UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE TECNOLOGIA

Angela Rosa Locateli de Godoy

Relationship between air pollutants and respiratory diseases: analysis by spatial data clustering and temporal association rules

Relação entre poluentes do ar e doenças respiratórias: análise por agrupamento de dados espacial e regras de associação temporais

Limeira
2023

Angela Rosa Locateli de Godoy

Relationship between air pollutants and respiratory diseases: analysis by spatial data clustering and temporal association rules

Relação entre poluentes do ar e doenças respiratórias: análise por agrupamento de dados espacial e regras de associação temporais

Thesis presented to the School of Technology of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Technology in Computer Science, in the area of Sistemas de Informação e Comunicação.

Tese apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutora em Tecnologia, na área de Sistemas de Informação e Comunicação.

Orientadora: Profa. Dra. Ana Estela Antunes da Silva

Este exemplar corresponde à versão final da Tese defendida por Angela Rosa Locateli de Godoy e orientada por Profa. Dra. Ana Estela Antunes da Silva.

Limeira
2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Tecnologia
Felipe de Souza Bueno - CRB 8/8577

G548r

Godoy, Angela Rosa Locateli, 1977-
Relationship between air pollutants and respiratory diseases : analysis by spatial data clustering and temporal association rules / Angela Rosa Locateli de Godoy. – Limeira, SP : [s.n.], 2023.

Orientador: Ana Estela Antunes da Silva.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Tecnologia.

1. Ar - Poluição. 2. Mineração de dados (Computação). 3. Doenças respiratórias. 4. Análise por agrupamento. 5. Mineração de regras de associação. I. Silva, Ana Estela Antunes da, 1965-. II. Universidade Estadual de Campinas. Faculdade de Tecnologia. III. Título.

Informações Complementares

Título em outro idioma: Relação entre poluentes do ar e doenças respiratórias : análise por agrupamento de dados espacial e regras de associação temporais
Palavras-chave em inglês:
Air - Pollution
Data mining
Respiratory tract diseases
Cluster analysis
Association rule mining
Área de concentração: Sistemas de Informação e Comunicação
Titulação: Doutora em Tecnologia
Banca examinadora:
Ana Estela Antunes da Silva [Orientador]
Danilo Covaes Nogarotto
Leila Droprinchinski Martins
Luiz Camolesi Júnior
Yara de Souza Tadano
Data de defesa: 17-03-2023
Programa de Pós-Graduação: Tecnologia

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-2858-5189
- Currículo Lattes do autor: http://lattes.cnpq.br/6221953390766236

Folha de Aprovação

Abaixo se apresentam os membros da comissão julgadora da sessão pública de defesa de dissertação para o Título de Doutora em Tecnologia na área de concentração Sistemas de Informação e Comunicação, a que se submeteu a aluna Angela Rosa Locateli de Godoy, em 17 de março de 2023 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

Profa. Dra. Ana Estela Antunes da Silva

Universidade Estadual de Campinas

Presidente da Comissão Julgadora

Prof. Dr. Danilo Covaes Nogarotto

Universidade Estadual de Campinas - UNICAMP / CPFL Energia

Profa. Dra. Leila Droprinchinski Martins

Universidade Tecnológica Federal do Paraná - UTFPR

Prof. Dr. Luiz Camolesi Júnior

Universidade Estadual de Campinas - UNICAMP

Profa. Dra. Yara de Souza Tadano

Universidade Tecnológica Federal do Paraná - UTFPR

Ata da defesa, será assinada pelos membros da Comissão Examinadora e constará no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-graduação da FT.

Acknowledgements

I thank God first of all, for giving me strength and wisdom at all times.

To all my family for their support, especially to my parents Maria Therezinha Siqueroli Locateli (in memorian) and Otávio Locateli, who, despite having little schooling, taught me the greatest values in life and the path to be followed. My eternal gratitude.

I would like to thank my children Gabriella and Eduardo for understanding and waiting for this very special moment. To my husband Henri for his unceasing support, sharing my goals and all his contributions during these years.

To my advisor, Prof. Ana Estela Antunes da Silva, for her teaching and sharing knowledge, constant support, and encouragement at all times during this work. I really appreciate the trust and opportunity given to me.

To the professors Dr. Simone Andrea Pozza and Dr. Guilherme Palermo Coelho for their valuable contributions and discussions.

To my friend Mirelle Candida Bueno for the partnership and friendship.

Finally, thank everyone who contributed directly or indirectly to execution this work.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Dedication

To my mother, my greatest example of simplicity, faith, and strength! Despite of her sore departure, the trust, she always placed in me, made me come here far. To you mother, my eternal and great love!!

Resumo

Bancos de dados espaço-temporais estão se tornando cada vez mais comuns e a busca por técnicas mais relevantes que considerem restrições de tempo e relações espaciais na mineração dos dados, torna a tarefa mais complexa. Embora alguns estudos tenham identificado padrões interessantes, eles não lidam de forma apropriada com intervalos de tempo, com diferentes durações e frequências, o que permite encontrar padrões mais significativos e estimar sua relevância ao longo do tempo. Para entender a situação atual da poluição do ar em todo o Estado de São Paulo, esta tese investiga padrões espaciais e temporais multipoluentes (MP_{10}, $MP_{2,5}$, NO_2, SO_2, O_3 e CO), envolvendo 56 estações com monitoramento de qualidade do ar, com cobertura de áreas urbanas, litoral e interior, o que torna o estudo mais abrangente e conclusivo, incluindo áreas menos citadas em outros trabalhos. O uso da técnica de agrupamento levou à descoberta de padrões interessantes, na busca por (dis)similaridades no tempo para encontrar padrões de comportamento dos poluentes em séries temporais. O agrupamento revelou disparidades na distribuição espacial dos poluentes e sazonalidades, permitindo caracterizar e identificar espacialmente a poluição atmosférica de toda região, com a qualidade avaliada pelo coeficiente de silhueta de 0,26 a 0,72. Na avaliação espacial, os grupos severamente poluídos localizavam-se na região metropolitana, no litoral e, algumas cidades do interior, por emissões industriais, veiculares, queimadas, agrícolas e outras. O agrupamento mostrou uma forte presença de O_3 e $PM_{2.5}$ em 65% e 72% das estações monitoradas em diversas regiões do Estado. Os grupos de PM_{10} e NO_2 estavam geograficamente distantes, enquanto $PM_{2.5}$, CO, SO_2 e O_3 mais próximos, sugerindo uma relação espacial de exposição. A partir dos grupos de séries temporais encontrados, regras de associação foram aplicadas para encontrar possíveis coocorrências entre episódios críticos de poluição e variáveis meteorológicas (temperatura do ar, umidade relativa do ar, velocidade do vento, índice pluviométrico e radiação solar global). As regras de associações permitiram identificar os elementos que ocorreram simultaneamente, com suporte e confiança significativos, superiores a 80%. As condições meteorológicas que contribuíram para os episódios críticos de poluição foram baixa temperatura e umidade, baixa pluviosidade e vento mais ameno associados ao aumento de concentração dos poluentes. Como as regras de associação clássicas não refletem a dimensão tempo, este trabalho implementa e valida o algoritmo ARMADA para: (i) busca de padrões associados a intervalos de tempo para identificar padrões mais frequentes e (ii) representar relações temporais por regras de associação baseadas nos padrões encontrados. A partir das regras de associação temporais geradas avalia-

se a relação de curto prazo entre os eventos críticos de poluição, para os poluentes de maior influência nas internações por doenças respiratórias. Nos resultados, as internações foram recorrentes na transição do verão para os períodos mais frios. Em aproximadamente 35% do total de dias com internação maior que a média anual, um ou mais poluentes tiveram alta concentração. As regras mostraram que os poluentes $PM_{2.5}$, PM_{10} e O_3 estão fortemente associados ao aumento de internações na cidade de São Paulo ($PM_{2.5}$ e PM_{10} com 38,5% de suporte e 77% de confiança), em Campinas ($PM_{2.5}$ com 66,1% de suporte e 94% de confiança) e o poluente O_3 com suporte máximo de 17,5%. No litoral (cidade de Santos), o SO_2 esteve relacionado às altas internações (43,85% de suporte e 80% de confiança). Com a duração de cada padrão, a Álgebra Intervalar de Allen (AIA) avaliou as relações temporais e indicou em quais intervalos de dias houve elevação no número de internações e por quanto tempo elas se sustentaram, bem como os efeitos na hospitalização após a exposição. As regras mais frequentes e robustas apresentaram que exposição a altas concentrações dos poluentes $PM_{2,5}$ e PM_{10} gerou internações no mesmo dia e novas admissões nos dias seguintes. Os poluentes que causaram o aumento das internações permanecem por três dias acima dos limites, oscilando em internações menores no 1º dia e novamente maiores no 2º e 3º dias. Paralelamente, mostra que menor concentração de todos os poluentes em dias seguintes ao pico de poluição, ainda mantém um alto índice de internações. Portanto, a pesquisa é de interesse amplo porque investiga a exposição da população à poluição do ar em uma grande massa de dados, que revela a representatividade espacial e padrões temporais de poluentes, bem como condições meteorológicas desfavoráveis à difusão, além de implementar algoritmos que exploram técnicas de mineração de dados por agrupamento de dados e regras de associação e evoluir cada um deles na abordagem espacial e temporal, abordando experimentos originais e passíveis de reprodução.

Palavras-chave: Poluentes Atmosféricos; Agrupamentos; Regras de Associação; Condições meteorológicas; Regras de Associação Temporais; Doenças respiratórias.

