Dissertação de Mestrado

Multivariate Modeling to handle Urban Air Pollution Data observed trough Vehicular Sensor Networks

de **Israel Loureiro Cavalcante Vasconcelos**

orientado por
Prof. Dr. André Luiz Lins de Aquino

Universidade Federal de Alagoas
Instituto de Computação
Maceió, Alagoas
29 de Janeiro de 2021

UNIVERSIDADE FEDERAL DE ALAGOAS
Instituto de Computação

MULTIVARIATE MODELING TO HANDLE URBAN AIR POLLUTION DATA OBSERVED TROUGH VEHICULAR SENSOR NETWORKS

Dissertação de Mestrado submetida ao Instituto de Computação da Universidade Federal de Alagoas como requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional de Conhecimento.

Israel Loureiro Cavalcante Vasconcelos

Advisor: Prof. Dr. André Luiz Lins de Aquino

Banca avaliadora:

Raquel da Silva Cabral	Prof$^{\underline{a}}$. Dr$^{\underline{a}}$., UFAL
Édler Lins de Albuquerque	Prof. Dr., IFBA

Maceió, Alagoas
29 de Janeiro de 2021

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 – 1767

V331m Vasconcelos, Israel Loureiro Cavalcante.
Multivariate modeling to handle urban air pollution data observed trough vehicular sensor networks / Israel Loureiro Cavalcante Vasconcelos. – 2021.
33 f. : il.

Orientador: André Luiz Lins de Aquino.
Dissertação (mestrado em Modelagem Computacional do Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2021.

Texto em inglês.
Bibliografia: f. 32-33.

1. Modelagem ambiental. 2. Rede de sensores - Veículos. 3. Análise multivariada. I. Título.

CDU: 004.414.23:504.064

Na extensa lista de nomes a quem devo gratidão, em primeiro lugar a Deus pois minha maior certeza está nessa força para os empreendimentos cada vez mais desafiadores sempre se originar da Sua infinita bondade, sozinho seria impossível. Imenso agredecimento ao suporte da família, avós Macário (*in memoriam*), Elisete, Antônio, Lígia. Pais, David e Sônia. Irmão, Gabriel. À minha prima Tati por me dar remédio.

Ao professor e amigo André Aquino por quem dirijo total respeito e lealdade, tamo junto é nois* obrigado por tudo sempre.

Agradeço à Lica por ser a melhor pessoa do mundo, por sempre ir na minha onda e me botar pra cima. Aos meus amados amigos Joãozinho (fala buzinho), Ricardo (ser-de-luz), Yasmin (chinês), Keyse (delegado), Ana Luísa (viva la revolución), Thacy, Teteu, Cristopher, Buxexa, Paulo, Pedro (morsa), Natália & Raíssa, Edu Bibiano, Rapha (emo), Yuri, Caio (rato branco), Serginho.

Pros minino* da firma: Fabrício, Rubens, Ivalbert, Roberto, Álvaro, Alexandre, Lucas, Henrique, Estácio, Rafael^2, Marx, Ueslei, Josafa, Caio, Gabriela, Ubiratan, Daniel, Adriana, Fábio, Wagner.

O tamanho dessa conquista é proporcional à quantidade de nomes colocados aqui, minha imensa gratidão a todos!

Do not be anxious about anything, but in everything, by prayer and petition, with thanksgiving, present your request to God. And the peace of God, which transcends all understanding, will guard your hearts and your minds in Christ Jesus. (Phillipians, 4:6-7)

Abstract

This work presents an interdisciplinary assessment that looks in-depth at the tracking of air quality in urban environments. This kind of application is well suited to be approached with wireless sensor networks' paradigm in their overall variations. Therefore a robust and diverse set of solutions have been developed following the technology capabilities advance. The proposed modeling application takes advantage of Vehicle Sensor Networks (VSN) by embedding sensor nodes to public transportation, addressing this study case with bus lines so that the mobiles spread the sampling activity through a large number of different places visited during the route. Simultaneously, it alleviates power management restrictions, packaging dimensions (size and weight), and general maintenance issues. We perform environmental modeling based on real data considering a temporal and spatial multivariate behavior on observed phenomena. We consider the city of São Paulo in our case study and parse the asserted data to create a multivariate map of samples, showing the behavior of five different air pollutants (CO, O_3, PM_{10}, NO_2 and SO_2) simultaneously while it also varies in time. The current development stage covers handling processes over input data that has unformatted or missing information due to being sourced from real sensors and creating the map mentioned above. Our methodology addresses: 1) the mentioned environmental simulation, 2) deploying mobile sensor nodes and perform sensing process, 3) implement network activity and delivery of collected data, 4) visualization of monitored environment based on gathered data using Voronoi Diagrams to fill blank data at non reached areas. Finally, some outspread of our work, in the research area, are the evaluation system-level performance and operational constraints through an event-based simulation, taking into account a detailed description of roads, bus lines, vehicle itineraries, and general traffic information.

Keywords: Environmental modeling, Vehicle Sensor Networks, Multivariate data analysis

Resumo

Este trabalho apresenta uma avaliação interdisciplinar analisando o monitoramento da qualidade do ar em ambientes urbanos. Este tipo de aplicação se enquadra adequadamente sob o paradigma das redes de sensores sem convencionais. No referido contexto, um conjunto robusto e diversificado de soluções vem sendo desenvolvido a medida que são aprimorados os recursos tecnológicos. A aplicação de modelagem proposta aproveita-se das Redes de Sensores Veiculares (VSN) ao incorporar nós sensores ao transporte público, utilizando no presente caso de estudo as linhas de ônibus disponíveis de modo que os veículos dispersem a atividade de amostragem pelos diferentes locais visitados durante seu percurso. Simultaneamente, as restrições de gerenciamento de energia, dimensões da embalagem (tamanho e peso) e problemas gerais de manutenção também são aliviadas. É realizada uma modelagem ambiental com base em dados reais considerando o comportamento multivariado temporal e espacial dos fenômenos observados. Consideramos a cidade de São Paulo em nosso estudo de caso e analisamos os dados meteorológicos para criar um mapa multivariado a partir das amostras, expondo o comportamento de cinco diferentes poluentes atmosféricos (CO, O_3, PM_{10} , NO_2 e SO_2) simultaneamente variando em função do tempo. O arcabouço de modelagem proposto contempla os processos de tratamento dos dados de entrada, que apresentam informações não formatadas ou ausentes devido a serem originados por sensores reais, além da criação do mapa mencionado acima. Nossa metodologia aborda: 1) a simulação ambiental e urbana anteriormente mencionada, 2) distribuição dos nós sensores móveis e realização do processo de detecção, 3) implementação das atividades de rede e entrega de dados coletados, 4) visualização do ambiente monitorado com base em dados coletados, usando Diagramas de Voronoi para preencher dados em branco em áreas não atingidas. Por fim, alguns desdobramentos de nosso trabalho, na área de pesquisa, são a avaliação de desempenho em nível de sistema e suas restrições operacionais por meio de uma simulação baseada em eventos, levando em consideração uma descrição detalhada de estradas, linhas de ônibus, itinerários de veículos e informações gerais de tráfego.

Palavras chave: Modelagem ambiental, redes de sensores veiculares, análise de dados multivariada.

List of Figures

List of Tables

Aggreg.	Aggregation
AQI	Air Quality Index
CO	Carbon Monoxide
CO_2	Carbon Dioxide
Compr.	Compression
CSN	Community Sensor Network
DDR	Dynamic Reporting Rate
EDGE	Efficient Data Gathering and Estimation
GPS	Global Position System
GSM	Global System for Mobile Communications
ID	Identifier
MDH	Multivariate Data Handling
MSB	Mobile Sensing Box
NA	Not Available
NO_2	Nitrogen Dioxide
NO_x	Nitrogen Oxide
O_3	Ozone
PM	Particulate Matter
PSD	Personal Sensing Device
Rep.	Reporting
SO_2	Sulfur Dioxide
SO_x	Sulfur Oxide
SQL	Structured Query Language
TOR	Time-constrained Opportunistic Relay
VOC	Volatile Organic Compounds
VSN	Vehicle Sensor Network
WSN	Wireless Sensor Network

Contents

Chapter 1

Introduction

1.1 Overview

The world around us has various phenomena monitored by devices provided with sensing, processing, and communication capabilities. While cooperatively working in an area of interest, such devices comprise a wireless sensor network (WSN) [1, 30].

This study evaluates a solution that considers a set of different physical phenomena observed by wireless sensor networks. In this context, the challenge of monitoring urban areas regarding subjects as air quality and meteorological conditions rises as notoriously relevant research opportunities [18, 29].

We refer to the resulting output from various phenomena sensing process as multivariate data. The samples of each monitored variable are simultaneously collected and stored by different sensors in the same node. To reach a better coverage area, we proposed to use embedding sensor nodes to public transportation [11], such as bus lines and trains. We spreading the sampling through a large number of different places visited during the route at the same time simplifies restrictions of power management, packaging dimensions (size and weight), and general maintenance issues.

Despite the advantages previously mentioned, these conditions aggravate redundancy issues due to the dynamics of urban traffic [29] (i.e., store repeatedly samples from the same place when the vehicle is in a traffic jam or high spatial similarity data at closer neighborhoods). An important aspect to highlight after taking a more in-depth look at available solutions is to realize the lack of an approach that **handles multivariate sampling** at distributed, noisy, and adverse behaved conditions, as typically seen under realistic urban environments. Simultaneously, data sampling techniques are well suited to solve redundancy problems such as discussed in VSN known constraints. Some of the mentioned techniques comprise variable reporting rate [6, 10, 28], node clustering [12, 14], data fusion [6, 10, 12, 14], and reconstruction of lost data [28].

Thus, the aspects mentioned above drive us to state the following research question:

"What is the impact of using a VSN-based solution to monitor the air quality that observes multiple phenomena simultaneously?". The solution must consider a multivariate data set as input and raise additional complexities compared to univariate ones.

We used a Spatio-temporal real dataset of available multivariate samples collected by ten stationary air quality stations in the experimental validation methodology. These samples are air pollution variables with some correlation with each other. With these data, we perform a multivariate interpolation to obtain a visualization covering the entire range of simulated environment at each unit of area in the field. An event-based simulation will put vehicle traffic over this previously generated field to evaluate the network behavior, restrictions, and parameters. The simulation strategy will make car-mounted sensors read the table with stated field data.

