Journal of Internet Services and Applications, 2023, 14:1, doi: 10.5753/jisa.2023.2378

Environmental Modeling and Traffic Simulation: A multivariate approach to monitor urban air pollutant agents.

Israel L. C.Vasconcelos [**Universidade Federal de Alagoas** | ***israel.vasconcelos@laccan.ufal.br***
André L L Aquino [**Universidade Federal de Alagoas** | ***alla@laccan.ufal.br***]

Campus A. C. Simões - Instituto de Computação, Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, Cep: 57072-970.

Received: 30 November 2021 • **Accepted:** 9 January 2023 • **Published:** 02 May 2023

Abstract This work presents an interdisciplinary assessment examining air quality tracking in urban environments. This application is well suited to be approached with wireless sensor networks' paradigm in their overall variations. The proposed modeling application takes advantage of Vehicle Sensor Networks (VSN) by embedding sensor nodes in public transportation, addressing this study case with bus lines so that the mobiles spread the sampling activity through many 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 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 from fossil-fueled vehicles (CO, O_3, PM_{10}, NO_2 and SO_2) simultaneously while it also varies in time. Furthermore, the experiment considers a detailed description of roads, bus lines, vehicle itineraries, and general traffic information. The input data that has unformatted or missing information due to being sourced from real sensors is handled to create the map mentioned above. Our methodology addresses the following: 1) the mentioned environmental simulation, 2) the deployment of mobile sensor nodes and performing sensing process, 3) the implementation of 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, our VSN-based approach improved 126 times lower error and 11 times higher coverage compared to conventional monitoring with air quality stations.

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

1 Introduction

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), Akyildiz *et al.* (2002). This study evaluates a solution considering different physical phenomena that wireless sensor networks observe. In this context, the challenge of monitoring urban areas regarding subjects such as air quality and meteorological conditions rises as notoriously relevant research opportunities, Rashid and Rehmani (2016); Yi *et al.* (2015).

We refer to the resulting output from various phenomena sensing processes as multivariate data. This result occurs because each monitored variable's samples are collected and stored simultaneously by different sensors in the same node. Generally, the air quality monitoring (sensors) is as close as possible to the main emitters. To reach a better coverage area, we proposed embedding sensor nodes in public transportation – Kaivonen and Ngai (2020) – such as bus lines and trains. Simultaneously, they visit many different places during the route. This mobility pattern simplifies the sampling, power management, number of packages, and other general maintenance issues. Additionally, vehicle sensors improve the identification of pollutants dispersion in the presence of wind or rain.

Despite the advantages previously mentioned, these conditions aggravate redundancy issues due to urban traffic dynamics, Yi *et al.* (2015) (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, as discussed in VSN (Vehicle Sensor Networks) known constraints. Some of the mentioned techniques comprise variable reporting rate: Devarakonda *et al.* (2013); Hu *et al.* (2011); Wang and Chen (2017), node clustering: Khedo *et al.* (2010); Ma *et al.* (2008), data fusion: Devarakonda *et al.* (2013); Hu *et al.* (2011); Khedo *et al.* (2010); Ma *et al.* (2008), and reconstruction of lost data: Wang and Chen (2017).

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 Spatiotemporal real data set 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 environments 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 5 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 Hu *et al*. (2009), Hu *et al*. (2011) and Wang and Chen (2017). 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 for execution on real scenarios, 3) to 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.

We can ensure a bottom line by developing a simulation environment that looks at urban pollution agents from a multivariate perspective. The main achievement of this study is that all referred researches work with univariate data, whereas we **propose to expand the evaluation to a multivariate domain**. We focus on observing the behavior of each monitored variable individually and the correlation among them. We consider real data input collected from air quality stations and assess the effort to handle a VSN application with this complex data type. We evaluate our model considering the absolute value of relative error and global field coverage metrics.

2 Urban Air Pollution Review

2.1 Health impacts and human effects

The addressed research in this work touches on an interdisciplinary subject since the case study's object relates to pollutant gases' harmful effects. First, 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 – U.S. Environmental Protection Agency (2018c) – 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, U.S. Environmental Protection Agency (2016a).

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 increased susceptibility to respiratory infections. In addition, conditions such 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, U.S. Environmental Protection Agency (2016b).

Chemical reactions between oxides of nitrogen (NO_x) and volatile organic compounds (VOC) generate 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, U.S. Environmental Protection Agency (2018a).

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, U.S. Environmental Protection Agency (2019).

Particulate matter (particle pollution, or PM) is 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 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, U.S. Environmental Protection Agency (2018b).

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, immune, reproductive, and developmental systems. Lead exposure is also very likely to affect the blood capacity for 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. U.S. Environmental Protection Agency (2017).

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 U.S. Environmental Protection Agency (2016c).

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 Al-Ali *et al.* (2010); Devarakonda *et al.* (2013); Kaivonen and Ngai (2020); Völgyesi *et al.* (2008); Wang and Chen (2017); Yi *et al.* (2015) and consists of a six-level scale containing reference values to pollutants concentration and risk descriptors 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. The AQI is the maximum value of I_p among all pollutants in a single packet of samples for a specific location.

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}} \times (C_p - BP_{Lo}) + I_{Lo} \qquad (1)$$

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};

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 handling multivariate data assessed in this research work.

Völgyesi *et al.* (2008) 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 (in 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.* (2008), the authors propose a sophisticated air pollution monitoring system that considers a wide range of aspects regarding urban areas, generating a broad set of information about the city. The experiment consists of a full network simulation alongside report guidelines to implement in the real environment with car-mounted and static sensors. They perform an entire simulation while guiding to apply 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. Similar and recent works are Puiu *et al.* (2022); Angelevska *et al.* (2021).

