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

Israel L. C. Vasconcelos[1] **and Andre L. L. Aquino**[1]

[1]Computer Institute – Federal University of Alagoas – Maceió, AL – Brazil

{israel.vasconcelos, alla}@laccan.ufal.br

Abstract. *This work presents an interdisciplinary assessment that looks in-depth at the tracking of air quality in urban environments. The proposed application takes advantage of Vehicle Sensor Networks (VSN) by embedding sensor nodes to public transportation, spreading the sampling activity through different places visited during the route. We perform environmental modeling based on real data collected from the city of São Paulo, considering the multivariate spatial 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. Finally, our VSN-based approach showed an improvement of 126 times lower error and 11 times higher coverage about conventional monitoring with air quality stations.*

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 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 [Rashid and Rehmani 2016, Yi et al. 2015].

We refer to the resulting output from various phenomena sensing processes 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 [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.

Despite the advantages previously mentioned, these conditions aggravate redundancy issues due to the dynamics of urban traffic [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 such as discussed in VSN 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 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 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 4 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 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.

We can assure a bottom line to develop 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. Related Work

WSN based air quality monitoring solutions are an already mature subject-matter in literature. In the following discussion, we present the papers with the closest relation to our research scope and address the open points regarding the handling of multivariate data assessed at this work.

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]. This case study takes advantage of bus lines as mobile sensing units, as seen in [Kaivonen and Ngai 2020], that ride through the city while collects the data. [Hu et al. 2009] and [Hu et al. 2011] present a standard VSN application implemented to perform micro-climate monitoring through CO_2 concentrations. This application integrates a map service to show the collected data from car-mounted sensors (GPS/Cellular based) running in a real environment. Adaptive reporting with data aggregation and V2V communication strategies to bandwidth management are considered as well. [Wang and Chen 2017] 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. This simulation comprises mobility and pollutant dispersion models, while advances on the methodology stated in [Hu et al. 2011].

Considering the overall aspects considered in the mentioned works, we observed a solid direction with a mature network simulation methodology that fits as a solid bottom line for our research.

Looking at Table 1, we highlight that our proposal **is the only one that handles multiple simultaneous phenomena** (MDH, Multivariate Data Handling), thereby taking into account the impact of spatial correlation between different sensed variables. Another characteristic that did not explore so far for this kind of solution is considering real data under a simulated environment. Besides that, we quantify the error and coverage improvement from a VSN-based application compared to conventional air quality stations, considering a realistic urban environment.

Table 1. Summary of related work (sensors and methods).

Article	Air Sensors/Indicators					Processing on Application Level			Experiment	
(per year)	CO_x	O_3	PM	NO_x	SO_x	Comp./Aggr.*	Adapt. Rep.*	MDH*	Real	Sim.*
PROPOSAL	x	x	x	x	x			x		x
[Völgyesi et al. 2008]	x	x		x		x			x	
[Ma et al. 2008]		x		x	x	x				x
[Hu et al. 2009]	x						x		x	
[Hu et al. 2011]	x					x	x			x
[Devarakonda et al. 2013]	x		x			x	x		x	
[Wang and Chen 2017]	x						x			x
[Kaivonen and Ngai 2020]	x			x					x	

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

3. Environmental Application Design

Let the overall behavior be denoted by

$$\begin{tikzcd} \mathcal{N} \arrow[r, "P"] & \mathbf{V}^* \arrow[r, "S"] \arrow[d, "R"] & V \arrow[r, "\omega"] & V' \arrow[d, "R"] \\ & D & & D' \end{tikzcd} \tag{1}$$

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 behavior for every phenomenon $P_i \mid i = \{1, \ldots, n\}$, where n is the number of 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, a sensor S_k is a **car-mounted mobile node** 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, we denote the global visualization of sensing activity resulting from the combination of all sets of phenomena covered by every sensor as $\mathbf{V} = \{V_{(1,1)}, \ldots, V_{(n,k)}\}$. Processes to reduce the size of collected data are not considered in this study so far.

