Revista Brasileira de Ciências Ambientais
Brazilian Journal of Environmental Sciences

Application of machine learning algorithms to $PM_{2.5}$ concentration analysis in the state of São Paulo, Brazil

Aplicação de algoritmos de aprendizado de máquina à análise de concentrações de $MP_{2,5}$ no Estado de São Paulo, Brasil

Angela Rosa Locateli Godoy[1], *Ana Estela Antunes da Silva*[2], *Mirelle Candida Bueno*[3], *Simone Andréa Pozza*[2], *Guilherme Palermo Coelho*[2]

ABSTRACT

Air quality monitoring data are useful in different areas of research and have varied applications, especially with a focus on the relationship between air pollution, respiratory problems, and other health hazards. The main atmospheric pollutants are: ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), and particulate matter (PM). PM is one of the main objects of study when one intends to protect people from exposure to pollutants. This study contributes to the analysis of $PM_{2.5}$ in 21 stations in the state of São Paulo monitored by the Environmental Company of São Paulo State (CETESB). It employs cluster analysis, a prominent data mining method for detecting patterns and discovering similarities which is important for assessing air pollution, especially in a geographically vast area such as that of the state of São Paulo, which does not follow a single pattern. Another data mining technique (association rules) supports the analysis of the relationship between pollutants and meteorological variables, as it allows identifying changes between elements that occur together, in a wide variety of data. Our objectives include determining stations with similar behaviors and exploring the temporal variety of the pollutant as it relates to the dominant meteorological factors in the periods of high concentration. The clustering algorithm automatically separates stations according to their monthly averages of $PM_{2.5}$ concentration between 2017 and 2019. The clusters of stations that showed the highest pollution rates essentially included urban centers with emissions by industries and vehicles, while those with the lowest rates were located further inland. A cyclical behavior in pollutant variation was also observed in the three years under study and for both clusters. For the months with the highest concentration of $PM_{2.5}$, association rule learning was applied to connect air temperature, relative humidity, and wind speed with $PM_{2.5}$ and carbon monoxide (CO) concentrations. The obtained results are useful to analyze the temporal and geolocation profiles of pollution by particulate matter, since they identify the behavior of the meteorological factors that predominate in periods of greater concentration.

Keywords: air pollutants; particulate matter; clustering; association rules; air quality; respiratory diseases.

RESUMO

Dados de monitoramento da qualidade do ar são úteis em diferentes áreas de pesquisa e aplicações, como por exemplo, no estudo da relação da poluição do ar com problemas respiratórios e outros prejuízos à saúde. Dentre os principais poluentes atmosféricos estão: ozônio (O3), dióxido de enxofre (SO2), monóxido de carbono (CO), dióxido de nitrogênio (NO2) e material particulado (MP). O MP é um dos principais objetos de estudos quando se pretende proteger as pessoas da exposição a poluentes. O presente trabalho contribui com a análise da concentração do poluente MP2,5, em 21 estações de monitoramento, observadas pela CETESB - Companhia Ambiental do Estado de São Paulo. Este estudo emprega mineração de dados por agrupamento, um método proeminente para reconhecer padrões e descobrir semelhanças, aspectos importantes para avaliar a poluição do ar, principalmente em uma área geograficamente vasta como o estado de São Paulo, que não segue um único padrão. A técnica de mineração por regras de associação, também aplicada, oferece suporte na análise da relação de poluentes com variáveis meteorológicas, por permitir identificar associações entre elementos que ocorrem juntos, em uma grande variedade de dados. Os objetivos incluem identificar estações com comportamentes semelhantes e explorar a variedade temporal do poluente relacionada aos fatores meteorológicos dominantes nos períodos de alta concentração. O algoritmo de agrupamento, separa de forma automática as estações a partir de médias mensais de concentração de MP2,5 nos anos de 2017 a 2019. Os grupos de estações com maiores índices encontrados do poluente foram os centros urbanos, com emissões por indústrias e veículos e, as estações com índices menores foram as localizadas mais ao interior do estado. Também houve a identificação de um ciclo sazonal nas variações do poluente nos três anos para os dois grupos. Para os meses de maior concentração de MP2,5 a técnica de regras de associação foi aplicada a fim de relacionar temperatura do ar, umidade relativa do ar e velocidade do vento, às concentrações dos poluentes MP2,5 e CO. Os resultados gerados são úteis na análise do perfil temporal e por geolocalização da poluição por material particulado e identifica o comportamento dos fatores meteorológicos que predominam nos períodos de maior concentração.

Palavras-chave: poluentes atmosféricos; agrupamentos; regras de associação; qualidade do ar; doenças respiratórias.

[1]Ph.D. student at the Graduate Program in Technology of Universidade Estadual de Campinas (Unicamp) – Campinas (SP), Brazil.

[2]Assistant professor at Unicamp – Campinas (SP), Brazil.

[3]Information Systems bachelor at Unicamp – Campinas (SP), Brazil.

Correspondence address: Angela Rosa Locateli Godoy – Paschoal Marmo St., 1888 – Jardim Nova Itália – CEP: 13484-332 – Limeira (SP), Brazil. E-mail: angelarl@unicamp.br

Conflicts of interest: nothing to declare.

Funding: none.

Received on: 05/01/2020. Accepted on: 08/11/2020

https://doi.org/10.5327/Z21769478782

Introduction

In the world population, nine out of 10 people breathe polluted air, according to the annual report of the World Health Organization (WHO). Every year, seven million people die worldwide by causes directly related to air pollution, but contamination levels remain high (WHO, 2019).

According to the Environmental Company of São Paulo State (CETESB, 2019), the main air pollutants regulated by the National Environment Council (CONAMA) are: coarse inhalable particles (PM_{10}), fine inhalable particles ($PM_{2.5}$), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), total suspended particles (TSP), smoke (SMO), and lead (Pb), the latter three being monitored only in specific situations. Studies on the effects of pollution on health (POLEZER *et al*., 2018; MACHIN; NASCIMENTO, 2018; SEINFELD; PANDIS, 2016; NODARI; SALDANHA, 2016) show that exposure to fine particulate matter ($PM_{2.5}$) can cause respiratory problems and even premature deaths, since it penetrates deeply into the respiratory system, reaching the pulmonary alveoli and the bloodstream.

Because it is associated with damage to human health and has impacts on climate and the environment, $PM_{2.5}$ was chosen as the study object in this research. PM are particles suspended in the atmosphere, solid or liquid, which can be generated by several sources, in different sizes and compositions (DIMITRIOU, 2016; ANDRADE *et al*., 2012; QUALAR, 2019). It is classified by its aerodynamic diameter (a_d): particles with $a_d \leq 2.5$ μm are named $PM_{2.5}$ (fine inhalable particulate matter) and those with $10 \geq a_d > 2.5$ μm, as PM_{10} (coarse inhalable particulate matter). These pollutants can come from several sources, such as vehicles, industries, power plants, and fires in general. Despite the PM origin, it may be transported by air masses between cities, by atmospheric circulation (NOGAROTTO, 2019).

Meteorological variables directly interfere with the concentration of atmospheric pollutants by controlling the dispersion process of substances that are toxic and carcinogenic or that potentiate harmful effects on the environment and health (YANAGI; ASSUNÇÃO; BARROZO, 2012). The relationship between pollutant concentration and meteorological variables such as: air temperature (TEMP), relative humidity (RH), wind speed (WS), wind direction (WD), precipitation (PRE), atmospheric instability, and others that vary during the year is well known (GUERRA; MIRANDA, 2011). Given this relationship, studies such as the one by Bisht and Seeja (2018), in India, predict next-day air quality from the previous day's pollutant concentration data (PM_{10}, $PM_{2.5}$, NO_2, CO, and O_3) and meteorological variables (RH, PRE, TEMP, WS, and WD), using regression models. Gonçalves *et al.* (2005), in a research study in the city of São Paulo, proved that during summer, hot and humid days favor the decrease of PM_{10}, SO_2, and O_3 concentrations.

In winter, air quality worsens, especially regarding PM and CO concentrations, since weather conditions in this season of the year are less favorable for their dispersion (SANTOS; CARVALHO; REBOITA, 2016; MORAES *et al*., 2019; CETESB, 2019). Therefore, the interaction between atmospheric conditions and sources of pollution defines air quality, which in turn determines the emergence of adverse effects on people's health.

A study by Abe and Miraglia (2018) shows a reduction of about 25.45% in $PM_{2.5}$ concentration in the city of São Paulo from 2000 to 2011, due to actions to contain the increase in the automotive fleet. Typically, in metropolitan regions, motor vehicles are a major cause of air pollution. A study by Andrade *et al.* (2012) states that vehicle emissions, biomass burning, and fuel combustion in industries explain at least 40% of $PM_{2.5}$ in six Brazilian states, including São Paulo.

