golubov, N., . . . Walden, J. (2015). Characterisation of extreme air pollution episodes in an urban valley in the Balkan Peninsula. *Air Quality, Atmosphere & Health, 9*, 129–141. doi:10.1007/s11869-015-0326-7

Baldasano, J. M., Soret, A., Guevara, M., Martínez, F., & Gassó, S. (2014). Integrated assessment of air pollution using observations and modelling in Santa Cruz de Tenerife (Canary Islands). *Science of theTotal Environment, 473-474*, 576-588. doi:10.1016/j.scitotenv.2013.12.062

Bedoya, J., & Martínez, E. (2009). Calidad del aire en el Valle de Aburrá Antioquia-Colombia. *Dyna, 76*(158), 7-15.

Betancur, M. S., Urán, O. A., & Stienen, Á. (2001). Cadenas productivas y redes de acción colectiva en Medellín y el Valle de Aburrá. *Economía, Sociedad y Territorio, III*(10), 221-259.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B, 26*(2), 211–252.

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

Carslaw, D. C., & Ropkins, K. (2012). openair --- an R package for air quality data analysis. *Environmental Modelling & Software., 27-28*, 52-61

Chiverton, A., Hannaford, J., Holman, I. P., Corstanje, R., Prudhomme, C., Hess, T. M., & Bloomfield, J. P. (2015). Using variograms to detect and attribute hydrological change. *Hydrology and Earth System Sciences, 19*, 2395–2408. doi:10.5194/hess-19-2395-2015

Cohen, A. J., Anderson, H. R., Ostra, B., Pandey, K. D., Krzyzanowski, M., Künzli, N., . . . Smith, K. (2005). The global burden of disease due to outdoor air pollution. *Journal of Toxicology and Environmental Health, Part A, 68*, 1-7. doi:10.1080/15287390590936166

Cressie, N. (1988). A graphical procedure for determining nonstationarity in time series. *Journal of the American Statistical Association, 83*(404), 1108-1116. doi:10.2307/2290143

Cressie, N. (1993). *Statistics for spatial data revised edition*. USA: John Wiley & Sons, Inc.

DANE, Departamento Administrativo Nacional de Estadística. (2007). Colombia. Proyecciones de Población Municipales por Área. Retrieved from www.dane.gov.co

Doan, T. K., Haslett, J., & Parnell, A. C. (2015). Joint inference of misaligned irregular time series with application to Greenland ice core data. *Advances in Statistical Climatology, Meteorology and Oceanography, 1*, 15–27. doi:10.5194/ascmo-1-15-2015

EEA. (2015). Air pollution. Retrieved from http://www.eea.europa.eu/soer-2015/europe/air