Abstract

Spatiotemporal databases are becoming more and more common and the search for more relevant techniques that consider time constraints and spatial relationships in data mining, makes the task more complex. Although some studies have identified interesting patterns, they do not properly deal with time intervals, with different durations and frequencies, which allows finding more significant patterns and estimating their relevance over time. To understand the current situation of air pollution throughout the State of São Paulo, this thesis investigates multi-pollutant spatial and temporal patterns (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3 and CO), involving 56 stations with air quality monitoring, with coverage of urban, coastal and inland areas, which makes the study more comprehensive and conclusive, including areas less cited in other works. The use of the clustering technique led to the discovery of interesting patterns, in the search for (dis)similarities in time to find behavior patterns of pollutants in time series. The grouping revealed disparities in the spatial distribution of pollutants and seasonality, allowing the spatial characterization and identification of atmospheric pollution throughout the region, with quality assured by the silhouette coefficient from 0.26 to 0.72. In the spatial evaluation, the severely polluted groups were located in the metropolitan region, on the coast and, in some cities in the interior, due to industrial, vehicular, burning, agricultural and other emissions. The grouping showed a strong presence of O_3 and $PM_{2.5}$ in 65% and 72% of the monitored stations in different regions of the State. The PM_{10} and NO_2 groups were geographically distant, while $PM_{2.5}$, CO, SO_2 and O_3 were closer, suggesting a spatial exposure relationship. From the groups of time series found, association rules were applied to find possible co-occurrences between critical episodes of pollution and meteorological variables (air temperature, relative humidity, wind speed, rainfall index and global solar radiation). The association rules allowed identifying elements that occurred simultaneously, with significant support and confidence, greater than 80%. The meteorological conditions that contributed to the critical episodes of pollution were low temperature and humidity, low rainfall and milder wind associated with increased concentration of pollutants. As the classic association rules do not reflect the time dimension, this work implements and validates the ARMADA algorithm to: (i) search for patterns associated with time intervals to identify more frequent patterns and (ii) represent temporal relationships by association rules based on the patterns found. From the temporal association rules generated to evaluate the short-term relationship between critical pollution events, for the pollutants with the greatest influence on hospitalizations due to respiratory diseases. In the

results, hospitalizations were recurrent in the transition from summer to colder periods. In approximately 35% of the total days with hospitalization greater than the annual average, one or more pollutants had a high concentration. The rules showed that the pollutants $PM_{2.5}$, PM_{10} and O_3 are strongly associated with the increase of hospitalizations in the São Paulo city ($PM_{2.5}$ and PM_{10} with 38.5% support and 77% confidence), in Campinas ($PM_{2.5}$ with 66.1% support and 94% confidence) and the pollutant O_3 with maximum support of 17.5%. On the coast (Santos city), SO_2 was related to high hospitalizations (43.85% support and 80% confidence). With the duration of each pattern, Allen's Interval Algebra (AIA) evaluated the temporal relationships and indicated in which intervals of days there was an increase in the number of hospitalizations and for how long they were sustained, as well as the effects on hospitalization after exposure. The most frequent and robust rules showed that exposure to high concentrations of $PM_{2.5}$ and PM_{10} pollutants generated admissions on the same day and new admissions on the following days. The pollutants that caused the increase in hospitalizations remain for three days above the limits, oscillating in smaller hospitalizations on the 1st day and again higher on the 2nd and 3rd days. At the same time, it shows that a lower concentration of all pollutants in the days following the pollution peak still maintains a high rate of hospitalizations. Therefore, the research is of broad interest because it investigates the population's exposure to air pollution in a large body of data, which reveals the spatial representativeness and temporal patterns of pollutants, as well as meteorological conditions unfavorable to the diffusion, in addition to implementing algorithms that explore data mining techniques by data clustering and association rules and evolve each one of them in the spatial and temporal approach, addressing original and reproducible experiments.

Keywords: Air pollutants; Association Rules; Meteorological conditions; Temporal Association Rules; Respiratory diseases.

Liste of Figures

List of Tables

List of Abbreviations and Acronyms

AI	Atmospheric Instability
AIA	Allen's Interval Algebra
AIH	Hospital Admission Authorizations
AQG	Global Air Quality Guidelines
AQLI	Air Quality Life Index
ARMADA	An algorithm for discovering richer relative temporal Association rules from temporal data
CEMADEN	National Center for Natural Disaster Monitoring and Alerts
CETESB	Environmental Company of the State of São Paulo
CID	International Classification of Diseases and Related Health Problems
CO	Carbon Monoxide
CONAMA	National Council for the Environment
Conf	Trust
DATASUS	Information Technology Department of the Brazilian Unified Health System
DV	Wind Direction
FMC	Smoke
GSR	Global Solar Radiation
IBGE	Brazilian Institute of Geography and Statistics
IEMA	Institute of Energy and Environment
INPE	National Institute for Space Research
MDB	Memory Database
MEMISP	MEMory Indexing for Sequential Pattern Mining
Minsup	Minimum support
NO_2	Nitrogen Dioxide
O_3	Ozone
PAHO	Pan American Health Organization
Pb	Lead
PIB	Gross National Product
PLUV	Rainfall
PM	Particulate Matter

$PM_{2.5}$	Fine Particulate Matter (aerodynamic diameter $d_a \leq 2{,}5$ μm)
PM_{10}	Coarse Inhalable Particulate Matter ($10 \geq d_a > 2{,}5$ μm)
PRE	Precipitation
PTS	Total Suspended Particles
QUALAR	Air Quality Platform
RH	Relative Humidity
RI	Rainfall Index
RMSP	Metropolitan Region of São Paulo
SO_2	Sulfur Dioxide
SMO	Smoke
Sup	Support
SUS	Unified Health System
TABNET	Tabulator of the Unified Health System databases
TARs	Temporal Association Rules
TEMP	Air Temperature
TSP	Total Suspended Particles
WD	Wind Direction
WHO	World Health Organization
WS	Wind Speed

Contents

1. Introduction

Air pollution is a threat to global health, responsible for more than 70% of all deaths in the world from causes directly related to high levels of pollution, equivalent to 41 million people, of which 85% in developing countries (NABIZADEH et al., 2019; CHEN and HOEK, 2020; PRANATA et al., 2020; WHO, 2021; POLEZER et. al, 2022). The Pan American Health Organization (PAHO, 2018) states that air pollution is still a significant challenge for Brazilian cities and states, increasingly urbanized, responsible for high death rates annually. In urban centers, there are high-level emissions, which include land, air, and water transport sources, industry and power generation, and biomass burning, among others, which explain at least 40% of the emission of fine particulate matter ($PM_{2.5}$) in six Brazilian states, including São Paulo (ANDRADE et al., 2012; CORÁ, LEIRIÃO, MIRAGLIA, 2020).

Although it is known that the air quality in the metropolises presents high levels of pollution, the countryside of the State of São Paulo has stood out in the search for urgent containment measures, which reflects industrial sources specific areas, intense traffic points or more general sources, such as fires and burning of agricultural residues, with resuspension (SOUZA, SCUR, HILSDORF, 2018; POLEZER et. al, 2022). Squizatto et. al (2021) states that 68% of the world's population lives in rural areas or small to medium-sized cities, impacted by distant sources of air pollution and Kawashima et al. (2020) demonstrate that for Brazil, atmospheric emissions from stationary sources play a fundamental role in the concentrations of atmospheric pollutants.

The World Health Organization (WHO), in September 2021 established new Global Air Quality Guidelines (AQG) what became more restrictive regarding safe levels, with even lower pollutant concentration values. This is an important milestone, considering that the previous update took place in 2005. However, many countries establish different standards, generally below the set limits (WHO, 2021; SANTANA et al., 2021). In the State of São Paulo, despite of the government implementing initiatives and setting targets to contain critical episodes of pollution, concentration levels still frequently exceed national limits (SÃO PAULO, 2013; CONAMA, 2018) and constantly international ones (WHO, 2021).

In the literature, 92% of monitoring stations indicates concentrations above the annual averages for $PM_{2.5}$ (CETESB, 2020; CORÁ, LEIRIÃO, MIRAGLIA, 2020). Abe and Miraglia (2018) report a reduction in the concentration of $PM_{2.5}$, by approximately 25.45%, in the city

of São Paulo, from 2000 to 2011 due to actions to contain the increase in the automotive fleet. However, according to the Institute of Energy and Environment - IEMA is still above the recommended by the WHO in the last 22 years, mainly due to particulates ($PM_{2.5}$ and PM_{10}), O_3, and NO_2 (IEMA, 2022). Andrade et al. (2017) presented the evolution of pollutants in the RMSP (Metropolitan Region of São Paulo) over the last thirty years. They indicated a small reduction in primary pollutants, such as CO, NO, SO_2, and PM_{10}, but the biggest challenge is the control of pollutants O_3 and $PM_{2.5}$.

The Environmental Company of the State of São Paulo - CETESB issues, daily, an air quality bulletin determined by the pollutant that has the highest daily concentration index, among the main atmospheric pollutants (WHO, 2021). However, the indices indicate when the concentrations reach a level where health is threatened. In a recent past, monitoring air quality was synonymous of managing it. However, management alone may not be enough. For the control and mitigation of high pollution levels, it is important to analyze the data at these levels, such as, for example, the analysis of pollutant levels that occurs simultaneously and the analysis of the relationship between pollutant emissions and diseases (GALVÃO et al., 2022).

Most studies focus on microdata, focused on the local topography of some cities and few pollutants, and most places of political relevance or industrial complexity (BELLINGER et al., 2017; RYBARCZYK, ZALAKEVICIUTE, 2018; SOMPORNRATTANAPHAN et al., 2020; BERGMANN et al., 2020). For this reason, this study considers most stations with automatic air monitoring by CETESB, which makes it more comprehensive, including less documented areas. With a multi-pollutant concentration analysis (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3 and CO), which integrates the effects of one or more pollutants based on concentration, with spatial parameters of latitude and longitude, this study focuses on distribution and not just on concentration thresholds.