In addition to associating air pollutants with meteorological variables, it is also possible to establish a relation between the behaviors of different air pollutants. Moisan, Herrera and Clements (2018) reported an association between car pollution and firewood burning as regards CO concentration in the atmosphere, noting that 54% of $PM_{2.5}$ concentration is composed of CO, which shows a direct relationship between these pollutants. They also found a strong negative correlation with the variables TEMP and WS, in addition to a positive relationship with RH. Saide *et al.* (2011) developed a CO forecasting system as a substitute for PM_{10} and $PM_{2.5}$, identifying a high correlation (of above 0.95) between these pollutants in Santiago (Chile), during winter nights. Therefore, by predicting CO, an estimate of PM could be obtained. The greatest benefit of the study was its ability to predict critical episodes up to 48 hours ahead. Reinhardt, Ottmar and Castilla (2011) observed that, in Brazil, the concentration levels of CO and particulate matter are correlated and that, during the burning season, CO levels in rural areas are comparable to those of urban centers, moderately polluted.

Considering this scenario, it is important to investigate the behavior of pollutants, in particular $PM_{2.5}$. Despite the fact that the problem is widely discussed in various spheres of the scientific community, the literature lacks studies whose assessment uses artificial intelligence techniques and involves knowledge about the associations between pollutants, emission sources, and their effects on air quality (AMEER *et al*., 2019). The analysis of the sources of pollution by $PM_{2.5}$ throughout the state of São Paulo is considered a zoning problem, zoning being the discovery of different regions with similar characteristics. Data clustering technique is a prominent method for recognizing new patterns, and it is applied in exploratory data analysis. It is a suitable solution when searching for similar patterns and behaviors in different regions, which leads to the discovery of previously unknown clusters (HAN; KAMBER; PEI, 2011; KWEDLO, 2011).

Research carried out in Brazil (NODARI; SALDANHA, 2016; GUIDETTI; PEREDA, 2018) and in other countries which applied clustering techniques identified regions with similar patterns of air pollution. A study in China (XIAO *et al*., 2020) performed cluster analysis to measure similarities in the characteristics of industrial emissions from 31 companies in different regions; results showed that pollution characteristics were similar for companies in the same cluster, which contributed to the development of specific measures for pollution control. Also in China, studies involving 13 sites with similar $PM_{2.5}$ concentration data resulted in the discovery of three clusters: two of industrial activities and another of agricultural and tourist activities (HUANG *et al.*, 2015).

In the United States, a research study clustered locations according to $PM_{2.5}$ levels and obtained clusters by regions with similar industrial activity (AUSTIN *et al.*, 2013). The study by Zou *et al.* (2014), conducted with data from the U.S. urban census, was used to investigate the population's exposure to air pollution, considering age, race, education level, and income. By applying a spatial clustering method, it was possible to show disparities in the spatial distribution of exposure to pollution throughout the territory.

Alternatively, clustering technique is also used as a preprocessing step for selecting attributes or applying other data mining algorithms. An example is the study by Du and Varde (2016), which applies association rules, clustering, and classification to identify relationships between particulate matter, pollution, and road traffic.

Another way to extract knowledge is by discovering relationships between different attributes in the database; the association rule algorithm has been efficient in this sense, given its applicability in several scenarios, such as the context of air pollution (NEIROTTI *et al.*, 2014; AGRAWAL; SRIKANT, 1994). Association rules also contribute to discovering unexpected rules with a high degree of interest in the context in which they are inserted. In our study, association rules looked for relationships between the behavior of $PM_{2.5}$ and meteorological variables, in the different clusters identified in the clustering step. They also attempted to verify whether $PM_{2.5}$ and CO were related.

Li *et al.* (2020) proposed, by using association rules, the analysis of data from various air monitoring stations in China and micro stations in the USA, considering the uneven distribution of environmental monitoring data and the characteristics of climate change, and obtained a correlation between pollutants which provides support for the treatment and prevention of air pollution. Souza and Rabelo (2016) applied association rules to identify a set of variables that often occur together: air pollutant concentrations and rates of respiratory problems. Sadat, Karimipour and Sadat (2014) explored, by association rules, the effect of air pollution on asthmatic allergies, indicating that distance from parks and roads, as well as pollutant concentrations of CO, PM_{10}, $PM_{2.5}$, and NO_2, are related to the prevalence of allergies in the most polluted month of the year, while SO_2 and O_3 have no effect on it.

This article proposes a data mining approach to analyze the air quality monitoring database provided by CETESB, between 2017 and 2019. Such analysis was carried out by applying machine learning techniques on two fronts:

- using the partitional clustering algorithm (K-medoids) to form clusters, based on the $PM_{2.5}$ concentrations of 21 stations in the state of São Paulo;
- applying the association rules algorithm (Apriori) to discover possible associations between meteorological variables that affect the increase in $PM_{2.5}$ concentration and investigate the seasonal relationship between $PM_{2.5}$ and CO.

These studies can generate knowledge that contributes to the management of air quality and provides information for an assessment of its impact on health and the environment.

Methods

The methodology used in this study will be presented as follows:

- a presentation of the places where the air pollution data were collected and how they were preprocessed so as to be used by machine learning algorithms;
- an explanation of clustering algorithms and association rules, as well as their respective validation metrics.

Study site

Diagnosis of air quality in the state of São Paulo is made by the network of monitoring stations of CETESB, which informs pollution concentrations, generating an air quality index that ranges between good, moderate, bad, very bad, and terrible. These scenarios are important in reporting the compliance with air quality standards set by law and making it possible to determine when these levels represent significant risks to human health.

Assessment is carried out based on the state's air quality standards (Table 1) established by State Decree no. 59,113 (SÃO PAULO, 2013) and by CONAMA Resolution no. 491 (BRAZIL, 2018). The national and state standards, both for air quality and critical episodes, are virtually the same.

Both the CONAMA Resolution and the State Decree define intermediate targets (IT) so that air pollution is gradually reduced based on the guidelines proposed by WHO. It can be observed (Table 1) that national values are well above the international quality standard.

To analyze the behavior of $PM_{2.5}$ in different areas of the state of São Paulo, we obtained data from all cities that have stations with pollutant monitoring. Altogether, there are 21 stations, listed in Table 2 along with their geolocation (Figure 1).

Database and preprocessing

The first database was obtained from the CETESB website, by the Air Quality platform (QUALAR, 2019), which contains data collected by automatic monitoring stations. Data on monthly average $PM_{2.5}$ concentration from January 1st 2017 to December 31st 2019 were used. They generated a set of 21 records (stations) and 36 columns (months) representing the three-year period.

On this first basis, preprocessing was carried out to identify months with missing values in $PM_{2.5}$ monitoring. To perform the study of time series, all values must be completed (CASTRO; FERRARI, 2016). Where values were missing in a given month, the last and next technique was adopted, which obtains an average between the previous and the next value of the missing attribute (PLAIA; BONDI, 2006), that is, when there is a missing value, it is replaced by the average between the previous and the next month.

In addition, the data were standardized using the Z-score technique, which modifies the original values for them to have an average of 0 and a standard deviation of 1, resulting in values that will be compared under the same scale (HAN; KAMBER, 2006; MITSA, 2010; BATISTA; CHIAVEGATTO, 2019).

To build the second database, used in the step of association rules extraction, we verified the stations that monitor PM$_{2.5}$ and that also provide monthly averages of the following meteorological variables: RH, TEMP, WS, in addition to CO concentration (QUALAR, 2019) between 2017 and 2019. Of the 21 stations

Table 1 – Comparison of international (WHO), national (CONAMA 491/2018), and state (State Decree 59,113/2013) air quality standards for PM$_{2.5}$.

Quality Standards	24 hours[1]	AAA[2]
WHO Standards	25	10
IT 1 (µg/m3)[3]	60[4]	20[4]
IT 2 (µg/m3)[3]	50	17
IT 3 (µg/m3)[3]	37	15
Final Standards (µg/m^3) [3]	25	10

[1] Average of 24 consecutive hours of sampling (should not exceed more than once a year); [2] annual arithmetic average; [3] national standards; [4] state standards; IT: intermediate targets; WHO: World Health Organization; CONAMA: National Environment Council; AAA: annual arithmetic average.Source: adapted from WHO (2019), Brazil (2018), and São Paulo (2013).

Table 2 – Cities and stations with PM$_{2.5}$ monitoring in the state of São Paulo.