In our case study, the process ω consists of assembling a Voronoi Diagram [Aurenhammer 1991] to fill left blank spaces from non visited areas. The Voronoi strategy used is

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

we consider the samples' location (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, 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. After that, it generates the reconstructed set $\mathbf{V}'$ from V, where we can use the same set of rules R to make decisions D'. The Diagram 1 is analogous to presented by Aquino et. al [Aquino et al. 2012].

The first rule considered to evaluate the performance at each scenario is the Absolute Value of Relative Error ($\hat{\epsilon}$) [Frery et al. 2010]

$$\hat{\epsilon} = \frac{1}{\mathcal{L}} \sum_{x,y}^{Z} \left| \frac{\mathbf{V}^*(x,y) - \mathbf{V}'(x,y)}{\mathbf{V}^*(x,y)} \right|, \tag{3}$$

where $\mathbf{Z}$ 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 $\mathbf{Z}$, $\mathbf{V}^*$ 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 $\mathbf{V}^*(i,j) \neq 0$.

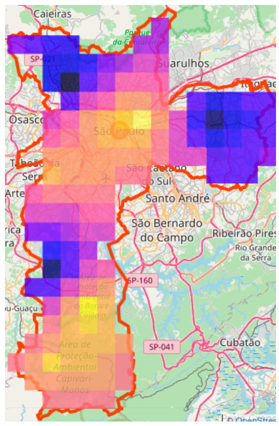


Figure 1. Matching city area with 25×25 pollution map.

4. Multivariate Pollutant Map Generation

The considered dataset contains real data collected from air quality stations placed in São Paulo, with samples 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 related to the burning of fossil-fueled vehicles: Carbon Monoxide (CO); Particulate Matter (PM10); Nitrogen Dioxide (NO_2); Ground-level Ozone (O_3) and Sulfur Dioxide (SO_2).

After a preliminary processing step to assert the unformatted raw data in good shape to perform subsequent operations, the date length was reduced to a range of one week, enough to run the simulations and allow feasible processing with available computing resources. Finally, **for validation purposes** we choose the interval of oct-15-2005 to oct-21-2005 arbitrarily through visual inspection, considering the occurrence behavior of N/A samples due to sensor fails or maintenance at this selected window.

4.1. Prediction

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 follows: (i) Assert each single column (Table 2) as a matrix (Table 3) with **hours** (0h–23h) $\times$ **days** (15–21); (ii) 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 $\times$ days matrix); (iii) Generate a merged normal curve parsing the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ and sample a random value from this distribution. (iv) After that, this procedure should deliver a Table with **NA** samples fixed for every variable (standalone pollutant sensor).

Table 2. Selecting gray to handle NA.

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 3. Entire 4-PM10 column asserted as hour $\times$ 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

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. Table 4 describes the arrangement of sensor availability, where variables colored on green shades are the missing ones. At the first turn, we interpolate

Table 4. 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

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 sequentially the new predicted samples at the previous turn.

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 3. We use the same technique to fix missing stations (Table 4). 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.

4.2. Traffic Simulation

The pollution map described in the previous Section (Table 5) is the general structure used as the baseline in our experiments. The subsequent stage consists of setting up the urban maps and traffic behavior. We use the Open Street Maps (OSM) API to get geographic information from the city by parsing latitude/longitude coordinates. With roads and highways structure appropriately in place, the subsequent step is to generate vehicle routes. For this task, we adopted the Simulator of Urban MObility (SUMO).

Table 5. Summary table representing $\mathbf{V}^*$.

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

The vehicle generation consists of bus stops and their respective defined lines for public transportation, available at fetched maps. During the simulation, the busses will loop on those designated 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. Three different traffic intensities are generated (referred respectively as LIGHT, AVERAGE and HEAVY traffic) in a range of 7 days. Visited coordinates at each bus line are stored to evaluate performance metrics. Collected data **is reported during the route** to a remote cloud-based service **using cellular network** attached with sensor nodes, following sample rate from Table 6.