Enzi, S., Bertolin, C., & Diodato, N. (2014). Snowfall time-series reconstruction in Italy over the last 300 years. *The Holocene*
24(3), 346-356. doi:10.1177/0959683613518590
EPA. (2015). Air Quality Trends. Retrieved from https://www3.epa.gov/airtrends/aqtrends.html
Gaviria, C. F., Muñoz, J. C., & González, G. J. (2012). Contaminación del aire y vulnerabilidad de individuos expuestos: un caso de estudio para el centro de Medellín. *Revista Facultad Nacional de Salud Pública, 30*(3), 316-327.
Gevers, M. (1985). On the use of variograms for the prediction of time series. *Systems Jr Control Letters, 6*, 15-21.
Gobernación de Antioquia. (2013). Evolución Demográfica de las Subregiones de Antioquia. Retrieved from https://www.dssa.gov.co/minisitio-dssa/
González-Duque, C. M., Cortés-Araujo, J., & Aristizábal-Zuluaga, B. H. (2015). Inflence of meteorology and source variation on airborne PM10 levels in a high relief tropical Andean city *Revista Facultad de Ingeniería, Universidad de Antioquia, 74*, 200-212.
Gu, J., Schnelle-Kreis, J., Pitz, M., Diemer, J., Reller, A., Zimmermann, R., . . . Cyrys, J. (2013). Spatial and temporal variability of PM10 sources in Augsburg, Germany. *Atmospheric Environment, 71*, 131-139. doi:10.1016/j.atmosenv.2013.01.043
Gur, S., Danielson, T., Xiong, Q., Hin, C., SreekanthPannala, Frantziskonis, G., . . . Daw, C. S. (2016). Wavelet-based surrogate time series for multiscale simulation of heterogeneous catalysis. *Chemical Engineering Science, 144*, 165-175. doi:10.1016/j.ces.2016.01.037
Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., & Kaufman, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental Health, 12*(43), 1-15.
Iaco, S. D., Palma, M., & Posa, D. (2013). Geostatistics and the Role of Variogram in Time Series Analysis: A Critical Review. In S. Montrone & P. Perchinunno (Eds.), *Statistical Methods for Spatial Planning and Monitoring*: Springer Verlag Italia.
Jalbert, J., Mathevet, T., & Favre, A.-C. (2011). Temporal uncertainty estimation of discharges from rating curves using a variographic analysis. *Journal of Hydrology, 397*, 83-92. doi:10.1016/j.jhydrol.2010.11.031
Ji, D., Li, L., Wang, Y., Zhang, J., Cheng, M., Sun, Y., . . . Miao, H. (2014). The heaviest particulate air-pollution episodes occurred in northern China in January, 2013: Insights gained from observation. *Atmospheric Environment, 92*, 546-556. doi:10.1016/j.atmosenv.2014.04.048
Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*: Academic press.
Laverde, L. A. (1988). Análisis Preliminar de Riesgos y Vulnerabilidad del Valle de Aburrá para Desastres. *Revista Investigacion y Educacion en Enfermeria, VI*(1), 97-109.
Londoño, L. A., Cañón, J. E., Villada, R. D., & López, L. Y. (2015). Caracterización espacial de PM10 en la ciudad de Medellín mediante modelos geoestadísticos. *Ingenierías USBmed, 6*(2), 26-35.
Ma, C. (2004). The use of the variogram in construction of stationary time series models. *Journal of Applied Probability, 41*(4), 1093-1103.
Matheron, G. (1963). Principles of Geostatistics. *Economic Geology, 58*(8), 1246-1266.

Ministerio de Ambiente, Vivienda y Desarrollo Territorial. (2010). *Política de Prevención y Control de la Contaminación del Aire*. Bogotá D.C.: Ministerio de Ambiente, Vivienda y Desarrollo Territorial.

Orduz, C. E., Toro, M. V., & Gómez, J. C. (2013). EPOC, bronquitis crónica y síntomas respiratorios, asociados a la contaminación por PM10 en la ciudad de Medellín (Colombia). *Revista Med, 21*(1), 21-28.

Padró-Martínez, L. T., Patton, A. P., Trull, J. B., Zamore, W., Brugge, D., & Durant, J. L. (2012). Mobile monitoring of particle number concentration and other traffic-related air pollutants in a near-highway neighborhood over the course of a year. *Atmospheric Environment, 61*, 253-264. doi:10.1016/j.atmosenv.2012.06.088

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

Poveda, G., Bedoya, M., Aristizábal, E., & Carmona, A. (2015). *Mountain Tropical Rainfall: Evidence of Phase-Locking between the Diurnal, Annual and Interannual Cycles in the Andes of Colombia*. Paper presented at the American Geophysical Union 2015 Fall Meeting, San Francisco, USA.

R Development Core Team. (2015). R: A Language and Environment for Statistical Computing. Viena, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., . . . Forastiere, F. (2013). Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Eff ects (ESCAPE). *The Lancet Oncology, 14*(9), 813–822. doi:10.1016/S1470-2045(13)70279-1

Rendón, A. M., Salazar, J. F., Palacio, C. A., Wirth, V., & Brötz, B. (2014). Effects of Urbanization on the Temperature Inversion Breakup in a Mountain Valley with Implications for Air Quality. *Journal of Applied Meteorology and Climatology, 53*, 840-858. doi:10.1175/JAMC-D-13-0165.1

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer.