Some studies apply statistical methods to interpret air pollutant data and probabilistic models (RYBARCZYK, ZALAKEVICIUTE, 2018; AMEER et al., 2019; REPRESA et al., 2019; BERGMANN et al., 2020). For example, the Bayesian confidence interval was used in Thailand to assess the $PM_{2.5}$ variation and road transport pollution from monitoring stations (THANGJAI et al., 2021). In China, the use of Poisson regression showed the association between lung cancer mortality and exposure to $PM_{2.5}$, PM_{10}, and NO_2 together with the occurrence of some meteorological factors (CHUNG et al, 2021).

The literature also indicates that data mining algorithms in atmospheric pollution generally treat time as a simple numerical attribute extracted from sequential data, ordered

linearly in time (KAM, FU, 2000; LAXMAN, SASTRY, 2006; WINARKO, RODDICK, 2007; MITSA, 2010). Few algorithms include time and space factors and consider time constraints and spatial relationships in extracting patterns in data intervals, making the execution more complex. Even so, some researchers have tried to adapt existing techniques for this (KAM, FU, 2000; MITSA, 2010; SHAHEEN, SHAHBAZ, GUERGACHI, 2013).

In addition to the lack of studies that take into account time and space factors in the analysis of pollutant data, most studies were carried out in the northern hemisphere, according to the systematic reviews by Bellinger et al. (2017) and Rybarczyk and Zalakeviciute (2018) and the least explored data mining techniques are data clustering (26%) and association rules (15%). Therefore, this study extends the existing literature by developing a temporal and spatial analysis and expands the application of the clustering techniques and association rules employed.

Temporal data can be represented by events (occurrence in time), time series (events based on regular intervals), or time intervals (different durations and frequencies) (WANG, SMITH, HYNDMAN, 2006; MITSA, 2010; WANG et al., 2018). When we deal with facts that last for a period of time, instead of analyzing the data as an instantaneous occurrence in chronological order, we consider the temporality of the data in time intervals, which allows finding more meaningful patterns and a better understanding of the relationships between intervals but requires computational complexity to estimate the relevance of a pattern over time. Most works opt for data chained in univariate time series, such as the maximum daily concentrations of a pollutant (KAM, FU, 2000; WINARKO, RODDICK, 2007; RAJ, PRASAD, BALAKRISHNAN, 2022).

The data clustering technique can be used to discover temporal patterns through the search for similarity in a given time interval, which allows for finding patterns of behavior in temporal series (KAM, FU, 2000; WINARKO, RODDICK, 2007; MITSA, 2010; AGHABOZORGI et al., 2015). In addition to the temporal aspect, the data mining technique by clustering allows viewing, by geolocation, in large data sets, via the use of spatial parameters such as latitude and longitude, the visual representation of the groups, which helps in understanding the characteristics of the groups data (HAN, KAMBER, PEI, 2011; AUSTIN et al., 2013; HUANG et al., 2015; AGHABOZORGI et al., 2015; JIN, HAN, 2017; ARCE et al., 2018; AMATO et al., 2020; GOVENDER, SIVAKUMAR, 2020; XIAO et al., 2020; YAO et al., 2020).

Research carried out in Brazil (NODARI, SALDANHA, 2016; GUIDETTI, PEREDA, 2018) and other countries that applied clustering techniques to the air pollution problem identified regions with similar air pollution patterns. In the USA, a survey grouped locations according to $PM_{2.5}$ levels and obtained groups by regions with similar industrial activity in these regions (AUSTIN et al., 2013). In China, studies involving 13 sites with similar $PM_{2.5}$ concentration data resulted in the discovery of three groups: two with industrial activities and others with agricultural and tourist activities (HUANG et al., 2015). A work by ZOU et al. (2014), carried out with USA urban census data, was used to investigate the population's exposure to air pollution, considering age, race, educational level, and income.

Therefore, in order to better understand the current situation of air pollution throughout the state of São Paulo, this work aimed to explore the data clustering technique (HAN, KAMBER, PEI, 2011; KWEDLO, 2011) to obtain spatial and temporal patterns of pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3 and CO), in groups with similar behaviors, involving stations with air quality monitoring throughout the State. Since previous studies are limited in exploring the spatial issue, the objective was reveal disparities in the spatial distribution of pollution throughout the territory, as well as the stations most impacted by each pollutant temporal variations.

At the same time, studies report that the behavior of pollutants in the atmosphere does not depend only on emissions but on the combination of unfavorable meteorological and topographical conditions that promote the concentration or dispersion of pollutants and potentiate the effects of pollution on the environment's health (AMATO, 2020; FEISTEL, HELLMUTH, 2021; CHIQUETTO et al., 2022; LEIRIÃO et al., 2022; LIU, 2022; SOBRINHO et al., 2023). In addition, there may be a contribution between cities through the long-distance transport of pollutants, which affects the climate, and the spread of pollution to neighboring regions (LI et al., 2017; NOGAROTTO, 2019; LIU et al., 2020; GALVÃO et al., 2022).

In the literature, most studies that use hypothesis generation apply mining by association rules, seeking to identify items that co-occur in a wide variety of data. Association rules show elements that occur together in a transaction. For example, an association rule can be represented as it follows: [temperature above average, wind speed below average] → [PM_{10} above average], with 98% support and 100% confidence (PAYUS et al., 2013; CAGLIERO et al., 2016; DU, VARDE, 2016; SOUZA, RABELO, 2016; BELLINGER et al., 2017; LI et al., 2020).

Studies present results with the association rules technique, such as Souza and Rabelo (2016), who applied the method to identify a set of variables that frequently occur together: concentration of air pollutants and rates of respiratory problems in Curitiba, Brazil. Sadat, Karimipour, and Sadat (2014) explored, through association rules, the effect of air pollution on asthmatic allergies. The study indicated that the distance to parks and roads, as well as the concentrations of pollutants 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 are not affected it. Du and Varde (2016) use association data mining algorithms, clustering, and classification to look for relationships between particulate matter, pollution, and vehicle traffic to make $PM_{2.5}$ predictions.

The temporal behavior of the pollutants identified by applying the clustering technique points to the variation in the concentration of pollutants over the years, shows a recurrent pattern. This group analysis was used as input for the association rules to identify the co-occurrence between critical pollution episodes and meteorological variables (air temperature, relative humidity, wind speed, rainfall index, and solar radiation global) predominant in the regions that were identified in the groups.

Short-term and long-term exposure to pollutants can induce severe health damage. In the short term, respiratory diseases may aggravate (JIANG, MEI, FENG, 2016; CHEN, HOEK, 2020), pre-existing cardiovascular diseases (PRANATA et al., 2020, BONT et al., 2022), and increased hospitalizations and emergency room visits. Cumulative and long-term exposure can aggravate chronic respiratory problems, cause cancer, and even cause the onset of diseases prematurely (SEINFELD, PANDIS, 2016; MACHIN, 2018; POLEZER et al., 2018; GONÇALVES et al., 2022; LEIRIÃO et al., 2020; WHO, 2021).

Epidemiological studies have been performed to investigate pollutants specific effects on mortality and cardiorespiratory morbidity. Gomes, Lucio, and Spyrides (2013) pointed out a significant association between the increase in the number of hospitalizations for asthma in children in 27 municipalities in greater São Paulo and the exposure to PM concentrations and meteorological factors, using the probability distribution of Poisson and interchangeable type correlation matrix. Nguyen et al. (2018) proposed a method that builds a network structure to encode relationships between sets of frequent items and discover patterns in the form of temporal association rules (TARs), where the rule is a particular type of cancer treatment and its consequent a set of co-occurring toxicities. Wang et al. (2018) included a temporal

relationship between the various items in association rules with a frequent item set tree based on segmentation, discretization, and time series grouping.

According to data from the State Department of Health of São Paulo, more than 25 million people are exclusive users of the SUS, representing about 58% of the 43 million inhabitants of the State (MENDES, 2018). An economic impact on health is associated with air pollution in Brazilian metropolitan regions (MIRAGLIA, GOUVEIA, 2014). Even with a lot of data available on health systems, research with machine learning methods still requires analytical effort for extracting and interpreting indicators and difficulty in comparing them in time and space (RYBARCZYK, ZALAKEVICIUTE, 2018; AMEER et al., 2019; REPRESA et al., 2019; BERGMANN et al, 2020).

This study implemented and validated the ARMADA algorithm (An algorithm for discovering richer relative temporal association rules from temporal data), the result of best practices already incorporated but tested only on synthetic datasets and little discussed in the literature. According to Winarko and Roddick (2007) and Mitsa (2010), it proved superior to the algorithms from which it was generated. It is more efficient for finding frequent temporal patterns in large databases, with time intervals of different durations and frequencies, allowing to find more meaningful patterns and a better understanding of the relationships between intervals and data. Among the few studies involving the ARMADA algorithm, Silveira et al. (2018) proposed a thematic space-time association rule extractor that uses concepts from the algorithm, but it is quite peculiar for analyzing time series of solar satellite images. João (2020) developed a new method for mining rules temporal associations involving continuous quantitative data. This work was used as an inspiration for the development of the current proposal.

One of the contributions of this thesis is to employ temporal association rules for short-term temporal analysis, in the co-occurrence between the concentration of multi-pollutants and hospitalizations for respiratory diseases. To obtain temporal association rules for multipollutant analysis (PM_{10}, $PM_{2.5}$, NO_2, O_3, SO_2 and CO), the algorithm considers the temporality and duration of critical events of high concentrations (when the WHO concentration limits are outdated) and reveals, through time intervals, the behavior of the pollutants with the most significant influence on hospitalizations for respiratory diseases (CID-10). Allen's Interval Algebra (AIA) was applied to these temporal relations, which exposed the short-term relation between temporal rules. Three cities in São Paulo State with different topographic

characteristics were considered (metropolitan region, coastal, and inland) to assess whether the behavior is similar.

