




Article

# Analysis of Atmospheric Pollutant Data Using Self-Organizing Maps

Emanoel L. R. Costa <sup>1,†</sup>, Taiane Braga <sup>2,†</sup>, Leonardo A. Dias <sup>3,†</sup> , Édler L. de Albuquerque <sup>4,\*,†</sup>   
and Marcelo A. C. Fernandes <sup>1,5,\*,†</sup> 

<sup>1</sup> Laboratory of Machine Learning and Intelligent Instrumentation, Federal University of Rio Grande do Norte, Natal 59078-970, RN, Brazil

<sup>2</sup> Federal Institute of Education, Science, and Technology of Bahia, Salvador 40301-015, BA, Brazil

<sup>3</sup> Centre for Cyber Security and Privacy, School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK

<sup>4</sup> Department of Industrial Processes and Chemical Engineering, Federal Institute of Education, Science and Technology of Bahia, Salvador 40301-015, BA, Brazil

<sup>5</sup> Department of Computer Engineering and Automation, Federal University of Rio Grande do Norte, Natal 59078-970, RN, Brazil

\* Correspondence: edler@ifba.edu.br (É.L.d.A.); mfernandes@dca.ufrn.br (M.A.C.F.)

† These authors contributed equally to this work.

**Abstract:** Atmospheric pollution is a critical issue in our society due to the continuous development of countries. Therefore, studies concerning atmospheric pollutants using multivariate statistical methods are widely available in the literature. Furthermore, machine learning has proved a good alternative, providing techniques capable of dealing with problems of great complexity, such as pollution. Therefore, this work used the Self-Organizing Map (SOM) algorithm to explore and analyze atmospheric pollutants data from four air quality monitoring stations in Salvador-Bahia. The maps generated by the SOM allow identifying patterns between the air quality pollutants (CO, NO, NO<sub>2</sub>, SO<sub>2</sub>, PM<sub>10</sub> and O<sub>3</sub>) and meteorological parameters (environment temperature, relative humidity, wind velocity and standard deviation of wind direction) and also observing the correlations among them. For example, the clusters obtained with the SOM pointed to characteristics of the monitoring stations' data samples, such as the quantity and distribution of pollution concentration. Therefore, by analyzing the correlations presented by the SOM, it was possible to estimate the effect of the pollutants and their possible emission sources.

**Keywords:** machine learning; atmospheric pollution; Self-Organizing Maps; Salvador-BA



**Citation:** Costa, E.L.R.; Braga, T.; Dias, L.A.; Albuquerque, É.L.d.; Fernandes, M.A.C. Analysis of Atmospheric Pollutant Data Using Self-Organizing Maps. *Sustainability* **2022**, *14*, 10369. <https://doi.org/10.3390/su141610369>

Academic Editors: José Carlos Magalhães Pires and Álvaro Gómez-Losada

Received: 12 July 2022

Accepted: 17 August 2022

Published: 20 August 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Air pollution is one of the crucial challenges of modern society. In recent years, pollution caused by industrial, vehicular, and toxic-chemical emission sources has increased significantly. This increase can be seen mainly in low- and middle-income countries, also called developing countries [1]. Despite the continuous pollution growth, awareness and pollution control programs are limited and receive little attention and financial resources from governments, international agencies, and philanthropic donors [1].

In addition, effectively managing regulations for controlling air pollution requires considerable knowledge about the costs and benefits. Currently, the primary efforts for measuring pollutants aim to avoid possible harm to people's health, such as respiratory or cardiovascular diseases that can result in hospitalizations and even death, usually affecting vulnerable groups of the population [2].

Complex mixtures of solid particles and gaseous pollutants contribute to air pollution. Among these are priority pollutants, commonly regulated by law and categorized as primary and secondary. The primary pollutants are substances that can be released directly

into the atmosphere, while the secondary pollutants are substances derived from the primary ones through photochemical reactions in the troposphere [3]. Regarding the gaseous pollutants, to be particularly mentioned are sulfur dioxide ( $\text{SO}_2$ ), nitrogen dioxide ( $\text{NO}_2$ ), carbon monoxide ( $\text{CO}$ ), volatile organic compounds (VOCs), solid materials or liquids suspended in the atmosphere due to their small size (called particulate matter (PM)), and the ozone ( $\text{O}_3$ ). The ozone is one of the major photochemical pollutants formed in the atmosphere by the reaction of nitrogen oxides ( $\text{NO}_x$ ) and hydrocarbons such as VOCs in the presence of sunlight, similarly to particulate sulfate and nitrate aerosols created from  $\text{SO}_2$  and  $\text{NO}_x$  [3].

The dispersion of atmospheric pollutants results from different elements such as temperature, relative humidity, atmospheric pressure, wind direction and speed, as well as topography [4]. Consequently, the complexity of analyzing and identifying pollutants and their primary sources in large-scale areas increases, which leads to the problem of positioning monitoring stations for data collection.

There are several emission sources of air pollutants, and a single source can emit several pollutants. For instance, the composition of fossil fuels used in motor vehicles can emit different pollutants during combustion and evaporation, or by the wear of tires and roads where vehicles run. Due to the increasing number of private vehicles, their emissions have become a dominant source of  $\text{CO}$ ,  $\text{CO}_2$ , VOCs,  $\text{NO}_x$  and PM. Meanwhile, industrial processes normally include pollutants such as  $\text{CO}$ , PM,  $\text{NO}_x$ , and  $\text{SO}_2$  [4–6].

Thus, monitoring the concentration of pollutants in the environment at specific points is essential. Identifying the main components enables understanding of the current condition of air pollution, variations, correlations, and possible emission sources, which leads to the development of public policies to raise awareness and reduce pollutants. Therefore, many researchers have proposed the analysis of environmental data mainly using multivariate statistical techniques [4,7,8].

Multivariate statistical methods such as correlation or cluster analysis [9–13], and principal component analysis [7,14,15] are commonly applied in various studies to identify the correlation among parameters that can influence air quality. Large databases that carry various information about air pollution require techniques to extract and identify characteristics inherent to the analyzed data.

In this context, machine learning has proved to be a great alternative to the traditional methods used [16,17]. A well-known algorithm that belongs to the group of unsupervised learning algorithms is self-organizing maps (SOM) [18]. The SOM supports data dimensionality reduction and clustering. In addition, the SOM does not need to make assumptions about the parameters' distribution, as it is capable of dealing with non-linear problems of great complexity and dimension and is effective in using noisy data [19].

The SOM algorithm is adopted in many applications to analyze data from atmospheric pollutants. For example, in Ref. [20], the SOM is used to analyze data regarding air quality. In Ref. [21], the SOM is used to identify the level of pollution during foundry and land mining. The study carried out in [22] used the SOM to highlight the impact on air quality caused by the circulation of different air types, which alters the concentration of pollutants in the atmosphere. For this purpose, it is essential to identify suitable placements for positioning monitoring stations, as shown in [23]. Finally, the SOM has also been used to obtain particulate-matter characteristics in the atmosphere by evaluating its concentration in both internal and external exposure and connecting them to human activities. According to [24], the SOM can also function as a pollution identifier by defining limits to classify regions with low or high concentrations of a specific pollutant, such as ozone, enabling the evaluation of pollution zones.

Therefore, this work proposes an SOM implementation to study and analyze atmospheric pollutants to identify their patterns and characteristics. The main contributions are:

- A machine-learning-based approach for analyzing the air quality of Salvador monitoring stations, using the Government of Bahia State database—to the best of our knowledge, this work is the first to analyze this data using machine-learning algorithms.

- We discuss the common factors among meteorological parameters and pollutants and their clusters' impact on each monitoring station.

## 2. Methodology

### 2.1. Case Study

Salvador city (State of Bahia) has a territorial area of 693.453 km<sup>2</sup> and a population of 2,675,656 people. Located in the northeastern region of Brazil, it has an urban core and rugged topography formed by several columns and valleys, with a rainy tropical climate with no dry season and an average annual temperature of 25 °C.

The Government of Bahia State, through CETREL S. A., the company that operated the air monitoring stations from 2011 to 2016, provided the air quality database for this work. It contains the air quality data of a monitoring network constituted of eight stations. Nonetheless, we used data from four stations: Barros Reis (BR), Campo Grande (CG), Dique do Tororó (DT), and Itaigara (IT), due to their inherent characteristics. Figure 1 illustrates the stations' distribution in Salvador and highlights the four chosen. It is important to mention that this is the first air quality monitoring network ever installed in the city of Salvador. Therefore, this work portrays the first analysis of the pollutant and meteorological parameters in the database provided.

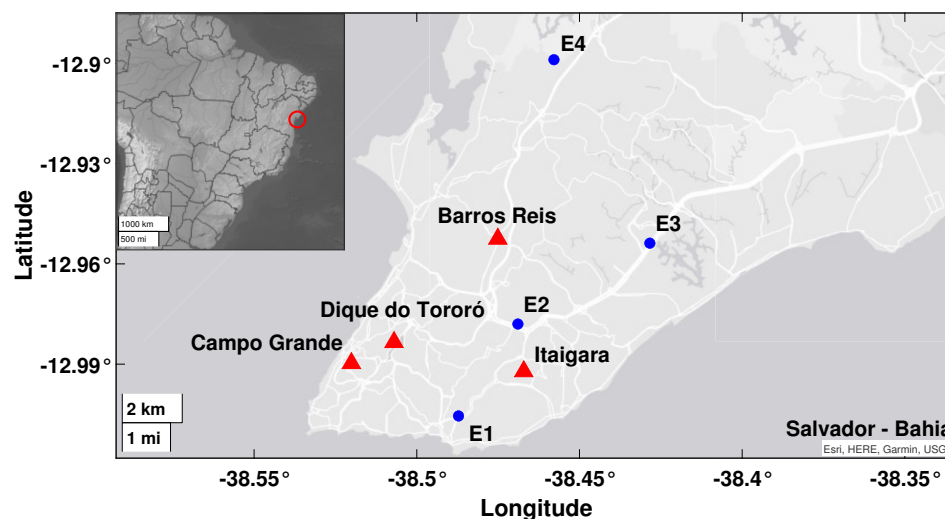


Figure 1. Location of the eighth air monitoring stations deployed in Salvador-BA.

### 2.2. Dataset

The dataset contains the hourly average of twelve features related to meteorological parameters and pollutants concentration. The meteorological parameters are wind speed (WS), ambient temperature (TEMP), relative air humidity (RH), the standard deviation of wind direction (STWD), rainfall, and wind direction. Meanwhile, the pollutants are SO<sub>2</sub>, CO, O<sub>3</sub>, particulate matter whose aerodynamic diameter is less than 10 µm (PM<sub>10</sub>) and the oxides of nitrogen NO<sub>2</sub> and NO. We removed the rainfall and wind direction variables due to the small amount of data available; thus, only ten features were used in our analysis.

We performed a data preprocessing step by removing the null lines, the measurement errors (identified by a specific terminology), and the outliers to improve the quality of the analysis. The outliers were removed by investigating the data dispersion and symmetry and, subsequently, using the quartile separatrix measure [25] to divide the dataset into three quartiles: Q<sub>1</sub>, Q<sub>2</sub> and Q<sub>3</sub>. Finally, based on the interquartile range (AIQ) [25], outliers with value greater than Q<sub>3</sub> + 3 × AIQ and less than Q<sub>1</sub> − 3 × AIQ, were removed from the database. We kept outliers with values greater than Q<sub>3</sub> + 1.5 × AIQ and less than Q<sub>1</sub> − 1.5 × AIQ to avoid a large reduction in the dataset. Table 1 presents the number of data samples for each monitoring station considered in our analysis and their period of operation.

