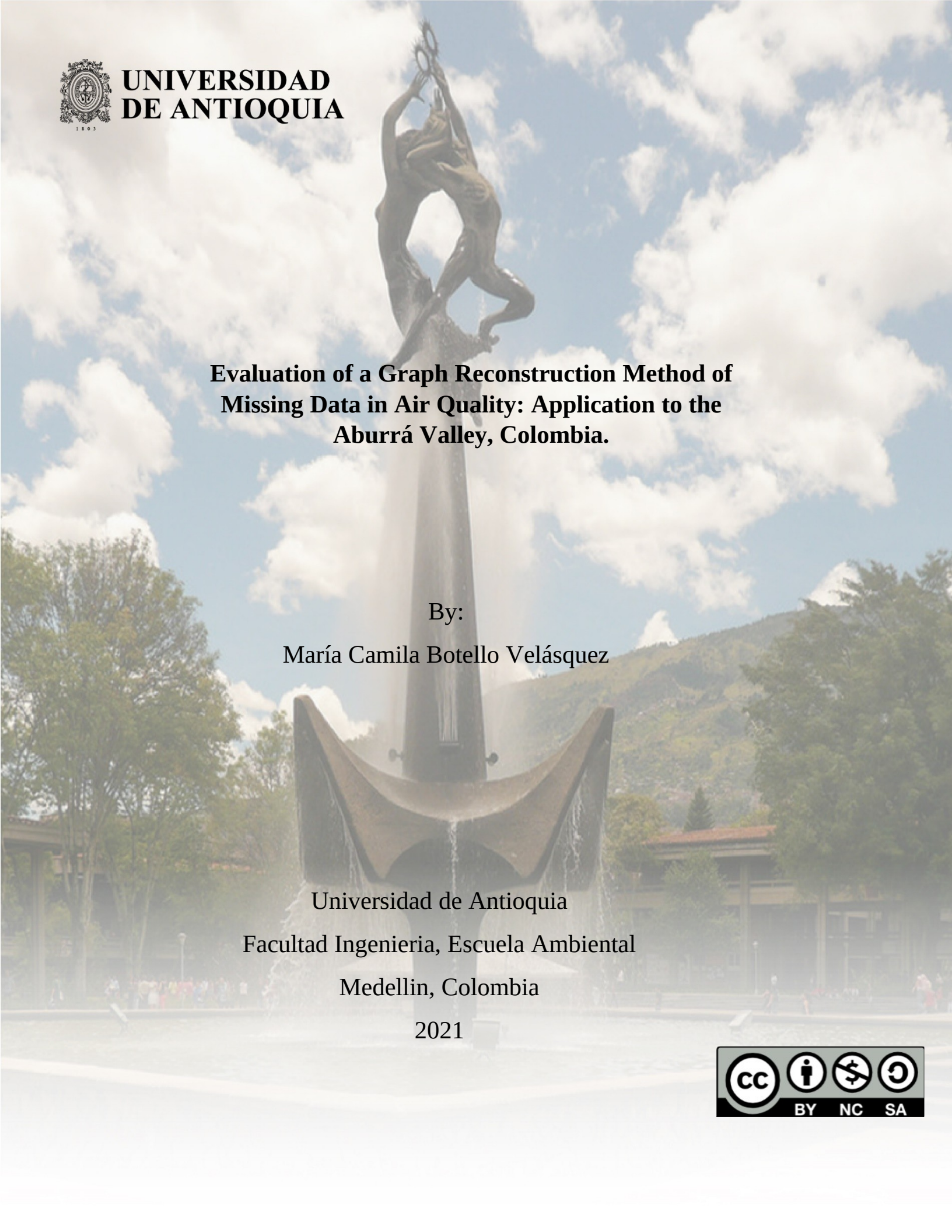

UNIVERSIDAD DE ANTIOQUIA

Evaluation of a Graph Reconstruction Method of Missing Data in Air Quality: Application to the Aburrá Valley, Colombia.

By:

María Camila Botello Velásquez

Universidad de Antioquia

Facultad Ingenieria, Escuela Ambiental

Medellin, Colombia

2021

UNIVERSIDAD DE ANTIOQUIA

Evaluation of a Graph Reconstruction Method of Missing Data in Air Quality: Application to the Aburrá Valley, Colombia.

by

María Camila Botello Velásquez

Advisor: Ph.D. Ángela María Rendón Pérez

A research work submitted in partial fulfillment for the degree of Bachelor of Science: Environmental Engineer

in the

Facultad de Ingeniería, Escuela Ambiental

Universidad de Antioquia

April 26, 2021

Dedicated to my beloved family.

"Enfrenta los obstáculos a medida que se presenten, no pierdas energía temiendo lo que pueda haber en el futuro."

Isabel Allende

Abstract

Air pollution is an environmental issue that concerns human health all around the world. The air quality is affected by human emissions, meteorological conditions, and topography. The measurement of pollutants is an important task to make better decisions for controlling high pollution concentrations. However, air quality sensing usually has problems due to machine failures, routine maintenance, among others. As a result, air quality datasets could have missing information that sometimes could represent more than 10% of the data. The correct reconstruction of these missing values plays an essential role in further environmental studies. In this work, we model the reconstruction of missing data as a problem of recovery of graph signals. Therefore, we evaluate the robustness of a graph signal reconstruction method in a dataset of Particular Matter ($PM_{2.5}$) in the Aburrá Valley, Colombia. We observe that 1) the model has better performance during dry months than in wet or transition seasons, and 2) the model could not follow pollution peaks because the algorithm assumes smooth changes in time. This model could be suitable to reconstruct data in the Aburrá Valley in dry seasons for other environmental studies.

Keywords: air quality, missing data, data reconstruction, graph signal processing.

Resumen

La contaminación atmosférica es un problema ambiental que afecta a la salud humana mundialmente. La calidad del aire se ve afectada por emisiones antropogénicas, por condiciones meteorológicas y por la topografía. La medición de contaminantes atmosféricos es una tarea importante para la toma de decisiones, por ejemplo, para controlar altas concentraciones de contaminación en una ciudad. Sin embargo, en la medición de la calidad del aire generalmente hay problemas debidos a fallas de los equipos, mantenimiento de rutina, entre otros. Como resultado, los conjuntos de datos de calidad del aire pueden tener información faltante, que a veces puede representar más del 10% de los datos. La reconstrucción de estos valores faltantes juega un papel importante en los estudios ambientales. En este trabajo, modelamos la imputación de datos faltantes como un problema de reconstrucción de señales gráficas. Evaluamos la robustez de un método de procesamiento de señales gráficas en un conjunto de datos de Material Particulado ($PM_{2.5}$) en el Valle de Aburrá, Colombia. Observamos que 1) el modelo tiene un mejor desempeño durante los meses secos que durante temporadas húmedas o de transición y 2) el modelo puede no predecir picos de contaminación dado que el algoritmo asume cambios suaves en el tiempo. Este modelo podría ser útil para reconstruir datos en el Valle de Aburrá en temporadas secas para ser utilizados en futuros estudios de calidad del aire.

Palabras clave: calidad del aire, datos faltantes, reconstrucción de datos, procesamiento de señales gráficas.

Acknowledgements

Firstly, I would like to express my gratitude to my advisor Prof. Ángela Rendón for the support in my bachelor thesis, for the trust she has had in me, and for her immense knowledge. Besides my advisor, I would like to thank Prof. Thierry Bouwmans and Jhony Giraldo for their insightful comments and encouragement.

I thank my friend Gisela for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last years. Also, I thank Jean-Christophe, Lucia, Newt, Wassila, and Kadija for their support and understanding these past months.

I would like to thank my mom for supporting me all my life and for her infinite love. Finally, I could not have completed this work without the support of Jhony. I thank him for always being there for me.

Contents

List of Figures

List of Tables

Section 1

Introduction

Air quality is a problem that concerns many cities around the world since air pollutants have negative impacts on human health. Air pollution causes an increase in mortality rate in several health problems such as strokes, heart diseases, chronic obstructive pulmonary diseases, lung cancer, and acute respiratory infections. More than 90% of people in the world breathe polluted air, and more than 7 million people die as a consequence of long-term exposure to polluted air every year [33]

Nowadays, measuring the air quality for monitoring the pollutant concentrations is an important task in urban areas. Monitoring with reliable datasets helps to identify 1) pollution concentration, 2) pollution hotspots, 3) pollution transport, and 4) extreme pollution events. Nevertheless, it is complicated to monitor, analyze, control, and manage air pollution in cities, especially in regions with complex topographic and meteorological conditions [8].

Often there are problems with air quality datasets as a result of missing data. Missing values can be due to 1) machine failures, 2) communication failures, 3) routine maintenance, and 4) human error, or other factors [42]. Lack of data can affect the performance of air pollution trend analyses, forecasting studies, and epidemiological studies, among others. Therefore, reconstructing the missing data becomes helpful to be able to use the information.

Researchers have proposed different methods to solve the problem of incomplete or unreliable data. Some studies omit data by removing rows or columns that contain missing values. However, this method can lead to omitting useful information that will increase the error in air quality analysis [18]. Instead of removing the missing information, data reconstruction methods keep the whole sample by filling missing values with imputed data. A variety of reconstruction approaches have been proposed and used.

For example, statistical imputation [2], machine learning-based imputation [23], neural networks [22, 24], and signal processing approaches [12, 42] are widely used methods in air pollution studies.