1.1 General objective and specific goals

The present work has the general objective of applying data mining techniques in order to discover patterns that relate the concentration levels of atmospheric pollutants with hospitalizations for respiratory diseases, taking into account the geolocation of the patterns and the time intervals in which the patterns happened. This objective can be systematized into specific goals highlighted below:

1. Explore how the clustering technique can contribute to the discovery of spatial and temporal patterns of pollutants in cities in the State of São Paulo;
2. Investigate, through association rules, considering the groups that present a high concentration of pollutants, co-occurrences between pollutant concentration, and meteorological variables in different regions;
3. Design and implement an algorithm for extracting temporal association rules for multivariate series in search of frequent patterns, which include temporal intervals and their relationships;
4. Analyze temporal rules that consider higher incidence multi-pollutants in the short-term relationship with the increase in hospitalizations for respiratory diseases, to track the behavior of evolutionary data in the discovery of patterns.

1.2 Thesis organization

The thesis was structured in an alternative format, as authorized by the Graduate Commission of the School of Technology of the University of Campinas.

The work is organized into seven chapters to develop the proposed research. Chapter 1 presents the introduction, contextualizing the subject, motivations, justifications, and research objectives. In the Chapter 2 are presented: the methodology employed, study site, data collection, exposition, and validation of the techniques in each experiment. In Chapters 3, 4, and 5, the articles resulting from the development of this research are arranged and organized according to the order of publication/submission, and, finally, in the Chapter 6, the conclusions of this research and future works.

2. Methodology

The study was carried out in the State of São Paulo (Brazil), involving 36 municipalities monitored by 62 automatic stations from CETESB (QUALAR, 2021), 29 stations in the RMSP and 33 stations in the countryside of the State and Coast in the period from January 2017 to October 2021 (1st and 2st steps: from January 2017 to December 2019 and 3st step: from January 2017 to October 2021). Data were obtained from three public databases at different steps of the research: (i) CETESB (QUALAR, 2021) for historical data on concentrations of pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, and O_3) and meteorological variables (air temperature, relative air humidity, wind speed, and global solar radiation); (ii) National Center for Monitoring and Natural Disaster Alerts (CEMADEN, 2020) for the historical rainfall index of monitored cities and (iii) TABNET/TabWin System of the Department of Information Technology of the Brazilian Unified Health System (DATASUS, 2021) with data from Hospital Admission Authorizations (AIH).

The proposed methodology and its evolution in three steps follow the processes presented in the scheme of Figure 2.1, which are discussed in detail in the following articles.

Published articles

- **Chapter 3:**

GODOY, ARL; SILVA, AEA; BUENO, MC; POZZA, SA; COELHO, GP (2021). Application of machine learning algorithms to $PM_{2.5}$ concentration analysis in the State of São Paulo, Brazil. Brazilian Journal of Environmental Sciences, 56, 152-165.

DOI: 10.5327/Z21769478782 (Published: Mar 2021)

- **Chapter 4:**

GODOY, ARL, SILVA, AEA (2022). Spatial patterns and temporal variations of pollutants at 56 air quality monitoring stations in the State of São Paulo, Brazil. Environmental Monitoring and Assessment 194, 910.

DOI: 10.1007/s10661-022-10600-z (Published: 18 October 2022)

Article subject

- **Chapter 5:**

GODOY, ARL, SILVA, AEA (2023). Short-term relation between air pollutants and hospitalizations for respiratory diseases: analysis by temporal association rules. Air Quality, Atmosphere and Health (Submitted: January 2023).

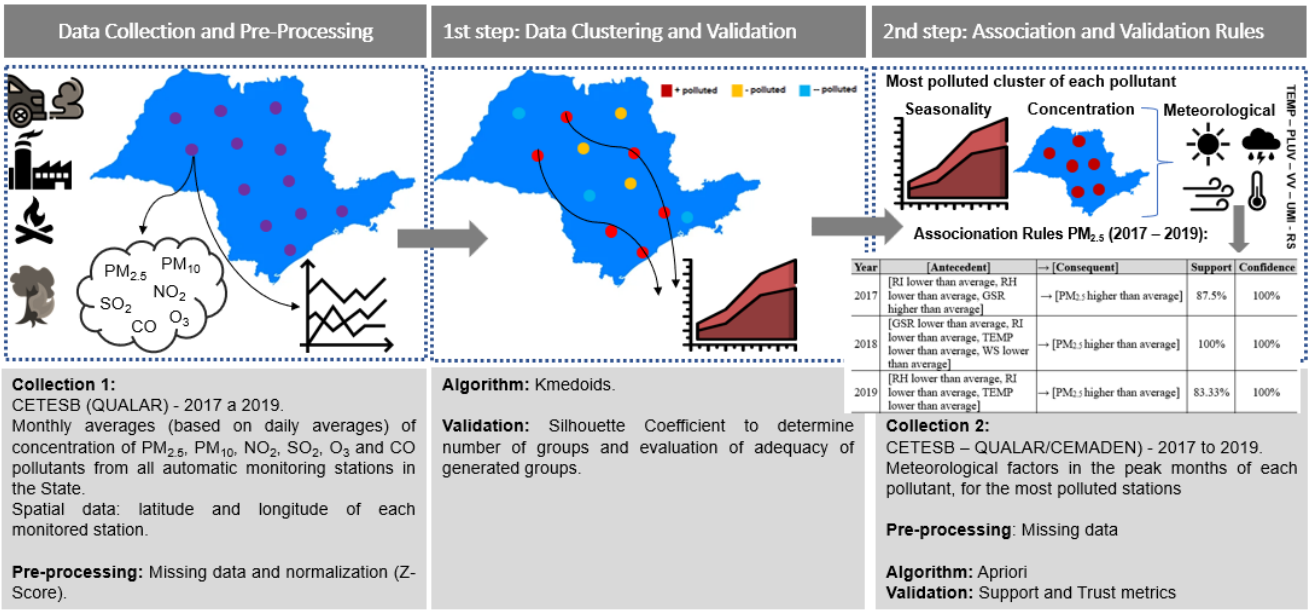


Figure 2. 1 - Methodology Overview (1st and 2nd step): Data Clustering and Association Rules

1st step: Clustering data mining implementation and analysis to discover frequent patterns of pollutants and their seasonal behavior. In the 1st article – chapter 3, the approach was applied experimentally only to the pollutant $PM_{2.5}$ to consolidate the techniques. Then this approach was expanded considering the other pollutants PM_{10}, NO_2, SO_2, O_3, and CO, from January 2017 to December 2019, which allowed an extension of the concept of spatiality and exploring the temporality of the groups generated in the 2nd article, that is, studying the spatial distribution of pollutants in periods of a higher concentration, to find out if, at the same time in other places, there was an increase in the same pollutant.

The basic clustering process consists of four main parts: dimensionality representation and reduction; similarity or distance measurement; clustering algorithm; and group evaluation (WANG et al., 2006; AGHABOZORGI et al., 2015). For similarity analysis, the K-medoids algorithm was implemented in the Python language, using open-source libraries specific for machine learning: Scikit-learn (PEDREGOSA et al., 2011), Pyclustering (NOVIKOV, 2019), Pandas (REBACK et al., 2020) and Geopy (KOSTYA, 2020). The method used as a decision criterion to determine the number of groups and assess their quality was the silhouette coefficient, which combines the concepts of cohesion and separation and consists of looking for the group with the highest silhouette value (KAUFMAN, ROUSSEEUW, 2005; MITSA, 2010).

The technique allows for identifying the behavior of the pollutants and knowing the groups formed for each one, focusing on the stations with the highest concentrations of pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, and O_3). The data obtained in this step allowed a seasonal analysis by pollutant, with similarities and variations between them and their groups. The groups are represented by geolocation through each monitoring station latitude and longitude parameters that reveal the geospatial distribution of pollutants in the State (GODOY et al. 2021; GODOY, SILVA, 2022a; GODOY, SILVA, 2022b).

Clustering time series is an approach widely used as an experimental technique for other mining algorithms, such as rule discovery, feature selection, classification, indexing, and anomaly detection, among others (MITSA, 2010; HAN et al., 2011). Therefore, the grouping result served as an input for the application of association rules in the 2nd step of the research, considering the clustering of stations categorized as severely polluted in all pollutants PM_{10}, $PM_{2.5}$, CO, NO_2, SO_2, and O_3.

2nd step: Implementation and investigation by association rules of the relationship between pollutants and meteorological variables (Temperature - TEMP, Wind Speed - VV, Relative Humidity - RH, Global Solar Radiation - GSR, and Pluviometric Index - IP) and identification of which meteorological conditions most influence each pollutant. The periods and groups with the highest values of concentration of pollutants obtained in the analysis of the groups were considered.

The process of searching for association rules was carried out in two phases. First, all possible combinations between attributes are considered to discover frequent itemsets with a minimum support value (minsup). Various rules are found at this step, but not all are interesting and useful. Then, rules that did not meet a minimum confidence threshold (minconf) were discarded (HAN et al., 2011; CASTRO, FERRARI, 2016). To obtain the association rules, the Apriori algorithm (AGRAWAL, IMIELINSKI, SWAMI, 1993) was used and implemented in the Python language using the "mlxtend" library (RASCHKA, 2018) and "Pandas" (REBACK et al., 2020). The suitability of the rule to evaluate the problem depends on the minimum support (minsup) and minimum confidence (minconf) values. Support and confidence value greater than 80% were considered. The proposal was to find all the patterns whose frequency was above a certain reliable limit of support and confidence.