4.3. Environment Assembly

The last stage of experimentation consists of put together the output from two previous ones (multivariate pollution maps from Section 4 and bus routes trace from Section 4.2). We base the environmental assembly application in R Statistical Language, where we handle the trace file to match coordinates with pollution maps.

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 municipality boundaries.

5. Results and Discussion

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 air quality stations or aided by a VSN network with sensor nodes mounted on public transportation (bus lines). Table 6 summarizes the parameter set for this experiment.

Table 6. 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 Intensity	Light (90k), Average (180k), Heavy (360k)
VSN Sample Rate	15 min.
Simulated Time	7 days (random seeds for each one)

5.1. Summary for Global Field Coverage

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

Table 7 shows the trace from bus lines in a 1-hour window, there is one instance of this table for each traffic intensity (see Table 6). It is assigned to the respective hour of day (Table 8). After that, we shape all visited coordinates as a plain list (eliminating repeated ones) and count how many are covered in relation to 132 squared units of entire map, generating the percentages seen at Table 9. 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 7. 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)

Table 8. 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 9. 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%	
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%	7.5%
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%	

Considering the presented strategy, after seven 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 6) covered by the map area and the known amount of 10 stations. **The regular monitoring system achieves a theoretical global field coverage of 7,5%**.

5.2. 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 3). Figure 2 shows the behavior of error metric for each pollutant, at bottom of Figure 3(c) map we see an information loss at continuous dark red area in comparison with 3(a) and 3(b) due to the lack of spatial reachability in stationary sensing, which was mitigated at VSN approach.

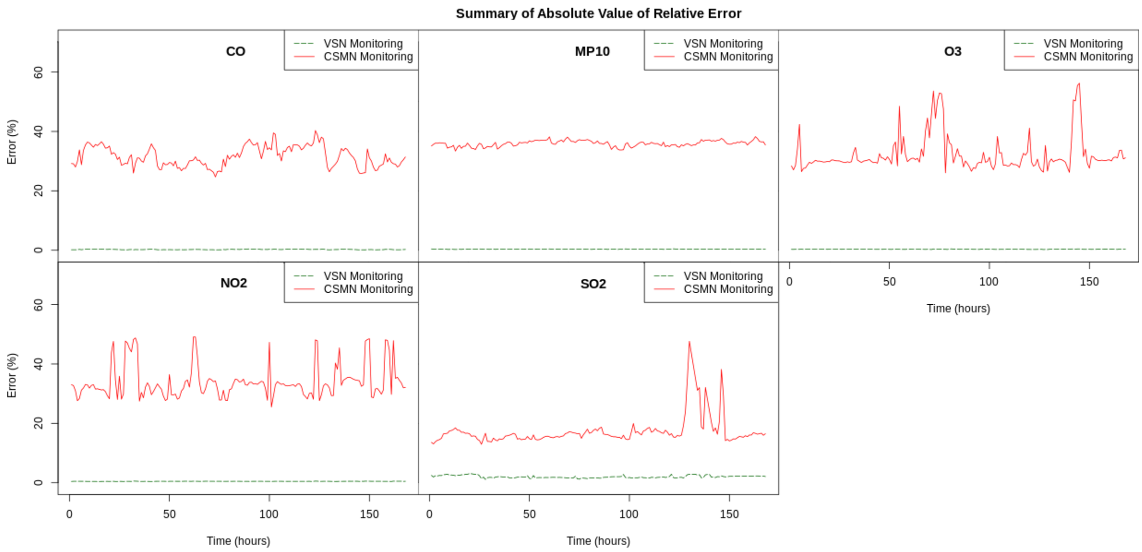
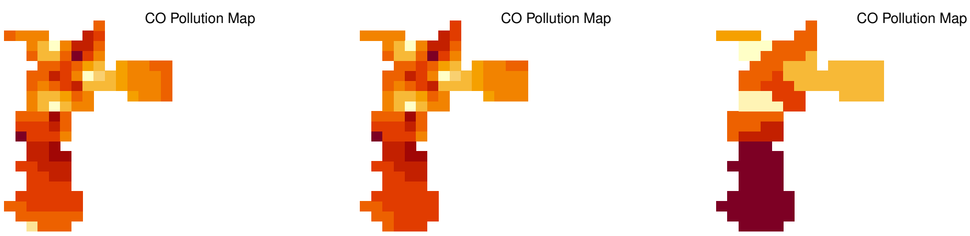