**Table 1.** The operation period for each monitoring station provided by CETREL S. A., and the number of data samples available in the dataset before and after the preprocessing step.

Station	Operation Start Date	Operation End Date	Number of Registered Samples	Number of Samples after Pre-processing
Barros Reis (BR)	8 November 2013	31 December 2016	27,584	21,559
Campo Grande (CG)	2 July 2011	31 December 2016	48,234	24,559
Dique do Tororó (DT)	19 June 2011	31 December 2016	48,550	42,037
Itaigara (IT)	18 October 2013	30 April 2016	22,203	15,535

In the meantime, Tables 2–5 present the dataset for the Barros Reis (BR), Campo Grande (CG), Dique do Tororó (DT), and Itaigara (IT) stations, respectively. As can be observed, all pollutants and atmospheric data are shown after the preprocessing step for each station in a concentration of pollutants in parts per billion (ppb).

**Table 2.** Descriptive statistics of pollutants and atmospheric data from the Barros Reis station ( $P = 21,559$  samples).

Parameters	Magnitude	Maximum	Mean	Average	Standard Deviation	Variation Coefficient
SO <sub>2</sub>	ppb	3.20	0.30	0.45	0.51	112.94%
CO	ppb	2180.00	570.00	601.60	335.70	55.80%
O <sub>3</sub>	ppb	22.70	4.80	5.47	3.80	69.36%
PM <sub>10</sub>	µg/m <sup>3</sup>	129.80	37.30	40.10	19.88	49.58%
NO	ppb	206.40	44.40	52.47	38.01	72.50%
NO <sub>2</sub>	ppb	49.20	13.30	14.15	7.61	53.84%
WS	m/s	10.80	2.20	2.62	1.75	67.00%
TEMP	°C	32.50	25.50	25.63	2.18	8.54%
RH	%	91.00	69.00	68.60	9.31	13.57%
STWD	°	73.30	31.30	31.61	11.61	36.73%

**Table 3.** Descriptive statistics of pollutants and atmospheric data from the Campo Grande station ( $P = 24,559$  samples).

Parameters	Magnitude	Maximum	Mean	Average	Standard Deviation	Variation Coefficient
SO <sub>2</sub>	ppb	1.70	0.20	0.32	0.31	97.20%
CO	ppb	1830.00	360.00	396.60	292.70	73.81%
O <sub>3</sub>	ppb	25.00	5.20	6.01	4.18	69.5%
PM <sub>10</sub>	µg/m <sup>3</sup>	77.30	19.30	21.10	12.53	59.38%
NO	ppb	139.00	25.10	28.03	23.37	83.38%
NO <sub>2</sub>	ppb	44.00	13.30	13.37	6.42	48.00%
WS	m/s	5.10	1.20	1.42	0.93	66.01%
TEMP	°C	34.30	26.50	26.72	2.31	8.66%
RH	%	94.00	72.00	71.12	9.54	13.41%
STWD	°	79.60	53.20	52.01	13.31	25.57%

**Table 4.** Descriptive statistics of pollutants and atmospheric data from the Dique do Tororó station ( $P = 42,037$  samples).

Parameters	Magnitude	Maximum	Mean	Average	Standard Deviation	Variation Coefficient
SO <sub>2</sub>	ppb	2.00	0.20	0.33	0.40	123.15%
CO	ppb	1000.00	220.00	239.40	163.90	68.44%
O <sub>3</sub>	ppb	34.30	7.20	8.15	5.37	65.88%
PM <sub>10</sub>	µg/m <sup>3</sup>	75.60	20.00	22.13	12.42	56.12%
NO	ppb	73.60	12.40	13.77	11.54	83.78%
NO <sub>2</sub>	ppb	31.30	8.20	8.67	5.02	57.92%
WS	m/s	6.90	1.50	1.63	1.01	61.72%
TEMP	°C	33.90	26.30	26.46	2.31	8.75%
RH	%	94.00	73.00	72.46	9.10	12.56%
STWD	°	78.80	33.00	38.45	15.34	39.90%

**Table 5.** Descriptive statistics of pollutants and atmospheric data from the Itaigara station ( $P = 15,535$  samples).

Parameters	Magnitude	Maximum	Mean	Average	Standard Deviation	Variation Coefficient
SO <sub>2</sub>	ppb	1.60	0.10	0.2502	0.33	131.89%
CO	ppb	1210.00	190.00	226.48	207.26	91.51%
O <sub>3</sub>	ppb	27.50	7.90	8.47	4.32	51.00%
PM <sub>10</sub>	µg/m <sup>3</sup>	67.40	13.60	16.16	10.98	67.94%
NO	ppb	70.70	11.40	15.50	13.45	86.77%
NO <sub>2</sub>	ppb	31.10	7.30	8.21	5.15	62.72%
WS	m/s	10.20	2.70	2.76	1.58	57.24%
TEMP	°C	33.40	25.00	25.04	2.27	9.06%
RH	%	93.00	71.00	71.43	9.08	12.71%
STWD	°	51.30	22.80	24.32	8.13	33.42%

As can be observed, the BR station presents a higher concentration of SO<sub>2</sub>, CO, and PM<sub>10</sub>. The SO<sub>2</sub> has a maximum of 3.20 ppb and an average of 0.45 ppb due to the burning of fuels with sulfur. Meanwhile, the CO has a maximum of 2180 ppb and an average of 601.6 ppb, produced by burning organic fuels. The PM<sub>10</sub> has an average of 40.10 ppb, almost double the value of other stations; it is a solid or liquid material that remains suspended in the atmosphere that can cause a significant impact on human health.

The CG station also has a high level of CO, with a maximum of 1830 ppb and an average of 396.6 ppb. Regarding the presence of nitrogen oxides (NO and NO<sub>2</sub>), the CG and DT stations present higher average and maximum concentrations due to the combustion processes and atmospheric chemical reactions. Concerning the O<sub>3</sub>, a secondary pollutant formed in the atmosphere indicating the presence of photochemical oxidants, it has its higher concentrations recorded at the DT and IT stations.

Therefore, the SO<sub>2</sub>, CO, and NO pollutants present the most significant variations in concentration. These pollutants are mainly generated from the burning of fossil fuels. Hence, the station location and the intensity of the vehicle's traffic around its region can lead to different concentration records at certain times of the day. The datasets comprise 24 h of daily data collection.

All stations show similar measured values regarding the meteorological parameters, except for wind speed which has a high average at BR and IT stations, and the standard deviation of wind direction at CG. Note that the values were rescaled from 0 to 1 to improve the SOM results. In addition, this work performed the z-score normalization and logarithmic transformation, obtaining data with null mean and unit variance and reducing the data scale, respectively.

### 2.3. Self-Organizing Maps (SOM)

The Self-Organizing Map (SOM) is a neural network model widely applied to data dimensionality reduction and clustering [18,26]. The map consists of  $M$  neurons commonly arranged in a two-dimensional array representing the incoming data by shifting the neurons' position towards it. The maps' topology can be rectangular, hexagonal, or square, among others [18].

The  $N$ -dimensional input data sample can be characterized as

$$\mathbf{x} = [x_1, x_2, \dots, x_N]. \quad (1)$$

Accordingly, each  $i$ -th neuron in the map is represented by a  $N$ -dimensional vector of weights expressed as

$$\mathbf{w}_i = [w_{i1}, w_{i2}, \dots, w_{iN}]. \quad (2)$$

Therefore, the topology of a two-dimensional map with  $M$  neurons can be expressed as  $(M_h \times M_v)$ , where  $M_h$  is the number of neurons in the horizontal and  $M_v$  is the number of neurons vertically; thus,  $M = M_h \times M_v$ .

The SOM algorithm iteratively molds the neurons' map to the input data topological form, based on a similarity metric, according to the following steps [18]:

1. Randomly initialize the  $M$  neurons' weight vectors.
2. Calculate the distance of each  $p$ -th input data sample,  $\mathbf{x}(p)$ , to all  $M$  neurons.
3. Define the winning neuron, also known as best matching unit (BMU); it is the  $j$ -th nearest neuron to the input data defined based on a distance metric as follows:

$$j = \arg \min_i \|\mathbf{x}(p) - \mathbf{w}_i\|, i = 1, 2, \dots, M. \quad (3)$$

4. Update the BMU neuron and its neighboring neurons' weights according to following

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \eta(t)h_{i,j}(t)(\mathbf{x}(p) - \mathbf{w}_i) \quad (4)$$

where  $\eta(t)$  is the learning rate (ranging from 0 to 1) and  $h_{i,j}(t)$  represents the BMU neighborhood function at the  $t$ -th iteration. The neighborhood function is described as

$$h_{i,j}(t) = \exp\left(-\frac{d_{i,j}^2}{2\sigma^2(t)}\right) \quad (5)$$

where  $d_{i,j}^2$  is the distance from the  $i$ -th neuron to the BMU ( $j$ -th neuron) and  $\sigma^2(t)$  is the neighborhood function size at the  $t$ -th iteration.

5. Repeat steps 2, 3 and 4 until the maximum number of iterations is reached, represented here by  $T$ .

The number of iterations must be enough to process the dataset samples several times; thus,  $T = b \times P$ , where  $b$  is the repetition number that every set of  $P$  samples is presented to the SOM. Moreover, increasing the iteration number ( $t$ ) decreases the radius of the neighborhood function,  $\sigma^2(t)$ . Consequently, the number of neurons nearby the BMU to be updated is reduced, strengthening their connection and similarities. After training the network, each  $p$ -th entry  $\mathbf{x}(p)$  is associated with a specific BMU in the output layer, and entries that share similar patterns will be associated with the same BMU or its neighbors, which can be understood as a grouping in the SOM.

We applied the SOM to each monitoring station shown in Table 1. Each  $p$ -th sample in the dataset has  $N = 10$  dimensions, 6 regarding atmospheric pollutants ( $\text{SO}_2$ , CO,  $\text{O}_3$ ,  $\text{PM}_{10}$ , NO, and  $\text{NO}_2$ ) and 4 concerning meteorological parameters (WS, TEMP, RH, and STWD). Therefore, the SOM enables analyzing the influence and characteristics of these variables.



#### 2.4. SOM Parameters

The map size is the first parameter to be defined. For this purpose, it is necessary to determine the number of neurons to be used during training; in addition, avoiding a large or small number of neurons is vital to prevent problems such as non-identification of characteristics and overfitting [27]. Commonly, the number of neurons can be determined using the following heuristic equation

$$M \approx 5\sqrt{P} \quad (6)$$

where  $P$  is the number of input data samples [27].

Subsequently, the map topology ( $M_h \times M_v$ ) was defined according to quality measures commonly used for the SOM network, the quantization error (QE) and topographic error (TE) [28,29]. For each station, different values of  $M_h$  and  $M_v$  were tested, in which  $M_h \times M_v = M$  (Equation (6)). Finally, to analyze the results, three different types of normalization were applied to the data: z-score, min-max, and logarithmic.