Hu *et al.* (2009) 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 in a real environment observing the CO_2 concentration. Sensor nodes have GPS and Cellular connections. 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 Hu *et al*. (2009), where the subject matter shows a detailed problem characterization, is presented at Hu *et al*. (2011). A network simulation evaluates applications' behavior by looking at estimated error and message traffic metrics. In our experimentation, we validate it through a complete network simulation that keeps the reports of implementation guidelines in realistic scenarios. Furthermore, 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**. Similar and recent works are Shakhov and Sokolova (2021); Shakhov *et al*. (2022).

The work addressed in Al-Ali *et al*. (2010) 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's top layer features the evaluation of the Air Quality Index and the interface with map service.

Khedo *et al*. (2010) 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 a simulation-only application of a traditional, static, and multi-hop-based WSN. In this paper, there's no explicit mention of air sensors employed. Instead, 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) describes the mobile schema, the cloud server application, and the 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 public transportation-based VSN and a Community Sensor Network (CSN) through a collaborative approach. Carbon Monoxide and Particulate Matter are the observed pollution indicators. Data aggregation is performed at the top-level base station to improve accuracy regarding the data processing, variable sampling concerning the vehicle, and spatial gradient speed.

Having a solid intersection with the subject addressed in this article, the survey from Yi *et al*. (2015) centralizes many relevant characteristics regarding the specific air pollution problem monitored through WSN technologies. They explain networks' differences, advantages, and drawbacks 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. Similar works are Pavani and Rao (2017); Yi *et al*. (2015).

Wang and Chen (2017) propose a probabilistic strategy to handle adaptive sampling of cars and balance 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 in methodology and metrics previously stated in Hu *et al*. (2011). Similar and recent works are Kumar *et al*. (2022); Barthwal and Acharya (2022)

Finally, Kaivonen and Ngai (2020) reports an experimental study with physical sensors attached to public transportation and describes this hardware prototype's development in detail. It discusses typical real-life challenges such as noise on measurements, numeric sensor precision, the efficiency of cellular connection and packet loss rate, and the 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. Similar and recent works are Shakhov *et al*. (2022); Qin *et al*. (2022).

Table 1 shows a side-by-side comparison (ordered by publication year) with the main aspects considered in related work. Looking at sensors and methods, abbreviations respectively for "Compression", "Aggregation", "Reporting", and "Multivariate Data Handling". We highlight that our proposal **handles multiple simultaneous phenomena (MDH, Multivariate Data Handling)** at the 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 overall aspects, it is relevant to highlight that **a combination of real data inputs and simulated environment for experimentation** is only noticed in our proposal. Another observation is that there's no explored opportunity to evaluate sensor-level sampling algorithms (also multivariate) in this kind of scenario.

4 Environmental Application Design

Let the overall behavior be denoted by

$$\begin{tikzcd}
\mathcal{N} \arrow[r, "P"] & \mathbf{V}^* \arrow[r, "S"] \arrow[d, "R"] & V \arrow[r, "\Psi"] & V' \arrow[r, "\omega"] & V'' \arrow[d, "R"] \\
& D & & & D'
\end{tikzcd} \tag{2}$$

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)Aquino *et al*. (2012); Vasconcelos *et al*. (2018).

Table 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
Völgyesi *et al.* (2008), (2008)	X	X	X		X			X		
Ma *et al.* (2008), (2008)			X		X	X		X		
Hu *et al.* (2009), (2009)		X							X	
Al-Ali *et al.* (2010), (2010)	X	X			X	X		X	X	
Khedo *et al.* (2010), (2010)	X									
Hu *et al.* (2011), (2011)		X					X		X	
Devarakonda *et al.* (2013), (2013)	X	X		X				X	X	
Wang and Chen (2017), (2017)	X	X							X	
Kaivonen and Ngai (2020), (2020)	X	X			X					

Overall aspects

Article	Experimentation		Input Source		Network Topology			
(per year)	Real Env.	Simulation	Real Data	Simulation	Static	Clustered	V2I	V2V
proposal		X	X				X	
Völgyesi *et al.* (2008), (2008)	X		X			X	X	
Ma *et al.* (2008), (2008)		X		X		X	X	X
Hu *et al.* (2009), (2009)	X		X				X	
Al-Ali *et al.* (2010), (2010)	X		X				X	
Khedo *et al.* (2010), (2010)		X		X	X	X		
Hu *et al.* (2011), (2011)		X		X			X	X
Devarakonda *et al.* (2013), (2013)	X		X				X	
Wang and Chen (2017), (2017)		X		X			X	X
Kaivonen and Ngai (2020), (2020)	X		X				X	

Replicate this overall behavior for every phenomenon $P_i \mid i = \{1, \ldots, n\}$, where n is the number of the different phenomena 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 the 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 dataset $(\mathbf{V}')$ obtained from a data sampling or fusion strategy $(\mathbf{\Psi})$ over its entire observed domain $(\mathbf{V})$. Then, 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'.

4.1 Area of Interest

The process described in Diagram 3 is the ”zero” step in 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 5.1 and 5.2**.

$$\mathcal{N} \xrightarrow{P} \qquad (3)$$

Initially, based on the raw data set, we generate the model for the area under experimentation and apply the library GeoBR, Lima *et al.* (2002) as support to import the requested maps. For this case study, we defined **the city of São Paulo as an area of interest**. Afterward, we perform general-purpose matrix/table handling features, available at R Platform – R Core Team (2014), implemented as scripted sequential procedures. The adopted data set is from Environmental Sanitation Technology Company[1] 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

[1] `https://cetesb.sp.gov.br/`

- P. D. Pedro II
- Pinheiros
- São Miguel Paulista
- Santana
- Santo Amaro

We set the arrangement for station coordinates following the real locations of the districts in which they are named. Figure 1 illustrates the map outputs containing these points. It also proposed a detailed description of roads, bus lines, vehicle itineraries, and general traffic information to achieve proper modeling at the network and application simulation stage.