Real data requires pre-processing to fix NA samples at the temporal axis at the modeling stage. On the other hand, looking at the spatial point of view, the lack of entire series for some variables at station coordinates (irregular data availability) requires a second additional pre-processing step to predict these missing points and perform the multivariate interpolation. This step involves a sequence of manual procedures and consumes significantly more implementation time. The methods described in section 3.1.1 discuss the adopted strategy to handle this data and prepare it for reconstruction.

The statement of **expected contributions** achieved at this research work **goes through a generalization purpose** addressing evaluation methods and experimentation scenarios featured closely at [9], [10] and [28]. Moreover, we intended 1) to provide a simulation framework that covers realistic use cases alongside a precise environmental model, 2) to raise a relevant subset of experimentation conditions and formulate guidelines to execution on real scenarios, 3) bring up side by side in comparison, considering the experiment results, the behavior of a classical strategy by static monitoring (air quality stations) alongside the presented VSN approach looking on its intrinsic operation principle.

From that, we can assure a bottom line to develop a simulation environment that **looks on urban pollution agents under a multivariate** point of view. We stated the starting **goals**, but not limited, summarized in the following bullets.

Overall: Implement a VSN-based solution that supports multivariate data processing.

Specifically: Evaluate the impact of multivariate observations on overall network behavior considering the following aspects

Hypothesis : Considering the global coverage ratio observed between the proposed VSN application and conventional stations, the equivalent error rate between the both strategies will decrease in a significantly higher rate in relation than coverage measurement on the most of cases.

The main achievement at this study is when we observe that all referred researches, including the previously mentioned articles in the current section, work with univariate data (CO_2 at [10], and Air Quality Index at [28]) whereas **we propose to expand the evaluation to a multivariate domain** while i) observe the behavior of each monitored variable individually, ii) the correlation between them, iii) consider as real input data collected from physical air quality stations, iv) assess the effort for handle a VSN application with this complex type of data and v) evaluate under metrics of absolute value of relative error and global field coverage.

1.2 Urban Air Pollution Review

1.2.1 Health impacts and human effects

The addressed research on this work touches an interdisciplinary subject since the case study's object relates to pollutant gases' harmful effects. To highlight the importance of awareness regarding the human body's adverse impact, we will briefly explain the most recurrently targeted pollutants in related work. After that, we will present the concentration thresholds for safe breathing [25] and the potential health damage in case of overexposure.

Critical organs such as the heart and brain receive a reduced amount of oxygen transported in the bloodstream, whereas breathing a high CO concentration in the air. Breathing air with a very high CO level can lead to confusion, dizziness, loss of consciousness, and death. It is more likely to occur in an enclosed environment, even though it could also happen outdoors.

When the CO levels elevate in an outdoor environment, they can be of critical concern for people with a specific heart condition, which reduces the blood oxygen transported to their hearts in situations, [19].

A group of highly reactive gases known as nitrogen oxide (NO_x) is composed of nitric acid, nitrous acid, and Nitrogen Dioxide (NO_2), commonly formed from burning fuel and used as the indicator for other nitrogen oxides. NO_2 at an elevated concentration can irritate the human respiratory system and aggravate respiratory diseases and symptoms such as coughing or difficulty breathing. Some symptoms may result in hospital admission. Long exposures to high concentrations of Nitrogen Dioxide can also cause serious effects such as the increased susceptibility to respiratory infections. Conditions as asthma and age-related ones present a greater risk if submitted to high concentrations of NO_2, which forms ozone and particulate matter in case of reacting with other chemicals in the air and can occur to other NO_x. Both reaction products may also cause harmful effects on the respiratory system, [20].

Chemical reactions between oxides of nitrogen (NO_x) and volatile organic compounds

(VOC) generate in tropospheric or ground-level ozone when pollutants emitted by vehicles and other sources chemically react submitted to sunlight. Hot sunny days in urban environments are most likely to reach unhealthy Ozone levels, even though it can also occur during colder weather. This gas can also be transported by wind, therefore reaching rural areas. Breathing ozone can trigger several health problems, including coughing and chest pain, which may harm lung tissue. Other effects include emphysema and asthma, leading to increased medical care, [23].

SO_2 is the component of the most significant concern, and we use it as the indicator for the larger group of gaseous sulfur oxides (SO_x). Other gaseous SO_x (such as SO_3) are found in the atmosphere at concentrations much lower than SO_2. Short-term exposure to SO_2 can harm the human respiratory system and make breathing difficult. People with asthma, particularly children, are sensitive to these effects of SO_2, [26].

Particulate matter (also called particle pollution, or PM) is the term for a mixture of liquid droplets in the air and solid particles. PM contains microscopic solids or liquid droplets that can be detected using an electron microscope. If being inhaled from a source such as construction sites and unpaved roads, they cause serious health problems. Particles with less than 2.5 micrometers in diameter can get deep into your lungs, and some may even get into the human bloodstream, [24].

We commonly found the most elevated air concentrations of lead around lead-smelters. Once inhaled, lead spreads through the blood and accumulates in the bones. Depending on the exposure level, it can adversely affect various systems, such as the cardiovascular system, immune system, and reproductive and developmental systems. Lead exposure is also very likely to affect the blood capacity of oxygen-carrying. The effects most iterant in modern populations are neurological in children and cardiovascular in adults. Infants and young children are susceptible to lead even in low concentrations, contributing to future behavioral problems, learning deficits, and lowered IQ. [22].

There are two categories into which we can divide mobile sources of air pollution: On-road vehicles, such as motorcycles and cars; and non-road ones and engines, such as aircraft, heavy equipment, marine vessels, and others: SMOG (Ground-level ozone), particle pollution, roadway air pollution zone, polluted air [21].

1.2.2 Air Quality Index

The Air Quality Index is a standard from U.S. Environmental Protection Agency, related literature widely adopt this standard as an evaluation metric [2, 6, 11, 27, 28, 29] and consists of a six-level scale containing reference values to pollutants concentration and risk descriptor for each level. A color scheme is also considered to ease the understanding, explained as follows:

Good – AQI $\leq$ 50 – Green: Outdoor air **is safe** to breathe.

Moderate – 51 $\leq$ AQI $\leq$ 100 – Yellow : Susceptible individuals should consider limiting prolonged or heavy outdoor exertion.

Unhealthy for sensitive groups – 101 $\leq$ AQI $\leq$ 150 – Orange : People with heart or lung disease (such as asthma), children, older adults, people who are active outdoors (including outdoor workers), people with specific genetic variants, and people with diets limited in certain nutrients **should reduce** prolonged or heavy outdoor exertion.

Unhealthy – 151 $\leq$ AQI $\leq$ 200 – Red: Sensitive people should **avoid prolonged or heavy exertion**; everyone else should reduce them.

Very unhealthy – 201 $\leq$ AQI $\leq$ 300 – Purple: Sensitive people should **avoid all outdoor exertion**; everyone else has to reduce outdoor exertion.

Hazardous – AQI $\geq$ 301 – Maroon: Everyone should avoid all outdoor exertion.

The pollutants considered at AQI evaluation are carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3) and particulate matter (PM 2.5 and PM 10). For each pollutant, the index I_p is given by Equation 1.1.The AQI is the maximum value of I_p among all pollutants in a single packet of samples for a specific location.

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}} \times (C_p - BP_{Lo}) + I_{Lo} \tag{1.1}$$

Where,

I_p = the index for pollutant p;
C_p = the truncated concentration of pollutant p;
BP_{Hi} = the concentration breakpoint that is greater than or equal to C_p;
BP_{Lo} = the concentration breakpoint that is less than or equal to C_p;
I_{Hi} = the AQI value corresponding to BP_{Hi};
I_{Lo} = the AQI value corresponding to BP_{Lo};

1.3 Related Work

WSN based air quality monitoring solutions are an already mature subject-matter in literature. In the following discussion, we give a solid bottom line to advance with proposed open points regarding the handling of multivariate data assessed at this research work.

Volgyesi et al. (2008) [27] present a simple vehicle network that consists of collected data and views in a web-based application. It describes the hardware architecture, memory employed, sensor specifications, and data traffic interfaces. The experiment consists of car-mounted sensors running in a real environment and evaluating AQI through monitoring of O_3, CO, and NO_2. These nodes have GPS and Bluetooth technologies. In this study case, the node sends the sampled data via Bluetooth to an intermediary gateway (on that instance, a standalone notebook) with an internet connection, responsible for forwarding data to the cloud server that handles it. The solution provides a map visualization to the final user/observer.

At Ma et. al (2011) [14], the authors propose a sophisticated air pollution monitoring system that considers a wide range of aspects regarding urban areas, generating a broad set of different information about the city. The experiment consists of a full network simulation alongside report guidelines to implement in the real environment with car-mounted sensors and static sensors. They perform an entire simulation while guiding to apply on the real situation with car-mounted sensors and static sensors that measure O_3, SO_2 and NO_x and Benzene. The application implements a hierarchical P2P network architecture formed by the mobile and stationary sensors, making full use of the roadside devices to fix the stationary nodes and the public vehicles to carry the mobile sensors. The application provides a SQL database. It centrally stores/maintains all the archived data: sensor samples and secondary information such as traffic and air quality data. Besides, they also implement node clustering and data fusion techniques.

Hu et. al (2009) [9] present a standard VSN application implemented to perform micro-climate monitoring. They report the design method of a hardware prototype node; this node is attached to the tested vehicle to network the mobile capability. The application integrates a map service to show the collected data. The experiment consists of car-mounted sensors running in a real environment observing the concentration of CO_2. Sensor nodes have GPS and Cellular connection. The prototype device issued at hand is decoupled in two parts, placed inside (GSM and GPS module) and outside (CO2 sensor) of the vehicle. There's a Zigbee-based intra-vehicle wireless exchange of the sensed data before uploading it to the server. At the sampling level, nodes perform an adaptive reporting rate based on the overall variance of CO2 concentration at sampling alongside a local data aggregation by a simple average.

An in-depth sequel of [9], where the subject matter shows a detailed problem characterization, is presented at Hu et. al (2011) [10]. A network simulation evaluates the behavior of applications looking at metrics of estimated error and message traffic. We validate it through a full network simulation that keeps the reports of implementation guidelines at realistic scenarios. They improve the network strategy by adding a V2V communication between the mobile sensors to save cellular bandwidth by performing data aggregation on forwarded sensor readings. Considering the overall aspects considered on it, the nature of this work **shows a mature simulation methodology** that fits as **an excellent bottom line at our research**.

The work addressed in [2] describes the hardware architecture, amount of memory employed, sensor specifications, and data traffic interfaces. It explains the cloud server, which centralizes the data and overall functional requirements for the application. It also approaches the layers of software architecture, highlighting the implementation level of physical layer functions. The application top layer features the evaluation of the Air Quality Index and the interface with map service.