Single imputation methods like mean [32, 38] and linear interpolation [31] are perhaps the easiest way to compute missing data. They fill the missing values using the corresponding mean or median of the observed values. Moreover, Nearest Neighbor (NN) algorithm fills the missing values using the value of the nearest neighbor (row) available, at the same time weighting the distances in proportion to the number of missing values in each row [3, 21, 44]. Alahamade et al. [2] combined different approaches to estimate the time series of missing pollutants. They used single and multiple imputation methods, such as NN imputation, Simple Moving Average (SMA), and Multivariate Imputation via Chained Equations (MICE) to impute ozone (O_3) concentration values.

Other imputation methods are more sophisticated such as machine learning-based imputation. Liu et al. [23] used a Low-Rank Matrix Completion (LRMC) algorithm to reconstruct PM_{10}, $PM_{2.5}$, NO_x, O_3, and SO_2 data. This approach reconstructs a matrix from the observed subset of its entries based on the low-rank property of the original matrix. However, LRMC assumes the original data is low-rank, which could not be true. That approach is limited since air pollution presents complex temporal and spatial relationships [42].

Unlike the methods described above, neural networks are widely used and have a good performance. Kalteh & Berndtsson [22] used neural networks to interpolate precipitation data. They used Multi-Layer Perceptron (MLP) and Self Organization Maps (SOM). Ma et al. [24] used an approach based on Long Short-Term Memory (LSTM) neural network, transfer learning, and iterative estimation to impute consecutive missing values. However, neural network methods require large amounts of data for neural network training.

Graph Signal Processing (GSP) is an emerging research field that analyzes signals living on irregular structures captured by graphs. GSP has many applications such as climate analysis [9, 11, 26] and sensor networks [40, 45]. GSP is a powerful tool when there are missing data problems. For example, this tool has been used to reconstruct missing data from the sea surface temperature, the global sea-level pressure, and the daily mean $PM_{2.5}$ datasets of California [37], to name a few.

There are some studies in the domain of GSP for reconstructing missing data. Qiu et al. [37] have demonstrated the excellent performance of GSP methods in data reconstruction with several experiments. For example, they used the California daily mean $PM_{2.5}$ concentration dataset with percentages of valid data ranging from 90% to 45%.

They proposed a batch method that was compared with Natural Neighbor Interpolation (NNI) [41], Low-Rank Matrix Completion [25], graph regularization [30], and graph-time Tikhonov [36]. When the sampling rate is 40%, they found that their method achieves better performance than all the compared methods.

Giraldo and Bouwmans [12] also used a dataset of the daily concentration of $PM_{2.5}$ in California and a dataset of sea surface temperature. They used a Sobolev reconstruction method that was compared with Qiu's method and NNI. They found that their method outperforms NNI in the sea surface temperature and $PM_{2.5}$ concentrations while having approximately the same performance that Qiu's method.

Unfortunately, most of these approaches have been applied in the domain of computer science. Few methods consider the complexity of the atmosphere and its different phenomenon. In general, data reconstruction methods evaluate air pollution as a problem in time and space. However, air pollution also depends on meteorological, climate, and geographical conditions [39].

Big cities located in complex terrains such as valleys, usually experience serious air pollution problems. The transport of air pollutants emitted from urban valleys depends on topographical conditions, wind fields, and the dynamics of the Atmospheric Boundary Layer (ABL) [14]. Topography affects pollutant concentrations because it acts as an impermeable barrier for atmospheric flows [43]. The ABL height determines the pollutant concentration because the Convective Boundary Layer (CBL) acts as an interface for exchanging momentum, water vapor, gases, and pollutants from the surface to the atmosphere [7]. Anthropogenic emissions build up and lead to critical air pollution episodes when there are unfavorable meteorological conditions [16].

Medellin is a city in the Aburrá Valley, a highly complex mountainous terrain located between the west and central Colombian mountains [16]. In this region, the air quality depends on emissions sources, meteorological conditions, and topographic barriers [39]. Since 2014 the air quality has been a relevant concern in this city. The local government has established an air quality monitoring network since 2013. It is a system that makes real-time monitoring of hydrological, meteorological, and air quality conditions. However, this system, like others around the world, presents missing values that could represent more than 10% of the data. Imputation or reconstruction of air pollutants data is an essential task, notably when the number of missing values is significant [23].

In this work, we model the air pollution data as a problem of reconstruction of time-varying graph signals. This model has the advantage of considering both spatial and temporal information and does not require a large amount of data, unlike previous approaches. The prior assumption of this model is that the data vary smoothly both in

space and in time. That is to say, the concentration of pollutants should be similar in nearby areas, and the change in time is gradual [12].

Our methodology will be applied to air pollution data in the Aburrá Valley in Colombia. The measurement, prediction and management of atmospheric pollution is a challenge in the region because it is affected by different phenomena and their interactions. We evaluate the performance of the graph model by using statistical metrics such as Root Mean Square Error (RMSE) and Coefficient of Correlation (R^2). The principal objective of this study is to test the robustness of the model to reconstruct properly the missing data involving $PM_{2.5}$ concentration changes due to meteorological variability, e.g., the case where there is a change from dry to rainy season. This research work is organized as follows. In Section 2, we describe the study area and its meteorological conditions. Section 3 presents a description of the model and the data used. Section 4 includes a description of the numerical model configuration and the experimental set-up. Then, in Section 5 we present the results of the graph model reconstruction. Finally, Section 6 presents the most important conclusions of the study.

Section 2

Study Area

The Aburrá Valley is a narrow valley with 64 km in length and 18.2 km wide. It is located in the Central Andes mountain (6°N - 6.5°N, and 75.3°W - 75.6°W); and its topography is irregular and sloping. The heights of mountains are between 1500 and 2800 m above sea level [15, 16, 27]. Figure 2.1 shows the geographical location, the topography of the Aburrá Valley, and the monitoring stations used in this work. The valley has a bimodal annual cycle of precipitation with two periods of higher rainfall (April to May and September to October), and two drier seasons (December to February and June to August) [5]

In the Aburrá Valley, the ABL height depends on meteorological conditions. The ABL height is higher during June and lower during March [15, 19]. This change in the ABL height can explain why during March and April the atmospheric pollution conditions are worse than in June. Moreover, low clouds form in the valley during March. These clouds decrease the radiative fluxes on the surface within the valley. As a result, the winds are calm at the local scale, and turbulent processes in the atmosphere are affected. These modifications lead to a low removal of atmospheric pollutants [10]. A similar phenomenon occurs during November where there is a transition between wet and dry seasons. However, the concentrations of $PM_{2.5}$ registered during this month are lower than during the first transitional period (March-April). This phenomenon is due to the behavior of the trade winds on the surface that is different during November [28].

The Metropolitan Area of the Aburrá Valley (AMVA) is a large urban agglomeration made up of ten municipalities. Medellin is the principal city in the valley, and is an important economic center. The AMVA is the second most important metropolitan area in Colombia, and has approximately 4 million inhabitants [8, 16]. Currently, it is the second-largest urban area and has one of the most important industrial centers in

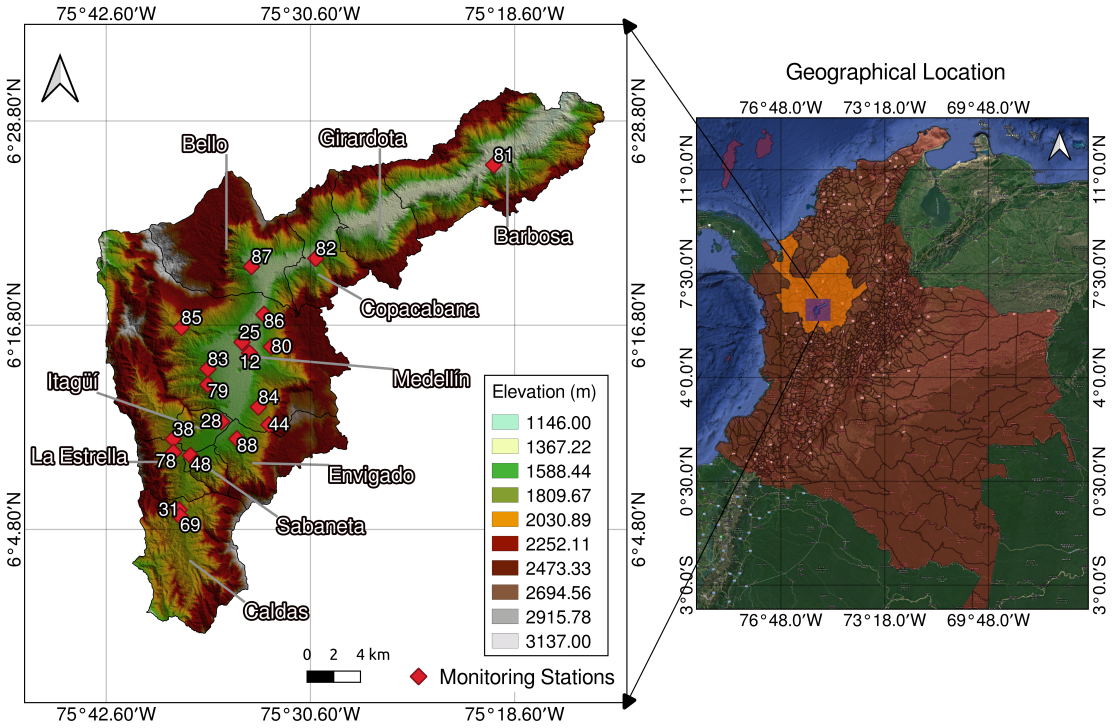

FIGURE 2.1: The geographical area of the Aburrá Valley, located in the department of Antioquia (in orange to the right), Colombia. The map shows (in red to the left) the location of the monitoring stations.

the country. As a consequence, industry and transportation produce a high quantity of air pollutants.