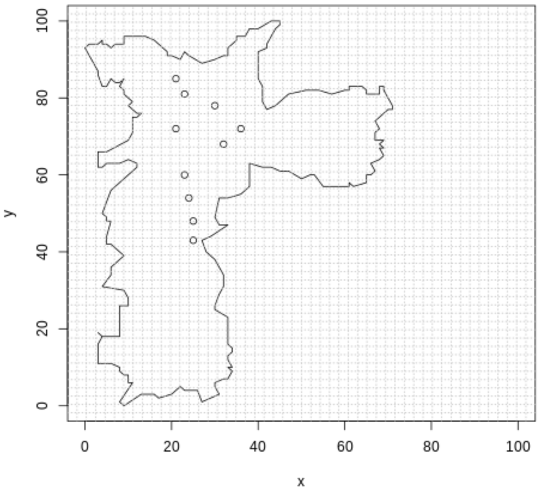


Figure 1. Exported map of São Paulo with air quality stations on real approximate locations.

4.2 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 exact location of stations. The first stage of the model construction is shown in the sub-process on Diagram 4

$$\begin{tikzcd} \mathcal{N} \arrow[r, "P"] & \mathbf{V}^* \arrow[d, "R"] \\ & D \end{tikzcd} \tag{4}$$

To achieve an adequate representation of the phenomenon under ideal sensing conditions ($\mathbf{V}^*$), we will perform a multivariate reconstruction: multivariate ordinary cokriging, Pebesma and Heuvelink (2016), 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 5.2**.

4.3 Network Settings

The network sub-process and sensing activity is covered at Diagram 5, 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.

$$\mathcal{N} \xrightarrow{P} \mathbf{V}^* \xrightarrow{S} V \tag{5}$$

Addressing the research background under urban zones, we can highlight alternative monitoring solutions approaching vehicle sensor networks: Yi *et al.* (2015); Hu *et al.* (2011); Devarakonda *et al.* (2013). The presented case study in this article takes advantage of bus lines as mobile sensing units, Kaivonen and Ngai (2020), that ride through the city while collecting the data.

The planned network setup considers a vehicle sensor network with a V2I communication, using a cellular network to upload the sensed data to a cloud server. This approach is recurrently observed in referenced work along with this article.

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

4.4 Sensing Data Processing

Following statements from Section 4, the sub-process of overall behavior (Diagram 2) that refers to the stage of data collection and processing is represented at Diagram 6 below:

$$V \xrightarrow{\Psi} V' \xrightarrow{\omega} V'' \tag{6}$$

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. $\mathbf{V}''$ is the rebuilt field after input $\mathbf{V}'$ to a technique which fills blank spaces ω.

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

4.5 Evaluation

Let the data reconstruction and evaluation rule set be denoted by

$$\begin{tikzcd} \mathcal{N} \arrow[r, "P"] & \mathbf{V}^* \arrow[r] \arrow[d, "R"] & \ldots \arrow[r] & V'' \arrow[d, "R"] \\ & D & & D' \end{tikzcd} \tag{7}$$

Where "..." represents the whole sub-process sequence from Diagram 6 (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 ($\hat{\epsilon}$) Frery *et al.* (2010), 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 belong to the internal area of Figure 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 input data that generates $\mathbf{V}^*$ was pre-processed to handle all NA measurements, it can always ensure the definition of $\hat{\epsilon}$ since $\mathbf{V}^*(i,j) \neq 0$.

We set the unavailable locations (blank spaces from non-visited areas) using Voronoi diagrams Aurenhammer (1991). The Voronoi diagram allows a fair comparison, considering the coverage estimation based on fixed sensors (Voronoi among monitoring stations location) and the mobile sensors (Voronoi among vehicles trajectory location). For example, without the Voronoi, the environment variables values used in the comparison will be only the monitoring station's location. The other ones will be zero.

Let the location of sensors (S) as a set of n points in an area. Then, the Voronoi diagram is the dominance of S_p over S_q is the subset (or sub-area) of the plane 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 composes $\mathbf{V}''$ for each monitoring approach.

4.6 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 5 details this procedure.

The second step is to set up the traffic simulation. At this stage, a set of sequential tasks 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, buses, and their respective routes are generated and executed with different traffic intensity levels. This step's primary outcome is to export the bus traces with visited map coordinates during the route. This procedure is detailed in Section 5.3.

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. This way, it is possible to evaluate the field coverage by looking at how many coordinates the solution covered over time. Section 5.4.1 details this procedure.

5 Multivariate Pollutant Map Generation

The pollutant input data is provided as a set of files containing real samples from air quality stations placed in São Paulo from Jan-01-2005 to Dec-31-2005. Those dates were considered according to their availability at the moment where the execution of the simulated environment happened, working as a validated proof of concept and expansible to apply more recent data in future works.

There are natural limitations to using vehicular sensors for air quality monitoring because it is not viable to use the same number of sensors used in a fixed environmental station in a vehicle. However, other applications use satisfactorily mobile sensors, from airplanes and boats to smartphones Devarakonda *et al.* (2013); Kaivonen and Ngai (2020). In our application, there are about 15 different variables reporting information, such as wind speed and direction, atmospheric stability, temperature, humidity, and other classes of pollutants are available. Thus, we reduce these variables from 15 to 5, considering the pollutant agents from burning fossil-fueled vehicles. 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: grouping all stations which have this sensor available on this list, where the columns are for miscellaneous information and measurements (resulting in five records with a similar structure as illustrated in table 2). On the other hand, rows repeat the date and time for samples on each listed station resulting in a noticeable redundancy amount.