Khedo et. al (2010) [12] discuss an air pollution monitoring system through a traditional static wireless sensor network. They propose a novel data aggregation technique called **Recursive Converging Quartiles** to avoid redundancies at the top level. They explain all network architecture components and highlight the duplication of elimination policy at routing and node clustering algorithms. The experiment consists of simulation only application of a traditional, static, and multihop-based WSN. In this paper, there's no explicit mention of air sensors employed. Instead of that, the variable of interest is a virtual abstraction of Air Quality Index samples. Besides, they also implement node clustering and data fusion techniques.

Following this line, the article by Devarakonda et. al (2013) [6] describes the mobile schema, the cloud server application, and involved costs of development for two different proposed hardware prototypes to measure pollution. Referred to as Mobile Sensing Box (MSB) and Personal Sensing Device (PSD), these equipment are respectively suited to be attached to the vehicle and carried by the user in the context of a collaborative sensor network. This application also offers a web-based portal that evaluates Air Quality Index and displays it as a heat map. They equip the sensor nodes with GPS, Cellular, and Bluetooth connection. Network topology comprises a public transportation based VSN and a Community Sensor Network (CSN) through a collaborative approach. Carbon Monoxide and Particulate Matter are the observed pollution indicators. Regarding the data processing, variable sampling concerning the vehicle, and spatial gradient speed, data aggregation is performed at the top-level base station to improve accuracy.

The design of physical WSN structures requires making decisions about a set of different requirements. The contribution of Hejlová and Vozenílek (2013) [8] addresses the criteria for the selection of hardware components. The perform a comprehensive review of available technologies, comparing the hardware suppliers in summary tables and classifying the components by communication capability, energy source, environment resistance, different technical aspects, dimensions, and price. The authors bring a relevant contribution with solid guidelines to present the available technologies suitable for air pollution monitoring problems.

Having a strong intersection with the subject addressed in this thesis, the survey from Yi et. al (2015) [29] centralizes a large number of relevant characteristics regarding the specific problem of air pollution monitored through WSN technologies. They introduce spatial coverage concepts versus temporal resolution and compare them with the overall cost of different network architectures. Explain the differences, advantages, and drawbacks of networks based on static nodes, community sensors, vehicle mobility, and conventional stationary base stations, alongside states of better suits for each purpose, such as cost efficiency, maintenance, and data quality. Moreover, it discusses the various air quality standards, comments available hardware, and the operating principles. In the same line, Pavani and Rao (2017) [15] addresses a similar issue as [29] in a shorter review.

Rashid and Rehmani (2016) [18] address the overall challenges of WSN in urban regions. Considering the choice of city environment as the region of interest in our research, they discuss an extensive range of sensor networks and allows us to bring up known issues and capabilities observed at related applications at a higher level.

Wang and Chen (2017) [28] propose a novel approach over a Vehicle Sensor Network, which consists of a probabilistic strategy to handle adaptive sampling of cars and balancing the trade-off between monitoring accuracy and communication cost with data traffic. Referred to as EDGE (Efficient Data Gathering and Estimation), it works with a dynamic grid partition based on the variation of pollutant concentration to compute and set the rate by consulting other nodes close to its current grid sector. This simulation comprises sophisticated mobility and pollutant dispersion models, and advances on methodology and metrics previously stated in [10].

Finally, Kaivonen and Ngai (2020) [11] report an experimental study with physical sensors attached to public transportation and describe this hardware prototype's development in detail. It explains the programming and deployment of sensors, data visualization, and evaluation of gathered samples with the Air Quality Index. Besides that, it shows the collected data. It discusses typical real-life challenges such as the noise on measurements, numeric sensor precision, the efficiency of cellular connection and packet loss rate, and limitation of coverage with fixed bus routes. Car-mounted sensor nodes provide communication with GPS and Cellular connection. We consider NO_2 and CO to assess the Air Quality Index while measuring temperature, humidity, and pressure as complementary data.

Tables 1.1 and 1.2 shows a side by side comparison (ordered by publication year) with main aspects considered at related work. Looking at Table 1.1, we highlight that our proposal **handles multiple simultaneous phenomena (MDH, Multivariate Data Handling)** at reconstruction step of environment, thereby taking into account the impact of spatial correlation between different sensed variables. Besides that, we also consider all components for evaluating the Air Quality Index. Regarding Table 1.2, is relevant to highlight that **a combination of real data inputs and simulated environment for experimentation** is only noticed at our proposal. Another observation is that there's no explored opportunity to evaluate sensor-level sampling algorithms (also multivariate) on this kind of scenario.

Table 1.1: Summary of related work (sensors and methods).

Article	Air Sensors/Indicators						Processing on Application Level			
(per year)	AQI	CO_x	O_3	PM	NO_x	SO_x	Compr.*	Aggreg.*	Adaptive Rep.*	MDH*
PROPOSAL		X	X	X	X	X				X
[27], (2008)	X	X	X		X			X		
[14], (2008)			X		X	X		X		
[9], (2009)		X							X	
[2], (2010)	X	X			X	X		X	X	
[12], (2010)	X									
[10], (2011)		X					X		X	
[6], (2013)	X	X		X				X	X	
[28], (2017)	X	X							X	
[11], (2020)	X	X			X					

*Abbreviations respectively for "Compression", "Aggregation", "Reporting", "Multivariate Data Handling".

Table 1.2: Summary of related work (overall aspects).

Article	Experimentation		Input Source		Network Topology			
(per year)	Real Env.	Simulation	Real Data	Simulation	Static	Clustered	V2I	V2V
PROPOSAL		x	x				x	
[27], (2008)	x		x			x	x	
[14], (2008)		x		x		x	x	x
[9], (2009)	x		x				x	
[2], (2010)	x		x				x	
[12], (2010)		x		x	x	x		
[10], (2011)		x		x			x	x
[6], (2013)	x		x				x	
[28], (2017)		x		x			x	x
[11], (2020)	x		x				x	

1.4 Document structure

We organize this document with the current Chapter 1 discussing the research subject's overall topics. Chapter 2 explain the methodologies, and tools, clarifying the flow in which we will perform the experiments will. Chapter 3 explores in-depth how the implementation happened.

At Chapter 4 the results and findings are discussed. Finally, Chapter 5 presents the conclusions and final remarks.

Chapter 2

Multivariate Air Quality Modeling and Application

2.1 Environmental Application Design

Let the overall behavior be denoted by

$$\begin{tikzcd} \mathcal{N} \arrow[r, "P"] & \mathbf{V}^* \arrow[r, "S"] \arrow[d, "R"] & V \arrow[r, "\Psi"] & V' \arrow[r, "\omega"] & V'' \arrow[d, "R"] \\ & D & & & D' \end{tikzcd} \tag{2.1}$$

where $\mathcal{N}$ denotes the environment and the process to be measured, P is the phenomenon of interest, and $\mathbf{V}^*$ is the time-space domain. If a complete and uncorrupted observation is possible, it can devise a set of rules (R), leading to ideal decisions (D).

Replicate this overall behavior for every phenomenon $P_i \mid i = \{1, \ldots, n\}$, where n is the number of different phenomenon under observation, thereby considering its multivariate manifestation.

Furthermore, $\mathbf{S}$ is the set of sensors where $S = \{S_1, \ldots, S_k\}$ and k is the number of sensors available on network. In this case, sensors are mobile and navigate through the monitored area. Each sensor provides measurements of the phenomenon and produces a report in the domain $V_{i,j} \mid 1 \leq i \leq n$ AND $1 \leq j \leq k$ (n is the number of different phenomenon under observation and k is the number of sensors, as mentioned previously). Thereby, the global visualization of sensing activity resulting from the combination of all sets of phenomena covered by every sensor, we denote as $\mathbf{V} = \{V_{(1,1)}, \ldots, V_{(n,k)}\}$.

Dealing with all the data is expensive regarding power, communication bandwidth, and storage capacity. Thus, usually, the application handles actions to output a reduced data-set ($\mathbf{V}'$) obtained from a data sampling or fusion strategy ($\mathbf{\Psi}$) over its entire observed domain ($\mathbf{V}$). After reconstructing the set $\mathbf{V}''$ from V' applying a reconstruction technique ω, we can use the same set of rules R to make decisions D'. The Diagram 2.1 is analogous to present by Aquino et. al [3].

2.1.1 Area of Interest

The process described in Diagram 2.2 is the "zero" step at our modeling approach, which refers to our environment ($\mathcal{N}$) definition. To perform this, we use the raw data-set to extract the environment $\mathcal{N}$ and phenomena of interest $P_i \mid i = \{1, ..., n\}$. **The practical implementation** of this step is described at

section 3.1.1 and 3.1.2.

$$\mathcal{N} \xrightarrow{P} \tag{2.2}$$

Initially, based on raw data-set, we generate the model for the area under experimentation and apply the library GeoBR [13] as support to import the requested maps. For this case study, we defined **the city of São Paulo as an area of interest**. After that, we perform a general-purpose matrix/table handling features, available at R Platform [17], implemented as scripted sequential procedures. The adopted data-set is from Environmental Sanitation Technology Company [5] (CETESB, in the Portuguese acronym) allows us to set 10 districts within the city with air quality stations and measurements available, listed below:

- Cambuci
- Centro
- Congonhas
- Horto Florestal
- Ibirapuera
- Lapa
- Moóca
- P. D. Pedro II
- Pinheiros
- São Miguel Paulista
- Santana
- Santo Amaro

We set the arrangement for stations coordinates following real locations of districts which they are named. Figure 2.1 illustrates the map outputs containing these points. It also proposed a detailed description of roads, bus lines, vehicle itinerary, and general traffic information to achieve proper modeling at the network and application simulation stage.

2.1.2 Phenomenon Observation

Considering the structure in mapped raw data, we identify a set of samples stored from each air quality station's sensed physical variable. This stream reports an observation at the same location of stations. The first stage from the model construction is showed at sub-process on Diagram 2.3

$$\begin{tikzcd}
\mathcal{N} \arrow[r, "P"] & \mathbf{V}^* \arrow[d, "R"] \\
& D
\end{tikzcd} \tag{2.3}$$

To achieve an adequate representation of phenomenon under ideal sensing conditions ($\mathbf{V}^*$), we will perform a multivariate reconstruction (multivariate ordinary cokriging, [16]) to interpolate non-monitored blank areas at this reference set. Thus, setting up a simulated environment while taking its behavior closer to a real one, since a continuous measurement of every single point inside a metropolitan area is an unfeasible task. Finally, snapshots of the field, showing the overall state at each instant, complete the physical process. **A practical instance for domain $\mathbf{V}^*$** was achieved following the process **described at section 3.1.2**.