The most critical episode of pollution was reported in 2016 (in February-March) in the Aburrá Valley. This episode was related to local meteorological conditions and atmospheric boundary layer variability [1, 16, 35]. Peláez et al. [35] found that in Medellin, between the years 2012-2017, the average annual concentrations of $PM_{2.5}$ exceeded the guideline value given by the World Health Organization (WMO). $PM_{2.5}$ has a bimodal behavior in the region because there is a peak in March and another in November. The maximum concentration of $PM_{2.5}$ is in March as a consequence of the transition between the dry and wet seasons. Moreover, this pollutant has a noticeable diurnal cycle. It has a peak of pollution in the morning (around 8:00) and another smaller peak at night (around 20:00) [17].

Section 3

Methodology and Data

This section presents the preprocessing steps of data and the model. Figure 3.1 shows the framework of this study. First, we have raw data of air quality from the Aburrá Valley. Afterward, a preprocessing and exploratory analysis is done in this dataset. The graph model is executed to reconstruct the missing data. Finally, three experiments are executed, and the results are analysed.

3.1 Air quality data

The air quality data come from the governmental institution Early Warning System of Medellín and the Aburrá Valley SIATA (Spanish acronym). SIATA provides air quality, meteorological, and hydrological data on their website `www.siata.gov.co`. Several researchers have used this dataset to identify: 1) what is the state of air quality in the metropolitan area, 2) what is the distribution of pollutants, and 3) what is the danger of pollutants in public health [1, 4, 8, 29].

In this study, we work with information between January 1st and December 31st, 2019 obtained from automatic stations of SIATA. The selected stations are distributed among nine of ten municipalities of the Aburrá Valley, and the data includes hourly measurements of $PM_{2.5}$. Table 3.1 shows the description of each station used in this work. $PM_{2.5}$

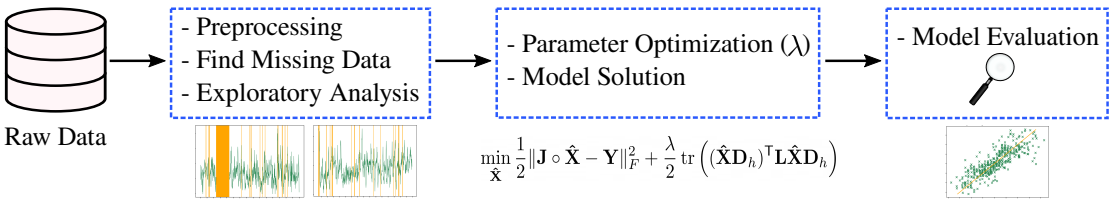

FIGURE 3.1: Methodology overview.

was selected due to its major contribution to the air pollution problems in the Aburrá Valley and its impacts on human health. We selected 2019 because of data availability and better meteorological characterization. The database was preprocessed as required by the graph model using Python.

TABLE 3.1: Summary of stations of the SIATA in the Aburrá Valley.

Code	Station
12	MED: Estación Tráfico Centro
25	MED: Centro-Occidente UNAL
28	ITA: Casa de la Justicia
31	CAL: Corp. Universitaria Lasallista
38	ITA: I.E Concejo de Itagüí
44	MED: El Poblado-Tanques de la Y
48	MED: Estación Tráfico Sur
69	CAL: E.U Joaquin Aristizabal
78	EST: La Estrella-Hospital
79	MED: I.E Pedro Octavio Amado
80	MED: Villa Hermosa
81	BAR: Torre Social
82	COP: Ciudadela Educativa La Vida
83	MED: Belén-I.E Pedro Justo Berrio
84	MED: INEM Sede Santa Catalina
85	MED: San Cristobal
86	MED: Aranjuez-I.E Ciro Mendia
87	BEL: I.E Fernando Vélez
88	ENV: E.S.E Santa Gertrudis
90	ITA: Estación de Policía los Gómez

MED: Medellín, ITA: Itagüí, SAB: Sabaneta, BAR: Barbosa, BEL: Bello, COP: Copacabana, ENV: Envigado, CAL: Caldas, EST: La Estrella.

3.2 Reconstruction model

A graph is usually represented as a set of nodes and a set of edges. This representation is encoded with the adjacency matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$, where N is the number of nodes. Similarly, the value $\mathbf{W}(i,j)$ inside the adjacency matrix represents the connectivity between the node i and j, i.e., if these nodes are connected the value $\mathbf{W}(i,j)$ is different from zero. Another important matrix is the diagonal degree matrix defined as $\mathbf{D}(i,i) = \sum_{j=1}^{N} \mathbf{W}(i,j)$. The Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Figure 3.2 shows the graph corresponding to the dataset of $PM_{2.5}$ in the Aburrá Valley. This graph was constructed with a k Nearest Neighbors (k-NN) with $k = 5$ [34].

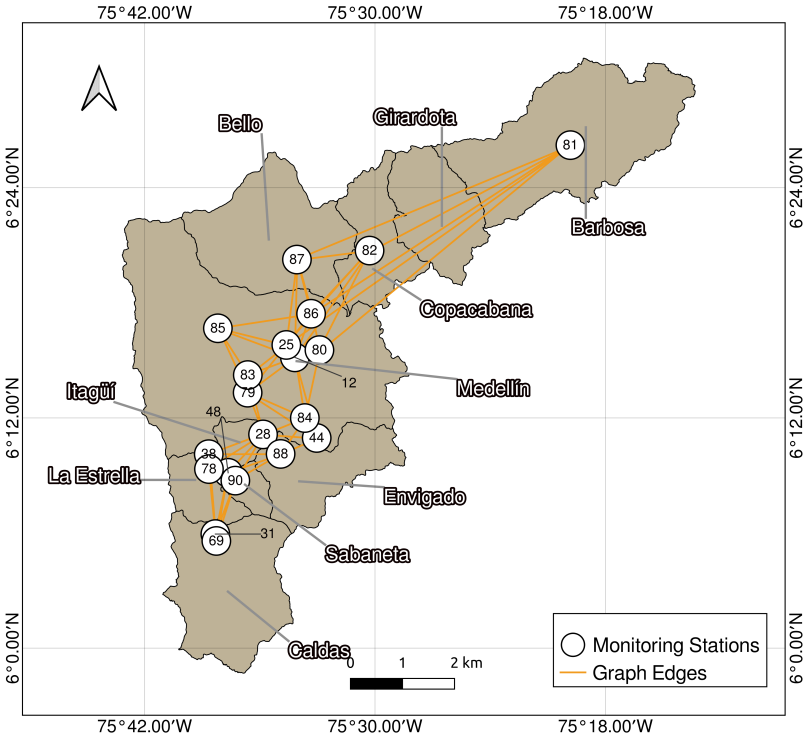

FIGURE 3.2: Graph representation of the $PM_{2.5}$ dataset in the Aburrá Valley.

In this work, the $PM_{2.5}$ is represented with a matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$, where N is the number of stations, and M is the number of data points in time. Similarly, the temporal differences operator was introduced by Qiu et al. [37] as a matrix $\mathbf{D_h} \in \{-1, 0, 1\}^{M \times (M-1)}$, and it is defined as follows:

$$\mathbf{D}_h = \begin{bmatrix} -1 & & & \\ 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{bmatrix} \in \{-1, 0, 1\}^{M \times (M-1)} \tag{3.1}$$

In the same way, the sampling matrix contains the information of whether a station at a specific time is failing, i.e., the sampling matrix tells us the location of missing data. This matrix $\mathbf{J} \in \{0, 1\}^{N \times M}$ is given as follows:

$$\mathbf{J}(i, j) = \begin{cases} 1 & \text{if } i \in \mathcal{S}_t, \\ 0 & \text{if } i \notin \mathcal{S}_t, \end{cases} \tag{3.2}$$

where S_t is the set of good nodes at time t. Finally, the reconstruction method solves the following optimization problem:

$$\min_{\hat{\mathbf{X}}} \frac{1}{2}\|\mathbf{J} \circ \hat{\mathbf{X}} - \mathbf{Y}\|_F^2 + \frac{\lambda}{2} \operatorname{tr}\left((\hat{\mathbf{X}}\mathbf{D}_h)^\top \mathbf{L}\hat{\mathbf{X}}\mathbf{D}_h\right) \tag{3.3}$$

where $\mathbf{Y} \in \mathbb{R}^{N \times M}$ is the matrix of observed values (the data that we know). This optimization problem is solved with conjugate gradient method [37].

3.3 Evaluation of the model performance

Residual methods are widely used for evaluation of environmental models, these methods calculate the difference between observed and modelled data points [6]. In this work, we used two residual metrics, frequently used in environmental data reconstruction studies [13, 20, 21, 23, 37, 38], to evaluate the performance of the reconstruction model: 1) Root Mean Square Error (RMSE) and 2) Mean Absolute Error (MAE). In addition, we used the Coefficient of determination (R^2) to test the ability of the model to preserve the pattern of data. The metrics are defined in Equations 3.4, 3.5 and 3.6.