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 5). Alongside the overall rearrangements, we also handle the format issues observed in raw data (i.e., comma instead of the dot at measurements representation and miscellaneous date format).

The date length was reduced to 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.

5.1 Dataset Handling

Finally, we provide an outline of the map from São Paulo at package **GeoBR**, Lima *et al.* (2002) that is, the input points to the prediction process alongside each air quality station's coordinates, placed at this step (as seen in Figure 1).

5.2 Prediction

Considering the data structure, we perform the prediction step in two directions, explained in the subsections below.

Table 2. Raw data.

Station ID	Date	Time (h)	Sample	Station Name
		Sensor (i.e.: O_3)		
1	01/01/2005	01:00	39,22	P. D. Pedro II
1	01/01/2005	02:00	23,85	P. D. Pedro II
1	01/01/2005	03:00	29,68	P. D. Pedro II
		...		
2	01/01/2005	01:00	30,85	Santana
		...		
47	31/12/2005	00:00	7,44	Horto Florestal

Table 3. Selecting gray to handle NA.

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	...	12.08	...	18.69
...	...	...	...	...	...	...	...
2005-10-19	02:00	0.79	31.39	...	NA	...	12.46
...	...	...	...	...	...	...	...
2005-10-21	23:00	0.62	4.32	...	13.94	...	16.19

Besides that, Figure 2 shows the ratio between the real data initially available (raw data) and samples predicted using the raw data as input with methods presented in subsections 5.2.1 and 5.2.2.

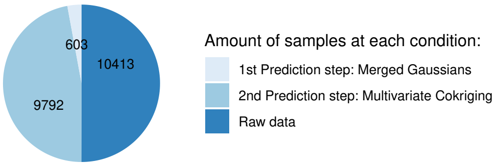


Figure 2. Amount of real and predicted samples.

5.2.1 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 described as follows.

1. Assert each single column (example at Table 3) as a matrix (example at Table 4) with **hours** (0h–23h) × **days** (15–22);
2. Test normality (Shapiro-Wilk) for everyone, evaluate mean μ and standard deviation σ by two times, for the 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 Table 5).

5.2.2 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 all points should be available on each station's coordinates.

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

	Station 4-PM10					
	15-Oct	16-Oct	...	19-Oct	...	21-Oct
01h	12.08	48.39	...	NA	...	14.06
02h	8.53	36.77	...	NA	...	9.93
03h	28.64	46.46	...	NA	...	16.22
...	...	...	...	...	...	...
23h	61.69	33.12	...	11.66	...	6.56
00h	60.20	16.20	...	11.95	...	13.94

Table 5. Summarized data after handling.

Date	Time	Station-Sensor$_{(1,\ldots,n)}$				
	(h)	1-CO	...	4-PM10	...	47-O_3
15-Oct	01:00	0.71	...	12.08	...	18.69
15-Oct	02:00	0.67	...	8.53	...	10.56
15-Oct	03:00	0.79	...	28.64	...	12.46
...	...	...	...	...	...	...
21-Oct	23:00	0.59	...	6.56	...	17.03
21-Oct	00:00	0.62	...	13.94	...	16.19

Table 6 describes the arrangement of sensor availability, where variables colored on green shades are missing. At the first turn, we interpolate the entire map based on CO, PM10, and O_3 at available stations $\{1, 3, 5, 16, 27\}$ and assign to missing stations the predicted data at respective coordinates. After that, repeat the similar process to NO_2 and SO_2 accumulating the new predicted samples at the previous turn sequentially.

With missing points fixed, we parse the data as input to multivariate ordinary cokriging (supported by R package Gstat, Pebesma and Heuvelink (2016)) to interpolate the entire map area, this assembled field represents $\mathbf{V}^*$ from Section 4. Finally, we use the same technique to fix missing stations (Table 6).

Finally, the achieved outcome is a set of five tables (one for each pollutant). Each table is assembled by placing columns with a list of valid coordinates from the map area. Each row in this table represents the entire map area in a particular timestamp.

In other words, each row in this table represents the entire map area in a particular timestamp, where x, y (two-

Table 6. Prediction of missing data.

Station	Sensors				
(ID)	**1st Turn Prediction**			**2nd**	**3rd**
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**

Table 7. 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

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 7 generates the visualization illustrated in the map of Figure 3.

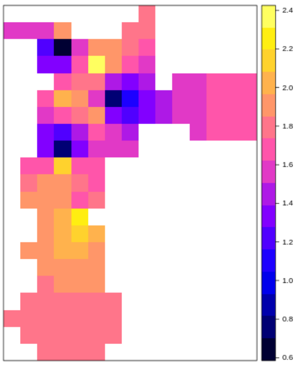


Figure 3. Multivariate pollution map of Carbon Monoxide (CO) at a certain timestamp.

5.3 Traffic Simulation

The pollution map described in the previous Section (Table 7) 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. To run this experimentation, we adopt the Simulator of Urban MObility – SUMO, Lopez *et al.* (2018). At this stage, we assess three sub-steps in the following subsections.

5.3.1 Fetch and build maps and roads

The starting point for building the simulation structure is to define a map tool compatible with SUMO. For this task, we use the Open Street Maps API[2], 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

Table 8. Raw trace file of bus routes generated at traffic simulation.

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

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 for 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 allow the parallelization of experiments, hence taking advantage of CPU multithreading.

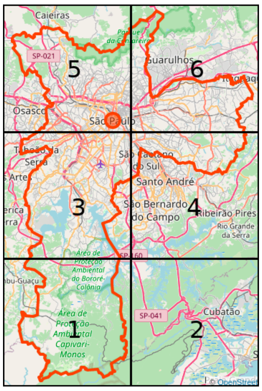


Figure 4. Illustration of map bounding boxes.

Furthermore, there is an intrinsic tricky detail in 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 and y=0. The offset matrix for coordinate fixing is illustrated in Table 8.