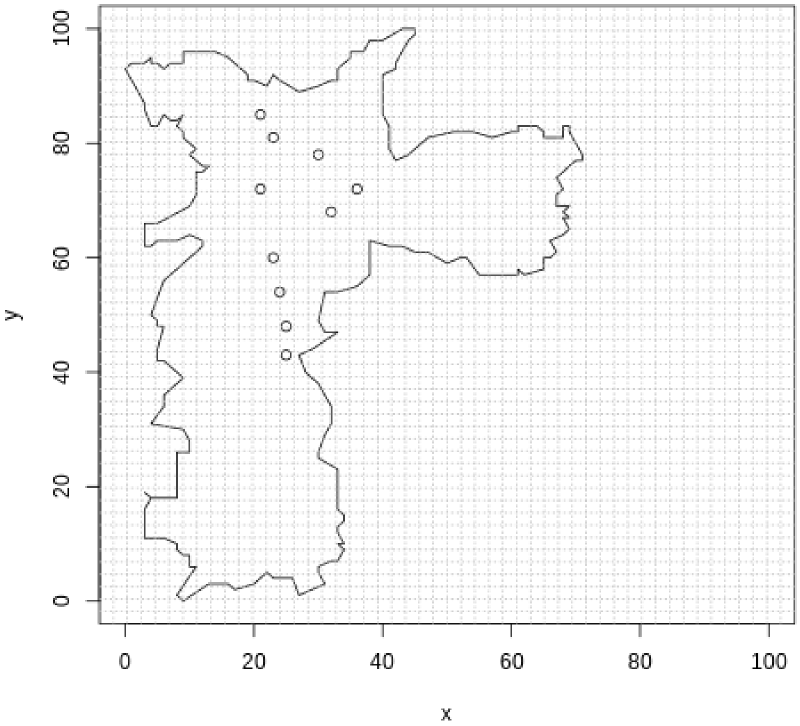


Figure 2.1: Exported map of São Paulo with air quality stations on real approximate locations.

2.2 Network Settings

The network sub-process and sensing activity is covered at Diagram 2.4, where $S = \{S_1, S_2, ..., S_k\}$ and k is the number of active nodes in the network. Considering the multivariate manifestation of monitored physical process, we assume a sensor S_k as $S_k(P)$ where $P = \{P_1, P_2, ..., P_n\}$ and n is the number of observed variables. As explained before, $\mathbf{V}$ is the domain that reports the resultant set from sensing activity.

$$\mathcal{N} \xrightarrow{P} \mathbf{V}^* \xrightarrow{S} V \tag{2.4}$$

Addressing the research background under urban zones, we can highlight alternative monitoring solutions approaching vehicle sensor networks [29, 10, 6]. This case study takes advantage of bus lines as mobile sensing units [11] that rides through the city while collects the data.

The planned network setup considers a vehicle sensor network with a combination of V2V communication performing data aggregation at the application layer alongside a cellular network to upload the sensed data to a cloud server, which is an approach recurrently observed at referenced work along with all this thesis.

In a succinct speak, a sensor S_k is embedded in a single device as a **car-mounted mobile node** and is able to navigate through the area and collect data from variables $\{P_1, P_2, ...P_n\}$.

2.3 Sensing Data Processing

In accordance with statements from Section 2.1, the sub-process of overall behavior (Diagram 2.1) that refers to the stage of data collection and processing is represented at Diagram 2.5 below:

$$V \xrightarrow{\Psi} V' \xrightarrow{\omega} V'' \tag{2.5}$$

Remembering that $\mathbf{V}$ is the domain that reports the resultant set from sensing activity, $\mathbf{V'}$ is a reduced instance of $\mathbf{V}$ after the action of reduction algorithm Ψ, and $\mathbf{V''}$ is the rebuilt field after input $\mathbf{V'}$ to technique which fills blank spaces ω.

In our case study, the process ω consists of assembly of an Voronoi Diagram and reduction algorithm Ψ is not applied, hence the step that comprises generation of $\mathbf{V'}$ was skipped.

2.4 Evaluation

Let the data reconstruction and evaluation rule set be denoted by

$$\begin{tikzcd}
\mathcal{N} \arrow[r, "P"] & \mathbf{V}^* \arrow[r] \arrow[d, "R"] & \ldots \arrow[r] & V'' \arrow[d, "R"] \\
& D & & D'
\end{tikzcd} \tag{2.6}$$

Where "..." represents the whole sub-process sequence from Diagram 2.5 (processing the sensed data). We apply the rules R over the reconstructed data. The first rule considered to evaluate the performance at each scenario is the Absolute Value of Relative Error ($\hat{\epsilon}$) [7], which is defined as follows:

$$\hat{\epsilon} = \frac{1}{\mathcal{L}} \sum_{x,y}^{S} \left| \frac{\mathbf{V}^*(x,y) - \mathbf{V}''(x,y)}{\mathbf{V}^*(x,y)} \right|,$$

where $\mathbf{S}$ is the set of (x,y) coordinates that belongs to internal area of Figure 2.1, parsed as valid inputs to reconstruction technique, $\mathcal{L}$ is the length of set $\mathbf{S}$, $\mathbf{V}^*$ is the field that represents the environment and was initially simulated; $\mathbf{V}''$ is the rebuilt field. Moreover, by the fact that by input data that generates $\mathbf{V}^*$ was pre-processed to handle all NA measurements, it can always ensure the definition of $\hat{\epsilon}$ since $\mathbf{V}^*(i,j) \neq 0$.

Besides that, as mentioned at Section 1.1 we also intend to formulate rules for assessment of **field coverage**, **bandwidth consumption** and **packet delivery delay**, as complementary evaluation metrics to be developed at subsequent steps of this work.

In order to fill left blank spaces from non visited areas, the adopted strategy consists of set those unavailable locations using Voronoi diagrams [4].

The Voronoi diagram is defined as follows: assuming the location of sensors (S) as a set of n points in an area, the dominance of S_p over S_q is the subset (or sub-area) of the plane that is closer to S_p than S_q. Formally,

$$dom(S_p, S_q) = \{x \in R^2 | \rho(x, S_p) < \rho(x, S_q)\},$$

where ρ represents the Euclidean distance function and x represents a given point in the R^2 plane. In this problem, the seeds in the diagram represent the locations visited by the busses (VSN strategy) and air quality station locations (Conventional monitoring strategy), and the dominance is the sub-areas (Voronoi cells) covered by each seed. Thus, this area is used to compose $\mathbf{V}''$ for each monitoring approach.

Chapter 3

Methodology and Implementation

This experiment comprises three main stages: i) generation of pollutant maps; ii) traffic simulation; iii) environment assembly. In the first step, we import the raw data-set to create the pollutant maps, so every coordinate from the city area has a well-defined sample for each timestamp. Section 3.1 details this procedure.

The second step is to set up the traffic simulation. At this stage, there is a set of sequential tasks that comprises a routine to fetch map files and split them into six bounding boxes. This action avoids scalability problems due to large file sizes. After that, cars, busses, and their respective routes are generated and executed with different traffic intensity levels. This step's main outcome is to export the bus traces with visited map coordinates during the route. This procedure is detailed at Section 3.2.

Finally, the third step is to assemble the overall environment by matching the measurements at each map coordinate and the exported trace with coordinates visited during the bus lines. In this way, it is possible to evaluate the field coverage by looking at how many coordinates the solution covered over time. Section 3.3 details this procedure.

3.1 Multivariate Pollutant Map Generation

3.1.1 Dataset Handling

It provided a set of files containing real data from air quality stations placed in São Paulo from Jan-01-2005 to Dec-31-2005. About 15 different variables reporting information such as wind speed and direction, atmospheric stability, temperature, humidity, and other classes of pollutants are available. To delimit the research scope, we reduce these variables from 15 to 5 considering the pollutant agents from burning of fossil-fueled vehicles. Thus, to evaluate our experiments, we selected as input data of Carbon Monoxide (CO); Particulate Matter (PM10); Nitrogen Dioxide (NO_2); Ground-level Ozone (O_3) and Sulfur Dioxide (SO_2).

We organize the raw data by variable (sensor) at each file: We group all stations which have available this sensor on this list, where the columns are for miscellaneous information and measurements (resulting in five records with a similar structure as illustrated at table 3.1). On the other hand, rows repeat the date and time for samples on each listed station resulting in a noticeable redundancy amount.

Table 3.1: Raw data.

		Sensor (i.e.: O_3)			
Station ID	Date	Time (h)	Sample	Validation	Station Name
1	01/01/2005	01:00	39.22	9	P. D. Pedro II
1	01/01/2005	02:00	23.85	9	P. D. Pedro II
1	01/01/2005	03:00	29.68	9	P. D. Pedro II
			...		
2	01/01/2005	01:00	30.85	9	Santana
			...		
47	31/12/2005	00:00	7.44	9	Horto Florestal

We handle these data according to the following steps: 1) filter by station ID, date, time, and samples; 2) Summarize all data in a single file whereas the columns are labeled by a combination of station ID (as a prefix) and each respective sensor, alongside date and timestamp (table 3.2. Alongside the overall rearrangements, we also handle the format issues observed at raw data (i.e., comma instead of the dot at measurements representation and miscellaneous date format).

The date length was reduced to a range of two weeks, enough to run the simulations and allow feasible processing with available computing resources. An important point to consider is the usual occurrence of unavailable data due to sensor fails or maintenance, which corroborates the reduction performed. Finally, for validation purposes, we choose the interval of oct-15-2005 to oct-22-2005 through visual inspection, and we select an appropriate subset.

Table 3.2: Sumamrized data after handling.

Date	Time (h)	Station-Sensor$_{(1,\ldots,n)}$					
		1-CO	1-PM10	...	5-NO_2	...	47-O_3
2005-10-15	00:00	0.71	8.52	...	NA	...	18.69
2005-10-15	01:00	0.67	33.37	...	38.74	...	10.56
2005-10-15	02:00	0.79	31.39	...	51.38	...	12.46
...	...	...	...	...	...	...	...
2005-10-21	23:00	0.62	4.32	...	25.39	...	16.19

Finally, we provide an outline of the map from São Paulo at package **GeoBR** [13] that is, the input points to the prediction process alongside each air quality station's coordinates, placed at this step (as seen in Figure 2.1).