RMSE is a metric for determining the error that summarises the difference between the observed and reconstructed concentration values. This metric expresses the error in the same units as the original data [6].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{3.4}$$

MAE is a metric similar to RMSE, except that the absolute value is used instead, thus, reducing the bias towards large events and is a more sensitive measure of residual error [6].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i| \tag{3.5}$$

Coefficient of determination R^2 is a square version of the Pearson correlation coefficient and ranges from 0 to 1. It describes the variance between the observed and the predicted concentrations [6].

$$R^2 = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\hat{x}_i - \tilde{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(\hat{x}_i - \tilde{x})^2}}\right)^2 \tag{3.6}$$

Section 4

Experimental Framework

Since we are unable to directly determine the performance of the reconstruction method on missing data, a portion of the actual data sets was randomly removed to estimate the ability of the model to reconstruct these data. The real corrupted data were not taken into account in the reconstruction method.

In this work, the graph reconstruction method was implemented in MATLAB. We performed three experiments (Figure 4.1). In the first experiment, we use the whole dataset, i.e., the 20 stations during all the year 2019. We compute the RMSE, MAE and R^2 for several missing data percentages in the set $\{0.1, 0.2, ..., 0.9\}$. Furthermore, a Monte Carlo cross-validation with three repetitions was performed.

For the second experiment, a Monte Carlo cross-validation with seven repetitions was performed. The RMSE was computed for several sampling densities, for each month of

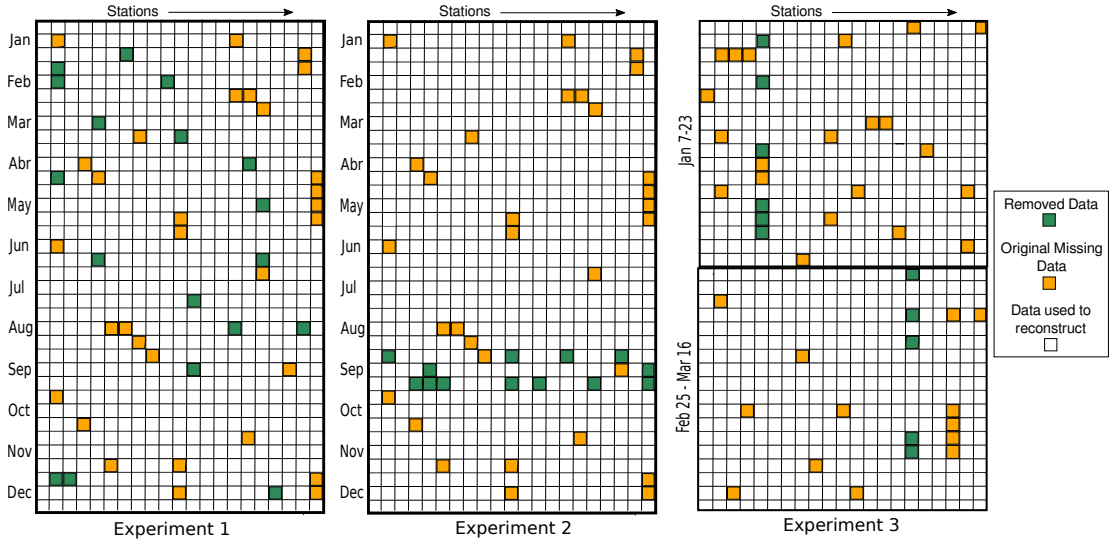

FIGURE 4.1: Experimental framework

the year 2019 in the dataset of the Aburrá Valley. In this case, we remove the samples from specific months while using the available data from the rest of the dataset. The sampling densities are in the set $\{0.1, 0.2, ..., 0.9\}$, where each sampling density means the amount of valid information. For example, when the sampling density is 0.9 the reconstructed data is 10%.

The third experiment also computes the RMSE, but in this case for the dry and first transition season. The dry season is between the days January 7 and 23, while the transition season (dry to wet) is between February 25 and March 16. In this experiment, the RMSE is performed for each station, and thus a Monte Carlo cross-validation with three repetitions is done. The percentage of reconstructed data in this case is in the set $\{1, 0.8, ..., 0.2\}$, where percentage of 1 (100%) means that the whole data in the selected window is missing for a specific station.

Section 5

Results and Discussion

This section presents the results of an exploratory data analysis of $PM_{2.5}$ in the Aburrá Valley. Moreover, this section shows the results of the different experiments performed of data reconstruction under different percentages of missing values.

5.1 Exploratory data analysis

An exploratory data analysis was performed to establish the quality of the time series of $PM_{2.5}$. This analysis determines the quantity and length of gaps of missing data. The amount of missing data in the dataset is 6.57%, with 6029 gaps of different time lengths. Table 5.1 shows the general information of missing data gaps. Most of the data gaps are in the interval of $g \leq 3h$, which means that the length of almost all gaps is 3 hours or less. Similarly, Table 5.1 shows that approximately 7% of the missing data is in the interval $3h < g \leq 12h$. Finally, we have less than 1% of missing gaps for time windows greater than 24 hours.

Table 5.1: General information of missing data gaps.

Gaps	Missing data (%)
g $\leq$ 3 h	92.52
3 h < g $\leq$ 12 h	6.68
12 h < g $\leq$ 24 h	0.48
24 h < g $\leq$ 36 h	0.08
36 h < g $\leq$ 48 h	0.03
48 h < g $\leq$ 60 h	0.00
60 h < g $\leq$ 72 h	0.00
g > 72 h	0.20

g: length of gap in time, h: hour

Table 5.2 shows detailed information about the missing data of the $PM_{2.5}$ database. We have the number of gaps of each station for every month in the year 2019. Furthermore, we show the maximum length of a gap in hours for every month. For example, in January the station 12 has 35 gaps of different lengths, and the maximum gap is 79 continuous hours. The maximum missing gap is 99 continuous hours in the whole dataset for station 81 in September. One of the worst cases is station 48 in May because it has a big gap of 73 continuous hours and 55 gaps of different lengths. Table 5.2 also shows that the least corrupted data is in station 69, where we do not have more than 20 gaps each month, and there is only a considerable gap in September.

TABLE 5.2: Detailed monthly information of missing data of $PM_{2.5}$ in the Aburrá Valley for each station.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Station	Number of gaps (number of hours of the maximum gap)											
12	**35 (79)**	31 (9)	40 (11)	54 (6)	55 (8)	45 (12)	41 (6)	28 (11)	26 (3)	35 (40)	31 (5)	22 (5)
25	25 (7)	16 (7)	24 (12)	38 (7)	53 (8)	69 (12)	69 (7)	46 (11)	73 (8)	67 (9)	53 (23)	42 (9)
28	**14 (74)**	19 (4)	15 (10)	18 (3)	8 (11)	18 (12)	26 (4)	21 (10)	18 (8)	24 (3)	11 (5)	16 (6)
31	**23 (77)**	23 (3)	21 (13)	45 (5)	46 (5)	42 (14)	35 (28)	18 (11)	35 (15)	38 (7)	48 (11)	37 (21)
38	17 (3)	21 (2)	28 (9)	19 (2)	38 (5)	**47 (77)**	21 (3)	21 (10)	16 (2)	22 (2)	16 (20)	26 (3)
44	22 (3)	20 (14)	37 (19)	33 (4)	**37 (74)**	44 (14)	28 (4)	24 (25)	25 (15)	37 (4)	37 (5)	34 (8)
48	55 (4)	26 (5)	19 (9)	55 (13)	**55 (73)**	83 (45)	95 (6)	61 (9)	74 (3)	66 (5)	50 (4)	53 (10)
69	16 (7)	14 (3)	7 (2)	11 (3)	12 (3)	11 (12)	16 (7)	14 (16)	**14 (76)**	11 (2)	18 (5)	17 (3)
78	25 (6)	16 (4)	9 (3)	15 (3)	17 (9)	18 (12)	22 (4)	21 (10)	**14 (75)**	19 (4)	18 (3)	17 (4)
79	15 (3)	13 (5)	19 (4)	20 (4)	18 (2)	21 (12)	23 (3)	30 (5)	17 (3)	12 (3)	17 (4)	16 (3)
80	20 (4)	11 (14)	12 (4)	23 (2)	23 (3)	32 (12)	38 (15)	22 (9)	21 (3)	25 (3)	25 (21)	37 (7)
81	25 (4)	15 (9)	11 (5)	21 (4)	48 (3)	38 (15)	38 (3)	28 (7)	**23 (99)**	31 (2)	36 (5)	29 (5)
82	15 (33)	17 (10)	8 (3)	18 (3)	22 (2)	28 (12)	15 (4)	23 (9)	17 (4)	22 (6)	19 (4)	22 (7)
83	18 (14)	11 (3)	7 (5)	16 (3)	12 (5)	24 (12)	22 (3)	20 (13)	14 (3)	13 (3)	10 (6)	13 (3)
84	21 (11)	10 (4)	8 (2)	24 (5)	19 (14)	27 (12)	20 (2)	23 (6)	17 (8)	18 (3)	21 (9)	22 (5)
85	18 (4)	12 (15)	11 (4)	16 (4)	21 (17)	24 (12)	20 (2)	23 (11)	19 (7)	20 (4)	21 (26)	19 (8)
86	16 (3)	19 (3)	8 (3)	10 (2)	13 (7)	19 (12)	13 (2)	23 (10)	15 (6)	14 (3)	17 (18)	26 (5)
87	**19 (75)**	19 (6)	14 (3)	19 (3)	18 (2)	26 (12)	45 (2)	50 (9)	24 (4)	20 (3)	31 (4)	38 (5)
88	24 (3)	13 (5)	8 (3)	16 (10)	23 (3)	24 (12)	16 (3)	26 (10)	**14 (95)**	17 (3)	28 (3)	11 (6)
90	17 (4)	15 (2)	11 (3)	12 (6)	17 (5)	27 (12)	25 (8)	22 (11)	16 (7)	15 (3)	22 (4)	**15 (79)**

Figure 5.1 illustrates some examples of different cases of gap distribution. There are distinct cases of missing data as shown in Figure 5.1. For instance, a) some months have several gaps and one or more of them have a big length, b) other months have few gaps but one or more of them are big gaps, c) other months also have numerous gaps but all of them are small, and d) the best case is where we have months with few gaps and these have a small length.