5.3.2 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, such as subways, city rails, and trolley cars, but we do not consider them so far.

The vehicle generation comprises bus stops and prior defined lines for public transportation. This information is available in maps built at the subsection 5.3.1. During the simulation, the buses 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. As a result, these vehicles disappear from the simulation after reaching the end of their routes.

It is relevant to highlight that the vehicle generation (Figure 5) 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 sub-

[2] `https://www.openstreetmap.org/`

section 5.4. 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.

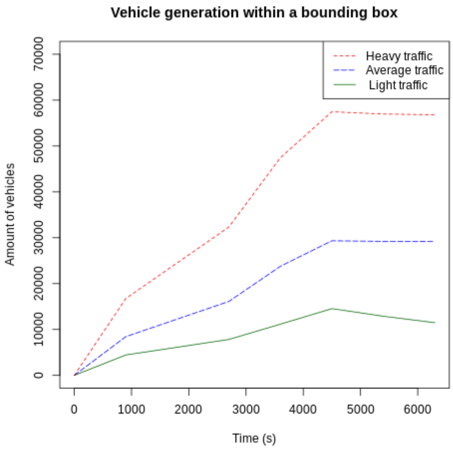


Figure 5. Vehicle generation at SUMO Simulator under different traffic conditions.

5.4 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 to 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.

5.4.1 Environment Assembly

The last stage of experimentation consists of putting together the output from two previous ones (multivariate pollution maps from Section 5 and bus routes trace from Section 5.3). 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 6.

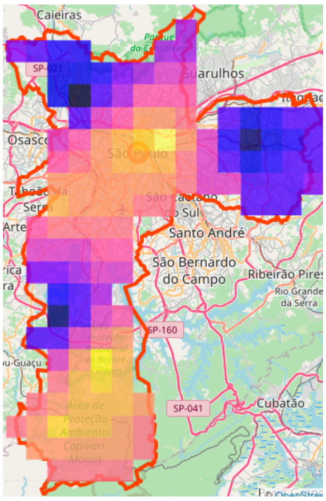


Figure 6. Matching the city area with scaled 25×25 pollutant map.

Initially, the raw trace is exported as a .xml file and looks like the Table 9. 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 it provides information about how the vehicles behave in different traffic conditions. Table 12 illustrates this structure.

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 9. 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)

6 Results

6.1 Preliminary Assumptions

Initially, we perform all the required data handling on traces and phenomena information. The following actions aim to achieve a performance assessment of overall field coverage

Table 10. 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)

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 11 shows the considered parameter set.

Table 11. 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)

6.2 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.

Table 12 shows the trace from bus lines in a 1-hour window. There is one instance of this table for each traffic intensity (see Table 11). It is assigned to the respective hour of the day (Table 13). After that, we shape all visited coordinates as a plain list (eliminating repeated ones) and count how many are covered to 132 squared units of the entire map, generating the percentages seen in Table 14. Global Coverage is achieved under this weighted sum $Light_{(s)} \times \frac{6}{24} + Average_{(s)} \times \frac{12}{24} + Heavy_{(s)} \times \frac{6}{24}$ (where s is the day/seed).

Table 12. Trace data from bus lines (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)
2700s	(7,8)	(3,9)	(5,7)	(7,7)	...	(4,6)	(6,8)
3600s	(7,8)	(3,9)	(5,8)	(7,7)	...	(4,7)	(6,8)

Considering the presented strategy, after a 7-day 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 11) covered by the map area and the available amount of 10 stations. **The regular monitoring system achieves a theoretical global field coverage of 7,5%**.

Table 13. Traffic intensity day times.

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	

Table 14. VSN summarized coordinates list to evaluate Global Coverage.

Day	Traffic Intensity ($\times$ Coords List)			Global Coverage	Global Coverage
(seed)	Light	Average	Heavy	(VSN)	(Stations)
1	87.12%	86.36%	87.12%	86.74%	7.5%
2	86.36%	86.36%	86.36%	86.36%	
3	86.36%	86.36%	86.36%	86.36%	
4	86.36%	86.36%	87.12%	86.55%	
5	87.12%	86.36%	87.12%	86.74%	
6	86.36%	86.36%	86.36%	86.36%	
7	87.12%	86.36%	87.12%	86.74%	

6.3 Summary for Absolute Value of Relative Error

In the current section, we assess the sampled data representativeness by evaluating the Absolute Value of Relative Error (detailed in Section 4).

Figures 7, 9, 13, 15, 11 show the behavior of error metric for each pollutant. Taking as the example at the bottom of Figure 8(c), we see an information loss at the continuous dark red area in comparison with 8(a) and 8(b). This behavior occurs due to the lack of spatial reachability in stationary sensing, which was mitigated by the VSN approach. We can also notice the same behavior at similar ones. Figure 8 illustrates side-by-side the performance for VSN and stationary sensing with reference data ($\mathbf{V}^*$), taking CO as an example. The figure shows lighter colors for higher measurements and darker for lower measurements from samples in a normalized scale from zero to one.

6.3.1 Summary for Carbon Monoxide

Looking at the Carbon Monoxide pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 7) is

- 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 8 shows the resulting behavior of each monitoring strategy.