3.1.2 Prediction

Considering the data structure, we perform the prediction step in two directions explained in the subsections below. Besides that, the Figure 3.1 shows the ratio between the real data initially available (raw data) and samples predicted using the raw data as input with methods presented at subsections 3.1.2 and 3.1.2.

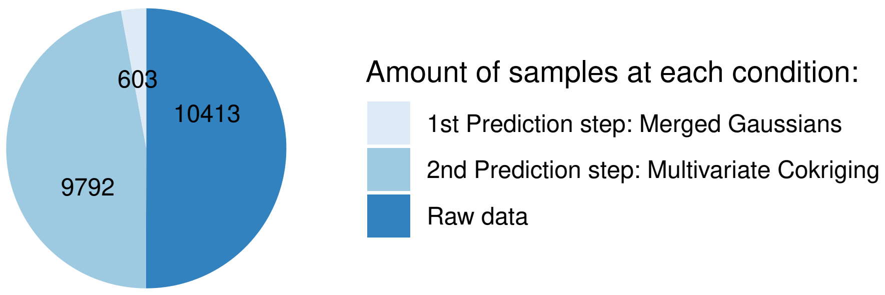


Figure 3.1: Amount of real and predicted samples.

About univariate time series

The first direction is to gather the variables one by one at each column (that shows samples of a sensor within a single station). This step aims to fill the blank spaces caused by **NA** occurrences keeping the overall behavior. For that, we take the actions as described as follows.

1. Assert each single column (example at Table 3.3) as a matrix (example at Table 3.4) with **hours** (0h–23h) × **days** (15–31);

Table 3.3: Selecting a column (gray highlight) to handle NA samples.

Date	Time (h)	Station-Sensor$_{(1,\ldots,n)}$					
		1-CO	1-PM10	...	4-PM10	...	47-O_3
2005-10-15	00:00	0.71	8.52	...	30.84	...	18.69
2005-10-15	01:00	0.67	33.37	...	35.54	...	10.56
2005-10-15	02:00	0.79	31.39	...	33.98	...	12.46
...	...	...	...	...	...	...	...
2005-10-21	23:00	0.62	4.32	...	13.94	...	16.19

Table 3.4: Entire **4-PM10** column asserted as a matrix hour × days (before prediction).

	15-Oct	16-Oct	17-Oct	18-Oct	19-Oct	20-Oct	21-Oct
01:00h	12.08	48.39	14.04	47.06	NA	9.59	14.06
02:00h	8.53	36.77	6.53	44.12	NA	20.96	9.93
03:00h	28.64	46.46	15.42	52.75	NA	18.18	16.22
04:00h	31.75	45.08	7.22	49.57	NA	16.59	10.04
05:00h	20.36	35.26	1.52	NA	NA	7.79	12.55
06:00h	36.69	37.38	12.13	NA	NA	12.28	19.94
07:00h	33.61	47.81	20.21	NA	NA	8.18	7.01
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
23:00h	61.69	33.09	33.12	NA	11.66	32.48	6.56
00:00h	60.2	16.2	47.48	NA	11.95	19.55	13.94

2. Test normality (Shapiro-Wilk) for everyone, evaluate mean μ and standard deviation σ by two times, for entire row and for the whole column that crosses on the current **NA** cell (at hour × days matrix);

3. Generate a merged normal curve parsing the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ and sample a random value from this distribution.

4. After that, this procedure should deliver a Table with **NA** samples fixed for every variable (standalone pollutant sensor, illustration at 3.5).

Table 3.5: Entire **4-PM10** column asserted as a matrix hour × days (after prediction).

	15-Oct	16-Oct	17-Oct	18-Oct	19-Oct	20-Oct	21-Oct
01:00h	12.08	48.39	14.04	47.06	21.19	9.59	14.06
02:00h	8.53	36.77	6.53	44.12	28.6	20.96	9.93
03:00h	28.64	46.46	15.42	52.75	55.92	18.18	16.22
04:00h	31.75	45.08	7.22	49.57	29.22	16.59	10.04
05:00h	20.36	35.26	1.52	5.7	2.7	7.79	12.55
06:00h	36.69	37.38	12.13	29.17	18.79	12.28	19.94
07:00h	33.61	47.81	20.21	55.06	8.4	8.18	7.01
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
20:00h	46.03	32.25	40.29	46.64	11.37	33.73	16.5
21:00h	53.54	39.73	27.42	30.22	15.45	16.69	13.9
22:00h	64.52	16.73	38.42	49.13	7.8	32.74	10.41
23:00h	61.69	33.09	33.12	2.14	11.66	32.48	6.56
00:00h	60.2	16.2	47.48	37.17	11.95	19.55	13.94

About overall multivariate measurements

All five pollutant sensors (CO, PM10, NO_2, O_3, SO_2) are not available on every station. For this reason, there is a lack of measurement at some input coordinates for reconstruction. This absence of data disturbs the prediction for the multivariate phenomena process so that all points should be available on each station coordinates.

We perform a partial pairing on intersections of stations with the same available sensor variables to estimate those missing points. We present a sequential grouped multivariate reconstruction to fill up all gaps.

The ID set considered for stations are $\{1, 2, 3, 4, 5, 8, 12, 16, 27, 47\}$, the arrangement of sensor availability (Table 3.6) is performed according with subsequent steps and color labels described below:

Available: Indicates that the time series for this variable at respective station is currently **available and ready to use**.

Missing: Indicates that the time series for this variable at respective station **is not available**.

Predicted: Indicates that a time series for this variable (which was **previously missing**) will be **predicted on this turn**, so that, will become displayed as **available** at next turn.

Table 3.6: Data availability.

Station ID	Sensors				
1	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
2	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
3	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
4	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
5	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
8	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
12	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
16	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
27	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
47	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$

1. Looking on data: CO, PM10 and O_3 are available at stations $\{1, 3, 5, 16, 27\}$. From that, predict missing points that should be at stations $\{2, 4, 47\}$ for CO, $\{8, 12, 47\}$ for PM10 and $\{4, 8, 12\}$ for O_3. In that way, fill up data according to directions below.

Table 3.7: Data availability at first turn of prediction.

Station ID	Sensors				
1	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
2	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
3	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
4	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
5	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
8	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
12	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
16	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
27	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
47	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$

2. On this partially fixed data, get the stations $\{5, 12, 27, 47\}$ to reconstruct and append NO_2 predictions on points of stations $\{1, 2, 3, 4, 8, 16\}$.

Table 3.8: Data availability at second turn of prediction.

Station ID	Sensors				
1	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
2	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
3	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
4	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
5	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
8	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
12	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
16	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
27	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$
47	**CO**	**PM10**	$\mathbf{O_3}$	$\mathbf{NO_2}$	$\mathbf{SO_2}$

3. On this next partially fixed data, get stations $\{5, 8\}$ to predict SO_2 append the evaluated points on stations $\{1, 2, 3, 4, 12, 16, 27, 47\}$

Table 3.9: Data availability at third turn of prediction.

Station ID	Sensors				
1	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
2	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
3	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
4	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
5	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
8	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
12	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
16	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
27	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**
47	**CO**	**PM10**	**O_3**	**NO_2**	**SO_2**

With missing points corrected after this procedure, we parse the data as input to overall multivariate kriging reconstruction (supported by R package GSTAT). This assembled field will be the starting point ($\mathbf{V}^*$ from Section 2.1, Diagrams 2.2 and 2.3) to run the vehicle network simulations.

Finally, the achieved outcome after data handling mentioned above is a set of five tables (one for each pollutant: CO, PM10, NO_2, O_3 and SO_2). Each table is assembled by placing lines as timestamps (day/hour) and columns as valid coordinates inside the map outline. The referred structure is displayed at Table 3.10.

In other words, each row in this table represents the entire map area in a certain timestamp, where x, y (two-dimensional) coordinates are disassembled in a vector shape (one-dimensional) to fit as table columns and, afterward, export them as a .CSV file. Note that the highlighted gray row at Table 3.10 generates the visualization illustrated at map of Figure 3.2.

Table 3.10: Pollutant summary table representing $\mathbf{V}^*$.

Pollutant P_i							
Date	Time (h)	Coordinates$_{(x,y)}$					
		(5,3)	(6,3)	(7,3)	...	(11,22)	(11,23)
2005-10-15	00:00	1.75	1.75	1.75	...	1.74	1.76
2005-10-15	01:00	1.76	1.77	1.77	...	1.75	1.77
2005-10-15	02:00	1.78	1.79	1.78	...	1.72	1.75
...	...	...	...	...	...	...	...
2005-10-21	23:00	1.57	1.58	1.58	...	1.37	1.41

3.2 Traffic Simulation

The pollution map described in the previous Section (Table 3.10) is the general structure used as the baseline in our experiments. The subsequent stage from building and executing our simulation framework consists of setting up the urban traffic behavior. We adopt the Simulator of Urban MObility (SUMO) to run this experimentation. At this stage, we assess three sub-steps at the following subsections.

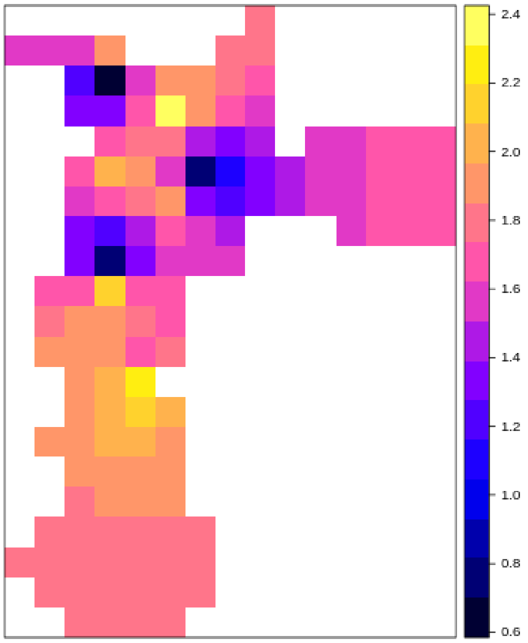


Figure 3.2: Multivariate pollution map of Carbon Monoxide (CO) at a certain timestamp.

3.2.1 Fetch and build maps and roads

The starting point for build the simulation structure is to define a map tool compatible with SUMO. For this task, we use the Open Street Maps (OSM) API that suits these requirements.

This map database can be reached through a web wizard through a visual user interface or integrated with a batch script to automate the download and building process. In our case study, the parsed boundaries cover the outline of the São Paulo area under the following coordinates:

South Latitude:	-24.01
South Longitude:	-46.83
North Latitude:	-23.35
North Longitude:	-46.36

Besides that, we split the downloaded map into six parts, making a grid with 3 × 2 shape. This action is necessary due to the unfeasible resource usage (CPU, RAM, and Disk) when we use a single huge map area as input to traffic simulation. An additional improvement is that every single part of the split map (hereafter referred to as **bounding box**) can be executed as independent instances and allowing the parallelization of experiments, hence taking advantage of CPU multi-threading.