Hence, all tests were performed with  $b = 500$ , a hexagonal topology, and the training algorithm was applied in two steps. Firstly, the learning rate and neighborhood function were initialized as  $\eta(0) = 0.5$  and  $\sigma^2(0) = \frac{M_h}{2}$ , respectively, and decreased over iterations. Secondly, these values were fixed as  $\eta = 0.05$  and  $\sigma^2 = 1$ . Tables 6–9 present the quality measures obtained for each test.

**Table 6.** SOM quality measures for Barros Reis Station data (best values in bold).

$(M_h \times M_v)$	$M$	z-Score		Min-Max		Logarithmic	
		QE	TE	QE	TE	QE	TE
(27 × 24)	648	1.4032	0.0649	0.2290	0.0636	0.7173	0.0606
(26 × 26)	676	1.3898	0.0687	0.2273	0.0661	0.7108	0.0616
(29 × 24)	696	1.3887	0.0668	0.2270	0.0668	0.7096	0.0607
(31 × 23)	713	1.3805	0.0631	0.2259	0.0607	0.7051	0.0593
(27 × 27)	729	1.3803	0.0612	0.2245	0.0653	0.7032	0.0629
(30 × 25)	750	1.3766	0.0660	0.2250	0.0667	0.7017	0.0616
(32 × 24)	768	1.3684	0.0649	<b>0.2232</b>	<b>0.0622</b>	0.6977	0.0601
(34 × 23)	782	1.3652	0.0701	0.2229	0.0644	0.6950	0.0644
(33 × 24)	792	1.3609	0.0655	0.2224	0.0673	0.6957	0.0587
(31 × 26)	806	1.3597	0.0658	0.2219	0.0663	0.6948	0.0622

**Table 7.** SOM quality measures for Campo Grande Station data (best values in bold).

$(M_h \times M_v)$	$M$	z-Score		Min-Max		Logarithmic	
		QE	TE	QE	TE	QE	TE
(31 × 23)	713	1.4216	0.0666	0.2384	0.0626	0.7346	0.0625
(27 × 27)	729	1.4187	0.0660	0.2382	0.0667	0.7294	0.0584
(30 × 25)	750	1.4131	0.0650	0.2369	0.0664	0.7277	0.0626
(32 × 24)	768	1.4082	0.0648	0.2360	0.0640	0.7253	0.0630
(28 × 28)	784	1.4099	0.0619	0.2352	0.0685	0.7230	0.0610
(31 × 26)	806	1.3994	0.0645	0.2341	0.0670	0.7215	0.0589
(34 × 24)	816	1.3948	0.0642	<b>0.2340</b>	<b>0.0624</b>	0.7193	0.0592
(33 × 25)	825	1.3949	0.0636	0.2334	0.0636	0.7173	0.0593
(35 × 24)	840	1.3925	0.0651	0.2336	0.0669	0.7163	0.0630
(36 × 24)	864	1.3898	0.0619	0.2324	0.0643	0.7146	0.0594

**Table 8.** SOM quality measures for Dique do Tororó Station data (best values in bold).

$(M_h \times M_v)$	$M$	z-Score		Min-Max		Logarithmic	
		QE	TE	QE	TE	QE	TE
(38 × 25)	950	1.2812	0.0686	0.2175	0.0668	0.6834	0.0630
(37 × 26)	962	1.2773	0.0684	0.2172	0.0670	0.6814	0.0641
(38 × 26)	988	1.2742	0.0668	0.2163	0.0679	0.6802	0.0621
(36 × 28)	1008	1.2687	0.0733	0.2157	0.0676	0.6798	0.0619
(32 × 32)	1024	1.2667	0.0736	0.2152	0.0679	0.6777	0.0659
(40 × 26)	1053	1.2628	0.0702	<b>0.2146</b>	<b>0.0678</b>	0.6745	0.0644
(39 × 27)	1040	1.2639	0.0695	0.2152	0.0688	0.6750	0.0611
(38 × 28)	1064	1.2581	0.0721	0.2141	0.0680	0.6747	0.0660
(37 × 29)	1073	1.2600	0.0706	0.2136	0.0728	0.6718	0.0632
(40 × 27)	1080	1.2609	0.0657	0.2136	0.0717	0.6730	0.0633

**Table 9.** SOM quality measures for Itaigara Station data (best values in bold).

$(M_h \times M_v)$	$M$	z-Score		Min-Max		Logarithmic	
		QE	TE	QE	TE	QE	TE
(24 × 23)	552	1.4306	0.0584	0.2428	0.0591	0.7736	0.0510
(26 × 22)	572	1.4237	0.0603	0.2422	0.0566	0.7709	0.0485
(24 × 24)	576	1.4210	0.0618	0.2421	0.0557	0.7704	0.0503
(27 × 22)	594	1.4192	0.0548	0.2403	0.0589	0.7684	0.0547
(25 × 24)	600	1.4152	0.0593	0.2412	0.0565	0.7659	0.0477
(27 × 23)	621	1.4126	0.0574	0.2400	0.0585	0.7654	0.0444
(25 × 25)	625	1.4063	0.0573	<b>0.2399</b>	<b>0.0553</b>	0.7625	0.0458
(27 × 24)	648	1.4086	0.0572	0.2381	0.0561	0.7595	0.0472
(26 × 26)	676	1.3945	0.0640	0.2371	0.0556	0.7553	0.0525
(27 × 26)	702	1.3861	0.0578	0.2363	0.0559	0.7516	0.0538

Considering both QE and TE measures, the lowest values were obtained using min-max normalization. Thus,  $M_h$  and  $M_v$  were chosen according to the best result, being highlighted in each table.

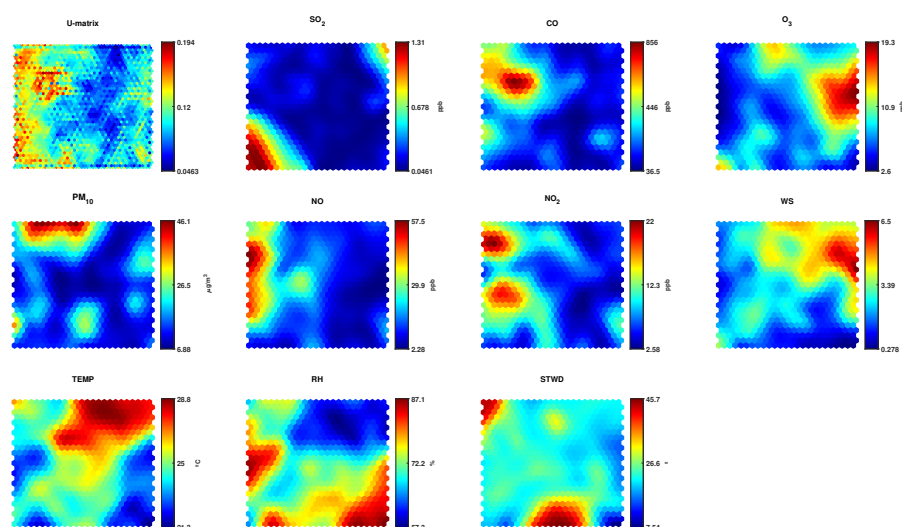
### 3. Results

#### 3.1. U-Matrix, Components Plane and Parameter Similarity

The SOM output can be represented by a unified distance matrix (U-matrix) and a component plane, both illustrated in Figure 2. The U-matrix provides a visualization of the relative distance between neurons in the map, which is evidenced through a color scale, and highlights the calculated distance between the adjacent neurons [18]. The closer the color approaches a dark blue in the U-matrix, the closer these neurons are, i.e., they have a more significant similarity. On the other hand, the closer the color approaches a dark red, the greater the distance between the neurons and their dissimilarity. In general, this form of representation allows us to consider that neurons with smaller distances form a cluster. In contrast, neurons with high distances can be considered as boundaries of a cluster.

The component plane shows the values of the weight vectors of each neuron through a color code, where the blue and red colors correspond to low and high values, respectively. This representation allows the recognition of parameter dependencies by comparing the patterns of each plane. The color gradient of a plane represents the parameters' value (component) for the analyzed samples. Each neuron is assigned a color according to the parameter value in that neuron; thus, it can be said that two or more parameters are related based on a comparison of their color gradients. A coherent gradient indicates a positive correlation, while an inverse gradient a negative correlation.





**Figure 2.** Unified distance matrix (U-matrix) and component planes of all analyzed variables ( $\text{SO}_2$ ,  $\text{CO}$ ,  $\text{O}_3$ ,  $\text{PM}_{10}$ ,  $\text{NO}$ ,  $\text{NO}_2$ ,  $\text{WS}$ ,  $\text{TEMP}$ ,  $\text{RH}$ , and  $\text{STWD}$ ) from the Itaigara station.

### 3.2. Itaigara Station

Analyzing the component planes in Figure 2, it is possible to note that the relative humidity ( $\text{RH}$ ) and temperature ( $\text{TEMP}$ ) planes display inverse gradients, indicating a negative correlation between these parameters—something already expected given their characteristics. For  $\text{CO}$ ,  $\text{NO}$ , and  $\text{NO}_2$  pollutants, their weight vectors present a dark red color on the left side of the components' plane, with a higher concentration of high values at the top left side; hence evidencing a certain similarity between them. These pollutants are generated by combustion, and incomplete burning of organic fuels, which are very common in cities with a large circulation of vehicles (the leading emitter) [5].

The  $\text{O}_3$  pollutant can be formed by the reaction of nitrogen oxides with VOCs. However, it presents a different pattern than  $\text{NO}_2$ , which contributes to the formation of photochemical oxidants such as  $\text{O}_3$ . As can be seen in the  $\text{O}_3$  component plane, its high-value region is concentrated on the right side, similar to the wind speed component plane. Therefore, it can be said that the  $\text{O}_3$  presence at the Itaigara Station probably came from another region carried by the wind, as it has a low concentration near traffic routes and is generated by photochemical reactions.

The  $\text{PM}_{10}$  showed a different pattern than the other pollutants. Its main concentration region, with high weight vector values, is in the upper part of the plane. Since its emission sources are diverse, such as vehicles, biomass burning, industries, and dust resuspension, it is difficult to identify the major contributor pollutant. However, its formation can also be carried out in the atmosphere through VOCs,  $\text{SO}_2$ , and nitrogen oxides.

The most distinct pattern presented was by  $\text{SO}_2$ , with high values and concentration in the lower left part, it does not resemble any other component plane. This pollutant is released mainly by heavy vehicles burning diesel oil in urban areas.

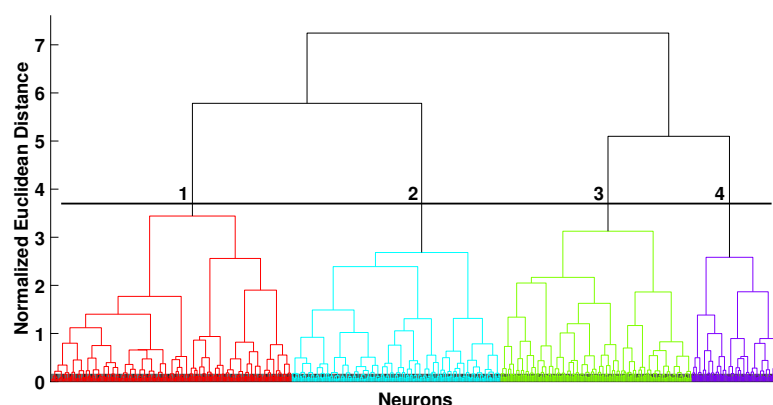
An SOM arranges similar patterns in the same neighborhood region, clustering the network's output. Hence, an investigation into the clustering of samples provides important information about the data.