City	Station
Campinas	Vila União
Guarulhos	Paço Municipal
Guarulhos	Pimentas
Osasco	Vila Quitaúna
Piracicaba	Campus FUMEP
Ribeirão Preto	Parque Ecológico Maurílio Biaggi
Santos	Ponta da Praia
São Bernardo do Campo	Centro
São José dos Campos	Jd. Satélite
São José do Rio Preto	Campo Atletismo Eldorado
São Paulo	Cidade Universitária (USP)
	Congonhas
	Grajau (Parelheiros)
	Ibirapuera
	Itaim Paulista
	Marginal Tietê (Ponte dos Remédios)
	Parque D. Pedro II
	Pico do Jaraguá (Serra da Cantareira)
	Pinheiros
	Santana
Taubaté	Parque Municipal “Eng. César A. C. Varejão”

FUMEP: Fundação Municipal de Ensino de Piracicaba; USP: Universidade de São Paulo.

whose data were obtained for the first database, seven met this new criterion (Table 3).

For this new dataset, all data must be categorical, since this is a restriction of the Apriori algorithm. Thus, each monthly average value was classified according to two categories: lower or higher than the annual average value of its respective meteorological variable or CO concentration. Table 3 represents an excerpt from the database, referring to the month of July 2018.

The algorithms applied in this study follow the unsupervised approach of machine learning, divided into two stages:

- application of the partitional clustering algorithm (*K-medoids*);
- association rules (Apriori).

The next sections discuss these algorithms.

Data clustering technique

Clustering algorithms can be either partitional or hierarchical. Their ability to cluster data based on intrinsic characteristics of the problem makes them interesting for studies. Such algorithms generate clusters formed by data samples that are similar to each other, according to some measure of similarity. Assuming, for example, a problem of clustering cities by the level of air quality, the clustering algorithms will map the cities and return clusters composed of those with similar pollution behavior. Within the cluster of partitional algorithms, the most common are K-means and K-medoids (JIN; HAN, 2017). The K-medoids algorithm uses objects from the database as the center of the clusters, called medoids, which have the lowest average dissimilarity compared to all other objects in the cluster. In the case of K-means, the centers of the clusters are calculated according to the average value of the objects in that cluster. In this case, outliers from the database can influence the formation of the clusters, since they contribute to the calculation of the central values of each cluster. This type of problem does not happen in the K-medoids algorithm, since the medoids correspond to real samples of the data and not averages (HAN; KAMBER, 2006), that is, the medoids are an element of the cluster itself and not a midpoint as occurs in K-means, which makes it less sensitive to outliers.

Both algorithms (K-means and K-medoids) were implemented in Python, using the open-source Scikit-Learn and PyClustering libraries, specific for machine learning (PEDREGOSA *et al.*, 2011).

To assess the quality of the clustering between the K-medoids and K-means algorithms, the silhouette coefficient was applied (KAUFMAN; ROUSSEEUW, 2005) to the results obtained by each algorithm. This coefficient measures the robustness of the partitions, helping to select the number of clusters, considering the internal similarity and external dissimilarity between them, that is, it combines cohesion (measures how well an element is within a cluster) and separation (measures how much the clusters are separated from each other). For example, supposing that the clustering algorithm returns two clusters, as in the previous example, the silhouette coefficient will verify whether all the elements of Cluster 1 are similar to each other and different from the elements of Cluster 2. An expected behavior would be that this hypothetical Cluster 1 would include cities with a high concentration of one pollutant and Cluster 2, cities with a low concentration of the same pollutant. Therefore, Cluster 1 and Cluster 2 would be cohesive, since they would have cities that show the same behavior, and also separated from each other for presenting an entirely different pattern.

The average value of the silhouette coefficient must be between -1 and 1, representing how well the clusters were formed. The ideal values are positive, with a silhouette coefficient close to 1. Equation 1 represents the average Silhouette calculation (S_p).

$$S_p = \sum_{1}^{n} \frac{s(x_i)}{n} \tag{1}$$

Where:

n = the number of objects in the database and the individual value of the silhouette coefficient of element x_i, given by $s(x_i)$, obtained by Equation 2:

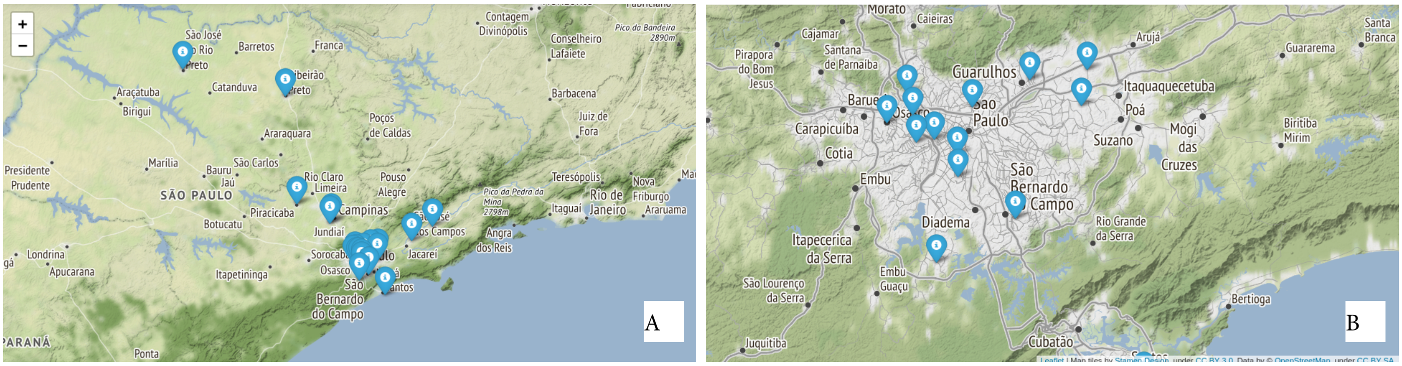

Figure 1 – (A) Map of location of the automatic $PM_{2.5}$ monitoring stations in the state of São Paulo and (B) $PM_{2.5}$ monitoring stations in the Metropolitan Region of São Paulo (MRSP).

$$s(x_i) = \frac{(b(x_i) - a(x_i))}{max\,\{a(x_i), b(x_i)\}} \qquad (2)$$

Where:

the values $a(x_i)$ and $b(x_i)$ = respectively, the average distance between x_i and all the objects in its cluster and the average distance of x_i to another cluster to which x_i does not belong.

The silhouette coefficient was also the evaluation metric chosen to determine which of the two algorithms (K-means and K-medoids) would be used in this study. Therefore, the database of monthly $PM_{2.5}$ averages was used and the two algorithms were applied to carry out this evaluation. The one that presented the best silhouette result was adopted for the clustering of stations. This experiment is presented in the Results section.

Association rules

The Apriori Association Rules algorithm aims to find frequent relationships in the datasets, that is, to generate rules of type X → Y, for which X and Y are items that belong to this dataset (AGRAWAL; SRIKANT, 1994). To analyze the possible patterns found in the months with the highest concentration of $PM_{2.5}$, the Apriori Association Rules algorithm was applied to find a subset of frequent parameters related to the database of $PM_{2.5}$.

The Apriori algorithm searches, from a transactional basis, which items are related. For example, in a hypothetical database that records the monthly values of the concentration of air pollutants and the number of hospital visits involving respiratory diseases, the association rules may return {$PM_{2.5}$, PM_{10}} → {increase in visits}, indicating that a high concentration of pollutants $PM_{2.5}$ and PM_{10}, causes, with a degree of certainty, an increase in hospital visits. This degree of certainty that measures the relevance and validation of the rules is provided by: support and confidence. Given the rule X → Y, the support (or coverage of the rule) represents the percentage of transactions in the database that contain the items of X and Y, indicating its relevance (CASTRO; FERRARI, 2016). The confidence or accuracy of a rule, in turn, corresponds to the number of rules in which the consequent (term after the →) of a rule appears in transactions in which the antecedent (term(s) preceding →) is also observed, that is, it is the conditional probability $P(Y|X)$ that given the consequent X of the rule, the antecedent Y also happens (MUELLER, 1995). In this study, the Apriori algorithm was implemented in Python, using the "mlxtend" library.

Results and Discussion

In the experiment to choose the clustering algorithm, the silhouette coefficient was used as the decision criterion, as it is a measure of quality for the entire structure of the partition. It was also used to choose the number of clusters (k), and, for this, 20 different cluster sizes, related to the number of cities, were tested.