Furthermore, there is an intrinsic tricky detail on the mentioned procedure. The six independent bounding boxes splitting the area imply an offset correction for each one since they will show relative coordinates starting from x=0, y=0. The offset matrix for coordinate fixing is illustrated at Table 3.11.

Table 3.11: Raw trace file of bus routes generated at traffic simulation.

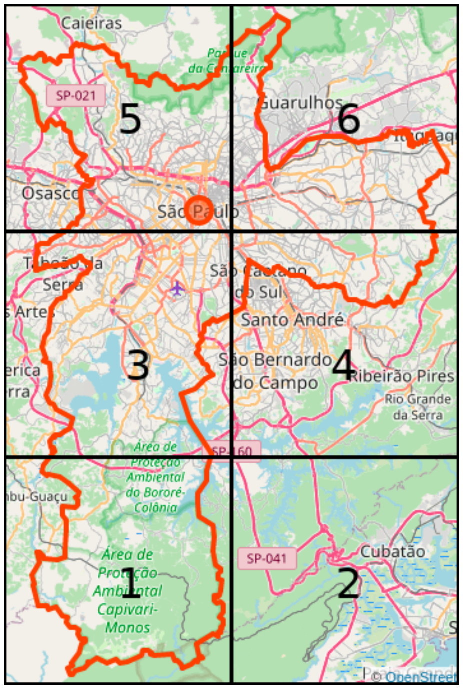


Figure 3.3: Illustration of map bounding boxes.

Bounding Box	x_{min}	y_{min}	x_{max}	y_{max}
1	0.00	0.00	24131.19	24378.38
2	24131.19	0.00	48233.20	24482.42
3	0.00	24378.38	24248.61	48987.30
4	24131.19	24482.42	48379.80	49342.09
5	0.00	48987.30	24280.14	73614.98
6	24131.19	48379.80	48358.66	72962.93

3.2.2 Generate description files for vehicles routes

When the roads and highway structures are appropriately in place, the subsequent step is to generate vehicle routes. SUMO simulator supports other vehicle types as subways, city rails, and trolley cars, but we do not consider them so far.

The vehicle generation consists of bus stops and their respective prior defined lines for public transportation. This information is available at maps built at the subsection 3.2.1. During the simulation, the busses will loop on those defined routes and be analyzed in realistic environments. On the other hand, for small passenger cars, the application performs an insertion with random routes and starting places for each one. These vehicles disappear from simulation after reach the end of their routes.

It is relevant to highlight that the vehicle generation (Figure 3.4) mostly happens during the graphic's ascending part. This behavior occurs because a parameter limits the maximum amount of vehicles to the set-points mentioned in subsection 3.2.3. The simulator only actuates to generate new vehicles when the older ones disappear after they finish their respective routes.

We consider each day as an independent seed that randomly sets the route, starting, and endpoint of each passenger vehicle ride (departure/arrival). In this way, we can meet adequate experimentation representativeness.

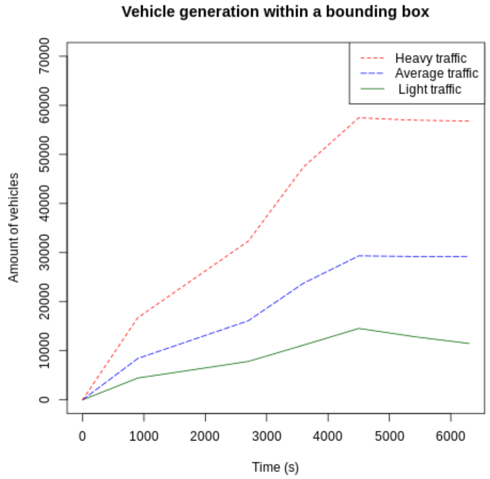


Figure 3.4: Vehicle generation at SUMO Simulator under different traffic conditions.

3.2.3 Run urban mobility simulation

After getting ready all previous settings, the main goal at this last sub-step is to generate the output traces, where the visited coordinates at each bus line will be displayed.

We generate three different traffic intensities (referred respectively as LIGHT, AVERAGE and HEAVY traffic), limiting each bounding box at 15000, 30000 and 60000 vehicles. Since there are six bounding boxes to cover the city outline fully, each scenario's resulting vehicle amount is 90000, 180000, and 360000. From that, approximately 12000 busses are running inside the entire map area.

Finally, we consider the range of 7 days adopted for the simulations with a phenomenon refresh rate of 1 sample per hour for generated data. We consider each day as an independent seed that randomly sets the route, starting, and endpoint of each passenger vehicle ride (departure/arrival). In this way, we can meet adequate experimentation representativeness.

3.3 Environment Assembly

The last stage of experimentation consists of put together the output from two previous ones (multivariate pollution maps from Section 3.1 and bus routes trace from Section 3.2). We base the environmental assembly application in R Statistical Language, where we handle the trace file to match coordinates with pollution maps.

We note an additional point of complexity since the downloaded maps from OSM API come with geo-referenced coordinates on WGS84 format (latitude/longitude or directly converted to an arbitrary x, y notation). To match the trace coordinates with multivariate pollution maps, we perform an intermediary step of scaling from the default OSM coordinate format to our 25×25 defined scale. This procedure is illustrated in Figure 3.5.

Initially, the raw trace is exported as a .XML file and looks like the Table 3.12. After the scaling described above, we generate from a halfway structure that suits two purposes: i) the previously mentioned data compatibility to calculate the overall field coverage; ii) evaluate how many sectors are visited by the busses concerning traffic intensities. This information can be assessed as a performance metric for our application since that it provides information about how the vehicles behave in different traffic conditions. Table 3.13 illustrates how this structure looks like.

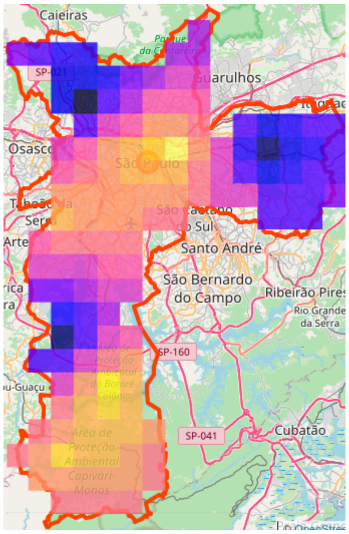


Figure 3.5: Matching the city area with scaled 25 × 25 pollutant map.

Finally, we match the map coordinates with the downloaded area under the following conditions: (i) If there is at least one bus inside a single sector from the 25 × 25 map, we consider it covered, (ii) we disregard points outside the area, i. e., only locations inside the city.

Table 3.12: Raw trace file of bus routes generated at traffic simulation.

Timestamp	Bus ID$_{(1,\ldots,n)}$	Coordinates$_{(x,y)}$
0s	1	(11140.66, 19814.56)
0s	2	(4667.23, 6571.91)
...	...	...
900s	1	(10125.04, 20243.59)
900s	2	(3024.51, 5135.84)
...	...	...
1800s	1	(9043.16, 18450.32)
...	...	...
2700s	1	(8486.62, 17651.08)
...	...	...
5400s	n	(850.87, 19199.05)

Table 3.13: Converted trace data with coordinates scaled as 25 × 25.

Timestamp	Bus ID$_{(1,\ldots,n)}$						
	0	1	2	3	...	$n-1$	n
0s	(5,8)	(3,8)	(4,6)	(5,8)	...	(6,8)	(7,7)
900s	(6,8)	(3,8)	(4,7)	(5,8)	...	(5,8)	(7,8)
1800s	(7,8)	(3,8)	(5,7)	(6,7)	...	(3,6)	(6,9)
...	...	...	...	...	...	...	...
5400s	(7,8)	(3,9)	(5,8)	(7,7)	...	(4,7)	(6,8)

Chapter 4

Results

4.1 Preliminary Assumptions

Initially, we perform all the required data handling on traces and phenomena information. The following actions aim to achieve a performance assessment looking at overall field coverage and error rate from measurements at both approaches: Sampling on conventional weather stations or aided by a VSN network with sensor nodes mounted on public transportation (bus lines).

Table 4.1: Simulation parameters

Parameter	**Values**
Pollutant Variables	CO, PM10, O_3, NO_2, SO_2
Pollutant Map Scale	25 × 25 size units
Map Area	132 squared size units
Number of Busses	12k
Traffic Density	Light (90k), Average (180k) , Heavy (360k)
VSN Sample Rate	900s (15 minutes)
Simulated Time	7 days (random seeds for each one)

4.2 Summary for Global Field Coverage

This performance assessment looks at the coordinates where each weather station is located or visited from each bus line. The obtained coordinates from this procedure are weighted under two directions: (i) concerning the broad set of map coordinates and (ii) about traffic intensities over the day.

From Table 3.13 we can summarize all visited coordinates from bus lines, the Timestamps column is matched with reference pollutant map according to the Table 4.2 to achieve the global coverage.

Table 4.2: VSN visited coordinates matching considering traffic intensity in relation to times of the day.

Time	Traffic Intensity		
(h)	Light	Average	Heavy
0h - 6h	x		
6h - 7h		x	
7h - 9h			x
9h - 11h		x	
11h - 13h			x
13h - 17h		x	
17h - 19h			x
19h - 0h		x	

After relating the covered areas from each traffic intensity with the respective time of the day, the assembled coverage matrix resulting from our VSN application looks like the Table 4.3 as follows.

Table 4.3: VSN summarized coordinates list to evaluate Global Coverage.

Day	Traffic Intensity (× Coords List)			Global Coverage
(seed)	Light	Average	Heavy	%
1	(3,5); (4,4); (4,5) ...	... (6,9); (7,10) ...	... (14,19); (15,16) ...	86.74%
2	(3,5); (4,4); (4,5) ...	... (6,10); (6,11) ...	... (15,18); (15,19) ...	86.36%
3	(3,5); (4,4); (4,5) ...	... (6,12); (6,13) ...	... (16,16); (16,17) ...	86.55%
...	...	...	...	...
7	(3,5); (4,4); (4,5) ...	... (6,14); (6,15) ...	... (16,18); (16,19) ...	86.74%

Considering the presented strategy, after a 7-days run with random and independent seeds for each day, **our VSN application achieved a global field coverage of 86.55%**, on average.