5.2 Results of the reconstruction model

Table 5.3 shows the performance of the analysis for different missing data percentages for the first experiment. The RMSE and the MAE decrease and the coefficient of determination increases, as the number of missing data decrease. We can see that the performance of the model is good for missing data less than 50% according to the R^2 in Table 5.3.

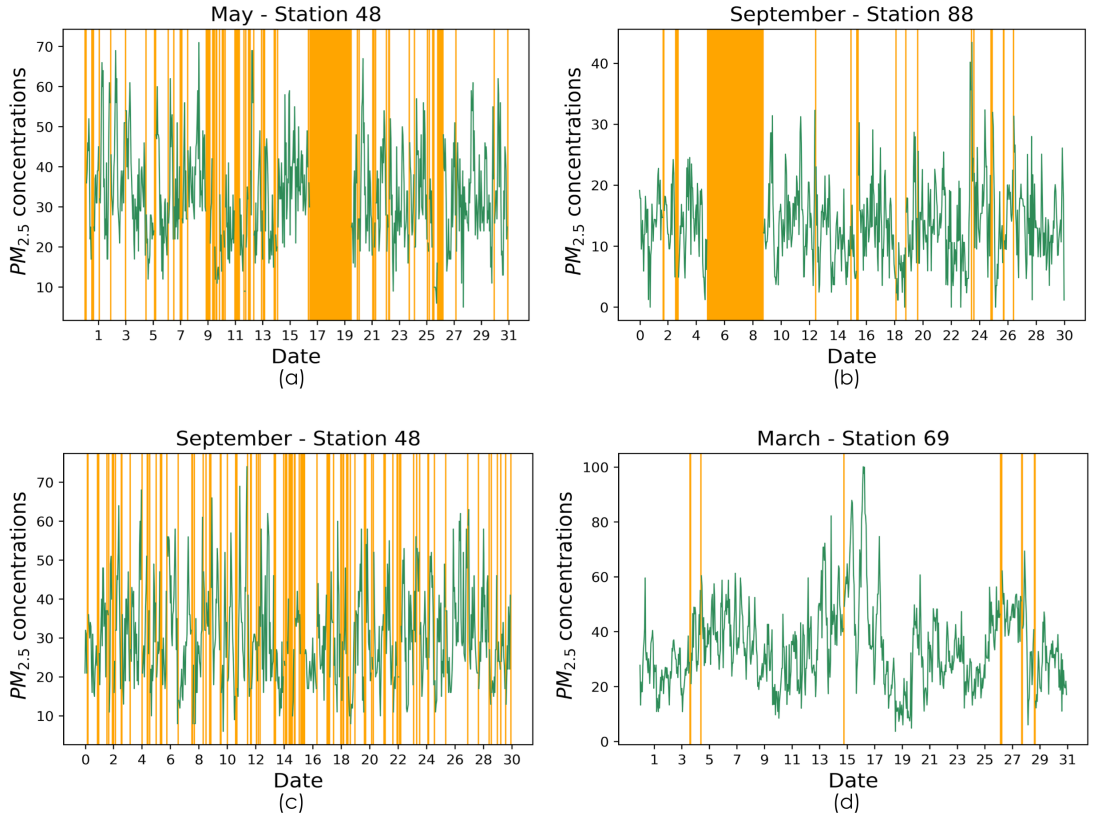

FIGURE 5.1: Some examples of the distribution of gaps in the dataset. (a) Many and big gaps, (b) few and big gaps, (c) many and small gaps, and (d) few and small gaps.

TABLE 5.3: Metrics for different missing data percentages $\{10\%, 20\%, ..., 90\%\}$. Root mean square error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2).

Indicators	Missing data percentage								
	90%	80%	70%	60%	50%	40%	30%	20%	10%
RMSE	10.17	7.66	6.79	6.32	5.98	5.76	5.59	5.46	5.36
MAE	5.45	4.21	3.68	3.34	3.08	2.87	2.69	2.54	2.40
R^2	0.62	0.77	0.82	0.85	0.87	0.89	0.90	0.91	0.92

Figure 5.2 shows the level of agreement between the original data and the reconstructed data in the first experiment. We can see that the scatter plots get skinnier as soon as the missing data percentage decreases. Figure 5.3 shows a part of the time series for the original and reconstructed data in the first 15 days of August in Station 69. Similar to the scatter plots in Figure 5.2, the level of agreement between the original data and the reconstructed data increases when the percentage of missing data decreases. However, the model cannot follow some pollution peaks since the model assumes smooth changes in time.

The reconstructed signal in Figure 5.3 is not exactly the same as the original, even when the data is not artificially corrupted, because the model minimizes the error, i.e., the

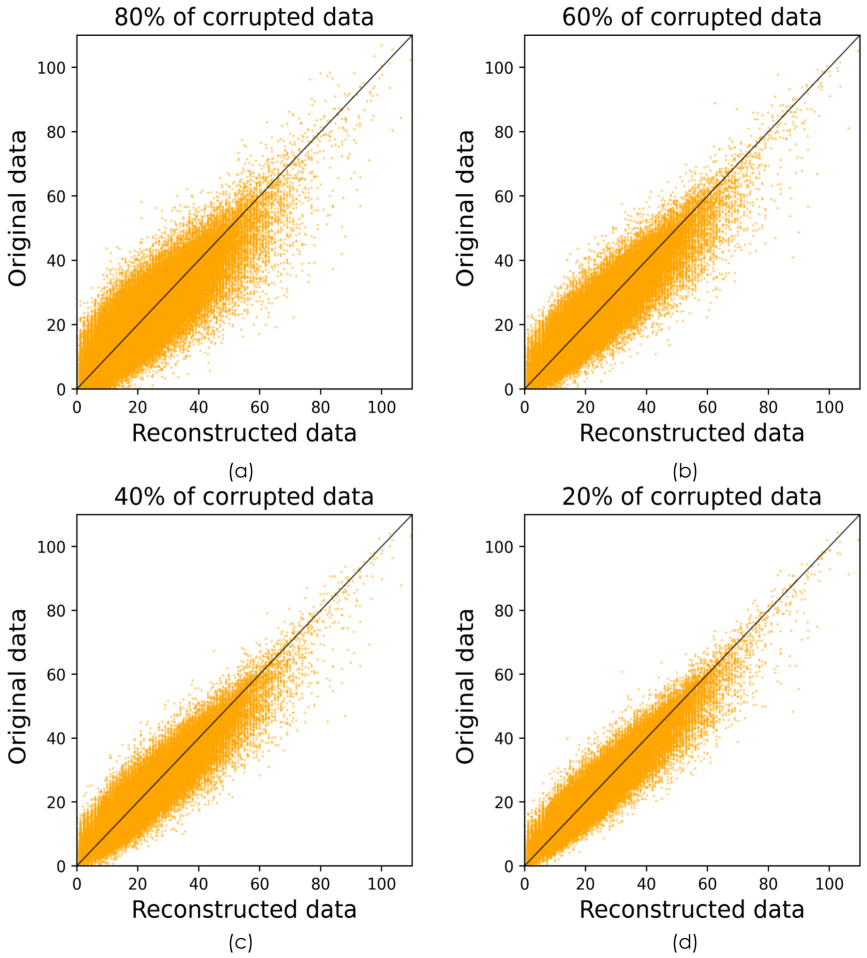

FIGURE 5.2: Scatter plots of the reconstruction results of value pair between original data and corresponding reconstructed values. Missing data percentage of (a) 80%, (b) 60%, (c) 40%, and (d) 20%.

model does not search for the exact same value to reduce noise. For example, if we have a peak of pollution and the model does not artificially corrupt that peak, the algorithm could consider that peak as noise. As a result, the model could smooth that peak.

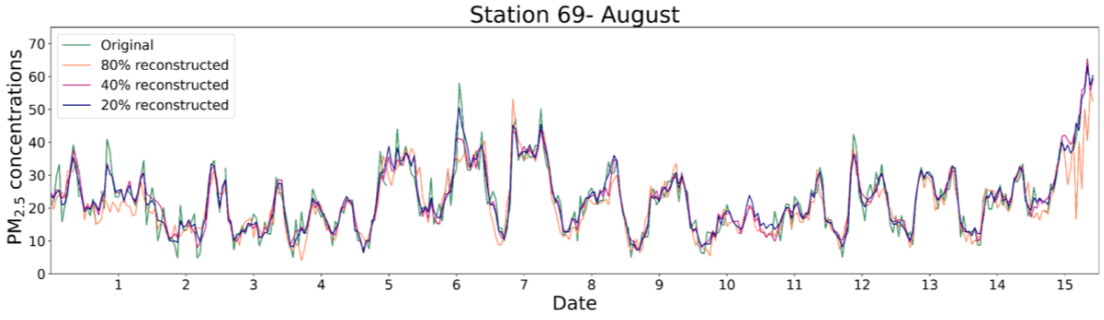

FIGURE 5.3: A part of the time series for the original (green) and reconstructed data for three percentages of missing data, 80% (coral), 40% (violet), and 20% (blue).