After 100 executions of the K-medoids algorithm, applied to the database of monthly averages of $PM_{2.5}$ concentration between 2017 and 2019, the average silhouette coefficient found was 0.26, while for the K-means algorithm, the average value was 0.28. Considering that the silhouette value can vary between -1 and 1, K-medoids was selected because it presents a better average silhouette value and is capable of handling outliers.

Figure 2 shows the relationship between the silhouette coefficient value corresponding to the number k of clusters. The best value corresponds to k = 2. Thus, the K-medoids algorithm was applied to obtain two clusters from the set of stations in the state of São Paulo, with $PM_{2.5}$ monitoring, and the clustering results were subsequently analyzed.

Table 3 – Example of the representation of the database in the month of July 2018, relating the stations that monitor $PM_{2.5}$ with meteorological variables TEMP, RH, WS, and CO concentration. The numerical values were transformed into a category, which may be higher or lower than the average.

	Station	TEMP	RH	WS	CO
0	Parque D. Pedro II	Below Average	Below Average	Below Average	Above Average
1	Pinheiros	Below Average	Below Average	Below Average	Above Average
2	Marg. Tietê-Pte	Below Average	Below Average	Below Average	Above Average
3	S. Bernardo-Centro	Below Average	Above Average	Below Average	Above Average
4	Guarulhos-Pimentas	Below Average	Below Average	Below Average	Above Average
5	S. José Campos - Jd	Below Average	Below Average	Below Average	Above Average
6	Taubaté	Below Average	Below Average	Below Average	Above Average
7	Ribeirão Preto	Below Average	Below Average	Below Average	Above Average

TEMP: temperature; RH: relative humidity; WS: Wind speed.

As a result of applying the K-medoids algorithm to the data, with a value of k = 2, the stations were divided into Clusters 1 and 2, shown in Table 4.

In the analyzed period, for all the stations monitored, the average annual concentrations of $PM_{2.5}$ were 16.43 µg/m^3 (standard deviation 6.45 µg/m^3) in 2017, 16.24 µg/m^3 (standard deviation 6.42 µg/m^3) in 2018, and 16 µg/m^3 (standard deviation 5.04 µg/m^3) in 2019, exceeding the annual threshold of 10 µg/m^3 established by WHO in all periods; note that the standard deviation remained constant in 2017 and 2018, and decreased in 2019. Analyzing each cluster, we can see differences:

- **Cluster 1:** 15 stations located mostly in metropolitan regions, more specifically in cities with an average annual global $PM_{2.5}$ concentration of 17.42 µg/m^3 and standard deviation of 4.72 µg/m^3;
- **Cluster 2:** 6 stations located in cities with relatively lower indexes, with an average annual global $PM_{2.5}$ concentration of 13.4 µg/m^3 and standard deviation of 4.83 µg/m^3.

Figure 3 shows that, between 2017 and 2019, higher concentrations of $PM_{2.5}$ predominate in Cluster 1 compared to Cluster 2, since the former consists of stations located in the Metropolitan Region of São Paulo (MRSP), as found in other studies (HUANG *et al.*, 2015; AUSTIN *et al.*, 2013). There is also a seasonal trend in the evolution of pollutant concentration and monthly peaks for both clusters in the same periods, suggesting a recurring pattern in the three years. Despite the similarity in seasonal behavior throughout the period, it is clear that in 2017 the month of greatest concentration is September, in 2018 it is July, and in 2019, June. In 2017, the peak concentration of the pollutant was lower than the peak in 2018, while in 2019, the $PM_{2.5}$ concentration level was below the one observed in previous years.

These cycles may be related to meteorological phenomena that have taken place over the period, which coincide with the data from CETESB's annual reports (CETESB, 2019), also identified in the literature (LI *et al.*, 2020; BISHT; SEEJA, 2018), and which were analyzed with the association rules algorithm (*Apriori*).

Figure 4 was generated for a better assessment of the physical proximity between the stations in the clusters, showing the geographical location of the stations in each cluster. Clusters 1 and 2 were identified by the colors red and blue, respectively, in Figures 4A and 4B.

The analysis on the map shows that most of the $PM_{2.5}$ monitoring stations present in Cluster 1 are in the Metropolitan Regions (MR) of São Paulo, Campinas, and Baixada Santista. Except for the Campinas region, which is also influenced by fires, the main source of pollutants in these MRs is fuel burning by the vehicle fleet and intense industrial emissions (CARDOSO *et al.*, 2017; HUANG *et al.*, 2015; YANAGI; ASSUNÇÃO; BARROZO, 2012). The stations with lower concentrations, represented by Cluster 2, are located further inland in the state and are more distant from each other, except for Ibirapuera station, which, despite being located in the city of São Paulo, is located farther from intense traffic routes.

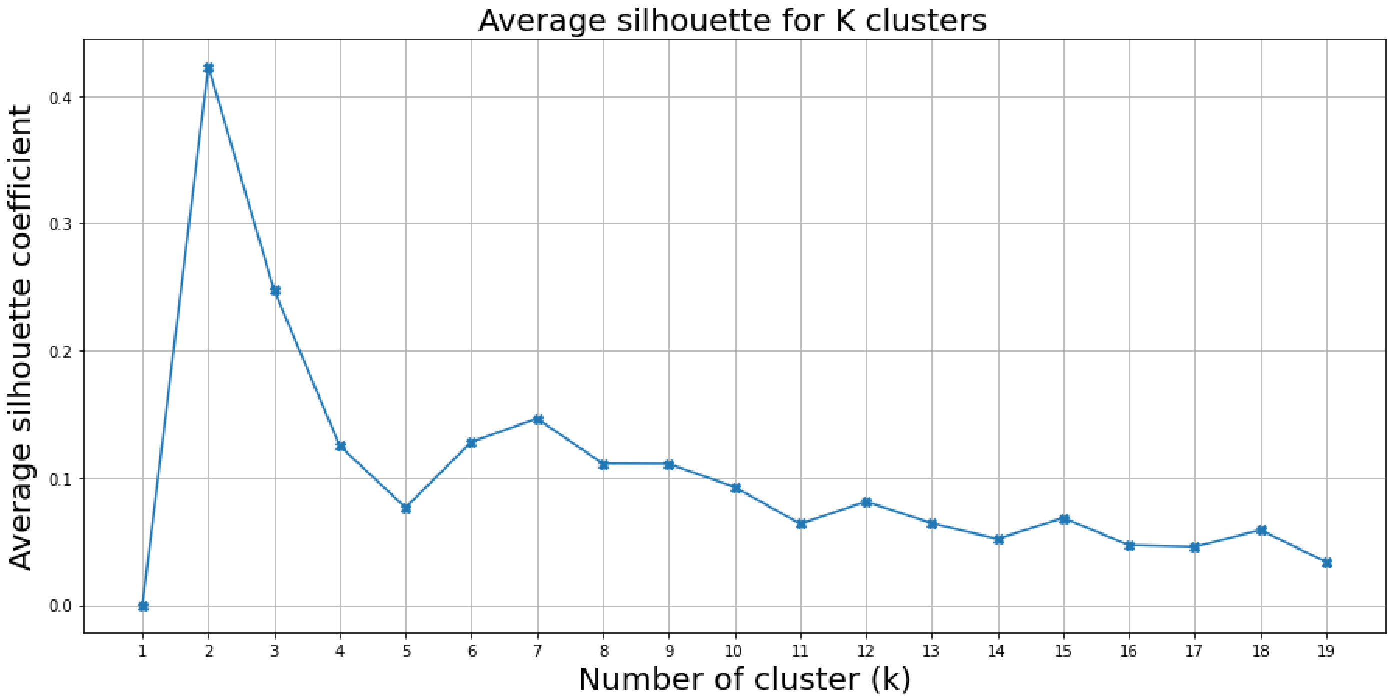

Figure 2 – Number k of clusters per silhouette coefficient value, obtained from the K-medoids algorithm, applied to the database of monthly averages of $PM_{2.5}$ concentration, between 2017 and 2019.

Comparing the results obtained, there is a correspondence between the clusters generated and other studies that investigate air pollution by $PM_{2.5}$ in the state of São Paulo: Araújo and Rosário (2020) identified from satellite data that the most polluted regions in the state are the MRs of São Paulo, Campinas, and Baixada Santista.

The analysis of the average monthly variation of $PM_{2.5}$ concentration in Clusters 1 and 2 indicates differences in pollutant concentrations between the two clusters, as can be seen in the boxplots in Figure 5. However, the interquartile ranges and maximum values (disregarding outliers) are similar.

Table 5 shows that, in 2017, the $PM_{2.5}$ concentration level increased from May to October, with a peak of about 29.8 µg/m^3 in September. Likewise, in 2018, the increase occurred from March to September, with a peak of 32.4 µg/m^3 in July, indicating an increase in the pollutant that year. The same behavior was repeated in 2019, from April to October, with a peak of 23.7 µg/m^3 in June, but with a reduction in the pollutant concentration.