Shumway, R. H., & Stoffer, D. S. (2011). *Time Series Analysis and Its Applications: With R Examples*: Springer Science+Business Media.

Steyn, D. G., DeWekker, S. F. J., Kossmann, M., & Martilli, A. (2013). Boundary Layers and Air Quality in Mountainous Terrain. In F. K. Chow, S. F. J. DeWekker, & B. J. Snyder (Eds.), *Mountain Weather Research and Forecasting*: Springer Science+Business Media B.V.

UNAL, Universidad Nacionalde Colombia Sede Medellín. (2015). *Análisis de Información 2014. Documento de Trabajo*. Medellín: Área Metropolitana del Valle de Aburrá.

UPB, Universidad Pontificia Bolivariana (2013). *Informe Final de Calidad del Aire en el Valle de Aburrá*. Medellín: Área Metropolitana del Valle de Aburrá.

WHO, World Health Organization. (2006). *Air Quality Guidelines. Global Update 2005*. Germany: Druckpartner Moser.

WHO, World Health Organization. (2014). Ambient (outdoor) air quality and health. Retrieved from http://www.who.int/mediacentre/factsheets/fs313/en/

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B), 73*(1), 3-36.

Zapata, C. E., Cano, N. A., Ramírez, M., Rubiano, C., & Jiménez, J. (2015). Influence of the extreme phases of the ENSO phenomenon (El Niño and La Niña) on air quality in the Metropolitan Area of the Aburr Valley (Colombia). *Sustainable Development, 2*, 663-675.

7 Appendix

7.1 Results of PM_{10} Variogram Analysis

Variogram Analysis for PM_{10} was performed using the same procedure as for $PM_{2.5}$ and NO_2. The theoretical semivariogram of the combined yearly and diurnal cycle was valid until a distance of 4 months/hours, whereby the semivariance until 4 hours was twice higher as the corresponding semivariance at a distance of 4 months. The resulting theoretical omnidirectional semivariogram and its parameters are presented in Figure A.1 and Table A.1

Table A.1: Parameters for the theoretical semivariogram of rPM_{10} diurnal/yearly cycles.

	rPM_{10}
Semivariogram model	Spherical (automatic)
Nugget [$(\mu g \cdot m^{-3})^{0.5}$]	0.0
Sill [$(\mu g \cdot m^{-3})^{0.5}$]	0.007
Range [(hour*month]	4.11
RMSE [$(\mu g \cdot m^{-3})^{0.5}$]	0.001

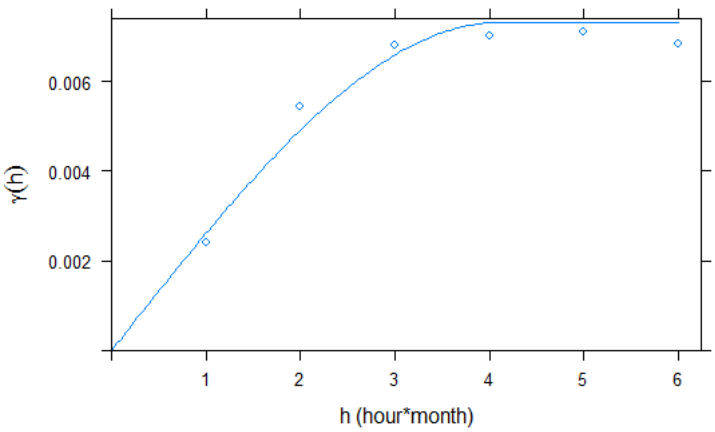


Figure A.1: Theoretical semivariogram model for rPM_{10} diurnal and yearly cycles. γ-value [$(\mu g \cdot m^{-3})^{0.5}$] and distance [hour*month]

The theoretical semivariogram for the combined diurnal and weekly cycle and its parameters (Figure A.2 and Table A.2) was valid until a distance of 2 hours/weekdays. Until this distance the semivariance of the diurnal cycle was equal to the semivariance of the weekly cycle