The U-matrix in Figure 2 illustrates how close or far the neurons are, showing their clusters. However, the cluster boundaries are not clearly represented, making it challenging to identify them. One of the methods for choosing the appropriate number of clusters is the so-called Davies–Bouldin index [30], an evaluation measure commonly used in SOM networks for validating clusters [31,32].

### 3.2.1. Sample Grouping with the SOM Algorithm

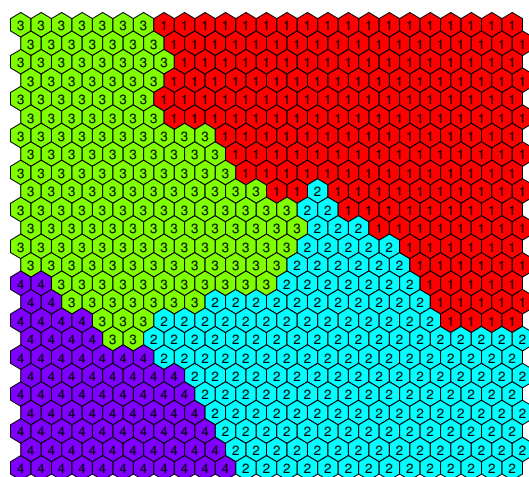
For the Davies–Bouldin index, the lowest value found indicates the best number of clusters for the analyzed problem. Thus, an experiment was conducted by varying the number of clusters from two to eight and observing the obtained values. The best result was achieved for a total of four clusters.

Aftwards, a hierarchical analysis was performed to define the neurons belonging to the four clusters. For this purpose, the Euclidean distance was used as the similarity metric and the Ward neuron linking criterion, illustrated by the dendrogram shown in Figure 3. A dendrogram threshold value is defined for that to which cluster each neuron belongs (horizontal line in Figure 3).



**Figure 3.** Hierarchical analysis of the neurons clusters using the Ward linkage method and Euclidean distance for the Itaigara station.

In addition, based on the hierarchical analysis, the SOM neurons were classified in four clusters, as shown in Figure 4. Therefore, the samples assigned to each cluster and its neurons present the characteristics of the distribution of pollutants and meteorological parameters. Table 10 shows the mean value of samples for each parameter and cluster.



**Figure 4.** SOM neurons grouped into four clusters obtained by the hierarchical analysis of the Itaigara station.

**Table 10.** Parameters' average values for every cluster formed by the SOM network for the Itaigara station.

Parameters	Parameter Average Value per Cluster			
	1	2	3	4
SO <sub>2</sub> (ppb)	0.18	0.09	0.17	0.89
CO (ppb)	153.18	126.03	443.43	230.86
O <sub>3</sub> (ppb)	11.93	7.38	5.45	7.61
PM <sub>10</sub> (µg/m <sup>3</sup> )	17.73	12.92	17.97	15.83
NO (ppb)	9.20	9.15	29.64	19.28
NO <sub>2</sub> (ppb)	6.10	6.81	12.29	9.10
WS (m/s)	4.15	1.83	2.26	2.20
TEMP (°C)	26.44	24.10	24.80	23.97
RH (%)	64.30	76.97	73.15	74.38
STWD (°)	21.59	26.16	26.39	23.43
#Samples	5240	4469	3799	2027

According to Table 10, cluster 1 samples exhibit, in general, a low concentration of air pollutants, except for O<sub>3</sub> and PM<sub>10</sub>, which have the highest average concentration. In addition, cluster 1 presents a wind speed and temperature considerably higher, and lower relative humidity. In total, about 34% of the data was assigned to cluster 1, thus sharing those characteristics.

Cluster 2, presented in Table 10, shows the lowest concentrations of SO<sub>2</sub>, CO, PM<sub>10</sub>, and NO pollutants, with intermediate values of O<sub>3</sub>, and NO<sub>2</sub>. It also presents the lowest average wind speed, intermediate temperature, and high relative humidity. In addition, cluster 2 is composed of 29% of the data, characterized by a low concentration of pollutants.

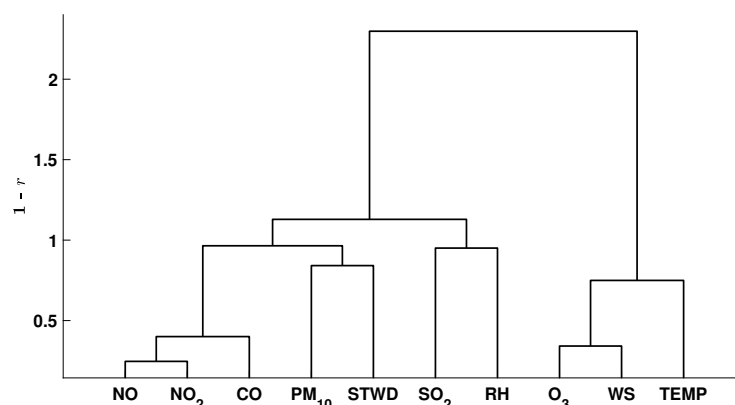
The highest concentrations of CO, PM<sub>10</sub>, NO, and NO<sub>2</sub> are found in cluster 3, as can be observed in Table 10. In contrast, SO<sub>2</sub> and O<sub>3</sub> show low values (with O<sub>3</sub> having the lowest total average among all clusters). The wind speed, temperature, and relative humidity have intermediate values. A total of 24% of the data was assigned to cluster 3, characterized by high pollutant concentration values.

Finally, cluster 4 is mainly characterized by the high concentration of the SO<sub>2</sub> pollutant compared to the others. The other pollutants present intermediate concentration values, as well as wind speed, temperature, and relative humidity. In addition, cluster 4 has the lower amount of samples; a total of 2027 (13%) were assigned here.

### 3.2.2. Parameter Correlation

The component planes allow an initial and preliminary analysis of parameters through their visual gradients which, in a certain way, can turn out to be subjective and discretionary. Thus, to carry out a more objective and effective analysis of the results, a correlation analysis was applied between the component planes seen in Figure 2. Figure 5 shows the similarity between the planes (parameters) using the Ward criterion and the Pearson correlation coefficient,  $r$ .

As can be observed in Figure 5, two main branches are seen in the correlation analysis. The first branch, on the left of the figure, includes all the pollutants studied but O<sub>3</sub>, whose origin is exclusively photochemical. Hence, O<sub>3</sub> is clustered with the wind speed and temperature.



**Figure 5.** Parameter correlation using Ward criterion and distance  $1 - r$ , where  $r$  is Pearson coefficient, for the Itaigara station.

The NO, NO<sub>2</sub>, and CO pollutants have a substantial similarity, probably due to a similar emission source such as vehicular, given the station allocation and the monitoring region. Those pollutants are correlated to PM<sub>10</sub>, which also has a vehicular origin. In addition, the PM<sub>10</sub> is connected to STWD, showing that intensive vertical turbulence (atmospheric instability), which is characterized by high STWD values, increases the PM<sub>10</sub> concentration. Thus, it can be said that the wind movement is dragging out PM<sub>10</sub> from other areas or causing the resuspension of particulate material at Itaigara station. In addition to vehicle influence, the particulate matter may also be dispersed by the existing vehicle movement, the wear of traffic lanes, and the vehicles' brake pads.

The similarity between RH and SO<sub>2</sub> shows the influence of RH on the formation or decomposition process of molecules during the heterogeneous procedure (liquid phase). In particular, the SO<sub>2</sub> can react with the air humidity and other oxidants in the atmosphere to form sulfuric acid H<sub>2</sub>SO<sub>4</sub> and ammonium sulfate [33].

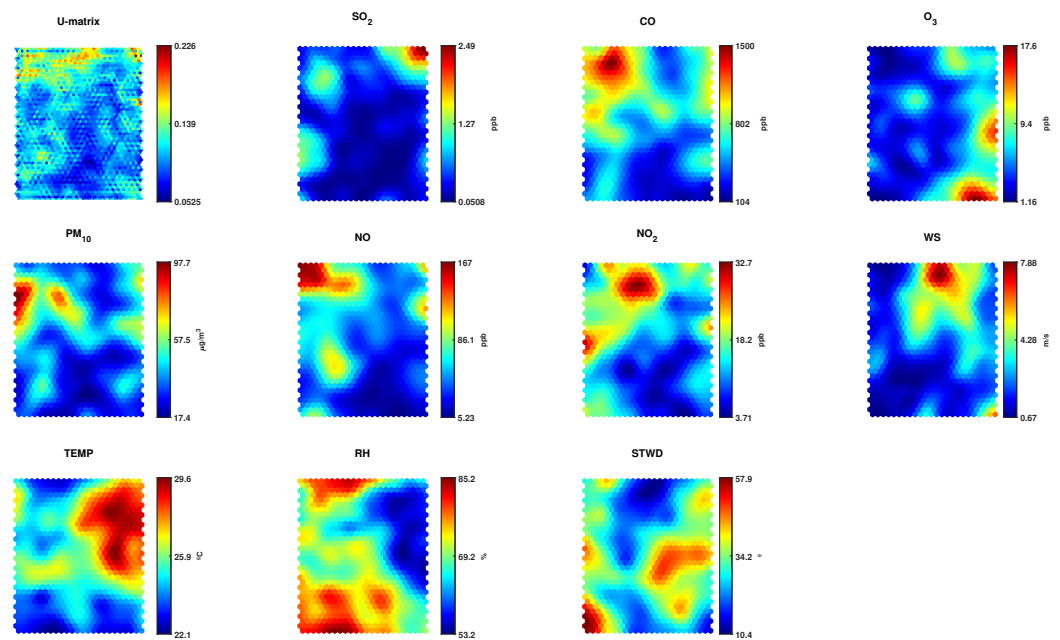
Meteorological parameters, such as wind speed, considerably influence the O<sub>3</sub> pollutant [24]. Given the similarity between O<sub>3</sub>, the wind speed, and temperature (Figure 5), we consider that O<sub>3</sub> is not generated at the monitoring station site but instead transported by winds along with other pollutants such as VOCs. The temperature may also be responsible, since high temperatures result from the increase in the speed of chemical processes, generating ozone in the region.

### 3.3. Barros Reis Station

In the BR station component planes (Figure 6), the weight vectors for the PM<sub>10</sub>, CO, NO, and NO<sub>2</sub> are displayed similarly across the map. The concentration of high values is on the upper left side, with average values in the nearby regions. The low values are located mainly in the lower right region of the map. All these pollutants can be formed from combustion processes, which shows the similarity obtained and, in particular, if they have a common source.

Unlike the pollutants discussed above, the O<sub>3</sub> component plane has its highest concentration at the bottom right of the map. O<sub>3</sub> is a secondary pollutant, i.e., its formation depends on atmosphere reactions from other pollutants, such as NO<sub>2</sub>. Still, its plane does not resemble the planes of primary pollutants. Similarly, PM<sub>10</sub> is also a secondary pollutant but is formed by SO<sub>2</sub>, and no similarity is seen in their plane. However, PM<sub>10</sub> can also be obtained from VOCs and nitrogen oxides, showing a relationship between their planes.