We compute the daily mean of the original and reconstructed data to evaluate the model

in another time scale, as shown in Figure 5.4. The model performs better in this case; even with a high percentage of corruption, the model fits quite well the original time series. In this case, the model can follow the peaks in pollution.

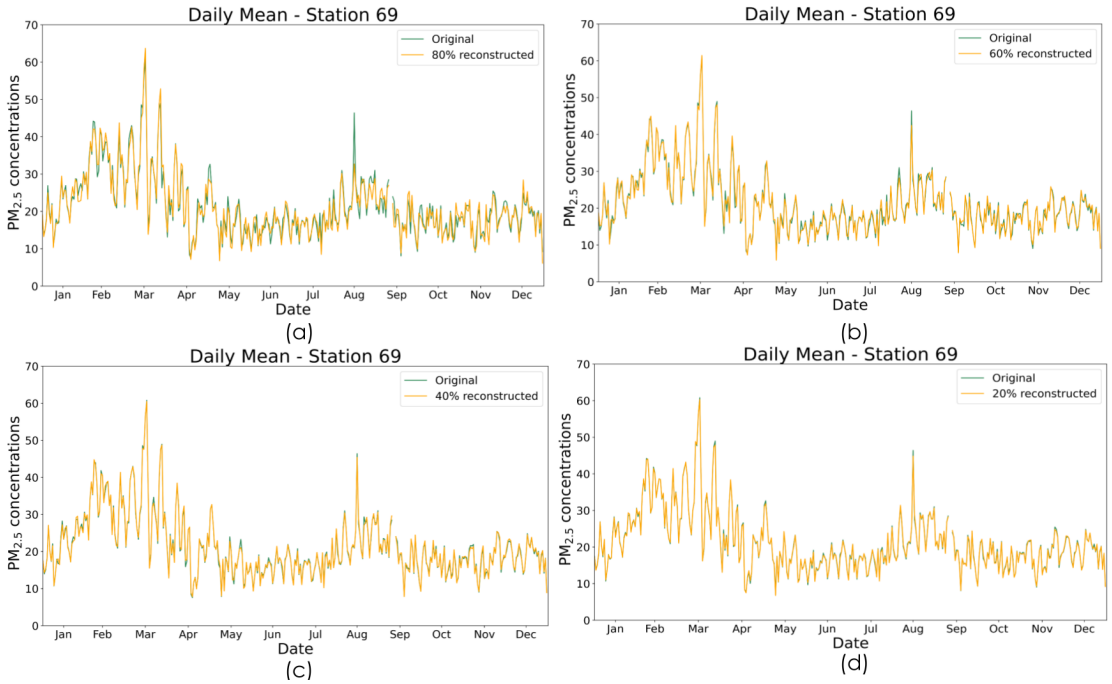

FIGURE 5.4: The daily mean of the original and reconstructed data in station 69. Missing data percentage of (a) 80%, (b) 60%, (c) 40%, and (d) 20%.

Figure 5.5 shows the reconstruction results for the second experiment. The reconstruction in some months has better results than others. During dry meteorological conditions like January, June, and July the model has a better performance. Nevertheless, the model does not perform as well as during March, April, May, and November (wet and transition seasons). The model has worse results during the transition seasons. Air pollution conditions in the Aburrá Valley are affected by weather conditions in all seasons. During dry months, the meteorological conditions favor the height of the ABL and the transport of pollutants. These conditions contribute to making better the air quality conditions. During wet seasons the meteorology affects the ABL height and transport of pollutants. During wet months air quality conditions are worse. Due to the graph reconstruction model not considering these meteorological variables its performance could be affected.

Table 5.4 shows the results of the third experiment. In this third experiment, we can see a similar result to the second one. The ability to reconstruct is better during the dry season than during the transition season. This is the same behavior observed in the first experiment, where February and March belong to the transition season. We can also notice that the increase in missing data percentage decreases the performance of

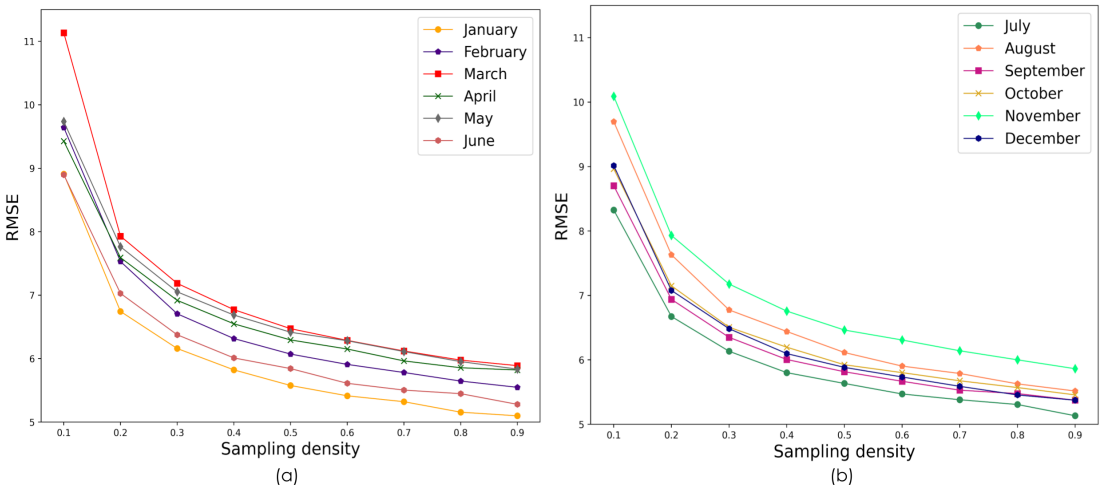

FIGURE 5.5: Reconstruction RMSE results for several months and several sampling densities with the graph-based algorithm. (a) First semester 2019, and (b) second semester 2019.

the reconstruction model. Similarly, the RMSE indicates that the general performance of the reconstruction method is fairly good when the missing information is less than 80%.

TABLE 5.4: Detailed monthly information of reconstruction. RMSE results of $PM_{2.5}$ in the Aburrá Valley for each station. During a dry season (7 - 23 January) and transition season (25 February - 16 March).

	Dry					Transition				
Station	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%
12	**14.64±0.00**	**9.64±0.27**	**7.50±0.23**	**6.93±0.55**	**6.69±0.82**	**12.84±0.00**	**9.67±0.67**	**8.77±0.59**	**7.83±0.56**	**6.47±0.26**
25	4.65±0.00	4.24±0.08	4.06±0.03	3.75±0.07	3.70±0.37	6.57±0.00	5.53±0.24	5.05±0.16	4.72±0.25	4.36±0.58
28	6.61±0.00	5.14±0.21	4.68±0.36	4.62±0.58	4.64±0.28	8.60±0.00	7.54±0.67	7.21±0.24	6.86±0.20	6.66±0.70
31	6.12±0.00	5.60±0.16	5.37±0.27	5.18±0.13	5.12±0.54	6.34±0.00	6.22±0.27	6.11±0.11	5.74±0.20	5.42±0.24
38	4.82±0.00	4.33±0.08	4.39±0.07	4.03±0.31	3.74±0.18	6.27±0.00	5.71±0.13	5.11±0.23	5.30±0.10	5.14±0.34
44	6.95±0.00	6.84±0.24	6.20±0.34	5.95±0.34	6.00±0.49	8.22±0.00	8.23±0.15	7.27±0.61	6.99±0.22	7.14±0.83
48	8.73±0.00	7.77±0.27	7.80±0.19	7.20±0.53	6.74±0.61	8.64±0.00	8.57±0.47	8.07±0.20	7.89±0.24	7.79±0.39
69	5.16±0.00	5.16±0.17	5.04±0.25	4.97±0.26	4.79±0.06	6.69±0.00	7.18±0.41	6.58±0.23	6.64±0.38	5.87±0.48
78	4.62±0.00	4.40±0.15	3.85±0.17	3.73±0.16	3.67±0.34	6.10±0.00	5.98±0.49	5.50±0.22	5.54±0.30	5.26±0.25
79	5.71±0.00	5.92±0.18	5.45±0.22	5.03±0.41	4.77±0.15	6.93±0.00	7.14±0.21	6.41±0.52	5.57±0.26	5.94±0.26
80	5.79±0.00	5.49±0.10	4.92±0.16	4.50±0.15	4.07±0.46	7.06±0.00	6.75±0.15	6.63±0.29	6.42±0.16	6.28±0.05
81	**14.99±0.00**	**5.34±0.09**	**4.97±0.08**	**4.62±0.31**	**4.40±0.47**	**31.68±0.00**	**7.26±0.18**	**6.75±0.17**	**6.58±0.28**	**6.04±0.51**
82	6.62±0.00	5.10±0.30	4.50±0.30	4.41±0.31	4.30±0.34	7.67±0.00	5.28±0.26	5.18±0.16	4.99±0.37	4.82±0.30
83	11.20±0.00	7.99±0.62	6.46±0.50	5.97±0.30	5.24±0.24	11.60±0.00	8.38±0.23	7.83±0.44	6.87±0.08	6.61±0.65
84	4.03±0.00	4.24±0.15	4.08±0.11	4.27±0.13	3.87±0.27	5.58±0.00	5.43±0.17	5.28±0.09	4.71±0.18	4.26±0.30
85	8.20±0.00	4.92±0.23	4.44±0.25	4.29±0.06	4.01±0.10	8.79±0.00	7.98±0.27	6.69±0.35	6.64±0.19	5.96±0.50
86	4.77±0.00	4.42±0.11	4.08±0.26	3.96±0.31	3.67±0.21	7.43±0.00	6.12±0.28	5.12±0.08	4.85±0.10	4.80±0.05
87	5.08±0.00	4.64±0.17	4.32±0.18	4.19±0.08	4.06±0.34	6.76±0.00	5.26±0.10	4.85±0.21	4.94±0.16	4.51±0.67
88	4.27±0.00	4.25±0.07	4.26±0.03	4.01±0.24	3.88±0.14	6.15±0.00	5.92±0.00	5.43±0.13	5.38±0.24	5.14±0.36
90	5.40±0.00	4.60±0.06	4.38±0.13	4.21±0.07	3.83±0.22	7.61±0.00	6.27±0.11	5.57±0.31	5.15±0.44	5.13±0.30
Mean	6.92±3.15	5.50±1.46	5.04±1.13	4.79±1.01	4.56±1.00	8.88±5.57	6.82±1.27	6.27±1.15	5.98±1.02	5.68±1.01