Studies show that meteorological factors such as TEMP, reduction in RH, and WS can impair the dispersion of $PM_{2.5}$, increasing health-related risks (INPE, 2019; CETESB, 2019). The studies by Santos, Carvalho and Reboita (2016) and Santos *et al.* (2019) confirm a significant difference between the concentration of $PM_{2.5}$ in dry and rainy periods, indicating the association between meteorological parameters and the pollutant.

To assess such a relationship, data of the months with the highest peaks (Figure 3 and Table 5), that is, September 2017, July 2018,

Table 4 – List of monitoring stations per clusters and their annual averages (2017 to 2019) of $PM_{2.5}$ concentration.

Monitoring stations	Monthly Averages of $PM_{2.5}$ Concentration		
CLUSTER 1	**2017**	**2018**	**2019**
Osasco	28.29	21.50	20.83
São Paulo – Marginal Tietê (Pte. Remédios)	19.50	19.92	20.00
Guarulhos – Paço Municipal	18.50	16.92	15.00
São Paulo – Santana	17.92	16.25	16.33
Guarulhos – Pimentas	17.83	21.08	19.75
São Paulo – Congonhas	17.83	18.42	17.67
São Paulo – Itaim Paulista	17.25	18.50	18.50
Campinas – Vila União	17.08	15.83	19.17
São Paulo – Grajau (Parelheiros)	17.00	18.67	16.92
São Paulo – Parque D. Pedro II	16.75	17.42	17.17
São Bernardo do Campo – Centro	16.17	16.00	16.17
São Paulo – Cidade Universitária (USP)	15.92	16.00	15.00
Santos – Ponta da Praia	15.58	14.08	14.42
São Paulo – Pinheiros	14.48	16.33	16.54
São Paulo – Pico do Jaraguá (Serra da Cantareira)	12.58	15.13	15.50
CLUSTER 2	**2017**	**2018**	**2019**
São Paulo – Ibirapuera	15.75	14.83	13.08
São José do Rio Preto	15.75	14.42	14.83
Taubaté	13.08	11.08	11.08
Ribeirão Preto	13.00	13.58	14.00
Piracicaba	12.67	13.33	13.00
São José dos Campos – Jd. Satélite	12.00	11.67	11.08

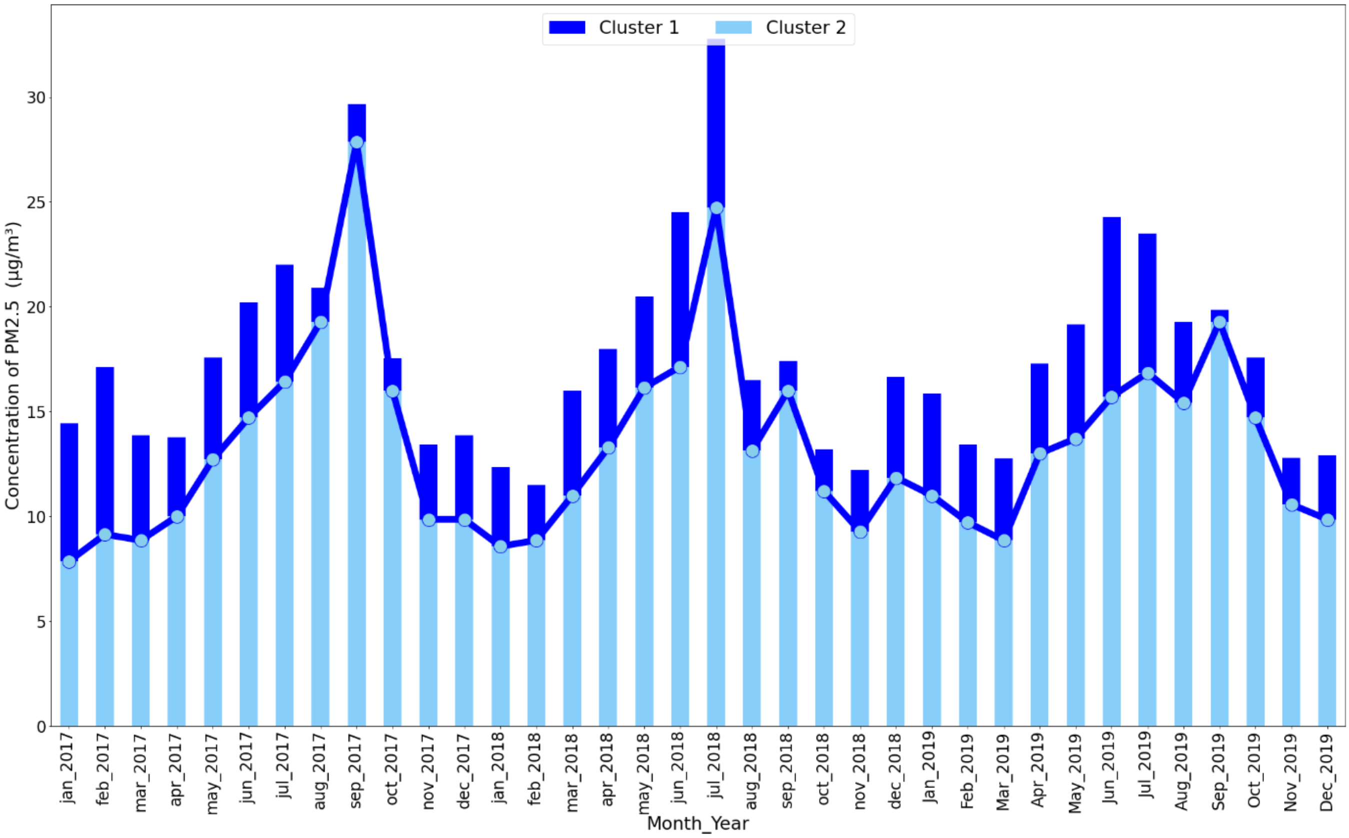

Figure 3 – Comparison of the monthly averages of PM$_{2.5}$ concentrations (μg/m³) between 2017 and 2019, in the cities of the state of São Paulo, with Cluster 1 being characterized mostly by the MRSP and Cluster 2, by inland cities.

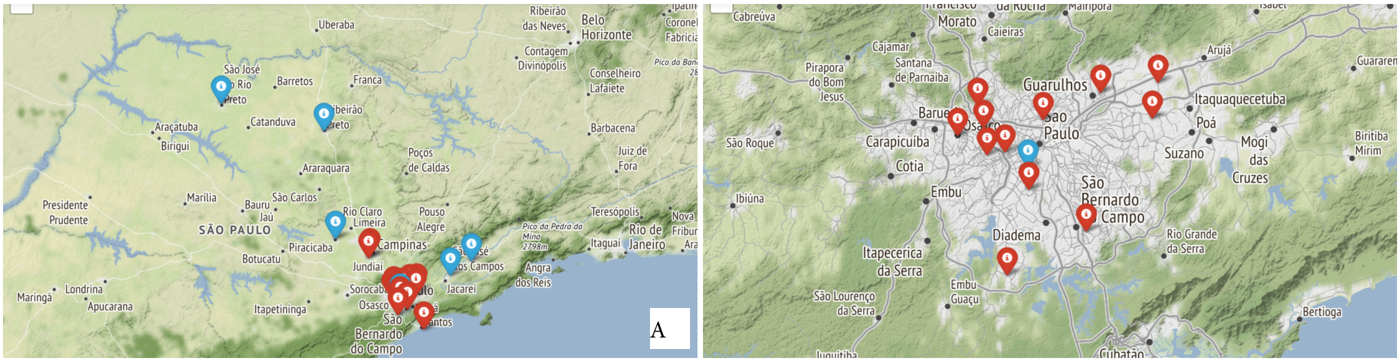

Figure 4 – (A) Visualization by geolocation of the clusters, created by the K-medoids algorithm; B) proximity of the elements of Cluster 1 on the map. Cluster 1 in red and Cluster 2 in blue.

and June 2019, were collected from the transactional base (containing the PM$_{2.5}$ concentration values for each station and the behavior of the meteorological variables) and submitted to the Apriori association rule algorithm. With that, we tried to find out which factors were more frequent in the three periods and how these meteorological factors were related.