With the rules, it is possible to identify the prevailing meteorological conditions when there are high concentrations of each pollutant in different regions. In addition to analyzing

meteorological factors and pollutants, a summary of the most frequent meteorological conditions during the entire period was carried out cities (GODOY et al. 2021; GODOY, SILVA, 2022a; GODOY, SILVA, 2022b).

Although exciting patterns have been identified, the Apriori algorithm is unsuitable for rule discovery over time intervals. As the temporality of data in intervals can present different durations and frequencies, which allows finding more significant patterns and a better understanding of the relationships between intervals (WINARKO, RODDICK 2007), research was carried out according to the 3rd step (Figure 2.2).

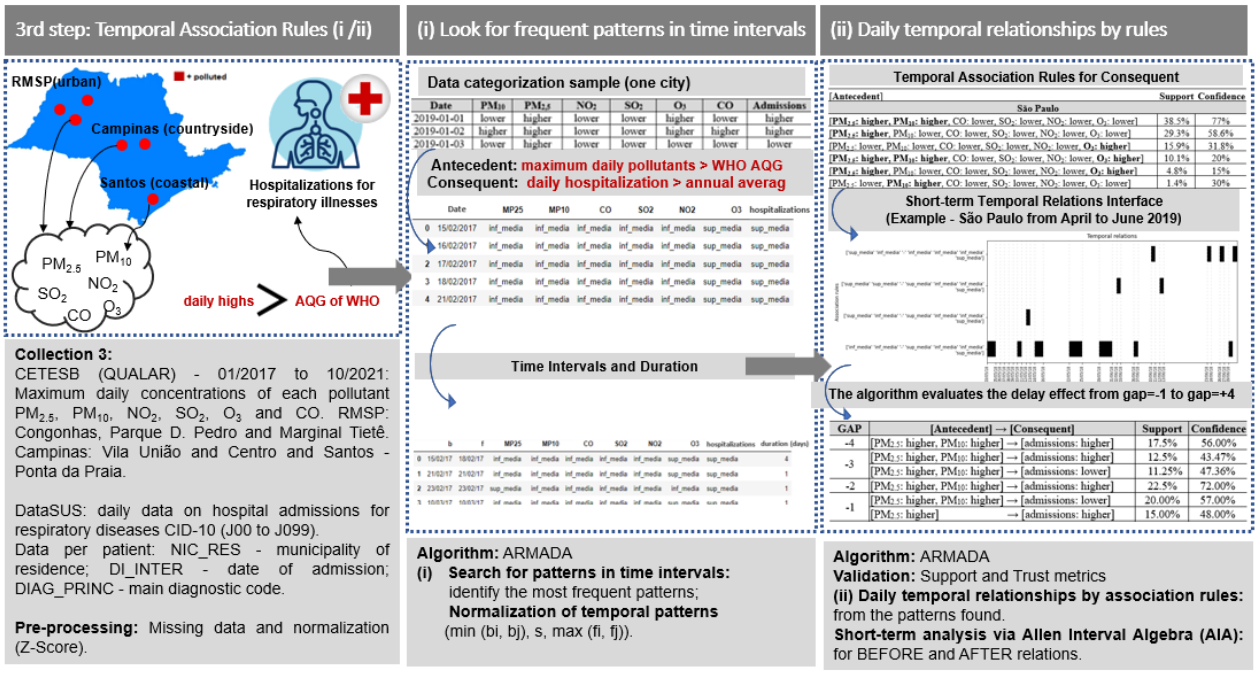


Figure 2. 2 - Overview of the Methodology (3rd step): Search for frequent patterns in time intervals and daily Time Relations by rules.

3rd step: Search for ideal techniques for exploring temporal data. Development and implementation of an algorithm that incorporates the mining of temporal association rules extending the ARMADA algorithm for temporal analysis of the impact on the number of hospitalizations for respiratory problems (CID-10: J00 to J99) from the frequent patterns due to the increase in air pollutants - PM_{10}, $PM_{2.5,}$ NO_2, SO_2, O_3 and CO.

At this stage, the study considers three cities in the State of São Paulo, which from January 2017 to October 2021 had the highest air pollution rates and the highest number of pollutants monitored by CETESB (QUALAR, 2021), but with different topographic characteristics: São Paulo located in the RMSP - Metropolitan Region of São Paulo, the coastal city of Santos and the city of Campinas, in the countryside of the State. Although the city of

Campinas presents expressive vehicle sources in some regions and some metropolitan characteristics, the choice was due to the availability of data for the entire period.

The ARMADA algorithm extends the MEMISP algorithm (MEMory Indexing for Sequential Pattern Mining) by Lin and Lee (2002), which explores frequent time patterns. MEMISP is the result of the combination of Apriori (AGRAWAL, IMIELINSKI, SWAMI, 1993) and GSP (AGRAWAL, SRIKANT, 1994) algorithms, based on the candidate generation and testing methodology and also on the FP-Grow (HAN, KAMBER, 2006) and PrefixSpan algorithms (HAN, KAMBER, PEI, 2011) that uses pattern growth. Based on MEMISP, which proved to be superior to the GSP (Generalized Sequential Patterns) and PrefixSpan algorithms from which was generated, ARMADA is the result of the best practices already incorporated, influenced by algorithms from both methodologies (pattern growth and candidate generation and testing).

Proposed by Winarko and Roddick (2007), ARMADA seeks to discover temporal patterns of data based on time intervals and generate temporal association rules. In tests on synthetic data sets, it proved superior to the algorithms from which was generated, and more efficient. For this research, the ARMADA algorithm was interpreted from its original proposal (WINARKO, RODDICK 2007) and implemented from its very generic pseudocode. The Python language was chosen due to its ease in handling unstructured data and scope of applicability. Open-source libraries, specific for machine learning, were also incorporated: Scikit -learn, Mlxtend, Pandas, Seaborn, Matplotlib, Apriori, Association_rules, and Numpy (PEDREGOSA et al., 2011; RASCHKA, 2018; REBACK et al., 2020). The specific objectives were:

(i) Search for frequent patterns recursively, with the help of an index table, which associates each pattern with the start and end times and considers the state of each variable in identifying the pattern. To find the patterns, the algorithm identifies time points in which the continuous attributes (pollutants and/or hospitalizations) present a state or behavior of interest, defined as a pollutant concentration higher than the limit established by the WHO.

The union of immediately neighboring time points allows the construction of time intervals. The algorithm employs notations proposed by Höppner (2001) that uses normalization of temporal patterns, having as reference the binary relations of Allen (1983). For this, the intervals are sorted according to their starting time points (*b1, s1, f1*), (*b2, s2, f2*), ..., (*bn, sn, fn*), ... where $bi \leq bi+1$ and $bi < fi$. Therefore, the sequential intervals are replaced

by a single gap for the pattern, resulting from the union of both, *(min (bi, bj), s, max (fi, fj))*. For a given period, the average pollutant concentration is considered the state <is>, represented by *<01/01/2019, 25µg/m3, 03/01/2019>*.

(ii) Represent the daily temporal relationships by association rules based on the patterns found. The process starts with the database of time intervals, when a new pattern is identified, the counter starts *(n=1)* and increments *(n=n+1)* when locating the same pattern in other time gaps until it registers all occurrences of the same pattern. The process is repeated for all the most frequent patterns, forming temporal association rules. All generated rules' support and confidence values are calculated to select the most frequent (the highest support) and the strongest (the highest confidence). In this experiment, the objective is to seek the rules with the greatest support and confidence.

Although temporal rules are sufficient to state which pollutants are associated with high hospitalization in different regions, a short-term analysis can still visualize the days and intervals in which each pattern occurred and the temporal relationships between them. The BEFORE applied Allen's Interval Algebra (AIA) and AFTER relations to describe the temporal relations. Such a contribution was not found in the literature and exposed the short-term relationship between temporal patterns.

Although it is relevant to understand the behavior of pollutants in the period when there is a rise in hospitalizations, it is equally important to investigate the temporal association between hospitalizations and pollution levels in the days before and later. The algorithm provides a time constraint (*maximum_gap*) applied by selecting one of the temporal association rules, which considers only relationships between intervals that meet the predefined *gap.* In this study, the *maximum_gap* = 4 was adopted, which considers four days before the discharge period (*gap=-1, gap=-2, gap=-3, gap=-4*) and four days after (*gap=1, gap =2, gap=3, gap=4*). The gap choice considers the period suggested in other epidemiological studies (MATOS et al., 2019; KHOSRAVI et al., 2020; NADALI et al., 2022).

Thus, in this research, the mining techniques of Data Clustering with spatial extension, Association Rules, and Temporal Association Rules were implemented in real databases of concentrations of atmospheric pollutants and occurrence of meteorological factors. The development of the methodology is in the form of scientific articles presented in sections 3 to 5 below, and the source code that comprises each step shown here is shared in a repository (https://github.com/angelalocateli)

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

3.1 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 (NODARI, SALDANHA, 2016; SEINFELD, PANDIS, 2016; MACHIN, NASCIMENTO, 2018; POLEZER et al., 2018) 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 (ANDRADE et al., 2012; DIMITRIOU, 2016; QUALAR, 2019). It is classified by its aerodynamic diameter (ad): particles with ad $\leq$ 2.5 μm are named $PM_{2.5}$ (fine inhalable particulate matter) and those with 10 $\geq$ ad > 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; CETESB, 20; MORAES et al., 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, and their effects on air quality (AMEER et al., 2019). The analysis 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 (AGRAWAL, SRIKANT, 1994; NEIROTTI et al., 2014). 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.

3.2 Data and Methods

The methodology used in this study will be presented as follows: (i) 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; (ii) an explanation of clustering algorithms and association rules, as well as their respective validation metrics.