Table A.2: Parameters for the theoretical semivariogram of rPM_{10} diurnal/weekly cycles

rPM_{10}	
Semivariogram model	Gaussian (automatic)
Nugget $[(\mu g \cdot m^{-3})^{0.5}]$	0.0007
Sill $[(\mu g \cdot m^{-3})^{0.5}]$	0.024
Range [(hour*month]	3.13
RMSE $[(\mu g \cdot m^{-3})^{0.5}]$	0.001

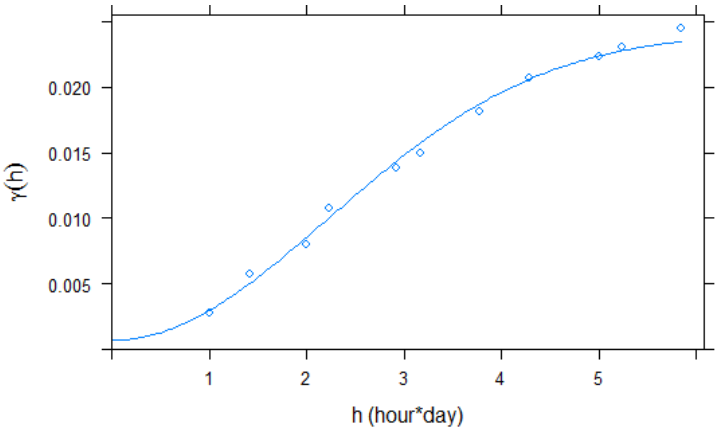


Figure A.2: Theoretical semivariogram model for rPM_{10} diurnal and weekly cycles. γ-value $[(\mu g \cdot m^{-3})^{0.5}]$ and distance [hour*week]

7.2 Kriging for PM_{10}

The kriging interpolation between diurnal and yearly cycle and diurnal and weekly cycle delivered very similar results to the $PM_{2.5}$ profiles. This was expected, given the strong correlation between both pollutants. The PM_{10} cycles are shown in Figure A.3

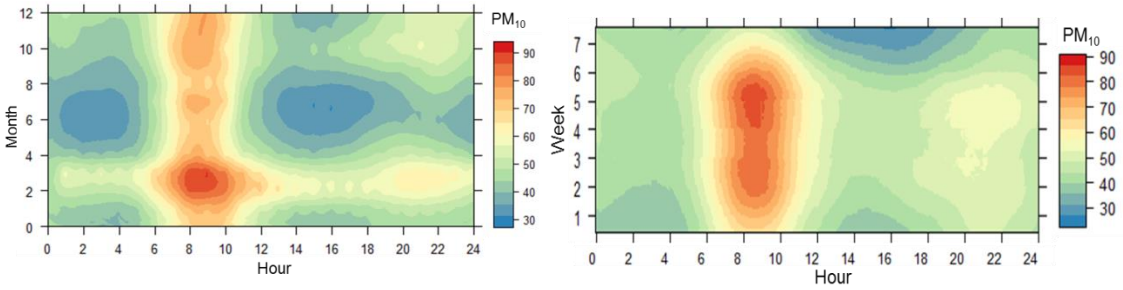


Figure A.3: Kriging interpolation of the yearly and diurnal cycle (left) and weekly and diurnal cycle (right) of PM_{10} [$\mu g \cdot m^{-3}$]. 2014 data used for the interpolation

The results of the kriging method for prediction of average diurnal profiles for every month of the year are summarized in Table A.3. Table A.4 presents the evaluation of kriging prediction for selected days in February-March.

Table A.3: Evaluation of the kriging model for prediction of PM_{10}. Results for average hourly values Jan-Dec 2014.