In the first run of Apriori, using September 2017 data, nine association rules were obtained, seven of which were repeated, that is, rules that had the same meaning. This takes place because the algorithm analyzes all the possibilities between the items. Therefore, the two main rules for this period are shown in Table 6. Support corresponds to the frequency with which the patterns occur throughout the database, indicating the percentage of occurrence of the transactions. Confidence measures the "strength" of rules, that is, it assesses whether transactions that satisfy the antecedent of the rules also satisfy their consequent. The rules that meet support and confidence are called "strong rules."

It can be concluded that, for the peak month of 2017, starting from Rule 1, a high concentration of $PM_{2.5}$, below-average RH, and above average CO concentration occur together with a frequency of 85%. This rule also informs that, when the concentration of CO is above the average, RH is below the average with a certainty of 100%. For Rule 2, at the peaks of $PM_{2.5}$ concentration, the factors that occur together with a 75% frequency are above average CO and above average TEMP. Regarding confidence, when CO is above average, temperature is above average with a certainty of 100%.

In the second run of Apriori, July 2018 data were used and 44 rules were obtained, and the three not repeated rules with greater support and confidence were chosen for analysis (Table 6).

For the high concentration of $PM_{2.5}$ in July 2018, Rule 1 identifies the following factors: below-average TEMP and below-average WS occur together with 100% frequency in the database. For Rule 2, the frequency of occurrence of the two factors is 87% and the probability of low WS given the occurrence of below-average RH is 100%. For Rule 3, three factors appear together with a frequency of 87% and 100% confidence, indicating that whenever the temperature becomes predominantly colder, the CO concentration increases and WS is below average, signaling that in colder seasons there is an increase in CO concentration, stimulated by the low dispersion of this pollutant.

In the last Apriori execution, June 2019 data were used and nine rules were obtained, two of which were the most representative (Table 6). The identified rules were similar to the rules of the previous year, with the predominant variables TEMP, RH, and WS below the average. Also, the months of high concentrations tend to be close from one year to the next.

According to the winter report of CETESB (2020), the winter of 2019 presented a predominance of a hot and dry air mass throughout the state of São Paulo, with low ventilation and absence of rains, making it difficult to disperse pollutants, which corroborates the rules obtained for 2019.

Considering that the periods with the highest concentration of $PM_{2.5}$ are the ones that present the greatest risk to the population and that meteorological factors have an influence on the increase in pollutant concentration, the rules presented in Table 6 could give warning indications for the increase in pollutant concentration. In Brazil, the studies by César *et al.* (2016) and Machin and Nascimento (2018) show the influence of the 5 $\mu g/m^3$ increase in the concentrations of $PM_{2.5,}$ resulting in increases between 20 and 38% in the risk of hospitalization due to pulmonary complications.

Thus, we can conclude that when the concentration of $PM_{2.5}$ increases, the measurements show the following behaviors: low RH and above-average TEMP. The results also indicate that high concentrations of $PM_{2.5}$ may be associated with below average TEMP, milder WS, and below-average RH. We observed an increase in CO, which suggests an association with the behavior of $PM_{2.5}$ in the winter months, also reported by Moisan, Herrera and Clements (2018) and Saide *et al.* (2011).

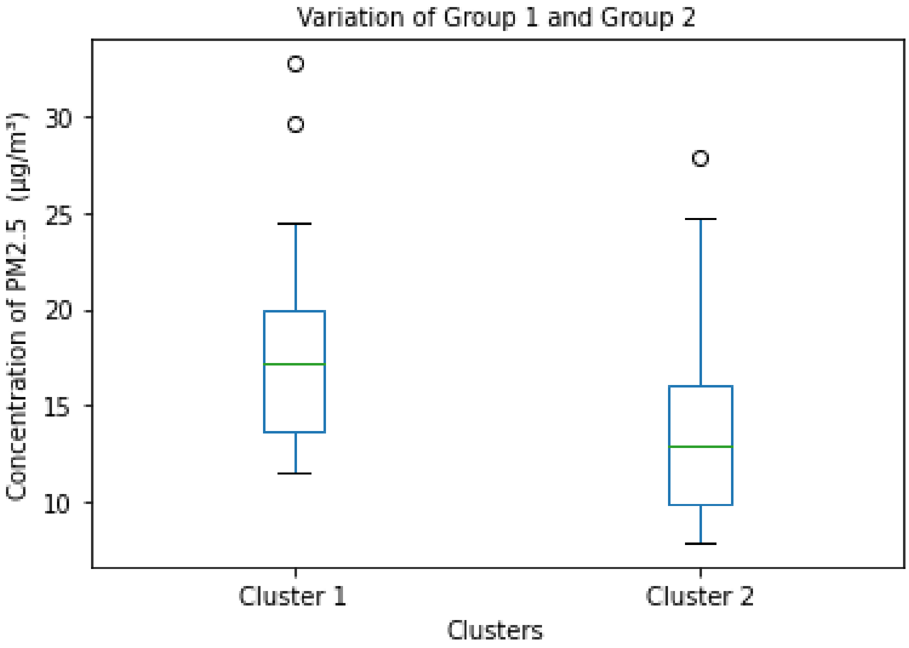

Figure 5 – Boxplots of Clusters 1 and 2, formed by the monthly averages of $PM_{2.5}$, between 2017 and 2019.

Table 5 – Monthly averages of $PM_{2.5}$ by clusters of stations and standard deviation of the clusters (in µg/m³), between 2017 and 2019. The highlighted months are the periods of greatest pollutant concentration in the three years, with emphasis on the peak months September/2017, July/2018, and June/2019.

Clusters	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
	2017											
Cluster 1 (C1)	14.1	16.4	13.6	13.5	17.1	19.6	21.5	20.6	29.5	17.3	13.2	13.7
Standard deviation	5.8	5.4	5.4	5.5	4.5	4.3	4.3	2.6	5.0	3.0	2.1	1.6
Cluster 2 (C2)	7.7	9.7	8.7	10.0	13.0	15.3	16.8	19.7	28.0	16.3	9.8	9.5
Standard deviation	1.6	2.1	1.5	1.3	1.5	1.5	2.5	3.7	5.6	4.1	1.9	2.3
	2018											
Cluster 1 (C1)	12.2	11.4	15.7	17.8	20.2	23.8	32.2	16.2	17.3	13.4	12.5	17.0
Standard deviation	1.5	1.5	2.1	3.2	3.1	4.9	6.3	2.2	2.1	1.4	1.9	4.8
Cluster 2 (C2)	8.3	8.7	10.8	13.0	16.2	17.7	24.8	13.3	16.2	10.5	8.2	10.2
Standard deviation	1.9	1.2	2.1	2.1	1.8	2.4	4.4	2.6	3.5	1.4	1.2	1.0
	2019											
Cluster 1 (C1)	15.9	13.7	12.8	17.3	18.9	23.7	23.0	18.9	19.5	17.5	13.0	13.1
Standard deviation	2.2	3.0	1.5	2.0	3.3	4.7	4.9	3.5	3.6	2.6	2.1	1.9
Cluster 2 (C2)	10.0	8.5	8.2	12.3	13.5	15.8	17.0	15.7	20.2	14.3	9.7	9.0
Standard deviation	1.1	0.5	0.4	0.5	1.9	2.0	1.5	3.5	6.8	3.3	1.4	1.5

Table 6 – Rules obtained by the Apriori algorithm and its respective Support and Confidence parameters*.

September 2017		
Rules (Antecedent → Consequent)	Support	Confidence
Rule 1. (Below-average RH) → (Above-average CO)	85%	100%
Rule 2. (Above-average TEMP) → (Above-average CO)	75%	100%
July 2018		
Rule 1. Below-average TEMP → Below-average WS	100%	100%
Rule 2. Below-average RH → Below-average WS	87%	100%
Rule 3: Above-average CO and below-average WS → Below average TEMP	87%	100%
June 2019		
Rule 1. Below-average TEMP → Below-average RH	62%	83%
Rule 2. Below-average WS → Below-average RH	50%	100%

*Annual averages for each meteorological variable; RH: relative humidity; TEMP: temperature; WS: Wind speed.

Conclusions

The analysis of $PM_{2.5}$ carried out in this study was done by the application of a clustering algorithm, which divided the values of measurements of $PM_{2.5}$ concentrations from 21 monitored stations, distributed over 36 months, between 2017 and 2019.

The experiments showed that the formation of two clusters is the most adequate. The results show that the stations belonging to the identified clusters have specific characteristics that lead to different pollution rates. The municipalities of the MRSP stand out as those with the highest concentration of $PM_{2.5}$, but cities inland, with a predominance of industrial and vehicular emissions, join these municipalities, forming one of the clusters. The stations of the other cluster, installed in less polluted locations, are in cities further inland, far from sources of pollution such as vehicle emissions and industrial processes.