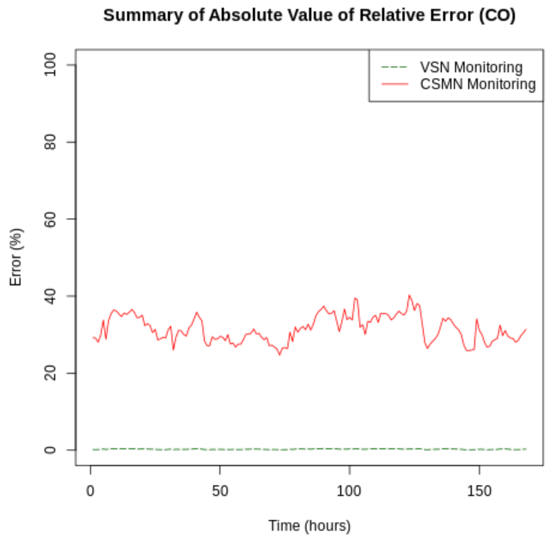
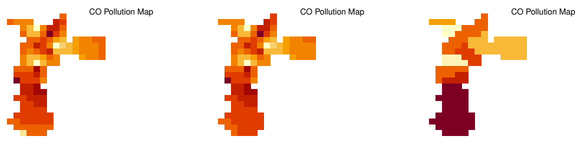


Figure 7. Error rate evaluated through 7 days from VSN and Conventional Stations (CO).



(a) Reference. (b) VSN. (c) Stations.

Figure 8. Side-by-side performance comparison between sampling with VSN and Conventional Stations for Carbon Monoxide.

6.3.2 Summary for Particulate Matter

Looking at the Particulate Matter pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 9) is

- 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 10 shows the resulting behavior of each monitoring strategy.

6.3.3 Summary for Nitrogen Dioxide

Looking at the Nitrogen Dioxide pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 11) is

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

This result means an improvement in 70 times lower error about the regular monitoring system. Figure 12 shows the resulting behavior of each monitoring strategy.

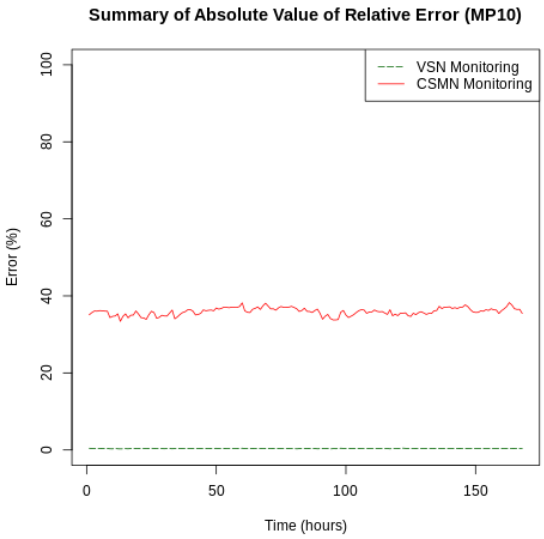
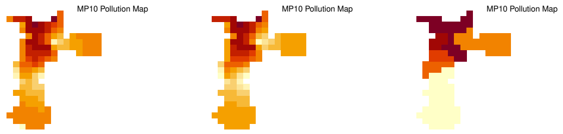


Figure 9. Error rate evaluated through 7 days from VSN and Conventional Stations (PM10).



(a) Reference. (b) VSN. (c) Stations.

Figure 10. Comparison between sampling with VSN and Conventional Stations for PM10.

6.3.4 Summary for Ground-level Ozone

Looking at the Nitrogen Dioxide pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 13) is

- 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 14 shows the resulting behavior of each monitoring strategy.

6.3.5 Summary for Sulfur Dioxide

Looking at the Sulfur Dioxide pollution map, the average Absolute Value of Relative Error along the entire time series (Figure 15) is

- 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 16 shows the resulting behavior of each monitoring strategy.

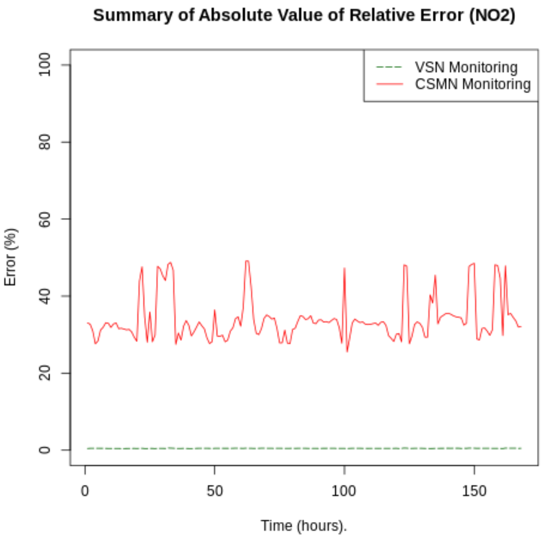
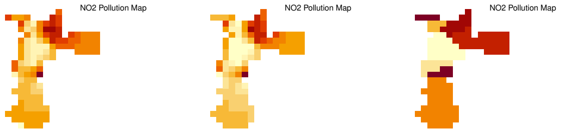


Figure 11. Error rate evaluated through 7 days from VSN and Conventional Stations (NO_2).



(a) Reference. (b) VSN. (c) Stations.

Figure 12. Comparison between sampling with VSN and Conventional Stations for NO2.

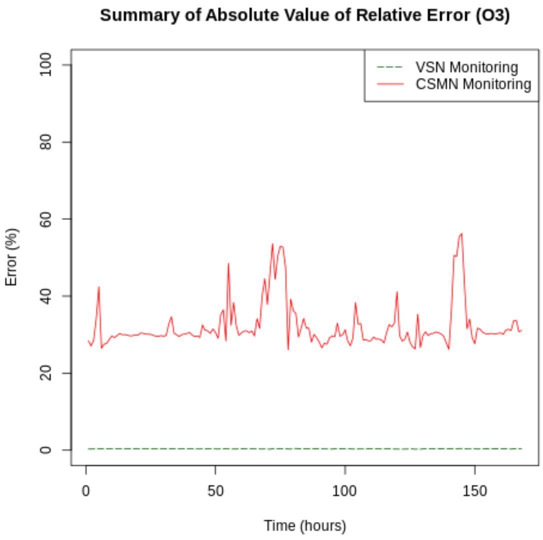
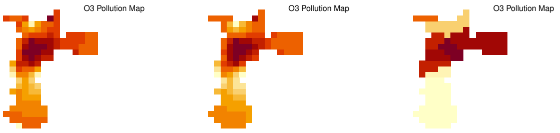


Figure 13. Error rate evaluated through 7 days from VSN and Conventional Stations (O_3).