Table 5.4 shows limitations in some stations, for example, the RMSE in stations 12 and 81 are clearly high in both seasons. The poor results in station 12 during the dry season could be partly explained by the big gaps of missing data in January (see Table 5.2). Similarly, for 100% of missing data the model heavily relies on spatial information, and therefore the reconstructed data in this case for station 12 is similar to its closest stations (25, 80, and 86)(see Figure 5.6). On the other hand, unlike the other cases, the

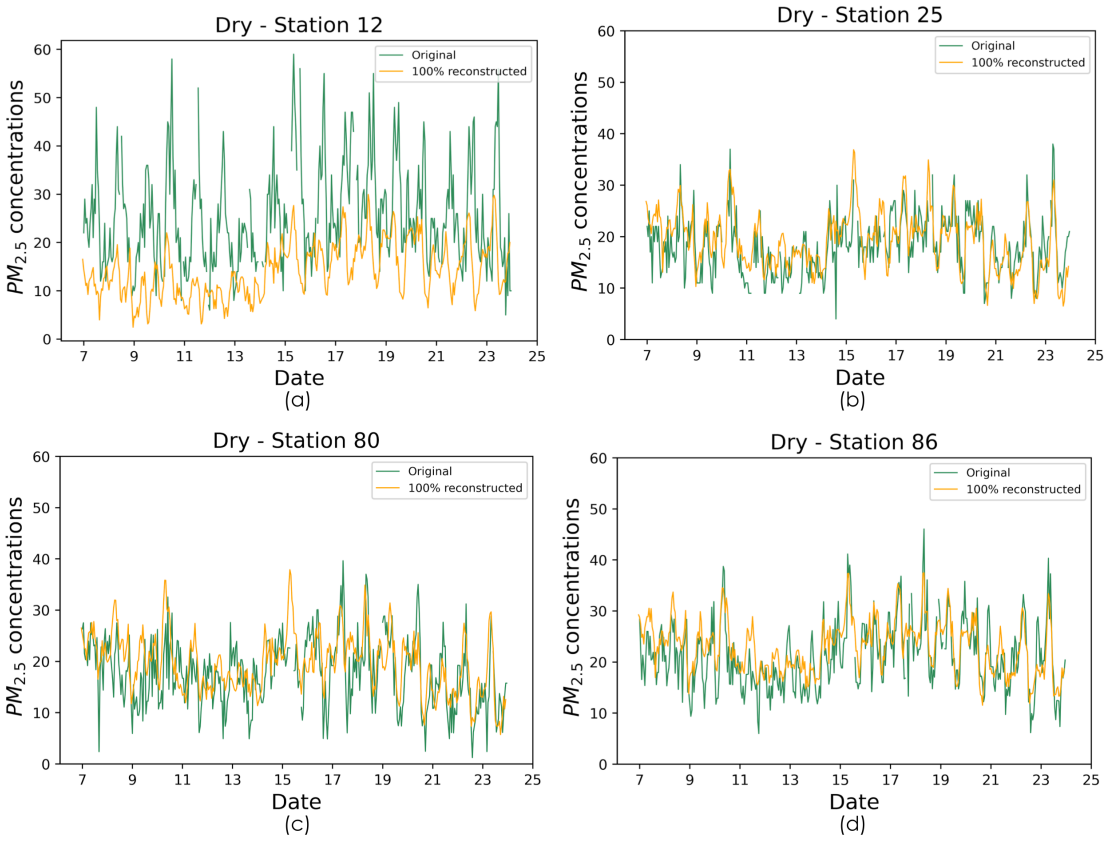

FIGURE 5.6: Spatial imputation results of various stations, each figure represent the real data (in green) and the reconstructed data (in yellow) when the missing data percentage was 100%.

station 12 shows better results in wet than in dry season, that is because we have less gaps in February-March than in January.

As shown in Table 5.4, the RMSE is high in station 81 for 100% of missing data for both seasons. Figure 3.2 shows that station 81 is geographically far away from the rest of stations. As a result, the reconstruction of 100% missing data relies entirely on the closest stations, and the spatial smoothness does not hold quite well in this case. However, the results improve as soon as we have less than 100% of missing data and we get some temporal information.

Figure 5.7 shows several Taylor diagrams of the different experiments. Figure 5.7 (a) and (b) shows the results in station 44 for dry and transition season, respectively. Figure 5.7 (c) and (d) show the Taylor diagrams in station 78. Every point represents the reconstructed time series for different percentages of missing data, and the black star represents the original time series. We observe that most of the points are under the black dotted line, this means that the reconstructed data set has less variation than the original data. The best results in terms of variance are at station 78 for the transition season. However, in this case the centered root mean square error is between 4 and 8

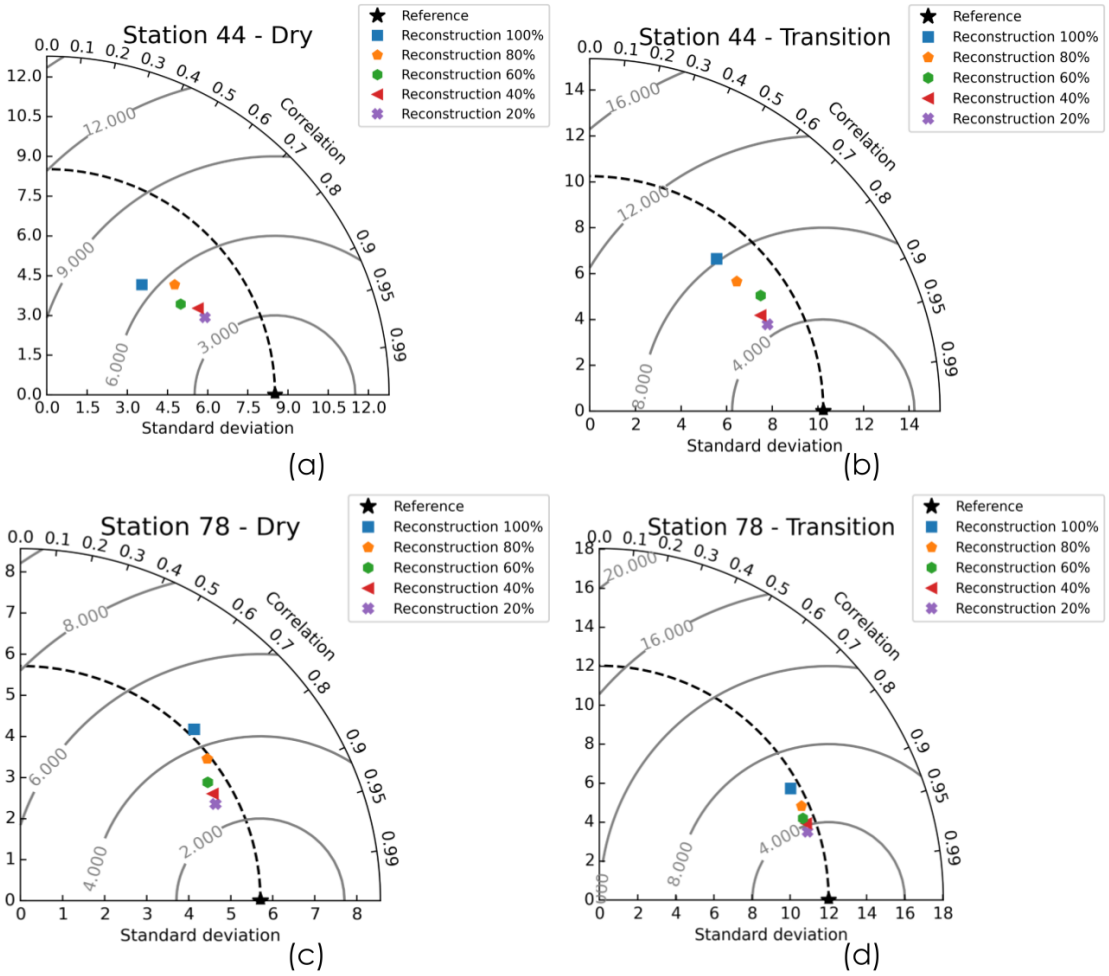

FIGURE 5.7: Taylor diagrams of $PM_{2.5}$ data for two stations. (a) and (b) station 44 for dry and transition season, respectively. (c) and (d) station 78 for dry and transition season, respectively.

for all the percentages of missing values, these results are higher than in the case of dry season in the same station. We can also see that the correlation increases as the percentage of missing data decrease. In cases of percentage of missing values lower than 40% the correlation is between 0.9 and 1. Therefore, there is a high level of agreement between original and simulated data. This result coincides with the analyzes of the other experiments.