Figure 2. Error rate evaluated through 7 days from VSN and Conventional Stations



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

Figure 3. Side-by-side performance comparison between sampling with VSN and Conventional Stations for Carbon Monoxide.

Considering the Carbon Monoxide pollution map, the Absolute Value of Relative

Error (AVRE) along the entire time series presented an average of 0.25% for VSN application 31.57% for Conventional Air Quality Stations. For Particulate Matter, AVRE indicates an average of 0.36% for VSN application and 35.93% for Conventional Air Quality Stations. For ground-level Ozone, AVRE indicates an average of 0.32% for VSN application and 32.27% for Conventional Air Quality Stations. For Nitrogen Dioxide, AVRE indicates an average of 0.48% for VSN application and 34.03% for Conventional Air Quality Stations. For Sulfur Dioxide, AVRE indicates an average of 2.02% for VSN application and 13.33% for Conventional Air Quality Stations.

This result means an improvement on 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. We can observe that the error difference between the two approaches decreases according to data availability displayed at Table 4. 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 shows a noticeable improvement (659% at the worst case from SO_2) on overall application behavior comparison to conventional strategy. Since only two sensors for an entire city is quite extreme restriction, any dataset with more sensors available is enough to mitigate this limitation.

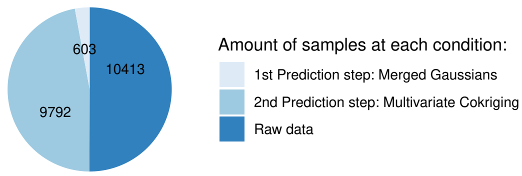


Figure 4. Sources from samples.

This result also shows that our model is robust even with extremely constrained input data. Highlighting all the procedures from Subsection 4.1 (i.e., NA fixing) that contributes with achieved result. Figure 4 details predicted samples from those procedures.

6. Conclusion and Final Remarks

The presented work has explored the air quality monitoring problem while taking an in-depth investigation into the modeling of complex environments. Beyond that, to deliver contributions that improve the view of how correlated multivariate phenomena behave.

The approached VSN supported by public transportation (bus lines) showed considerably higher performance than the regular system based on air quality stations, behaving with error rates near zero and about 11.5 times higher global coverage, which figures as a solid contribution what is observed at literature so far.

Thereby, overall observed performance indicates that the proposed application on this case study is suitable to be executed in real-world scenarios and can predict realistic behaviors of correlated physical processes accurately. Future directions consider evaluating different data processing algorithms and improving environmental modeling with variables that did not consider at last turn, such as wind speed/direction, temperature, humidity data on evaluation, and finally, find out novel insights. The complete project is available at link `https://github.com/isrvasconcelos/VSN-MultivariateAirPollution`.

References

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

[Aquino et al. 2012] 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.

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

[Devarakonda et al. 2013] 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 *Proceedings of the 2nd ACM SIGKDD int. workshop on urban computing*.

[Frery et al. 2010] 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.

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

[Hu et al. 2011] 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.

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

[Khedo et al. 2010] 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.

[Ma et al. 2008] 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.

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

[Rashid and Rehmani 2016] 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.

[Völgyesi et al. 2008] 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 (ipsn 2008)*.

[Wang and Chen 2017] 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.

[Yi et al. 2015] 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.