(a) Reference. (b) VSN. (c) Stations.

Figure 14. Comparison between sampling with VSN and Conventional Stations for O_3.

6.3.6 Discussion

In summary, we can see an improvement in the order of 126, 99.8, 100, 70, and 6.59 times lower error, respectively, for CO, PM10, O_3, NO_2 and SO_2 concerning the regular monitoring system. Furthermore, we can observe that the error difference between the two approaches decreases according to data availability displayed in Table 6. Note that the SO_2 sensor is available at only two stations (IDs 5 and 8), and this limited number of input samples causes an artificial homogeneity on predicted data for this variable, which pulls the AVRE measurements closer with other sensors.

Even with this limitation of data availability that disturbs the prediction of variables with few inputs, our proposal delivers a noticeable improvement (659% at the worst case from SO_2) on overall application behavior comparison to conventional strategy. Moreover, since only two sensors for an entire city is quite an extreme restriction, any data set with more sensors available is enough to mitigate this limitation, showing that the proposed model is robust even with extremely constrained input data.

7 Conclusion and Final Remarks

The presented article has explored the problem of air quality monitoring while taking a more in-depth investigation into the modeling of complex environments. Beyond that, it delivers contributions that improve the view of how correlated multivariate phenomena behave.

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

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. Overall observed performance indicates that the proposed application in this case study is suitable for real-world scenarios.

As future directions, we consider evaluating different classes of data processing algorithms and improving environmental modeling with variables not considered at the last turn, such as wind speed/direction, temperature, and humidity data on evaluation.

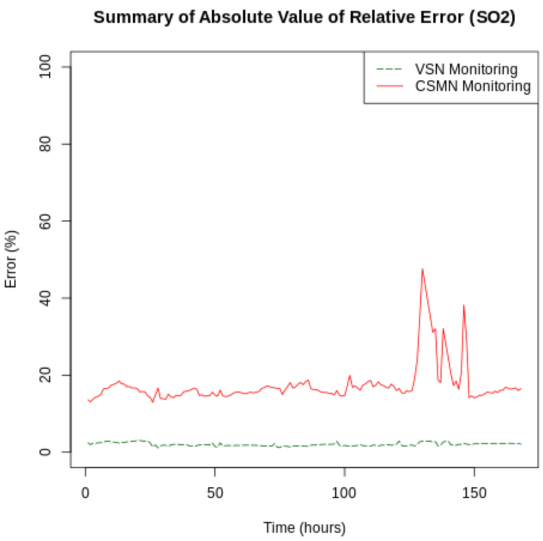
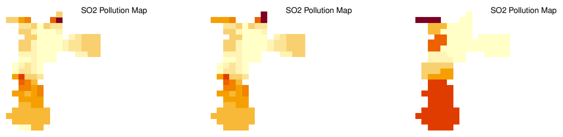


Figure 15. Error rate evaluated through 7 days from VSN and Conventional Stations (SO_2).



(a) Reference. (b) VSN. (c) Stations.

Figure 16. Comparison between sampling with VSN and Conventional Stations for SO_2.

Acknowledgements

Declarations

Authors' Contributions

All authors contributed to the writing of this article, read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Data can be made available upon request.

References

Akyildiz, I., Su, W., Sankarasubramaniam, Y., and Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer Networks*, 38(4):393–422. DOI: 10.1016/S1389-1286(01)00302-4.

Al-Ali, A.-R., Zualkernan, I., and Aloul, F. (2010). A mobile GPRS-sensors array for air pollution monitoring. *IEEE Sensors Journal*, 10(10):1666–1671. DOI: 10.1109/JSEN.2010.2045890.

Angelevska, B., Atanasova, V., and Andreevski, I. (2021). Urban air quality guidance based on measures categorization in road transport. *Civil Engineering Journal-Tehran*, 7(2):253–267. DOI: 10.28991/cej-2021-03091651.

Aquino, A., Junior, O., Frery, A., Albuquerque, E., and Mini, R. (2012). Musa: multivariate sampling algorithmfor wireless sensor networks. *IEEE Transactions on Computers*, 63(4):968–978. DOI: 10.1109/TC.2012.229.

Aurenhammer, F. (1991). Voronoi diagrams: A survey of a fundamental data structure. *ACM Computing Surveys*, 23:345–405. DOI: 10.1145/116873.116880.

Barthwal, A. and Acharya, D. (2022). Performance analysis of sensing-based extreme value models for urban air pollution peaks. *Modeling Earth Systems And Environment*, 8(3):4149–4163. DOI: 10.1007/s40808-022-01349-y.

Devarakonda, S., Sevusu, P., Liu, H., Liu, R., Iftode, L., and Nath, B. (2013). Real-time air quality monitoring through mobile sensing in metropolitan areas. In *ACM International Workshop on Urban Computing*. DOI: 10.1145/2505821.2505834.

Frery, A., Ramos, H., Alencar-Neto, J., Nakamura, E., and Loureiro, A. (2010). Data driven performance evaluation of wireless sensor networks. *Sensors*, 10(3):2150–2168. DOI: 10.3390/s100302150.

Hu, S.-C., Wang, Y.-C., Huang, C.-Y., and Tseng, Y.-C. (2009). A vehicular wireless sensor network for co^2 monitoring. In *IEEE Sensors*. DOI: 10.1109/ICSENS.2009.5398461.