Section 6

Conclusions

In this work, we tested a graph model for the reconstruction of missing data in a dataset of air quality in the Aburrá Valley. The methodology is composed of a preprocessing stage, exploratory data analysis, execution of the model, and analysis of the results. We conclude that the performance of the model improves as soon as the percentage of missing data decrease. However, the model cannot follow the pollution peaks since the model assumes smooth changes in time. The model also minimizes the error and reduces the noise; accordingly, the reconstructed dataset is not precisely the same as the original data, even in the values not artificially corrupted.

We also concluded that the model works well in general settings, but its performance decreased when there are extreme meteorological events. The model has a better performance during dry seasons than during wet and transition seasons. During the dry months the meteorological conditions favor the dispersion of pollutants whereas during wet months the meteorological conditions contribute to worse air quality conditions. These weather events were not included in the model, therefore the model performance is affected.

To future work it could be interesting to analyze the performance of the model in the daily cycle of pollution, to see if the daily cycle of precipitation affects the results of the reconstruction method. In addition, an interesting future direction of this work is to include meteorological information into the model.

Bibliography

[1] Aguiar-Gil, D., Gómez-Peláez, L. M., Álvarez-Jaramillo, T., Correa-Ochoa, M. A., and Saldarriaga-Molina, J. C. (2020). Evaluating the impact of $PM_{2.5}$ atmospheric pollution on population mortality in an urbanized valley in the American tropics. *Atmospheric Environment*, 224:117343.

[2] Alahamade, W., Lake, I., Reeves, C. E., and De La Iglesia, B. (2020). Clustering imputation for air pollution data. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 585–597. Springer.

[3] Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., Aziz, N. A. A., Azaman, F., Latif, M. T., and Zainuddin, S. F. M. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, & Soil Pollution*, 225(8):1–14.

[4] Bedoya, J. and Martínez, E. (2009). Calidad del aire en el Valle de Aburrá Antioquia-Colombia. *Dyna*, 76(158):7–15.

[5] Bedoya-Soto, J. M., Aristizábal, E., Carmona, A. M., and Poveda, G. (2019). Seasonal shift of the diurnal cycle of rainfall over Medellin's Valley, central Andes of Colombia (1998-2005). *Frontiers in Earth Science*, 7:92.

[6] Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., and Perrin, C. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, 40:1–20.

[7] Betts, A. (1973). Non-precipitating cumulus convection and its parameterization. *Quarterly Journal of the Royal Meteorological Society*, 99(419):178–196.

[8] Cabrera, B. et al. (2016). A geostatistical method for the analysis and prediction of air quality time series: application to the Aburrá Valley region.

[9] Chen, S., Sandryhaila, A., Moura, J. M., and Kovacevic, J. (2014). Signal denoising on graphs via graph filtering. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 872–876. IEEE.

[10] Cuervo López, C. M. (2017). Caracterización del comportamiento del vapor de agua y energía potencial convectiva disponible precedente a eventos de precipitación sobre el Valle de Aburrá. *Escuela de Geociencias y Medio Ambiente*.

[11] Dong, X., Thanou, D., Frossard, P., and Vandergheynst, P. (2015). Laplacian matrix learning for smooth graph signal representation. In *2015 IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3736–3740. IEEE.

[12] Giraldo, J. H. and Bouwmans, T. (2020). On the minimization of Sobolev norms of time-varying graph signals: Estimation of new Coronavirus disease 2019 cases. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

[13] Hadeed, S. J., O'Rourke, M. K., Burgess, J. L., Harris, R. B., and Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of The Total Environment*, 730:139140.

[14] Henao, J. J., Rendón, A. M., and Salazar, J. F. (2020). Trade-off between urban heat island mitigation and air quality in urban valleys. *Urban Climate*, 31:100542.

[15] Herrera Mejía, L. (2015). Caracterización de la capa límite atmosférica en el Valle de Aburrá a partir de la información de sensores remotos y radiosondeos. *Escuela de Geociencias y Medio Ambiente*.

[16] Herrera-Mejía, L. and Hoyos, C. D. (2019). Characterization of the atmospheric boundary layer in a narrow tropical valley using remote-sensing and radiosonde observations and the WRF model: the Aburrá Valley case-study. *Quarterly Journal of the Royal Meteorological Society*, 145(723):2641–2665.

[17] Isaza Uribe, A. (2018). Evaluación de la variabilidad temporal de la estructura termodinámica de la atmósfera y su influencia en las concentraciones de material particulado dentro del Valle de Aburrá. *Escuela de Geociencias y Medio Ambiente*.

[18] Izonin, I., Tkachenko, R., Logoyda, M., Mishchuk, O., Kynash, Y., et al. (2019). Sgd-based wiener polynomial approximation for missing data recovery in air pollution monitoring dataset. In *International Work-Conference on Artificial Neural Networks*, pages 781–793. Springer.

[19] Jiménez Mejía, J. F. (2016). Altura de la capa de mezcla en un área urbana, montañosa y tropical. caso de estudio: Valle de Aburrá (Colombia).

[20] Junger, W. and De Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, 102:96–104.

[21] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18):2895–2907.

[22] Kalteh, A. M. and Berndtsson, R. (2007). Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). *Hydrological sciences journal*, 52(2):305–317.

[23] Liu, X., Wang, X., Zou, L., Xia, J., and Pang, W. (2020). Spatial imputation for air pollutants data sets via low rank matrix completion algorithm. *Environment international*, 139:105713.

[24] Ma, J., Cheng, J. C., Ding, Y., Lin, C., Jiang, F., Wang, M., and Zhai, C. (2020). Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series. *Advanced Engineering Informatics*, 44:101092.

[25] Ma, S., Goldfarb, D., and Chen, L. (2011). Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1):321–353.

[26] Mei, J. and Moura, J. M. (2015). Signal processing on graphs: Estimating the structure of a graph. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5495–5499. IEEE.

[27] Mejía-Echeverry, D., Chaparro, M. A., Duque-Trujillo, J. F., Chaparro, M. A., and Castañeda Miranda, A. G. (2018). Magnetic biomonitoring as a tool for assessment of air pollution patterns in a tropical valley using Tillandsia sp. *Atmosphere*, 9(7):283.

[28] Montoya, E. (2018). Caracterización de la concentración de contaminantes del aire a partir del estudio de la dinámica atmosférica en el Valle de Aburrá. *Escuela de Geociencias y Medio Ambiente*.

[29] Murillo-Escobar, J., Sepulveda-Suescun, J. P., Correa, M. A., and Orrego-Metaute, D. (20219). Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia. *Urban Climate*, 29:100473.

[30] Narang, S. K., Gadde, A., Sanou, E., and Ortega, A. (2013). Localized iterative methods for interpolation in graph structured data. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 491–494. IEEE.

[31] Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., and Ramli, N. A. (2015). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. In *Materials Science Forum*, volume 803, pages 278–281. Trans Tech Publ.

[32] Norazian, M. N., Shukri, Y. A., Azam, R. N., et al. (2008). Estimation of missing values in air pollution data using single imputation techniques.

[33] Organization, W. H. (2020). *WHO global strategy on health, environment and climate change: the transformation needed to improve lives and wellbeing sustainably through healthy environments.* World Health Organization.

[34] Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., and Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828.

[35] Peláez, L. M. G., Santos, J. M., de Almeida Albuquerque, T. T., Reis Jr, N. C., Andreão, W. L., and de Fátima Andrade, M. (2020). Air quality status and trends over large cities in South America. *Environmental Science & Policy*, 114:422–435.

[36] Perraudin, N., Loukas, A., Grassi, F., and Vandergheynst, P. (2017). Towards stationary time-vertex signal processing. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3914–3918. Ieee.

[37] Qiu, K., Mao, X., Shen, X., Wang, X., Li, T., and Gu, Y. (2017). Time-varying graph signal reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):870–883.

[38] Quinteros, M. E., Lu, S., Blazquez, C., Cárdenas-R, J. P., Ossa, X., Delgado-Saborit, J.-M., Harrison, R. M., and Ruiz-Rudolph, P. (2019). Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. *Atmospheric environment*, 200:40–49.

[39] Rendón, A. M., Salazar, J. F., Palacio, C. A., Wirth, V., and 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(4):840–858.

[40] Shi, X., Feng, H., Zhai, M., Yang, T., and Hu, B. (2015). Infinite impulse response graph filters in wireless sensor networks. *IEEE Signal Processing Letters*, 22(8):1113–1117.

[41] Sibson, R. (1981). A brief description of natural neighbour interpolation. *Interpreting multivariate data*.

[42] Yu, Y., Li, V. O., and Lam, J. C. (2020). Missing air pollution data recovery based on long-short term context encoder. *IEEE Transactions on Big Data*.

[43] Zardi, D. and Whiteman, C. (2013). Diurnal mountain wind systems. mountain weather research and forecasting, FK chow, SFJ DeWekker, and B. Snyder, Eds.

[44] Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., and Li, T. (2015). Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2267–2276.

[45] Zhu, X. and Rabbat, M. (2012). Graph spectral compressed sensing for sensor networks. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2865–2868. IEEE.