The SO<sub>2</sub> plane displays a unique pattern, with its highest values concentrated in the upper right region of the map, showing no similarity with the other pollutants. The component planes referring to meteorological parameters showed different distributions, with a negative correlation between TEMP and RH. At the same time, the high WS values are concentrated in the upper central region, and STWD with values dispersed throughout the map.

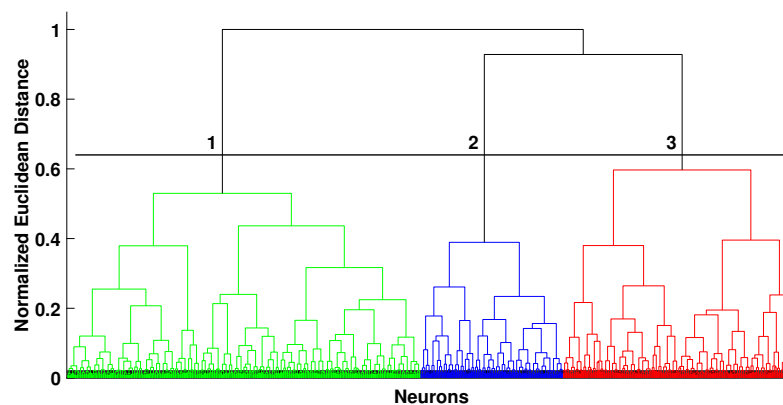


**Figure 6.** Unified distance matrix (U-matrix) and component planes of all analyzed variables ( $\text{SO}_2$ ,  $\text{CO}$ ,  $\text{O}_3$ ,  $\text{PM}_{10}$ ,  $\text{NO}$ ,  $\text{NO}_2$ ,  $\text{WS}$ ,  $\text{TEMP}$ ,  $\text{RH}$  and  $\text{STWD}$ ) from the Barros Reis station.

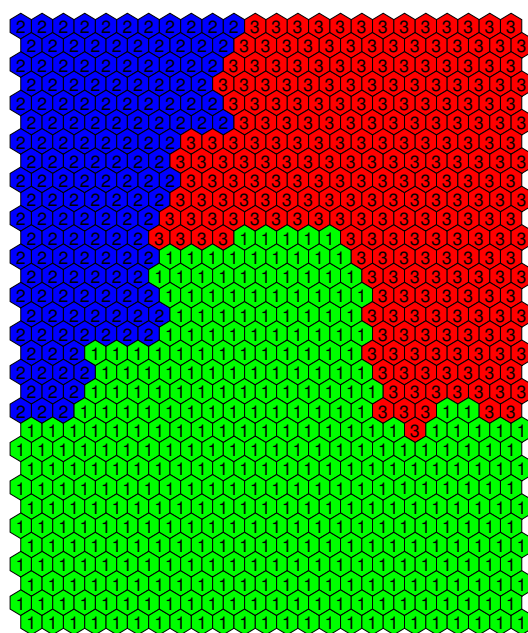
### 3.3.1. Sample Grouping with the SOM Algorithm

Figure 6 presents the clusters through the U-matrix, representing the neurons with their distance to adjacent neurons. The cluster number was defined with the Davies–Bouldin index by varying it from two to eight, reaching the best result for three clusters.

Subsequently, a hierarchical analysis was performed to define the neurons belonging to the three clusters. Thereupon, the Ward criterion and the Euclidean distance were used as similarity metrics. Figure 7 displays the dendrogram obtained with the threshold value used for segregation. Meanwhile, Figure 8 shows how the clusters were arranged on the map.



**Figure 7.** Hierarchical analysis of the neurons clusters using the Ward linkage method and Euclidean distance for the Barros Reis station.



**Figure 8.** SOM neurons grouped into three clusters obtained by the hierarchical analysis of the Barros Reis station.

The samples are linked to a particular neuron belonging to one of the three clusters, allowing the analysis of the sample's distribution regarding the clusters.

Table 11 shows the average values of every parameter according to the cluster. As can be seen, cluster 1 represents the samples with the lowest pollutant concentration, except for O<sub>3</sub> which has a median value among the others. Meteorological parameters such as wind speed, temperature, and relative humidity also have low values. In total, the cluster has 10,599 samples with these characteristics, corresponding to 49.16% of the station data.

**Table 11.** Parameters average values for every cluster formed by the SOM network for the Barros Reis station.

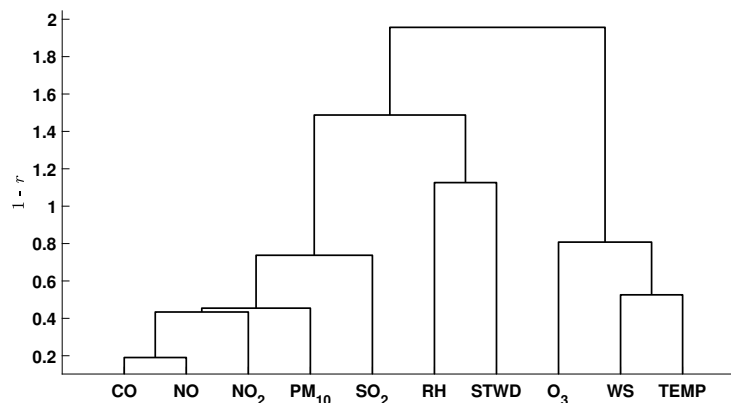
Parameters	Parameter Average Value per Cluster		
	1	2	3
SO <sub>2</sub> (ppb)	0.28	0.64	0.59
CO (ppb)	442.11	973.00	621.71
O <sub>3</sub> (ppb)	5.42	3.03	7.07
PM <sub>10</sub> (µg/m <sup>3</sup> )	33.31	54.00	42.13
NO (ppb)	37.54	91.22	51.88
NO <sub>2</sub> (ppb)	10.74	20.23	15.68
WS (m/s)	1.93	1.93	4.12
TEMP (°C)	24.66	24.74	27.68
RH (%)	72.64	72.19	60.05
STWD (°)	32.73	31.04	30.19
#Samples	10,599	4183	6777

In the meantime, cluster 2 exhibits the highest concentration of pollutants, displaying a considerable difference from the values of other clusters except for O<sub>3</sub>, which has the lowest average value obtained. Similar to cluster 1, the wind speed, temperature, and relative humidity also have low values. Cluster 2 has 4183 samples, equivalent to 19.40% of the data.

Finally, the samples assigned to cluster 3 have an intermediate value of pollutants concentration, with average values between the clusters 1 and 2 range, except for O<sub>3</sub> which has the highest average concentration recorded. In addition, cluster 3 has 31.44% of the station data with the highest wind speed and the lowest relative humidity.

### 3.3.2. Parameter Correlation

The component planes, shown in Figure 6, present the correlation between parameters. Meanwhile, Figure 9 presents the parameters similarity obtained using the Ward linking method and the Pearson correlation coefficient.



**Figure 9.** Parameter correlation using Ward criterion and distance  $1 - r$ , where  $r$  is Pearson coefficient, for the Barros Reis station.

As shown in Figure 9, there is a substantial similarity between CO and NO pollutants. Given the BR station characteristics (located in between two avenues), it can be said that motor vehicles are the primary emission source of those pollutants. Likewise, the NO<sub>2</sub> and PM<sub>10</sub> pollutants are also emitted by combustion in vehicles; in addition, they can be formed secondarily by photochemical processes. Regarding SO<sub>2</sub>, it can be said that the primary emission source is the burning process of fuels, such as diesel and gasoline, from heavy vehicles such as trucks, buses, microbuses, and light vehicles.

Unlike other pollutants, the O<sub>3</sub> showed a clear relationship with meteorological parameters such as wind speed and temperature, similar to the Itaipara station. Nonetheless, this relationship with meteorological parameters is not strong as in other stations.

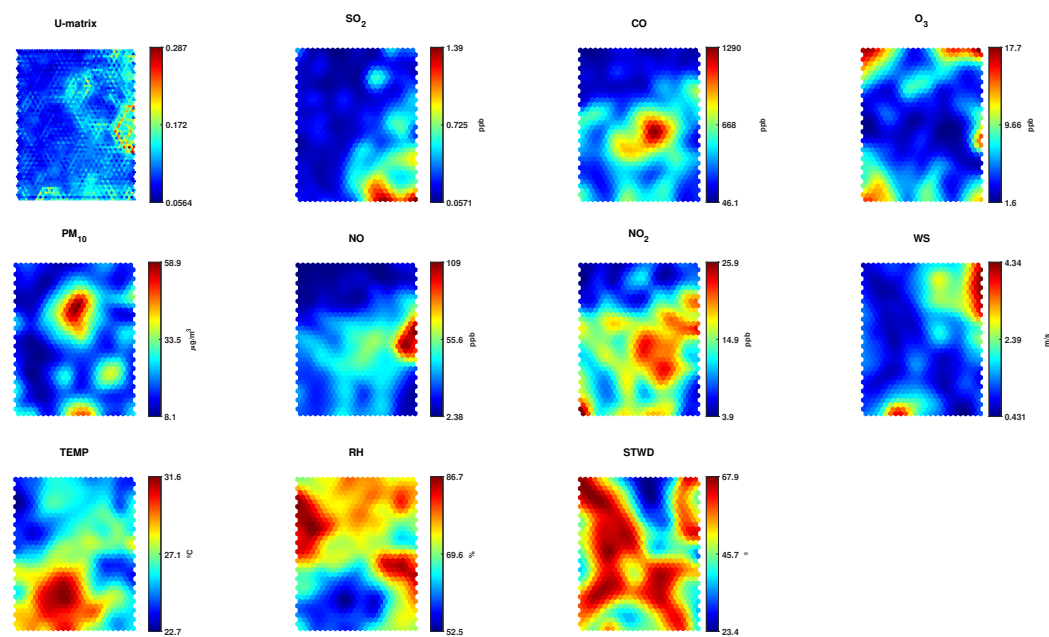
The STWD indicates the local atmospheric stability. Its inverse relationship with RH can be related to the regions' water molecules' dissipation. Hence, the data regarding pressure and heat could improve the analysis precision by demonstrating the influence of the wind direction. The RH and STWD present a negative relationship with the other pollutants, consequently leading to the non-contribution or reduction in the present concentrations.

### 3.4. Campo Grande Station

Figure 10 illustrates the component planes for the CG station. Concerning the planes of nitrogen oxide, a significant similarity between NO<sub>2</sub> and CO can be observed, with high values concentrated in the central part of the map. The NO plane is also similar to the CO and NO<sub>2</sub>, but the high values are concentrated in the region to the right, while median values are concentrated in the map center. The emission source of these pollutants is fuel combustion, especially from vehicles.

The SO<sub>2</sub> has high values concentrated in the lower right region of the map. The PM<sub>10</sub>, on the other hand, did not show significant pattern similarities with other planes, having a higher concentration in the upper part of the map and moderate concentration in the lower part, equivalent to small regions of the SO<sub>2</sub> and NO<sub>2</sub> planes. Likewise, the O<sub>3</sub> pollutant also shows no similarity with other component planes. Despite its formation, resulting from the reaction between NO<sub>2</sub> and VOCs, its concentration of high values is located at the map edges, having similarities with the concentration regions of high values of meteorological parameters, such as WS, TEMP, RH, and STWD.