3.2.1 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 3.1) 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.

Table 3. 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

Note: [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).

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 3.1) 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 3.2 along with their geolocation (Figure 3.1).

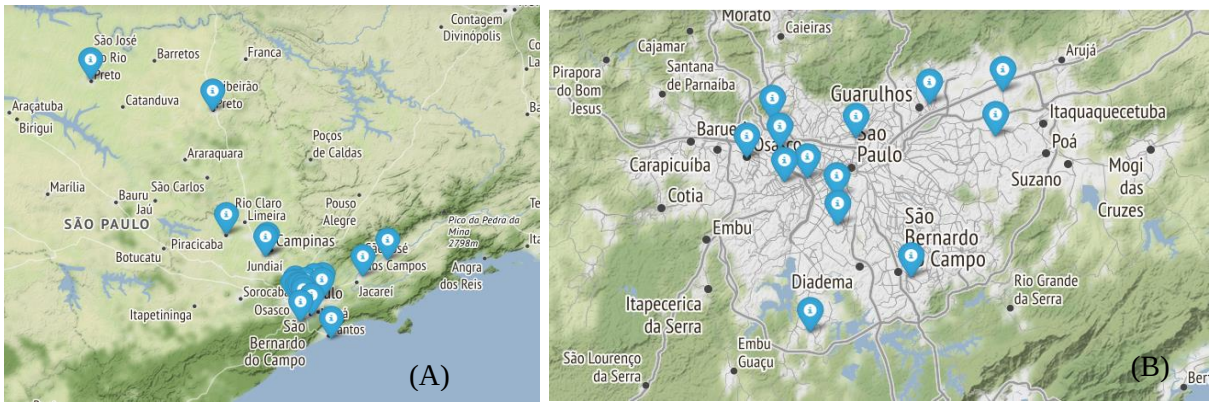


Figure 3. 1 - (A) Map of location of the automatic $PM_{2.5}$ monitoring stations in the state of São Paulo and (B) $PM_{2.5}$ monitoring stations in the Metropolitan Region of São Paulo (RMSP).

Table 3. 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”

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

3.2.2 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 whose data were obtained for the first database, seven met this new criterion (Table 3.3).

Table 3. 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 hig 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

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

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 (from each of the years) value of its respective meteorological variable or CO concentration. Table 3.3 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 steps: (i) application of the partitional clustering algorithm (K-medoids); (ii) association rules (Apriori). The next sections discuss these algorithms.

3.2.3 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 3.1 represents the average Silhouette calculation (S_p).

$$S_p = \sum_{1}^{n} \frac{s(x_i)}{n} \quad (3.1)$$

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 3.2:

$$s(x_i) = \frac{\left(b(x_i) - a(x_i)\right)}{max\,\{a(x_i), b(x_i)\}} \quad (3.2)$$

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.

3.2.4 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.

3.3 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 to 1, both algorithms presented significant and very close mean silhouette values, but the *k-medoids* algorithm was selected for being is capable of handling outliers.

Figure 3.2 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, on the x-axis of figure 3.2 - number of clusters.

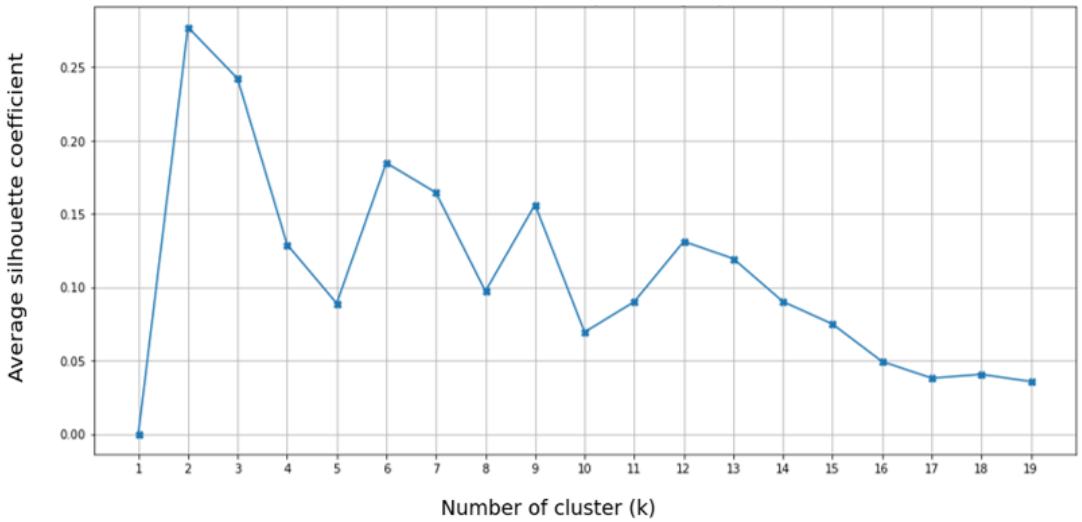


Figure 3. 2 - Number of clusters (k) per silhouette coefficient value, obtained from the K-medoids algorithm, applied to the database of monthly averages of $PM_{2.5}$ concentration, between 2017 and 2019.

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 3.4.

Table 3. 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

In the analyzed period, for all the stations monitored, the average annual concentrations of $PM_{2.5}$ were 16.43 $\mu g/m^3$ (standard deviation 6.45 $\mu g/m^3$) in 2017, 16.24 $\mu g/m^3$ (standard deviation 6.42 $\mu g/m^3$) in 2018, and 16 $\mu g/m^3$ (standard deviation 5.04 $\mu g/m^3$) in 2019, exceeding the annual threshold of 10 $\mu g/m^3$ 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 $\mu g/m^3$ and standard deviation of 4.72 $\mu g/m^3$;

- **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 3.3 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 (RMSP), as found in other studies (AUSTIN et al., 2013; HUANG et al., 2015). 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.

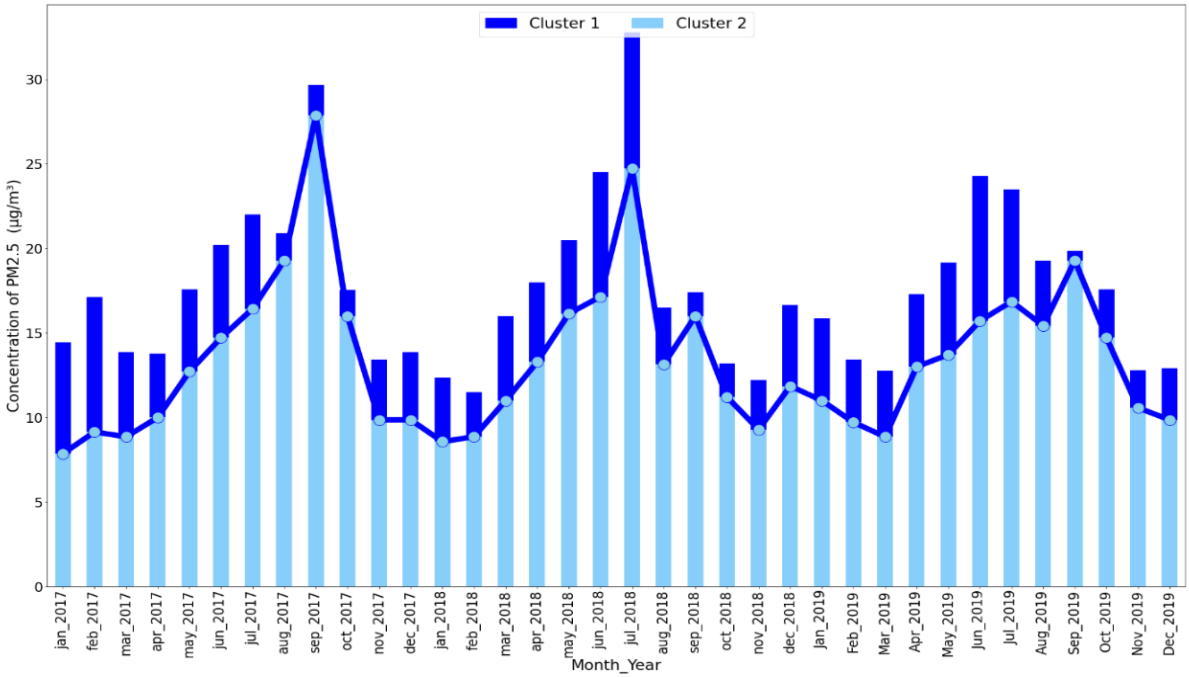


Figure 3. 3 - Comparison of the monthly averages of $PM_{2.5}$ concentrations (μg/m^3) between 2017 and 2019, in the cities of the state of São Paulo, with Cluster 1 being characterized mostly by the RMSP and Cluster 2, by inland cities.

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 (BISHT, SEEJA, 2018; LI et al., 2020), and which were analyzed with the association rules algorithm (Apriori).

Figure 3.4 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 3.4(A) and 3.4(B).