On the other hand, the conventional stations are only aware of phenomenon data on their current sector, taking into account the 132 sectors (see Table 4.1) covered by the map area and the known amount of 10 stations. **The regular monitoring system achieve a theoretical global field coverage of 7,5%**.

4.3 Summary for Carbon Monoxide

In the following sections, we assess the sampled data representativeness through the evaluation of Absolute Value of Relative Error (detailed at Section 2.4)

Looking on Carbon Monoxide pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 4.1) is described as follows:

- VSN average error: 0.25%
- Conventional Stations average error: 31.57%

This result means an improvement on the order of 126 times lower error about the regular monitoring system. Figure 4.2 shows the resulting behavior of each monitoring strategy.

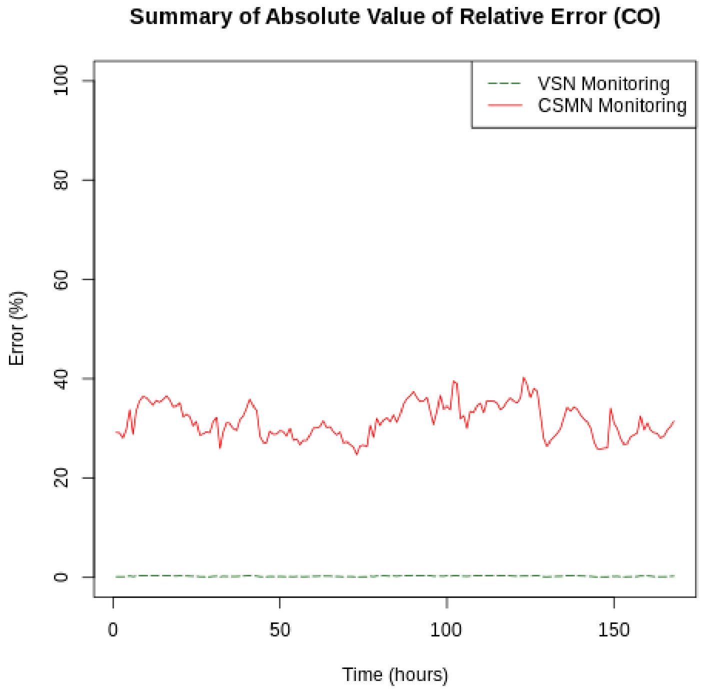
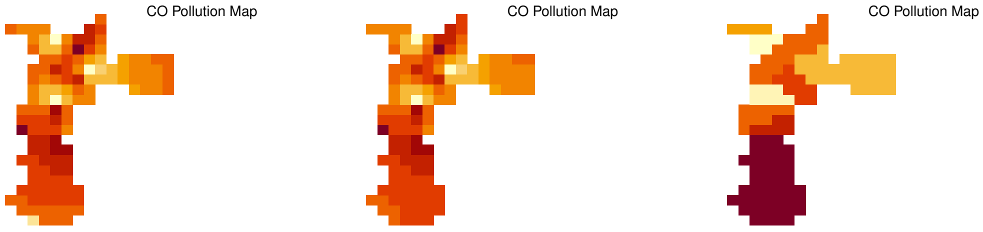


Figure 4.1: Error rate evaluated through 7 days from VSN and Conventional Stations (CO).



(a) Reference multivariate data. (b) Samples from VSN. (c) Samples from Stations.

Figure 4.2: Comparison between sampling with VSN and Conventional Stations for CO.

4.4 Summary for Particulate Matter

Looking on Particulate Matter pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 4.3) is described as follows:

- VSN average error: 0.36%
- Conventional Stations average error: 35.93%

This result means an improvement on the order of 99.8 times lower error about the regular monitoring system. Figure 4.4 shows the resulting behavior of each monitoring strategy.

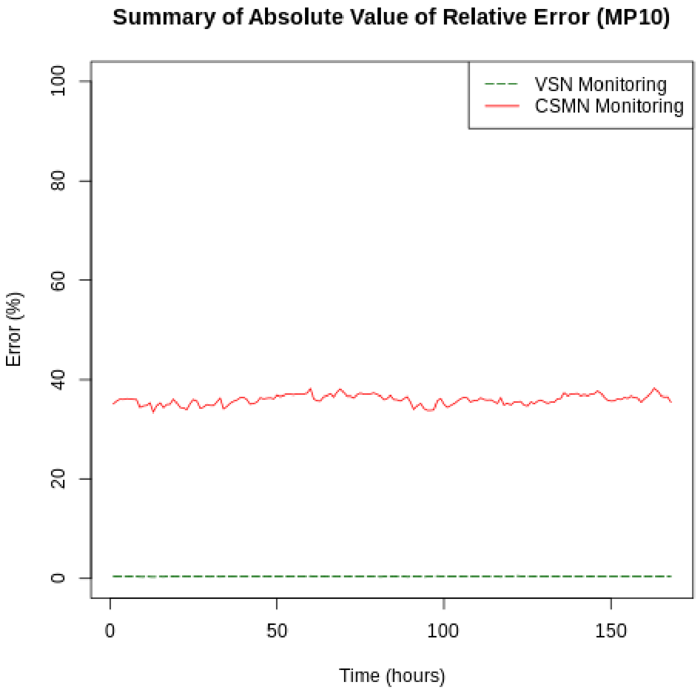
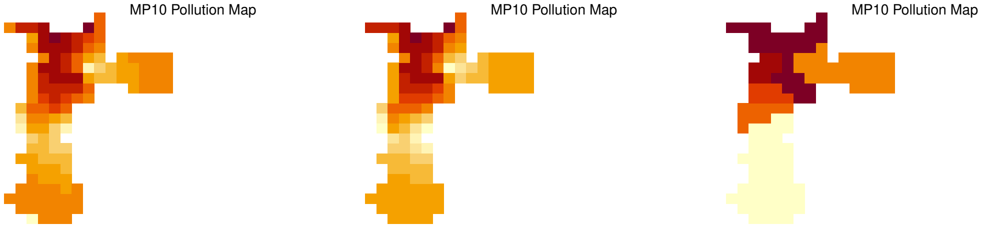


Figure 4.3: Error rate evaluated through 7 days from VSN and Conventional Stations (PM10).



(a) Reference multivariate data. (b) Samples from VSN. (c) Samples from Stations.

Figure 4.4: Comparison between sampling with VSN and Conventional Stations for PM10.

4.5 Summary for Nitrogen Dioxide

Looking on Nitrogen Dioxide pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 4.5) is described as follows:

- VSN average error: 0.48%
- Conventional Stations average error: 34.03%

This result means an improvement in the order of 70 times lower error about the regular monitoring system. Figure 4.6 shows the resulting behavior of each monitoring strategy.

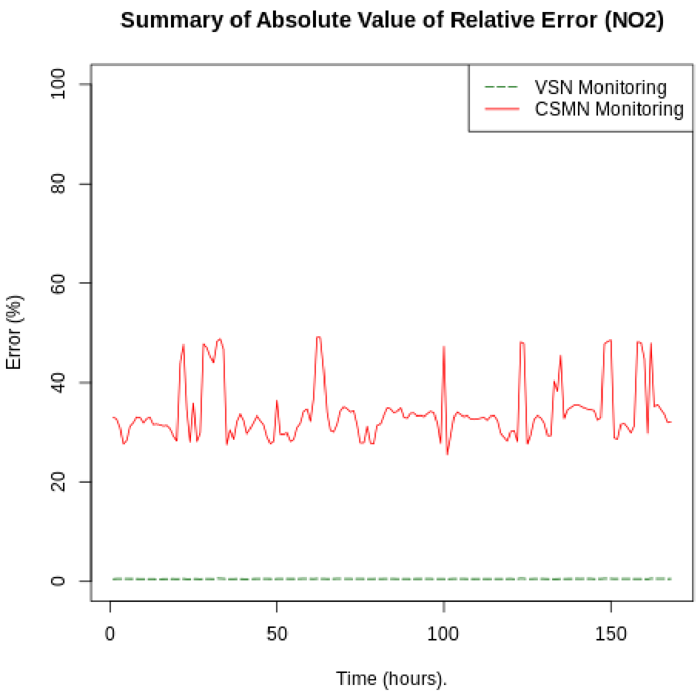
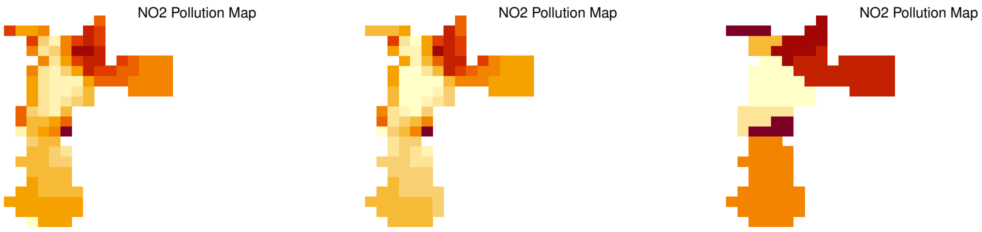


Figure 4.5: Error rate evaluated through 7 days from VSN and Conventional Stations (NO_2).



(a) Reference multivariate data. (b) Samples from VSN. (c) Samples from Stations.

Figure 4.6: Comparison between sampling with VSN and Conventional Stations for NO2.

4.6 Summary for Ground-level Ozone

Looking on Nitrogen Dioxide pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 4.7) is described as follows:

- VSN average error: 0.32%
- Conventional Stations average error: 32.27%

This result means an improvement on the order of 100 times lower error about the regular monitoring system. Figure 4.8 shows the resulting behavior of each monitoring strategy.