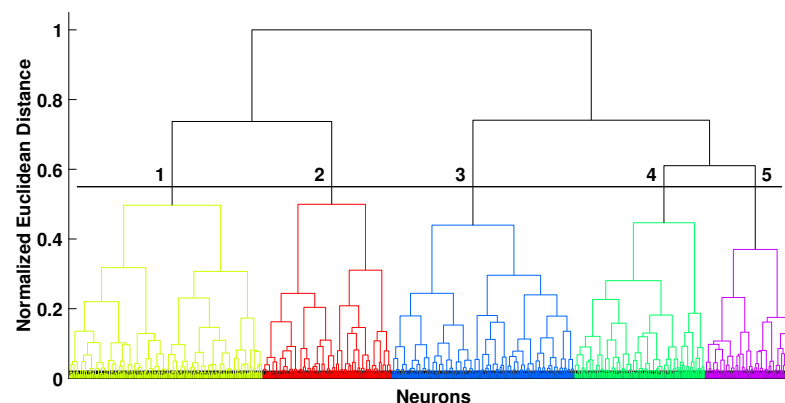




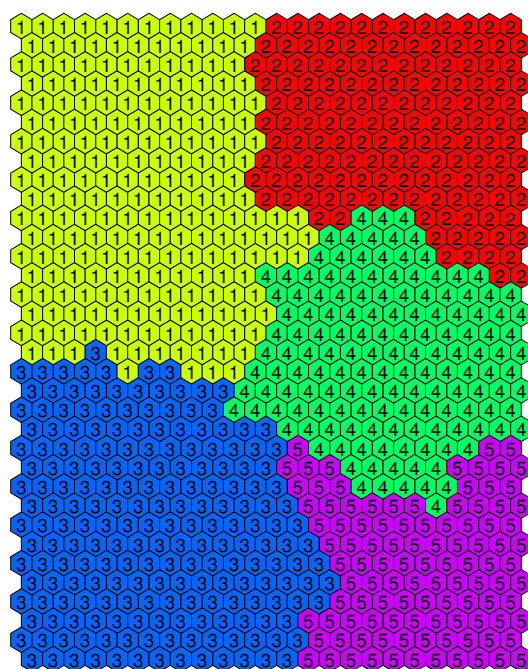
**Figure 10.** Unified distance matrix (U-matrix) and component planes of all analyzed variables ( $\text{SO}_2$ ,  $\text{CO}$ ,  $\text{O}_3$ ,  $\text{PM}_{10}$ ,  $\text{NO}$ ,  $\text{NO}_2$ ,  $\text{WS}$ ,  $\text{TEMP}$ ,  $\text{RH}$  and  $\text{STWD}$ ) from the Campo Grande station.

#### 3.4.1. Sample Grouping with the SOM Algorithm

To identify the CG station clusters through the U-matrix, illustrated in Figure 10, the Davies–Bouldin was used and the cluster number varied from two to eight. The best result was obtained for five clusters. Aftward, the neurons belonging to each cluster were obtained according to a hierarchical analysis defined based on the *Ward* method and Euclidean distance. Figure 11 shows the resulting dendrogram and the segregation threshold. Meanwhile, Figure 12 displays the neurons distribution regarding the clusters.



**Figure 11.** Hierarchical analysis of the neurons clusters using the Ward linkage method and Euclidean distance for the Campo Grande station.



**Figure 12.** SOM neurons grouped into five clusters obtained by the hierarchical analysis of the Campo Grande station.

Each CG station dataset sample was integrated into the cluster with the neuron it most resembles. Thus, an analysis was performed regarding the samples' distribution by cluster based on the average values of parameters, as shown in Table 12.

**Table 12.** Parameters average values for every cluster formed by the SOM network for the Campo Grande station.

Parameters	Parameter Average Value per Cluster				
	1	2	3	4	5
SO <sub>2</sub> (ppb)	0.14	0.26	0.23	0.38	0.82
CO (ppb)	250.98	269.62	456.78	655.67	404.37
O <sub>3</sub> (ppb)	6.12	7.07	6.59	4.15	5.72
PM <sub>10</sub> (µg/m <sup>3</sup> )	20.01	23.71	17.64	24.45	22.21
NO (ppb)	15.07	18.97	30.47	56.49	24.42
NO <sub>2</sub> (ppb)	10.34	11.12	14.33	19.15	13.12
WS (m/s)	0.89	2.70	1.40	1.36	0.94
TEMP (°C)	25.21	25.73	29.22	26.26	26.90
RH (%)	77.31	75.14	61.52	73.56	68.58
STWD (°)	57.82	42.83	53.05	50.50	52.28
#Samples	6640	4166	6223	4229	3301

According to Table 12, cluster 1 has the lowest average values of concentration for the SO<sub>2</sub>, CO, NO, and NO<sub>2</sub> pollutants, while the PM<sub>10</sub> and O<sub>3</sub> show intermediate values. Moreover, the wind speed and temperature are the lowest of all. Cluster 1 consists of 6640 data samples, equivalent to 27.04% of the dataset.

Cluster 2 also presents low average values of the concentrations of the pollutants, with values slightly higher than those obtained in cluster 1, except for the O<sub>3</sub> pollutant, which has a higher concentration average. Similar behavior can be seen for the meteorological parameters except for the wind speed, which shows the highest average among all clusters. In total, 16.96% of the data was assigned to cluster 2.

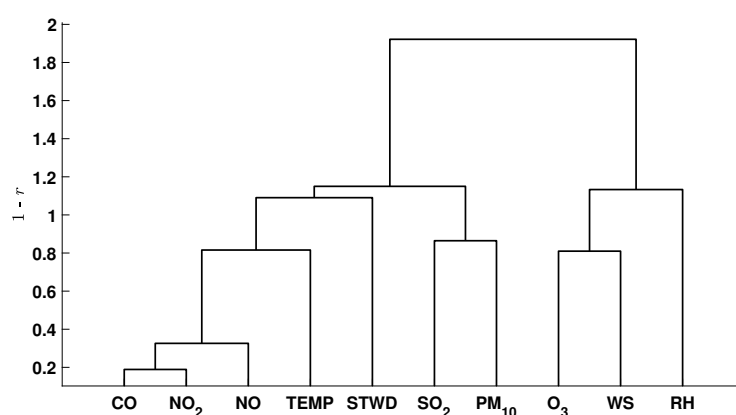
The samples assigned to cluster 3 present intermediate values for all pollutants concentration and meteorological parameters, where the temperature has the highest average and the relative humidity the lowest. This cluster has 25.34% of the data.

The highest concentrations of CO, PM<sub>10</sub>, NO, and NO<sub>2</sub> are found in cluster 4, with an intermediate concentration of SO<sub>2</sub> and the lowest concentration of O<sub>3</sub>. Meanwhile, all meteorological parameters showed intermediate values compared to other clusters. Cluster 4 has a total of 17.22% of the data.

Meantime, cluster 5 stands out with the highest average concentration of the SO<sub>2</sub> pollutant. The other pollutants, as well as the meteorological parameters, present intermediate average values. In total, 7.44% of the data was assigned to cluster 5.

### 3.4.2. Parameter Correlation

The hierarchical representation for the CG station was obtained with the *Ward* method and the *Pearson* correlation coefficient. Figure 13 presents the parameters' correlation obtained.



**Figure 13.** Parameter correlation using *Ward* criterion and distance  $1 - r$ , where  $r$  is *Pearson* coefficient, for the Campo Grande station.

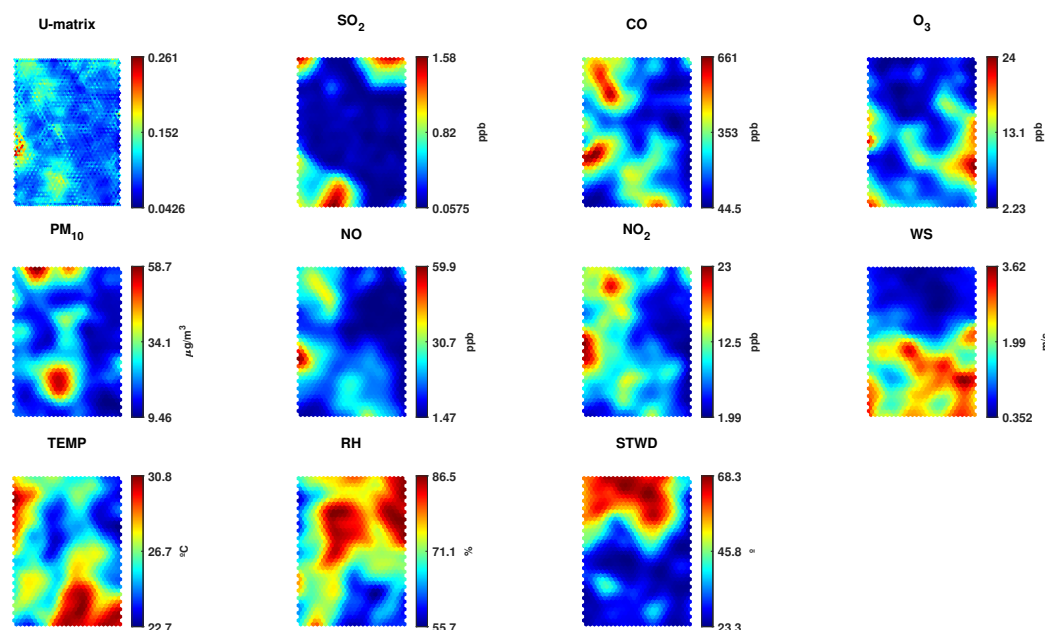
First of all, the similarity between CO, NO, and NO<sub>2</sub> pollutants can be seen. These pollutants are emitted in urban areas mainly by motor vehicles, and their similarity validates the idea of a potential common emission source. The temperature is also similar to those three pollutants as it contributes to chemical processes that form them—for example, the NO<sub>2</sub> results from the sunlight action on NO. Thus, the temperature can impact the amount of those pollutants present in every season.

The PM<sub>10</sub> is a primary and secondary pollutant, and it is correlated to SO<sub>2</sub>. Thus, its atmospheric formation can be linked to gases turning into particles due to chemical reactions in the air, such as sulfur dioxide. The SO<sub>2</sub> is generated from the burning of fuels with sulfur in its composition, such as diesel oil or industrial fuel oil, and it appears to be related to the PM<sub>10</sub> due to motor vehicle emissions, among other processes.

The photochemical oxidant, O<sub>3</sub>, has a certain correlation with the wind speed, but with a much lower similarity than that presented by the Itaigara station. In addition, there is no apparent relationship with the temperature. The RH has a negative relationship with O<sub>3</sub> and wind speed, which may be a consequence of solar radiation; low RH concentrations are related to a high solar incidence and, therefore, a greater disposition to the O<sub>3</sub> formation.

### 3.5. Dique do Tororó Station

The SOM network component planes for the DT station are shown in Figure 14. The pollutants that are mainly emitted by combustion processes, such as CO, NO, NO<sub>2</sub>, and PM<sub>10</sub> showed similar distribution patterns of values, with the highest concentration from the left side to the upper left side of the map. In contrast, the PM<sub>10</sub> has higher values at the bottom of the map, similar to the temperature and wind speed.