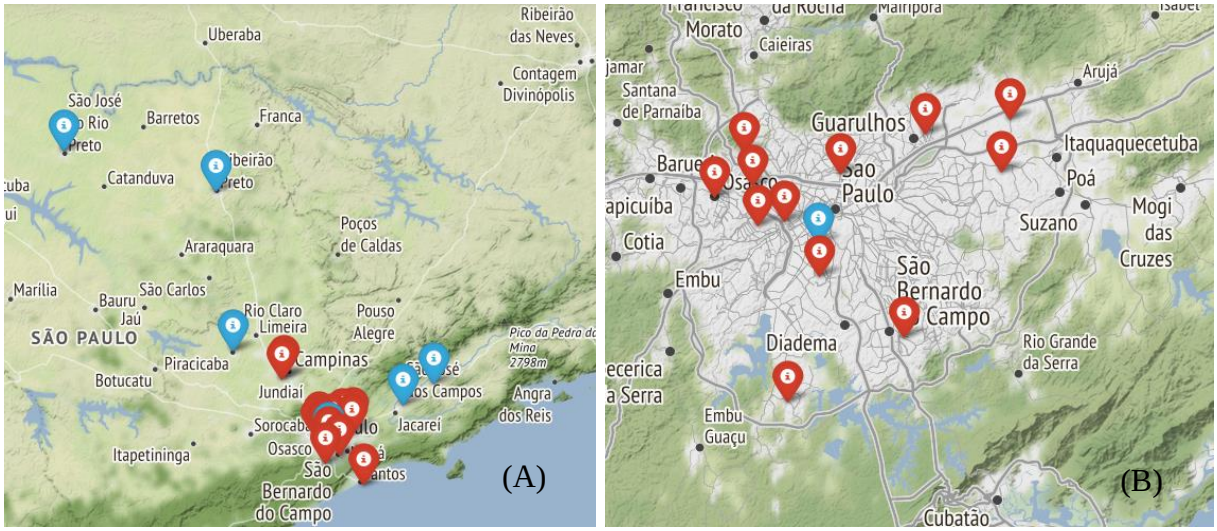


Figure 3. 4 - (A) Visualization by geolocation of the clusters, created by the K-medoids algorithm; (B) proximity of the elements of Cluster 1 on the map. Cluster 1 in red and Cluster 2 in blue.

The analysis on the map shows that most of the $PM_{2.5}$ monitoring stations present in Cluster 1 are in the Metropolitan Regions (RM) 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 RMs is fuel burning by the vehicle fleet and intense industrial emissions (YANAGI, ASSUNÇÃO, BARROZO, 2012; HUANG et al., 2015; CARDOSO et al., 2017). 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 farther from intense traffic routes and the afforestation mitigates the effects of pollution.

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 RMs 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 3.5. However, the interquartile ranges and maximum values (disregarding outliers) are similar.

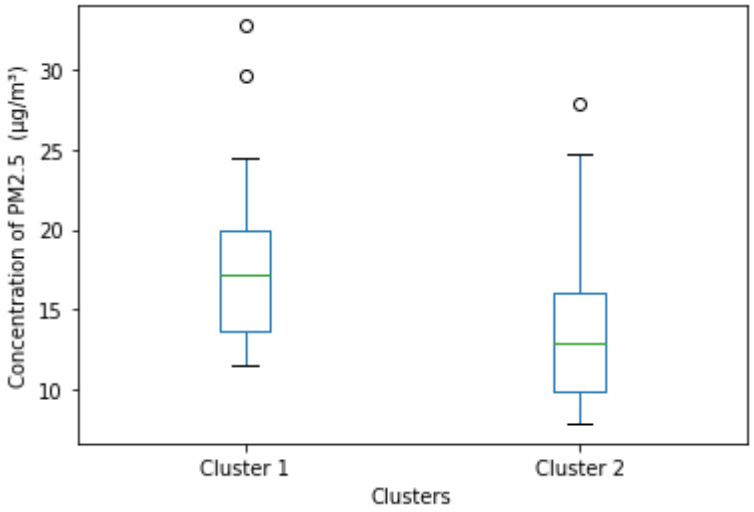


Figure 3. 5 - Boxplots of Clusters 1 and 2, formed by the monthly averages of $PM_{2.5}$, between 2017 and 2019.

Table 3.5 shows that, in 2017, the $PM_{2.5}$ concentration level increased from May to October, with a peak of about 29.8 $\mu g/m^3$ in September. Likewise, in 2018, the increase occurred from March to September, with a peak of 32.4 $\mu g/m^3$ 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 $\mu g/m^3$ 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 3.3 and Table 3.5), that is, September 2017, July 2018, 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 3.6. 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."

Table 3. 5 - Monthly averages of $PM_{2.5}$ by clusters of stations and standard deviation of the clusters (in $\mu g/m^3$), 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

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 3.6).

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.

Table 3. 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%

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

In the last Apriori execution, June 2019 data were used, and nine rules were obtained, two of which were the most representative (Table 3.6). 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 3.6 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).

3.4 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 RMSP 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 $PM_{2.5}$, 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.

4. Spatial patterns and temporal variations of pollutants at 56 air quality monitoring stations in the state of São Paulo, Brazil

4.1 Introduction

Large-scale urbanization, economic development, and rapid industrialization have contributed to the deterioration of air quality, by increasing the emission and concentration of atmospheric pollutants, in addition to inducing climate change, generating additional risks to public health and the environment. Air pollution is considered a risk factor for chronic diseases such as diabetes, cancer, and cardiovascular diseases and has caused premature diseases and deaths, being responsible for more than 70% of all deaths in the world, equivalent to 41 million people, of which 85% in low and middle-income countries due to high volumes of emissions (POLEZER et al, 2018; WHO, 2021; LEIRIÃO et al., 2020). However, the behavior of pollutants in the atmosphere does not depend only on local emissions, but on the combination of meteorological and topographical conditions unfavorable to dispersion. Besides, there may be a contribution between cities, through the transport of long-range pollutants, which affect climate and the spread of pollution to neighboring regions (LI et al., 2017; LIU et al. 2020). Therefore, air quality can still change depending on the weather conditions present, such as temperature and relative humidity, rainfall, wind speed and direction, global solar radiation, and other parameters that directly interfere with the dispersion and reduce the concentration of pollutants, helps in the dissolution of gases, and produces adverse effects (SANTOS et al., 2016; LIU et al., 2019; AMATO, 2020).

Given the relationship between concentration of pollutants and meteorological conditions, several studies prove that meteorology promotes the formation or dispersion of pollutants, altering their load. In China, a sharp and short-term increase in the concentration of pollutants in several regions occurred during the winter and meteorological conditions (strong wind, higher humidity, and increased pressure) favored the accumulation of particulate matter generating great fog (LIU et. al, 2021). A study in Austria concluded that reductions in NOx emissions resulted in lower peaks in O_3 concentrations in rural areas, but this effect was greater with solar radiation, more present in summer than in spring and daytime rather than nighttime cycles (STAEHLE et al., 2022). Studies such as Bisht and Seeja (2018), in India, predict the air quality of the next day based on pollutant concentration data from the previous day (PM_{10}, $PM_{2.5}$, NO_2, CO, and O_3) and meteorological variables (humidity relative, rainfall, temperature,

wind speed and direction). In Brazil, Australia, United States, and other countries, the frequency and intensity of fires have impacted the quality of the air. A study classified the risk of fires by combining spatial data and meteorological estimation principles applied to the burning of vegetation, among the meterological factors are the accumulation of consecutive days without rain, low humidity, and high temperatures, which can be aggravated by strong wind (GALVÃO et al., 2022). Reinhardt, Ottmar, and Castilla (2011) observed that 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 in urban centers, moderately polluted.

The increase in air pollution also contributes to fluctuations in the climate system that result in meteorological data that diverge from the average predicted for a given period. Climate change has been debated internationally and Brazil, as well as China, the United States, and other countries, is considered one of the largest emitters of greenhouse gases (Carbon Brief, 2021). Most emissions are from fuel combustion and industrial emissions, but also from deforestation and land use changes that increase the release of greenhouse gases and contribute significantly to global warming. Feistel and Hellmuth (2021) demonstrate in their study that the hydrological cycle is the dominant component of the climate system and ocean evaporation is directly related, so that small changes in the oceanic evaporation rate would be enough to offset or double the warming effect. atmospheric greenhouse. As a result, we have long periods of droughts and prolonged droughts, increased temperature, and water insufficiency, with prediction of being even more intense in the future, but already quite frequent in several regions of Brazil and several countries (GALVÃO et al., 2022).

It is also possible to perceive that meteorological conditions potentiate the effects of pollution on health and the seasonal variation of pollutants exposes this relationship. Studies report that the high concentration of atmospheric pollutants and meteorological factors such as low relative humidity and lack of precipitation, contributes to the increased risk of hospitalizations and deaths caused by respiratory diseases, such as the emergence of flu, rhinitis, bronchitis, asthma, and the growth of viruses, bacteria, molds, fungi, and allergens (LAM et al., 2016; MORAES et al., 2019; ESCOBAR, 2020). Gomes et al. (2013) showed a significant association between the increase in the number of hospitalizations for asthma in children in 27 cities in the state of São Paulo, Brazil, due to exposure to concentrations of particulate matter and the action of meteorological factors. The Pan American Health Organization (PAHO, 2018) reports that air pollution is still a major challenge for Brazilian cities and states, accounting for more than 51,000 deaths annually. According to reports, the state of São Paulo considered the

largest and most populous economy in Brazil (IBGE, 2021) has been facing problems arising from air pollution. The Institute for Energy and the Environment reported that in some cities the overshoot was four times the WHO recommended (SANTANA et al., 2021).

Despite State Decree No. 59.113 (São Paulo, 2013) setting goals for the state of São Paulo to contain critical episodes of pollution and containment measures, concentration levels still frequently exceed national limits (SÃO PAULO, 2013; CONAMA, 2018) and constantly exceed international limits (WHO, 2021).