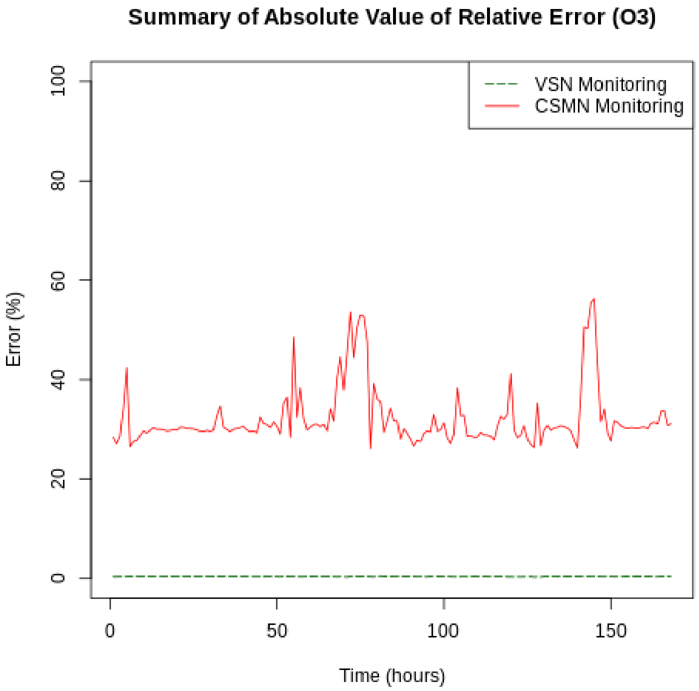
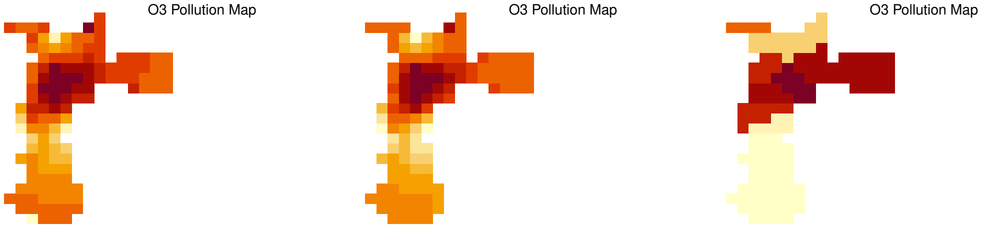


Figure 4.7: Error rate evaluated through 7 days from VSN and Conventional Stations (O_3).



(a) Reference multivariate data. (b) Samples from VSN. (c) Samples from Stations.

Figure 4.8: Comparison between sampling with VSN and Conventional Stations for O_3.

4.7 Summary for Sulfur Dioxide

Looking on Sulfur Dioxide pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 4.9) is described as follows:

- VSN average error: 2.02%
- Conventional Stations average error: 13.33%

This result means an improvement on the order of 6.59 times lower error concerning the regular monitoring system. Figure 4.10 shows the resulting behavior of each monitoring strategy.

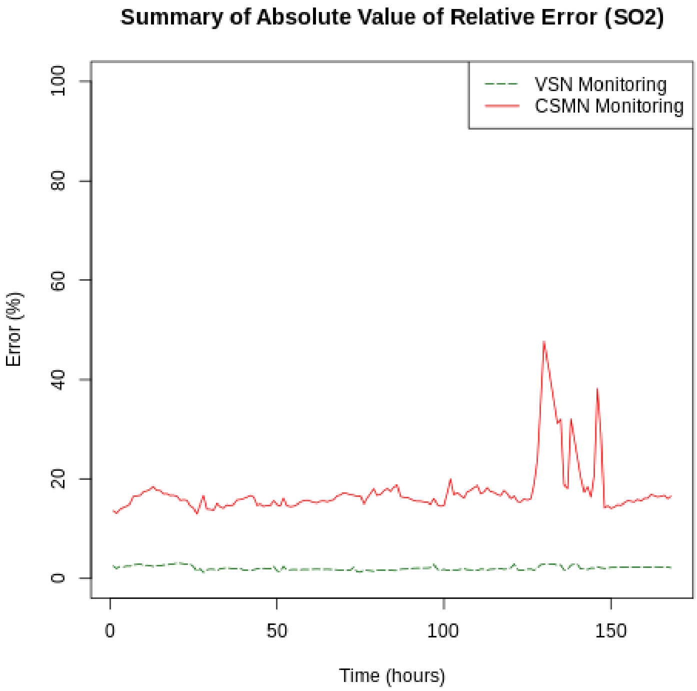
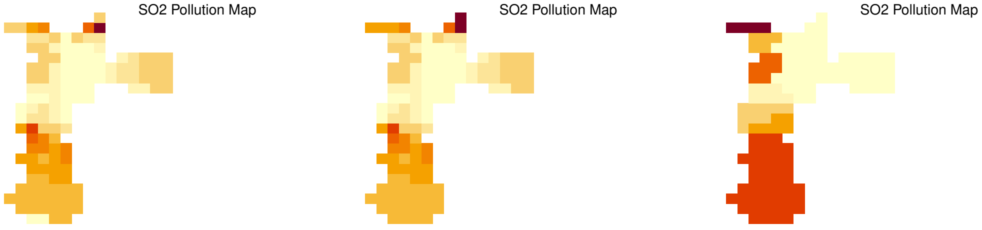


Figure 4.9: Error rate evaluated through 7 days from VSN and Conventional Stations (SO_2).



(a) Reference multivariate data. (b) Samples from VSN. (c) Samples from Stations.

Figure 4.10: Comparison between sampling with VSN and Conventional Stations for SO_2.

Chapter 5

Conclusion and Final Remarks

The presented thesis has explored the problem of air quality monitoring while takes a more in-depth investigation on modeling of complex environments. Beyond that, to deliver contributions that improves the view of how correlated multivariate phenomena behaves.

A network simulation environment was developed to validate the consistency of proposed modeling and methodologies, thereby assess the obtained research outcomes as close as possible with real-life scenarios. The achieved results shows that approaching of multivariate-based processing techniques is a trustworthy viable path to accurately predict realistic behaviors of correlated physical processes.

Besides that, the approached Vehicle Sensor Network supported by public transportation (bus lines) showed considerably higher performance than the regular monitoring system based on conventional air quality stations, behaving with low error rates and about 11.5 times higher global coverage. Thereby, overall observed performance indicates that the proposed application on this case study is suitable to be executed in real world scenarios.

Future directions consider evaluate different classes of data processing algorithms and improve environmental modeling with variables that were not considered at last turn, such wind speed/direction, temperature, humidity data on evaluation and finally, find out novel insights.

References

[1] Akyildiz, I., Su, W., Sankarasubramaniam, Y., e Cayirci, E. Wireless sensor networks: a survey. *Computer Networks 38*, 4 (2002), 393–422.

[2] Al-Ali, A.-R., Zualkernan, I., e Aloul, F. A mobile gprs-sensors array for air pollution monitoring. *IEEE Sensors Journal 10*, 10 (2010), 1666–1671.

[3] Aquino, A., Junior, O., Frery, A., Albuquerque, E., e Mini, R. Musa: multivariate sampling algorithmfor wireless sensor networks. *IEEE Transactions on Computers 63*, 4 (2012), 968–978.

[4] Aurenhammer, F. Voronoi diagrams: A survey of a fundamental data structure. *ACM Computing Surveys 23* (1991), 345–405.

[5] Companhia Ambiental do Estado de São Paulo, 2005.

[6] Devarakonda, S., Sevusu, P., Liu, H., Liu, R., Iftode, L., e Nath, B. Real-time air quality monitoring through mobile sensing in metropolitan areas. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (2013).

[7] Frery, A., Ramos, H., Alencar-Neto, J., Nakamura, E., e Loureiro, A. Data driven performance evaluation of wireless sensor networks. *Sensors 10*, 3 (2010), 2150–2168.

[8] Hejlová, V., e Vozenílek, V. Wireless sensor network components for air pollution monitoring in the urban environment: criteria and analysis for their selection. *Wireless Sensor Network 2013* (2013).

[9] Hu, S.-C., Wang, Y.-C., Huang, C.-Y., e Tseng, Y.-C. A vehicular wireless sensor network for co 2 monitoring. In *SENSORS, 2009 IEEE* (2009).

[10] Hu, S.-C., Wang, Y.-C., Huang, C.-Y., e Tseng, Y.-C. Measuring air quality in city areas by vehicular wireless sensor networks. *Journal of Systems and Software 84*, 11 (2011), 2005–2012.

[11] Kaivonen, S., e Ngai, E. Real-time air pollution monitoring with sensors on city bus. *Digital Communications and Networks 6*, 1 (2020), 23–30.

[12] Khedo, K., Perseedoss, R., Mungur, A., e others. A wireless sensor network air pollution monitoring system. *International Journal of Wireless Mobile Networks 2*, 2 (2010), 31–45.

[13] Lima, P., Câmara, G., e Queiroz, G. Geobr: Intercâmbio sintático e semântico de dados espaciais. Relatório técnico, Instituto Nacional de Pesquisas Espaciais, 2002.

[14] Ma, Y., Richards, M., Ghanem, M., Guo, Y., e Hassard, J. Air pollution monitoring and mining based on sensor grid in london. *Sensors 8*, 6 (2008), 3601–3623.

[15] Pavani, M., e Rao, T. Urban air pollution monitoring using wireless sensor networks: a comprehensive review. *International Journal of Communication Networks and Information Security 9*, 3 (2017), 439–449.

[16] Pebesma, E., e Heuvelink, G. Spatio-temporal interpolation using gstat. *RFID Journal 8*, 1 (2016), 204–218.

[17] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2014.

[18] Rashid, B., e Rehmani, M. H. Applications of wireless sensor networks for urban areas: A survey. *Journal of network and computer applications 60* (2016), 192–219.

[19] U.S. Environmental Protection Agency. Basic information about carbon monoxide (co) outdoor air pollution, 2016.

[20] U.S. Environmental Protection Agency. Basic information about no2, 2016.

[21] U.S. Environmental Protection Agency. How mobile source pollution affects your health, 2016.

[22] U.S. Environmental Protection Agency. Basic information about lead air pollution, 2017.

[23] U.S. Environmental Protection Agency. Ground-level ozone basics, 2018.

[24] U.S. Environmental Protection Agency. Particulate matter (pm) basics, 2018.

[25] U.S. Environmental Protection Agency. Technical assistance document for the reporting of daily air quality – the air quality index (aqi). Relatório técnico, U.S. Environmental Protection Agency, 2018.

[26] U.S. Environmental Protection Agency. Sulfur dioxide basics, 2019.

[27] Völgyesi, P., Nádas, A., Koutsoukos, X., e Lédeczi, Á. Air quality monitoring with sensormap. In *2008 International Conference on Information Processing in Sensor Networks (ipsn 2008)* (2008).

[28] Wang, Y.-C., e Chen, G.-W. Efficient data gathering and estimation for metropolitan air quality monitoring by using vehicular sensor networks. *IEEE Transactions on Vehicular Technology 66*, 8 (2017), 7234–7248.

[29] Yi, W., Lo, K., Mak, T., Leung, K. S., Leung, Y., e Meng, M. L. A survey of wireless sensor network based air pollution monitoring systems. *Sensors 15*, 12 (2015), 31392–31427.

[30] Yick, J., Mukherjee, B., e Ghosal, D. Wireless sensor network survey. *Computer Networks 52*, 12 (2008), 2292–2330.