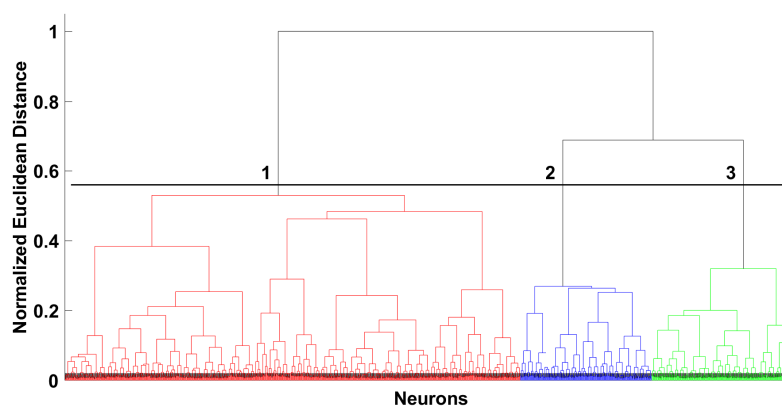


**Figure 14.** Unified distance matrix (U-matrix) and component planes of all analyzed variables ( $\text{SO}_2$ ,  $\text{CO}$ ,  $\text{O}_3$ ,  $\text{PM}_{10}$ ,  $\text{NO}$ ,  $\text{NO}_2$ ,  $\text{WS}$ ,  $\text{TEMP}$ ,  $\text{RH}$  and  $\text{STWD}$ ) from the Dique do Tororó station.

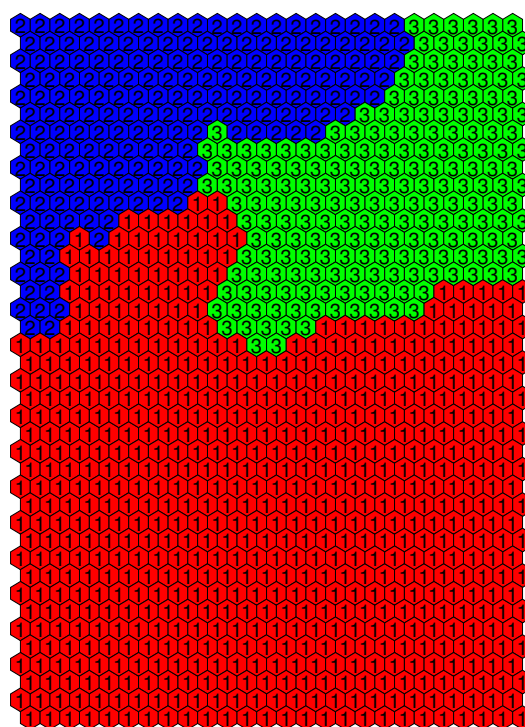
As in the other stations, the  $\text{SO}_2$  showed a different pattern from the other pollutants, with regions of high values concentration at the edges of the map. However, one of the high-concentration edges slightly coincides with those of the  $\text{CO}$ ,  $\text{NO}$ , and  $\text{NO}_2$ . Lastly, the  $\text{O}_3$  displays high values at the lower right region of the map, with a similar distribution to the wind speed plane. The other planes, such as relative humidity and  $\text{STWD}$  (which can influence the concentration of pollutants), showed patterns with well-defined regions at the top of the map.

### 3.5.1. Sample Grouping with the SOM Algorithm

The map neurons, represented by their respective distances to adjacent neurons in the U-matrix (Figure 14), were used to visualize and determine the clusters. For this purpose, the Davies–Bouldin index was used, and the number of clusters varied from two to eight, resulting in the best amount with three clusters. Again, hierarchical analysis was carried out using the Ward criterion and Euclidean distance. Figure 15 illustrates the dendrogram, and Figure 16 the segregation borders of the map.



**Figure 15.** Hierarchical analysis of the neurons clusters using the Ward linkage method and Euclidean distance for the Dique do Tororó station.



**Figure 16.** SOM neurons grouped into three clusters obtained by the hierarchical analysis of the Dique do Tororó station.

Each cluster was assigned a certain number of samples according to their characteristics. Table 13 presents the concentration averages of each pollutant according to the cluster.

**Table 13.** Parameters average values for every cluster formed by the SOM network for the Dique do Tororó station.

Parameters	Parameter Average Value per Cluster		
	1	2	3
SO <sub>2</sub> (ppb)	0.30	0.36	0.34
CO (ppb)	245.23	342.03	130.06
O <sub>3</sub> (ppb)	9.68	5.18	6.05
PM <sub>10</sub> (µg/m <sup>3</sup> )	21.23	29.66	18.23
NO (ppb)	15.14	20.06	3.91
NO <sub>2</sub> (ppb)	8.71	12.13	5.45
WS (m/s)	2.23	0.67	0.66
TEMP (°C)	26.90	27.22	24.40
RH (%)	69.65	71.88	81.69
STWD (°)	29.18	60.49	47.53
#Samples	26,101	7511	8425

As can be seen in Table 13, cluster 1 represents the samples with the highest mean value of O<sub>3</sub> and intermediate values of the other pollutants (SO<sub>2</sub>, CO, NO, NO<sub>2</sub>, and PM<sub>10</sub>). The highest concentration value is the wind speed, while relative humidity and STWD are the lowest. Cluster 1 has 26,101 samples sharing its characteristics, equivalent to 62.09% of the station data.

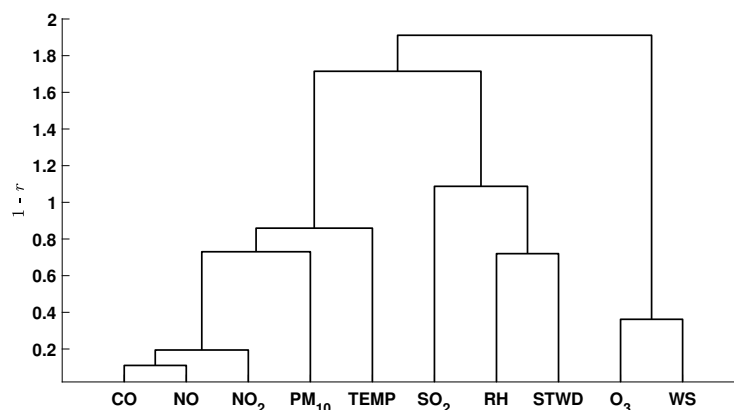
The pollutants in cluster 2 had the highest average concentration, except for O<sub>3</sub> which showed the lowest concentration among all clusters. The wind speed presents low values, and the temperature parameter is the highest. In total, 17.87% of data constitutes this cluster.

Finally, the pollutants in cluster 3, that is, CO, NO, NO<sub>2</sub>, and PM<sub>10</sub>, had the lowest average concentrations, with SO<sub>2</sub> and O<sub>3</sub> showing intermediate values. The temperature

and wind speed parameters have the lowest values found and the relative humidity the highest. Cluster 3 represents 20.04% of the station data with 8425 samples.

### 3.5.2. Parameter Correlation

The DT station component planes, shown in Figure 14, presents the parameters correlation. Meanwhile, Figure 17 illustrates the parameter similarity obtained through the *Ward* criterion and the *Pearson* correlation coefficient.



**Figure 17.** Parameter correlation using *Ward* criterion and distance  $1 - r$ , where  $r$  is *Pearson* coefficient, for the Dique do Tororó station.

As can be seen in Figure 17, the CO, NO, and NO<sub>2</sub> pollutants have the most significant similarity, a characteristic also observed for other stations. All stations are located in urban centers with a large flow of vehicles, leading to the possibility of a common emission source of these pollutants, mainly coming from the local vehicular fleet. The PM<sub>10</sub> also showed a certain similarity with those pollutants, indicating a possible emission from fuel burning. The temperature parameter at the DT is also related to the mentioned pollutants, different from other stations where it is associated with O<sub>3</sub>. In addition, the temperature can contribute to NO<sub>2</sub> formation and PM<sub>10</sub> in secondary processes.

Like Barros Reis station, the RH and the STWD at the DT station are somewhat similar but with a positive coefficient. The RH and STWD can be influenced by atmospheric parameters such as pressure and heat and, consequently, the wind conditions and water particles.

The SO<sub>2</sub>, different from the Itaigara station, is not correlated to either the PM<sub>10</sub> or RH, as it is probably being generated by an independent source and not reacting to other pollutants.

Given that O<sub>3</sub> is a secondary pollutant, it was only correlated with wind speed, with no apparent similarity with temperature or nitrogen oxides. Therefore, its concentration at the DT station may be transported by the wind accompanied by other pollutants.

## 4. Discussion

The SOM implementation presented in the previous sections identifies the correlation among different air quality parameters for many monitoring stations. The SOM component planes provide a visual representation of the similarities between pollutants and meteorological parameters, simplifying their analysis and highlighting peculiarities.

Usually, the CO, NO, and NO<sub>2</sub> pollutants were related, showing higher similarities. On the other hand, the meteorological parameters differed from PM<sub>10</sub> and SO<sub>2</sub>. The RH and STWD parameters at Barros Reis station showed a negative correlation, unlike at Dique do Tororo station, where a positive correlation was presented. At Itaigara station, the influence of atmospheric stability was identified through the relationship between STWD and PM<sub>10</sub>. Meanwhile, Campo grande station shows some degree of similarity between PM<sub>10</sub> and SO<sub>2</sub>. These relations are essential to identify the influence of meteorology on air-pollutants concentrations and information employed to create strategies for mitigating air-pollution critical episodes.

Unlike other pollutants, the  $O_3$  presents a more significant link with meteorological parameters such as WS, as seen at Itaigara, Dique do Tororó and Campo grande stations. Thus, we can infer that the wind is mainly responsible for the transport of  $O_3$ . In addition, the correlation of the TEMP, WS, and  $O_3$  parameters at Barros Reis and Itaigara stations indicates an increase in  $O_3$  resulting from chemical processes, probably due to the influence of solar radiation.

The data of Dique do Tororo and Barros Reis stations were grouped only into three clusters, with their cluster 1 emphasizing a large number of samples with higher concentrations of  $O_3$ . In contrast, the other clusters present a sample distribution with intermediate to high concentrations for the CO, NO,  $NO_2$ ,  $PM_{10}$ , and  $SO_2$  pollutants. Meanwhile, the Itaigara station has four clusters, with one mainly characterized by the  $SO_2$  pollutant; the remaining clusters are defined by higher concentrations of CO, NO,  $NO_2$ ,  $PM_{10}$ , and  $O_3$ . Similar to Itaigara, the Campo Grande station has one cluster (out of five) where  $SO_2$  is predominant, while the other clusters display low and high concentrations.

Commonly, studies about atmospheric pollutants rely on methods such as principal component analysis (PCA) and hierarchical analysis to define clusters based on similarity. For example, the studies carried out by [8,34] describe the clusters' characteristics according to the percentage of their main components' variance, thus, indicating which variables have more significance for their definition. Meanwhile, by applying a hierarchical classification on the SOM neurons, we can obtain the variables' concentration value and influence on defining each cluster.

In the meantime, in [13,35], the number of clusters is fixed for all monitoring stations, and the k-nearest neighbors provide a relationship between the defined clusters of each station. However, the SOM also allows an individual characteristic analysis of each pollutant, like in [35].

Thereby, the SOM enables finding similarities and estimating the link between parameters more deeply. As described in this work, the SOM can obtain data patterns and cluster characteristics and demonstrate the parameters' influence, which is not trivial in other techniques. Additionally, it can also deal with the non-linearity complexity of air pollution data [36], simplifying the analysis process and increasing its precision; this shows the advantage of using a machine-learning-based approach compared to traditional methods.