Two very characteristic clusters were formed, with variations in pollutant concentration that followed a pattern throughout each year. A seasonal behavior was observed in the temporal study, which is repeated in every period, in both clusters. There is a higher incidence of $PM_{2.5}$ in winter, which peaked (September 2017, July 2018, and June 2019) in critical months, when the meteorological variables (TEMP, RH, WS) contribute to the increase in pollutant concentration.

From the clustering results, another algorithm was applied to meteorological data related to September 2017, July 2018, and June 2019, to find associations with the meteorological factors mentioned above in the periods of greatest concentration of $PM_{2.5}$. The results showed that, in September 2017, the predominant meteorological factors were low RH and above average TEMP. In July 2018 and June 2019, the rules showed that below average TEMP and RH and milder WS were the main meteorological factors that occurred during the period with the highest average pollutant concentration. Finally, we also observed a direct relationship between the concentrations of CO and $PM_{2.5.}$

The rules found can be useful in creating warning signs for possible increases in the concentration of PM, since the results confirm a relationship between episodes of high concentration and atmospheric conditions in the region, providing subsidies for managing air quality in the state of São Paulo.

Acknowledgments

This study was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Financing Code 001.

The authors thank Espaço da Escrita – Pró-Reitoria de Pesquisa – UNICAMP – for the language services provided.

Contribution of authors:

Godoy, A.R.L.: Conceptualization, Methodology, Investigation, Formal analysis, Validation, Writing — original draft, Data curation. Silva, A.E.A.: Methodology, Writing — review & editing. Bueno, M.C.: Conceptualization, Methodology, Investigation, Formal analysis, Software, Validation, Writing — original draft, Data curation. Pozza, S.A.: Conceptualization, Writing — review & editing. Coelho, G.P.: Methodology, Writing — review & editing.

References

ABE, K.; MIRAGLIA, S. Avaliação de impacto à saúde do programa de controle de poluição do ar por veículos automotores no município de São Paulo, Brasil. *Revista Brasileira de Ciências Ambientais (Online)*, n. 47, p. 61-73, 2018. https://doi.org/10.5327/Z2176-947820180310

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. *In*: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994. *Proceedings*… 1994. p. 487-499.

AMEER, S.; SHAH, M. A.; KHAN, A.; SONG, H.; MAPLE, C.; ISLAM, S. U.; ASGHAR, M. N. Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities. *IEEE Access*, v. 7, p. 128325-128338, 2019. https://doi.org/10.1109/access.2019.2925082

ANDRADE, M.; MIRANDA, R. M.; FORNARO, A.; KERR, A.; OYAMA, B.; ANDRE, P. A.; SALDIVA, P. Vehicle emissions and PM2.5 mass concentrations in six Brazilian cities. *Air Quality, Atmosphere and Health*, v. 5, p. 79-88, 2012. https://doi.org/10.1007/s11869-010-0104-5

ARAÚJO, J.; ROSÁRIO, N. Poluição atmosférica associada ao material particulado no estado de São Paulo: análise baseada em dados de satélite. *Revista Brasileira de Ciências Ambientais (Online)*, v. 55, n. 1, p. 32-47, 2020. https://doi.org/10.5327/Z2176-947820200552

AUSTIN, E.; COULL, B. A.; ZANOBETTI, A.; KOUTRAKIS, P. A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. *Environment International*, v. 59, p. 244-254, 2013. https://doi.org/10.1016/j.envint.2013.06.003

BATISTA, A. F. M.; CHIAVEGATTO, A. D. P. Machine Learning aplicado à Saúde. Workshop: Machine Learning. *In*: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADO À SAÚDE, 19., 2019. *Proceedings*... Sociedade Brasileira de Computação, 2019. Available at: <https://sol.sbc.org.br/livros/index.php /sbc/catalog/view/29/95/245-1>. Accessed on: Jul. 20, 2020.

BISHT, M.; SEEJA K.R. Air Pollution Prediction Using Extreme Learning Machine: A Case Study on Delhi (India). *In*: SOMANI, A.; SRIVASTAVA, S.; MUNDRA, A.; RAWAT, S. (eds.). *Proceedings of First International Conference on Smart System, Innovations and Computing*. Smart Innovation, Systems and Technologies. Singapore: Springer, 2018. v. 79. p. 181-189.

BRAZIL. Ministério do Meio Ambiente. Conselho Nacional do Meio Ambiente. *Resolução nº 491, de 19 de novembro de 2018*. Brasil, 2018. Available from: <http://www2.mma.gov.br/port/conama/legiabre.cfm?codlegi=740>. Accessed on: Jun. 10, 2019.

CARDOSO, K. M.; PAULA, A.; SANTOS, J. S.; SANTOS, M. L. P. Uso de espécies da arborização urbana no biomonitoramento de poluição ambiental. *Ciência Florestal*, v. 27, n. 2, p. 535-547, 2017. https://doi.org/10.5902/1980509827734

CASTRO, L. N.; FERRARI, D. G. *Introdução a Mineração de Dados*. Conceitos Básicos, Algoritmos e Aplicações. São Paulo: Saraiva, 2016. 351 p.

CÉSAR, A. C. G.; NASCIMENTO, L. F. C.; MANTOVANI, K. C. C.; VIEIRA, L. C. P. Fine particulate matter estimated by mathematical model and hospitalizations for pneumonia and asthma in children. *Revista Paulista de Pediatria*, v. 34, n. 1, p. 18-23, 2016. https://doi.org/10.1016/j.rppede.2015.12.005

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (CETESB). *Relatório de Qualidade do Ar no estado de São Paulo*. São Paulo: Governo do Estado de São Paulo / Secretaria do Meio Ambiente / Companhia Ambiental do Estado de São Paulo, 2019. Available from: <https://cetesb.sp.gov.br/ar/

wp-content/uploads/sites/28/2019/05/Relat%C3%B3rio-de-Qualidade-do-Ar-2017.pdf>. Accessed on: May 8, 2019.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (CETESB). *Winter Operation Report*. Available at: <https://cetesb.sp.gov.br/ar/wp-content/uploads/sites/28/2020/03/Relatório-Operação-Inverno-2019.pdf>. Accessed on: Apr. 12, 2020.

DIMITRIOU, K. Upgrading the estimation of daily PM10 concentrations utilizing prediction variables reflecting atmospheric processes. *Aerosol and Air Quality Research*, v. 16, n. 9, p. 2245-2254, 2016. https://doi.org/10.4209/aaqr.2016.05.0214

DU, X.; VARDE, A. S. Mining PM2.5 and traffic conditions for air quality. *In*: INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION SYSTEMS, 7., 2016. *Proceedings*… ICICS, 2016. p. 33-38. https://doi.org/10.1109/IACS.2016.7476082

GONÇALVES, F. L. T.; CARVALHO, L. M. V.; CONDE, F. C.; LATORRE, M. R. D. O.; SALDIVA, P. H. N.; BRAGA, A. L. F. The efects of air pollution and meteorological parameters on respiratory morbidity during summer in São Paulo City. *Environment International*, v. 31, n. 3, p. 343-349, 2005. https://doi.org/10.1016/j.envint.2004.08.004

GUERRA, F. P.; MIRANDA, R. M. Influência da meteorologia na concentração do poluente atmosférico PM2,5 na RMRJ e na RMSP. *In*: CONGRESSO BRASILEIRO DE GESTÃO AMBIENTAL, 2., 2011. *Proceedings*... 2011.

GUIDETTI, B.; PEREDA, P. *Air Pollution Consequences in São Paulo:* Evidence for Health. 2018. 20 p.

HAN, J.; KAMBER, M. *Data Mining:* Concepts and Techniques. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining:* Concepts and Techniques. 3ª ed. Burlington: Morgan Kaufmann, 2011.

HUANG, P.; ZHANG, J.; TANG, Y.; LIU, L. Spatial and temporal distribution of PM2.5 pollution in Xi'an city, China. *International Journal of Environmental Research and Public Health*, v. 12, n. 6, p. 6608-6625, 2015. https://doi.org/10.3390/ijerph120606608

INSTITUO NACIONAL DE PESQUISAS ESPACIAIS (INPE). *Boletins de Informações Climáticas do CPTEC/INPE*, ano 24, n. 1-12, 2019. Available from: <http://infoclima1.cptec.inpe.br>. Accessed on: May 8, 2019.

JIN, X.; HAN, J. K-Medoids Clustering. *In*: SAMMUT, C.; WEBB, G. I. (Eds.). *Encyclopedia of Machine Learning and Data Mining*. Boston: Springer, 2017. p. 697-700. https://doi.org/10.1007/978-1-4899-7687-1_432

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data:* An Introduction to Cluster Analysis. New York: Wiley Series in Probability and Statistics, 2005.