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PM_{10}	Mean observed value [$\mu g \cdot m^{-3}$]	53.3	66.2	67.5	53.0	48.2	45.8	46.1	47.1	52.7	56.1	54.5	55.4
	Mean simulated value [$\mu g \cdot m^{-3}$]	44.8	50.3	62.2	63.9	49.4	44.4	41.7	42.9	43.7	48.6	52.0	50.8
	Coeff. determ. R^2	0.85	0.92	0.82	0.88	0.98	0.98	0.99	0.97	0.96	0.91	0.94	0.97
	RMSE [$\mu g \cdot m^{-3}$]	9.5	16.3	7.2	11.5	2.6	2.6	5.5	5.8	9.4	8.1	3.9	5.0
	Index of Agreement	0.81	0.65	0.90	0.78	0.98	0.99	0.95	0.94	0.83	0.87	0.97	0.95

Table A.4: Evaluation of the kriging model for prediction of PM_{10}. Results for 8 selected days in February-March 2014-2015

pollutant	Coefficient of determination R^2	RMSE [$\mu g \cdot m^{-3}$]	Index of Agreement
PM_{10}	0.46	19.3	0.82

A graphical comparison between observed and predicted PM_{10} values is presented in Figure A.4 and A.5. The kriging method was capable of accurately predict the diurnal variability of PM_{10}

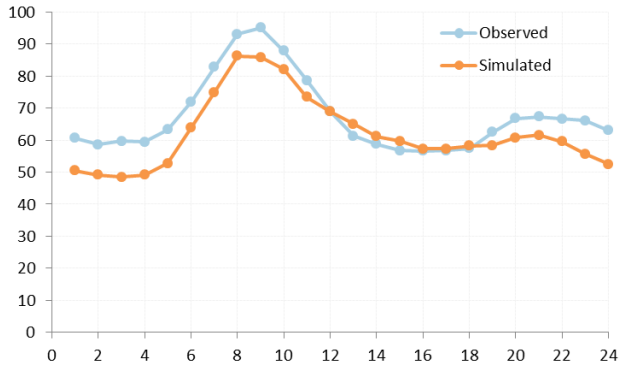


Figure A.4: Observed and predicted values March 2014. PM_{10} [$\mu g \cdot m^{-3}$] hourly averages

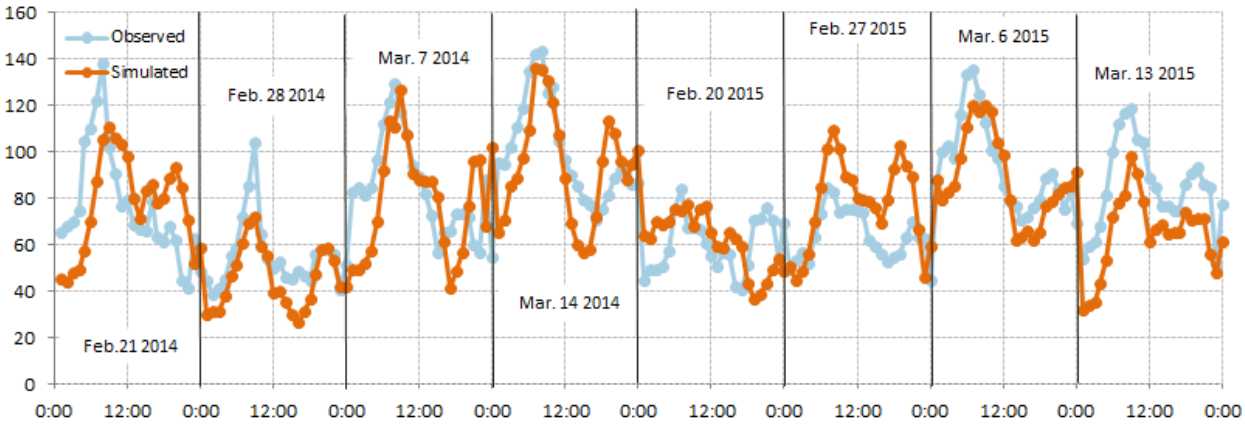


Figure A5: Observed and simulated values by PM_{10} prediction. Hourly values for every day of analysis

7.3 Data Transformation of $PM_{2.5}$ and NO_2

Lilliefors-Tests were applied to check for normal distribution in the $PM_{2.5}$ and NO_2 data. The untransformed data presented a non-normal distribution. A power transformation (fourth root) was applied to obtain normally-distributed data. After the transformation the data showed normal distribution. The fourth root transformed data was used for the calculation of semivariograms. The results of the transformation are summarized in Figure A.6 and Table A.5