## 5. Conclusions

We implemented an SOM to analyze the air-quality data of four stations in the monitoring network of Salvador, Brazil. A detailed discussion regarding pollutants and their correlation with meteorological parameters is provided, assisting in estimating possible common emission sources and the influence of meteorological parameters. The latter permits the establishment of relations between meteorology and pollutants concentration, which is vital for developing, for example, alert systems to identify critical episodes of air pollution or for assisting in developing strategies to improve air quality.

The SOM outputs enabled the identification of data particularities concerning the parameters analyzed. For example, the data samples' concentration of Dique do Tororo and Barros Reis stations showed a cluster with a high concentration of  $O_3$ . In contrast, the other clusters presented well-defined contributions of remaining pollutants. The Itaigara and Campo Grande stations presented a more detailed definition regarding the clusters of (1) CO, NO,  $NO_2$ ; (2)  $MP_{10}$ ; (3)  $O_3$ ; and (4)  $SO_2$ . Thus, the SOM also allows an analysis of the particularities of each cluster.

The results showed that the SOM could identify characteristics, describe similarities, recognize patterns, and define clusters of air-pollution problems. Unlike traditional methods, the SOM proved to be a good tool for studying atmospheric pollutants, providing several aspects that can contribute to and improve discussions in this area. To the best of our knowledge, this is the first study to analyze Salvador's air-quality monitoring database. Therefore, the tool developed and the results presented and discussed here can assist further studies and aid in the development of public policies for pollution management.



**Author Contributions:** All the authors have contributed in various degrees to ensure the quality of this work (e.g., E.L.R.C., T.B., L.A.D., É.L.d.A. and M.A.C.F. conceived the idea and experiments; E.L.R.C., T.B., L.A.D., É.L.d.A. and M.A.C.F. designed and performed the experiments; E.L.R.C., T.B., L.A.D., É.L.d.A. and M.A.C.F. analyzed the data; E.L.R.C., T.B., L.A.D., É.L.d.A. and M.A.C.F. wrote the paper. É.L.d.A. and M.A.C.F. coordinated the project). All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)—Finance Code 001.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors wish to acknowledge the financial support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for their financial support. The authors want to thank the “CETREL S. A. Company and Bahia State Government” for the availability of the monitoring data in Salvador.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Landrigan, P.J.; Fuller, R.; Acosta, N.J.R.; Adeyi, O.; Arnold, R.; Basu, N.N.; Baldé, A.B.; Bertollini, R.; Bose-O'Reilly, S.; Boufford, J.I.; et al. The Lancet Commission on pollution and health. *Lancet* **2017**, *391*, 462–512. [\[CrossRef\]](#)
- Zivin, J.G.; Neidell, M. Air pollution's hidden impacts. *Science* **2018**, *359*, 39–40. [\[CrossRef\]](#) [\[PubMed\]](#)
- Turner, M.C.; Andersen, Z.J.; Baccarelli, A.; Diver, W.R.; Gapstur, S.M.; Pope, C.A., III; Prada, D.; Samet, J.; Thurston, G.; Cohen, A. Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *CA A Cancer J. Clin.* **2020**, *70*, 460–479. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhang, J.; Zhang, L.; Du, M.; Zhang, W.; Huang, X.; Zhang, Y.; Yang, Y.; Zhang, J.; Deng, S.; Shen, F.; et al. Identifying the major air pollutants base on factor and cluster analysis, a case study in 74 Chinese cities. *Atmos. Environ.* **2016**, *144*, 37–46. [\[CrossRef\]](#)
- Zhang, K.; Batterman, S. Air pollution and health risks due to vehicle traffic. *Sci. Total Environ.* **2013**, *450–451*, 307–316. [\[CrossRef\]](#)
- Bai, L.; Wang, J.; Ma, X.; Lu, H. Air Pollution Forecasts: An Overview. *Int. J. Environ. Res. Public Health* **2018**, *15*, 780. [\[CrossRef\]](#)
- Núñez-Alonso, D.; Pérez-Arribas, L.V.; Manzoor, S.; Cáceres, J.O. Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region. *J. Anal. Methods Chem.* **2018**, *2019*, 9753927. [\[CrossRef\]](#)
- Tian, D.; Fan, J.; Jin, H.; Mao, H.; Geng, D.; Hou, S.; Zhang, P.; Zhang, Y. Characteristic and Spatiotemporal Variation of Air Pollution in Northern China Based on Correlation Analysis and Clustering Analysis of Five Air Pollutants. *J. Geophys. Res. Atmos.* **2020**, *125*, e2019JD031931. [\[CrossRef\]](#)
- Manimaran, P.; Narayana, A.C. Multifractal detrended cross-correlation analysis on air pollutants of University of Hyderabad Campus, India. *Phys. A Stat. Mech. Its Appl.* **2018**, *502*, 228–235. [\[CrossRef\]](#)
- Bai, Y.; Jin, X.; Wang, X.; Wang, X.; Xu, J. Dynamic Correlation Analysis Method of Air Pollutants in Spatio-Temporal Analysis. *Int. J. Environ. Res. Public Health* **2020**, *17*, 360. [\[CrossRef\]](#)
- Zhao, S.; Yu, Y.; Yin, D.; He, J.; Liu, N.; Qu, J.; Xiao, J. Annual and diurnal variations of gaseous and particulate pollutants in 31 provincial capital cities based on in situ air quality monitoring data from China National Environmental Monitoring Center. *Environ. Int.* **2016**, *86*, 92–106. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yin, D.; Zhao, S.; Qu, J. Spatial and seasonal variations of gaseous and particulate matter pollutants in 31 provincial capital cities, China. *Air Qual. Atmos. Health* **2016**, *10*, 359–370. [\[CrossRef\]](#)
- Li, C.; Wang, Z.; Li, B.; Peng, Z.; Fu, Q. Investigating the relationship between air pollution variation and urban form. *Build. Environ.* **2019**, *147*, 559–568. [\[CrossRef\]](#)
- Periš, N.; Buljac, M.M.B.; Buzuk, M.; Brinić, S.; Plazibat, I. Characterization of the Air Quality in Split, Croatia Focusing Upon Fine and Coarse Particulate Matter Analysis. *Anal. Lett.* **2015**, *48*, 553–565. [\[CrossRef\]](#)
- Wang, C.; Zhao, L.; Sun, W.; Xue, J.; Xie, Y. Identifying redundant monitoring stations in an air quality monitoring network. *Atmos. Environ.* **2018**, *190*, 256–268. [\[CrossRef\]](#)
- Ran, Z.Y.; Hu, B.G. Parameter Identifiability in Statistical Machine Learning: A Review. *Neural Comput.* **2017**, *29*, 1151–1203. [\[CrossRef\]](#)
- Capizzi, G.; Sciuto, G.L.; Monforte, P.; Napoli, C. Cascade Feed Forward Neural Network-based Model for Air Pollutants Evaluation of Single Monitoring Stations in Urban Areas. *Neural Comput.* **2015**, *61*, 327–332. [\[CrossRef\]](#)
- Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2001.
- Asan, U.; Ercan, S. An Introduction to Self-Organizing Maps. In *Computational Intelligence Systems in Industrial Engineering: With Recent Theory and Applications*; Atlantis Press: Paris, France, 2012; pp. 295–315. [\[CrossRef\]](#)

20. Pearce, J.L.; Waller, L.A.; Chang, H.H.; Klein, M.; Mulholland, J.A.; Sarnat, J.A.; Sarnat, S.E.; Strickland, M.J.; Tolbert, P.E. Using self-organizing maps to develop ambient air quality classifications: A time series example. *Environ. Health* **2014**, *11*, 56. [\[CrossRef\]](#)
21. Zhong, B.; Wang, L.; Liang, T.; Xing, B. Pollution level and inhalation exposure of ambient aerosol fluoride as affected by polymetallic rare earth mining and smelting in Baotou, north China. *Atmos. Environ.* **2017**, *167*, 40–48. [\[CrossRef\]](#)
22. Jiang, N.; Scorgie, Y.; Hart, M.; Riley, M.L.; Crawford, J.; Beggs, P.J.; Edwards, G.C.; Chang, L.; Salter, D.; Virgilio, G.D. Visualising the relationships between synoptic circulation type and air quality in Sydney, a subtropical coastal-basin environment. *Int. J. Climatol.* **2017**, *37*, 1211–1228. [\[CrossRef\]](#)
23. Moosavi, V.; Aschwanden, G.; Velasco, E. Finding candidate locations for aerosol pollution monitoring at street level using a data-driven methodology. *Atmos. Meas. Tech.* **2015**, *8*, 3563–3575. [\[CrossRef\]](#)
24. Li, D.; Liao, Y. Pollution zone identification research during ozone pollution processes. *Environ. Monit. Assess.* **2020**, *192*, 591. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Fávero, L.P.L.; Belfiore, P.P. *Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel, SPSS e Stata*, 1st ed.; Elsevier: Rio de Janeiro, Brazil, 2017.
26. Kohonen, T.; Oja, E.; Simula, O.; Visa, A.; Kangas, J. Engineering applications of the self-organizing map. *Proc. IEEE* **1996**, *84*, 1358–1384. [\[CrossRef\]](#)
27. Vesanto, J.; Alhoniemi, E. Clustering of the self-organizing map. *IEEE Trans. Neural Netw.* **2000**, *11*, 586–600. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Pözlbauer, G. Survey and Comparison of Quality Measures for Self-Organizing Maps. In *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*; Elfa Academic Press: Vysoké Tatry, Slovakia, 2004; pp. 67–82.
29. Kiviluoto, K. Topology preservation in self-organizing maps. In *Proceedings of the Proceedings of International Conference on Neural Networks (ICNN'96)*, Washington, DC, USA, 3–6 June, 1996; Volume 1, pp. 294–299. [\[CrossRef\]](#)
30. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *1*, 224–227. [\[CrossRef\]](#)
31. Li, T.; Sun, G.; Yang, C.; Liang, K.; Ma, S.; Huang, L. Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes. *Sci. Total. Environ.* **2018**, *628–629*, 1446–1459. [\[CrossRef\]](#)
32. Li, Y.; Wright, A.; Liu, H.; Wang, J.; Wang, G.; Wu, Y.; Dai, L. Land use pattern, irrigation, and fertilization effects of rice-wheat rotation on water quality of ponds by using self-organizing map in agricultural watersheds. *Agric. Ecosyst. Environ.* **2019**, *272*, 155–164. [\[CrossRef\]](#)
33. Turalioğlu, F.S.; Nuhoglu, A.; Bayraktar, H. Impacts of some meteorological parameters on SO<sub>2</sub> and TSP concentrations in Erzurum, Turkey. *Chemosphere* **2005**, *59*, 1633–1642. [\[CrossRef\]](#)
34. Dominick, D.; Juahir, H.; Latif, M.T.; Zain, S.M.; Aris, A.Z. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmos. Environ.* **2012**, *60*, 172–181. [\[CrossRef\]](#)
35. Iizuka, A.; Shirato, S.; Mizukoshi, A.; Noguchi, M.; Yamasaki, A.; Yanagisawa, Y. A Cluster Analysis of Constant Ambient Air Monitoring Data from the Kanto Region of Japan. *Int. J. Environ. Res. Public Health* **2014**, *11*, 6844. [\[CrossRef\]](#)
36. Yeganeh, B.; Motlagh, M.; Rashidi, Y.; Kamalan, H. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. *Atmos. Environ.* **2012**, *55*, 357–365. [\[CrossRef\]](#)