Hu, S.-C., Wang, Y.-C., Huang, C.-Y., and Tseng, Y.-C. (2011). Measuring air quality in city areas by vehicular wireless sensor networks. *Journal of Systems and Software*, 84(11):2005–2012. DOI: 10.1016/j.jss.2011.06.043.

Kaivonen, S. and Ngai, E. (2020). Real-time air pollution monitoring with sensors on city bus. *Digital Communications and Networks*, 6(1):23–30. DOI: 10.1016/j.dcan.2019.03.003.

Khedo, K., Perseedoss, R., Mungur, A., *et al.* (2010). A wireless sensor network air pollution monitoring system. *International Journal of Wireless Mobile Networks*, 2(2):31–45. DOI: 10.48550/arXiv.1005.1737.

Kumar, G. J. R., Agbulu, G. P., Rahul, V, T., Natarajan, V, A., and Gokul, K. (2022). A cloud-assisted mesh sensor network solution for public zone air pollution real-time data acquisition. *Journal Of Ambient Intelligence And Humanized Computing*. DOI: 10.1007/s12652-022-03704-4.

Lima, P., Câmara, G., and Queiroz, G. (2002). Geobr: Syntactic and semantic interchange of spatial data (in portuguese). Available at: `http://www.dpi.inpe.br/gilberto/papers/geobr_geoinfo2002.pdf`.

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., and Wießner, E. (2018). Microscopic traffic simulation using sumo. In *21st IEEE International Conference on Intelligent Transportation Systems*, pages 2575–2582. IEEE. DOI: 10.1109/ITSC.2018.8569938.

Ma, Y., Richards, M., Ghanem, M., Guo, Y., and Hassard, J. (2008). Air pollution monitoring and mining based on sensor grid in london. *Sensors*, 8(6):3601–3623. DOI: 10.3390/s80603601.

Pavani, M. and Rao, T. (2017). Urban air pollution monitoring using wireless sensor networks: a comprehensive review. *International Journal of Communication Networks and Information Security*, 9(3):439–449. DOI: 10.17762/ijcnis.v9i3.2708.

Pebesma, E. and Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *RFID Journal*, 8(1):204–218. DOI: 10.32614/RJ-2016-014.

Puiu, S., Udristioiu, M. T., and Velea, L. (2022). Air pollution management: A multivariate analysis of citizens' perspectives and their willingness to use greener forms of transportation. *International Journal Of Environmental Research And Public Health*, 19(21). DOI: 10.3390/ijerph192114613.

Qin, X., Do, T. H., Hofman, J., Bonet, E. R., La Manna, V. P., Deligiannis, N., and Philips, W. (2022). Fine-grained urban air quality mapping from sparse mobile air pollution measurements and dense traffic density. *Remote Sesing*, 14(11). DOI: 10.3390/rs14112613.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available at:http://www.R-project.org/.

Rashid, B. and Rehmani, M. H. (2016). Applications of wireless sensor networks for urban areas: A survey. *Journal of network and computer applications*, 60:192–219. DOI: 10.1016/j.jnca.2015.09.008.

Shakhov, V., Materukhin, A., Sokolova, O., and Koo, I. (2022). Optimizing urban air pollution detection systems. *Sensors*, 22(13). DOI: 10.3390/s22134767.

Shakhov, V. and Sokolova, O. (2021). On modeling air pollution detection with internet of vehicles. In *15th International Conference on Ubiquitous Information Management and Communication*. DOI: 10.1109/IMCOM51814.2021.9377350.

U.S. Environmental Protection Agency (2016a). Basic information about carbon monoxide (co) outdoor air pollution. Available at:https://www.epa.gov/co-pollution/basic-information-about-carbon-monoxide-co-outdoor-air-pollution#Effects.

U.S. Environmental Protection Agency (2016b). Basic information about no2. Available at:https://www.epa.gov/no2-pollution/basic-information-about-no2#Effects.

U.S. Environmental Protection Agency (2016c). How mobile source pollution affects your health. Available at: https://www.epa.gov/mobile-source-pollution/how-mobile-source-pollution-affects-your-health.

U.S. Environmental Protection Agency (2017). Basic information about lead air pollution. Available at:https://www.epa.gov/lead-air-pollution/basic-information-about-lead-air-pollution#health.

U.S. Environmental Protection Agency (2018a). Ground-level ozone basics. Available at:https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#effects.

U.S. Environmental Protection Agency (2018b). Particulate matter (pm) basics. Available at: https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#effects.

U.S. Environmental Protection Agency (2018c). Technical assistance document for the reporting of daily air quality – the air quality index (aqi). Available at:https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf.

U.S. Environmental Protection Agency (2019). Sulfur dioxide basics. Available at:https://www.epa.gov/so2-pollution/sulfur-dioxide-basics#effects.

Vasconcelos, I., Martins, I., Figueiredo, C., and Aquino, A. (2018). A data sample algorithm applied to wireless sensor network with disruptive connections. *Computer Networks*, 146:1–11.

Völgyesi, P., Nádas, A., Koutsoukos, X., and Lédeczi, Á. (2008). Air quality monitoring with sensormap. In *2008 International Conference on Information Processing in Sensor Networks*. DOI: 10.1109/IPSN.2008.50.

Wang, Y.-C. and Chen, G.-W. (2017). Efficient data gathering and estimation for metropolitan air quality monitoring by using vehicular sensor networks. *IEEE Transactions on Vehicular Technology*, 66(8):7234–7248. DOI: 10.1109/TVT.2017.2655084.

Yi, W., Lo, K., Mak, T., Leung, K. S., Leung, Y., and Meng, M. L. (2015). A survey of wireless sensor network based air pollution monitoring systems. *Sensors*, 15(12):31392–31427. DOI: 10.3390/s151229859.