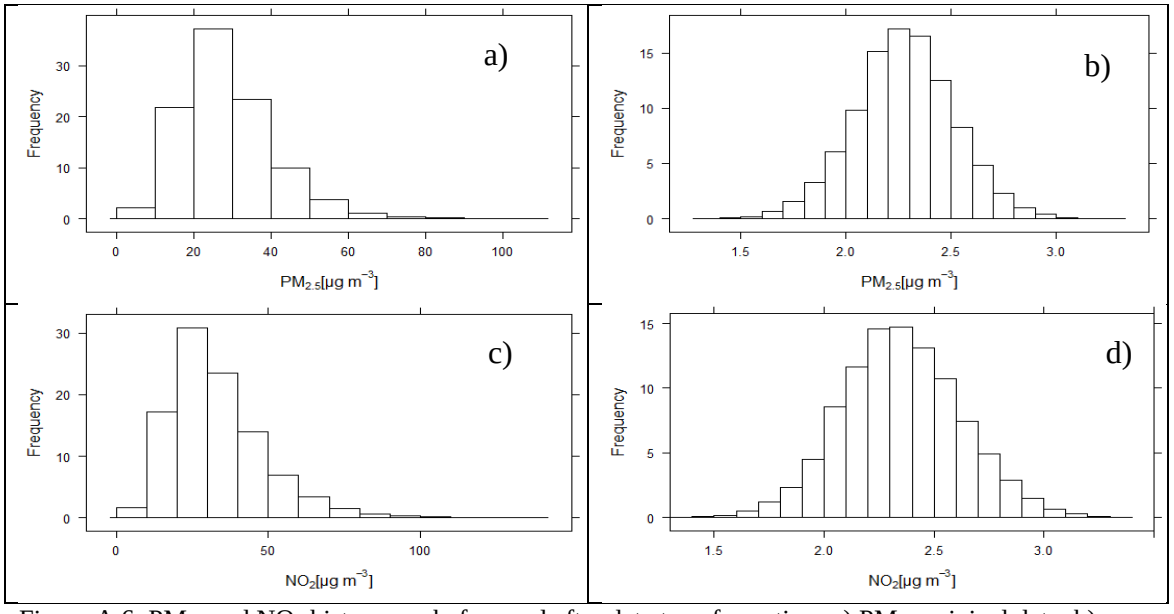


Figure A.6: PM_{25} and NO_2 histograms before and after data transformation: a) $PM_{2.5}$ original data; b) $PM_{2.5}$ data fourth root transformed; c) NO_2 original data; d) NO_2 data fourth root transformed

Table A.5: p-values of Lilliefors test for PM_{25} and NO_2. Original data and fourth root transformation

pollutant	p-value Original Data	p-value 4th-root transformation
PM_{25}	1.363e-08	0.4858
NO_2	7.219e-14	0.4187

Acknowledgments

This work was supported by the following people; without their collaboration its development would have been impossible.

First of all, I would like to express my highest consideration to my main supervisor Professor Dr. Karl Auerswald, Grassland Gruoup, Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt, TUM, whose support, advices and constructive criticism have helped me throughout the project.

I am equally grateful to Professor Dr. Uwe Schlink, Department of Urban and Environmenatal Sociology, Helmholtz Centre for Environmental Research - UFZ, who was my supervisor at this institution. His guidance and expertise in atmospheric research, geostatistical analysis and modelling were of great importance for the development of the research study.

In addition, I would like to thank the team of REDAIRE, Universidad Nacional de Colombia, Sede Medellín, for providing the air pollution data used in the project, as well as for their support regarding its analysis and interpretation.

I am also very thankful to my colleagues Abdul, Daniel and Max, who were always happy to help me and made the stressful time during the thesis much more enjoyable.

Finally, the biggest gratitude to the people I consider my family: Sofía, who has been like a second mother to me; my girlfriend Andrea, who has always been by my side and gives me hope in the most difficult moments; and my uncle René and my aunts Rose and Teresa, who I know are thinking about me all the time in Chile. Thank you all very much!