KWEDLO, W. A clustering method combining differential evolution with the K-means algorithm. *Pattern Recognition Letters*, v. 32, n. 12, p. 1613-1621, 2011. https://doi.org/10.1016/j.patrec.2011.05.010

LI, Z.; ZHOU, W.; LIU, X.; QUIAN, Y.; WANG, C.; XIE, Z.; MA, H. Research on Association Rules Mining of Atmospheric Environment Monitoring Data. *In*: HONG, W.; LI, C.; WANG, Q. (eds.). *Technology-Inspired Smart Learning for Future Education*. NCCSTE 2019. Singapore: Springer, 2020. (Communications in Computer and Information Science, v. 1216.) https://doi.org/10.1007/978-981-15-5390-5_8

MACHIN, A. B.; NASCIMENTO, L. F. C. Efeitos da exposição a poluentes do ar na saúde das crianças de Cuiabá, Mato Grosso, Brasil. *Cadernos de Saúde Pública*, v. 34, n. 3, p. 1-9, 2018. https://doi.org/10.1590/0102-311X00006617

MITSA, T. Temporal data mining. *In*: MITSA, T. *Temporal Data Mining*. New York: Chapman and Hall, 2010. p. 46-48. https://doi.org/10.1201/9781420089776

MOISAN, S.; HERRERA, R.; CLEMENTS, A. A dynamic multiple equation approach for forecasting PM2,5 pollution in Santiago, Chile. *International Journal of Forecasting*, v. 34, n. 4, p. 566-581, 2018. https://doi.org/10.1016/j.ijforecast.2018.03.007

MORAES, S. L.; ALMENDRA, R.; SANTANA, P.; GALVANI, E. Meteorological variables and air pollution and their association with hospitalizations due to respiratory diseases in children: A case study in São Paulo, Brazil. *Cadernos de Saúde Pública*, v. 35, n. 7, p. 1-16, 2019. https://doi.org/10.1590/0102-311x00101418

MUELLER, A. *Fast sequential and parallel algorithms for association rule mining: a comparison*. Thesis (M.S.) – Department of Computer Science, University of Maryland, College Park, 1995.

NEIROTTI, P.; MARCO, A.; CAGLIANO, A. C.; MANGANO, G.; SCORRANO, F. Current trends in smart city initiatives: Some stylised facts. Cities, v. 38, p. 25-36, 2014. https://doi.org/10.1016/j.cities.2013.12.010

NODARI, A. S.; SALDANHA, C. B. Episódios críticos de Poluição Atmosférica no município de Porto Alegre/RS. *In*: INTERNATIONAL SYMPOSIUM ON ENVIRONMENTAL QUALITY, 10., 2016. Available at: <http://www.abes-rs.uni5.net/centraldeeventos/_arqTrabalhos/trab_20160910113702000000650.pdf>. Accessed on: Feb. 20, 2019.

NOGAROTTO, D. C. *Avaliação de modelos de regressão de trajetórias para a previsão de poluentes atmosféricos*. 145f. Thesis (Doctoring) – Faculdade de Tecnologia, Universidade Estadual de Campinas, Limeira, 2019. Available at: <http://www.repositorio.unicamp.br/handle/REPOSIP/334421>. Accessed on: May 22, 2020.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E.. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, n. 85, p. 2825-2830, 2011. Available from: <http://www.jmlr.org/papers/v12/pedregosa11a.html>. Accessed on: Mar. 5, 2020.

PLAIA, A., BONDI, A. L. Single imputation method of missing values in environmental pollution datasets. *Atmospheric Environment*, v. 40, n. 38, p. 7316-7330, 2006. https://doi.org/10.1016/j.atmosenv.2006.06.040

POLEZER, G.; TADANO, Y. S.; SIQUEIRA, H. V.; GODOI, A. F. L.; YAMAMOTO, C. I.; ANDRÉ, P. A.; PAULIQUEVIS, T.; ANDRADE, M. F.; OLIVEIRA, A.; SALDIVA, P. H. N.; TAYLOR, P. E.; GODOI, R. H. M. Assessing the impact of PM2.5 on respiratory disease using artificial neural networks. *Environmental Pollution*, v. 235, p. 394-403, 2018. https://doi.org/10.1016/j.envpol.2017.12.111

QUALAR (2019). *Qualidade do Ar*. Dados meteorológicos. CETESB. Available from: <https://cetesb.sp.gov.br/ar/qualar>. Accessed on: May 8, 2019.

REINHARDT, T. E.; OTTMAR, R. D.; CASTILLA, C.; Smoke Impacts from Agricultural Burning in a Rural Brazilian Town. *Journal of the Air & Waste Management Association*, v. 51, n. 3, p. 443-450, 2011. https://doi.org/10.1080/10473289.2001.10464280

SADAT, Y. K.; KARIMIPOUR, F.; SADAT, A. K. Investigating the relation between prevalence of asthmatic allergy with the characteristics of the environment using association rule mining. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, v. 40, n. 2W3, p. 169-174, 2014. https://doi.org/10.5194/isprsarchives-XL-2-W3-169-2014

SAIDE, P. E.; CARMICHAEL, G. R.; SPAK, S. N.; GALLARDO, L.; OSSES, A.; MENA-CARRASCO, M.; PAGOWSKI, M. Forecasting urban PM10 and PM2. 5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model. *Atmospheric Environment*, v. 45, n. 16, p. 2769-2780, 2011. https://doi.org/10.1016/j.atmosenv.2011.02.001

SANTOS, F. S.; PINTO, J. A.; MACIEL, F. M.; HORTA, F. S.; ALBUQUERQUE, T. T. A.; ANDRADE, M. F. Avaliação da influência das condições meteorológicas na concentração de material particulado fino (MP2,5) em Belo Horizonte, MG. *Engenharia Sanitária e Ambiental*, v. 24, n. 2, p. 371-381, 2019. https://doi.org/10.1590/s1413-41522019174045

SANTOS, T. C.; CARVALHO, V. S. B; REBOITA, M. S. Avaliação da influência das condições meteorológicas em dias com altas concentrações de material particulado na Região Metropolitana do Rio de Janeiro. *Engenharia Sanitária e Ambiental*, v. 21, n. 2, p. 307-313, 2016. https://doi.org/10.1590/s1413-41522016139269

SÃO PAULO. *Decreto nº 59.113, de 23 de abril de 2013*. Estabelece novos padrões de qualidade do ar e dá providências correlatas. Com retificações posteriores. São Paulo, 2013. Available from: <https://www.al.sp.gov.br/repositorio/legislacao/decreto/2013/decreto-59113-23.04.2013.html>. Accessed on: Dec., 2019.

SEINFELD, J. H.; PANDIS, S. N. *Atmospheric Chemistry and Physics from Air Pollution to Climate Change*. 3rd ed. New York: Wiley, 2016.

SOUZA, F. T.; RABELO, W. S. A data mining approach to study the air pollution induced by urban phenomena and the association with respiratory diseases. *In*: INTERNATIONAL CONFERENCE ON NATURAL COMPUTATION, 2016. *Proceedings*… 2016. p. 1045-1050. https://doi.org/10.1109/ICNC.2015.7378136

WORLD HEALTH ORGANIZATION (WHO). *Nine out of ten people worldwide breathe polluted air, but more countries are taking action*. WHO, 2019. Available from: <https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action>. Accessed on: May 8, 2019.

XIAO, C.; CHANG, M.; GUO, P.; YUAN, M.; XU, C.; SONG, X.; XIONG, X.; LI, Y.; LI, Z. Characteristics analysis of industrial atmospheric emission sources in Beijing–Tianjin–Hebei and Surrounding Areas using data mining and statistics on different time scales. *Atmospheric Pollution Research*, v. 11, n. 1, p. 11-26, 2020. https://doi.org/10.1016/j.apr.2019.08.008

YANAGI, Y.; ASSUNÇÃO, J. V.; BARROZO, L. V. The impact of atmospheric particulate matter on cancer incidence and mortality in the city of São Paulo, Brazil Influência do material particulado atmosférico na incidência e mortalidade por câncer no Município. *Cadernos de Saúde Pública*, v. 28, n. 9, p. 1737-1748, 2012. https://doi.org/10.1590/S0102-311X2012000900012

ZOU, B.; PENG, F.; WAN, N.; MAMADY, K.; WILSON, G. J. Spatial cluster detection of air pollution exposure inequities across the United States. *PLoS One*, v. 9, n. 3, e91917, 2014. https://doi.org/10.1371/journal.pone.0091917