In several Brazilian states, including São Paulo, even with initiatives implemented to contain the increase in the number of motor vehicles and with restrictive emission limits, light, and heavy vehicles are still the main sources of pollution in urban areas, although emissions from the processes industries are also relevant in several areas (ABE, MIRAGLIA, 2018; IAP, 2020; WRI Brasil, 2021). Urban transport is one of the main sources of air pollution between various urban areas and requires short and long-term mitigation and control measures (ANGELEVSKA et al., 2021). In the Metropolitan Area of São Paulo (RMSP), the evolution of pollutant concentrations over the last thirty years indicates an increase in secondary pollutants O_3 and $PM_{2.5}$ and a small reduction in primary pollutants CO, NO_2, SO_2, and PM_{10}. Primary pollutants from vehicular emissions, industrial emissions, and biomass burning, under certain meteorological conditions, become secondary pollutants in the atmosphere, such as aerosols and fine particles, which can be moved over great distances. On the coast, the city of Cubatão, characterized by a large industrial and port complex, has high concentrations of atmospheric pollutants, influenced by the local topography that makes dispersion difficult and has low wind speed (ANDRADE et al., 2017; ABE, MIRAGLIA, 2018; CORÁ et al., 2020). Although it is known that the air quality in the RMSP presents high levels of pollution, Environmental Company of the State of São Paulo (CETESB) has highlighted that the countryside of the state seeks urgent measures to contain pollution levels, especially in periods of low humidity, in which the extraction and transport of raw materials are the main sources of particulate matter emissions in several regions (SOUZA, SCUR, HILSDORF, 2018; CETESB, 2020).

Some studies apply statistical methods to interpret data on atmospheric pollutants, as well as probabilistic models, based on prior knowledge of the problem, such as the Bayesian approach. For example, the Bayesian confidence interval was used in Thailand to assess the $PM_{2.5}$ variation and road transport pollution, from monitoring stations in urban areas (THANGJAI et al., 2021). In London to optimize monitoring of ground-level air pollution, the

temporal concept was included with temporal modeling (HELLAN et al., 2022). Poisson regression showed the association between lung cancer mortality and exposure to $PM_{2.5}$, PM_{10}, and NO_2 and weather, being more significant in periods of low temperature, low humidity, and wind speed (CHUNG et al., 2021). In China, one work applied Poisson regression and results showed the association between air pollution and hospitalization for hypertension. A 10μg/m^3 increase in $PM_{2.5}$, PM_{10}, SO_2, and NO_2 concentrations or 1 mg/m^3 increment in CO was significantly associated with relative risks of hospital admissions due to hypertension (LIU, DONG, ZHAI, 2022).

Although the number of studies with machine learning approaches has grown, most were carried out in the northern hemisphere (Asia, Europe, and North America) according to systematic reviews by Bellinger et al., 2017 and Rybarczyk and Zalakeviciute, 2018, the most applied data mining techniques are regression and classification in 59% of the surveys, 26% with data clustering and 15% of the studies used association rules. This study not only expands the existing literature by developing a temporal and spatial study but also expands the application of clustering techniques and association rules.

Therefore, a broad investigation of the temporal and spatial behavior of multipollutant is essential, considering sectoral differences in air pollution that have local and regional influences over time, to intervene assertively to improve air quality. It is possible to present the large-scale distribution of the main pollutants and their relationships, in addition to revealing the critical periods of exposure combined with the associated meteorological background. However, most studies on air pollution focus on microdata, the local topography of some cities and low pollutants, largely of political relevance or industrial complexity (WANG et al., 2018; AMEER et al., 2019; LIU et al., 2019; SANTOS et al., 2019). The present work combines the analytical power of data clustering for the recognition of spatial and temporal patterns of pollutants and identifies by association rules the meteorological conditions related to periods of high concentrations. It is essential to consider that, as it is a time series, this analysis incorporates the temporal aspect of the mining. It involves a large volume of data, with studies still scarce in Brazil and few researchers in South America (RYBARCZYK, ZALAKEVICIUTE, 2018; REPRESA et al., 2019; SOMPORNRATTANAPHAN et al., 2020).

Air quality monitoring data is a time series composed of many data points at regular intervals, representing events that recur from time to time. Although each time series consists of many data points, it can also be viewed as a single object (AGHABOZORGI et al., 2015). The clustering of these objects leads to the discovery of interesting temporal patterns and the

search for (dis)similarity in time to find patterns of time series behavior (KAM, FU, 2000; WINARKO, RODDICK, 2007; MITSA, 2010). As a result, the cluster data mining technique allows visualizing, by geolocation, in large data sets, the visual representation of the clusters, which helps in understanding the characteristics of the data, with spatial parameters such as latitude and longitude (JIN, HAN, 2017; GOVENDER, SIVAKUMAR, 2020; XIAO et al., 2020). 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. In the literature, most studies that use hypothesis generation apply association rule mining, seeking to identify items that co-occur in a wide variety of data (CAGLIERO et al., 2016; BELLINGER et al., 2017; LI et al., 2019).

Therefore, to understand the current situation of air pollution in the entire state of São Paulo, this work aims to explore the data clustering technique to obtain the spatial and temporal patterns of pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3, and CO), in groups with similar behaviors, involving the 56 stations with air quality monitoring throughout the state. By choosing periods of regular seasonal changes for each pollutant, for groups of severely polluted stations, the association rules technique identifies the meteorological parameters (air temperature, relative humidity, wind speed, rainfall, and radiation global solar) associated with the critical periods of concentration of each pollutant. Since previous studies are limited in space and time, a deeper investigation is needed to reveal disparities in the spatial distribution of pollution throughout the territory, as well as the stations most impacted by each pollutant, their temporal variations and antecedent meteorological conditions associated with peak concentrations in different regions.

This article is organized as follows: the “Data and Methods” presents the study site, the database, and the research methodological procedures; in the “Results and Discussion”, the analysis of the clusters and the association rules obtained are presented, in addition to the debate and implications. Finally, conclusions and future perspectives.

4.2 Data and methods

4.2.1 Study site and database

The study was carried out in the state of São Paulo (Brazil), from January 2017 to December 2019, with automatic air quality monitoring data from CETESB (QUALAR, 2019)

composed of 62 stations that monitor 36 municipalities, with 29 stations in the municipalities of the Metropolitan Area of São Paulo (RMSP), while 33 stations are installed in the countryside of the state and on the coast.

The data were generated considering the daily averages of each pollutant. Based on the daily averages, we obtained the monthly averages of pollutants PM_{10}, $PM_{2.5}$, CO, NO_2, SO_2, and O_3. For the meteorological variables, we used the following values: Temperature (TEMP) highest daily value; Wind Speed (WS) daily average; Relative Humidity (RH) lowest daily value; Global Solar Radiation (GSR) daily average. These values were provided by CETESB (QUALAR, 2019). The rainfall index (RI) was obtained by the National Center for Monitoring and Alerts for Natural Disasters (CEMADEN, 2020) with the values of daily averages, based on the hourly average of accumulated rainfall, in millimeters. This approach has previously been applied experimentally for the pollutant $PM_{2.5}$ and some meteorological factors (GODOY et al., 2021).

The logic the choice of data and order of execution of the experiment was: (i) applied the algorithm for clustering in the database by pollutant, we have the geographic domain represented in clustering according to the temporal variation of pollutant concentration for the period (2017 to 2019), (ii) we then categorized each cluster into pollution levels (G1: severe, G2: moderate, or G3: less polluted), (iii) we identified the seasonality of pollutants for each year from 2017 to 2019. To explain the meteorological influence on pollution levels, (iv) for each pollutant we selected the G1 (group of stations with the highest concentration indices) and respective months of peak concentration, (iv) we collected daily meteorological data (TEMP, WS, RH, GSR and RI) selected and generated the monthly averages of each parameter, (v) we apply the algorithm of association rules in the database by pollutant, obtaining the most frequent antecedent meteorological conditions associated in the most polluted stations and critical periods of concentration.

It is essential to highlight that the pollutants Smoke and lead were not included in the database, as they are monitored in specific areas and situations. The choice of meteorological parameters was defined after certifying in the literature their specific effects for each pollutant, in addition to the availability of data for analysis. Initially, there was no RI available at CETESB and we noticed in the 1st experiment that RH and TEMP could be better discussed with this index, later provided by CEMADEN.

In the pre-processing step, 56 monitoring stations were selected, using the limit of up to 10% of missing data for each pollutant in the entire period (2017 to 2019). Missing values

represented an average of 8.3% of records in the database. As the use of time-series requires having all the values filled in, we applied the last and next technique, the average between the values of the day before and the day after, to fill in the missing data. In the adjacent intervals of two missing days, we chose the average of the two previous days with the two subsequent days, and, in superior adjacent intervals, we adopted the average of the current month of the station (PLAIA, BONDI; 2006; CASTRO, FERRARI, 2016; PINTO et al., 2018).

Then, the data were submitted to a standardization process with the Z-score technique, which resizes the original values to have an average equal to 0 and a standard deviation equal to 1, which resulted in values compared on the same scale (HAN, KAMBER, 2006; MITSA, 2010).

4.2.2 Clustering of Temporal Data

The clustering of entire time-series is considered a clustering of a set of individual time-series in relation to their similarity, where the objects are time-series. The basic clustering process consists of four main parts: representation and reduction of dimensionality; measurement of similarity or distance; clustering algorithm; and evaluation definition (WANG et al., 2006; AGHABOZORGI et al., 2015).

In the reduction of dimensionality, the most representative characteristics of the database are identified. From there, the time-series data are converted into simple static objects (Table 4.1), which will serve as input for conventional clustering algorithms. Then, clustering algorithm is applied to the objects using a distance measure for similarity calculation. Finally, there are cluster evaluations, which enable analysis of the results, to ensure greater reliability of the clustering process (AGHABOZORGI et al., 2015).

Table 4. 1 - Sample of conversion of the daily time-series into monthly static values, of the pollutant concentration values of each monitoring station, in this example, for PM_{10}.

$PM_{10}(\mu g/m^3)$	
Station: São José dos Campos	
Date	Daily value
01/01/2017	21
01/02/2017	14
…	…
01/31/2017	33

$PM_{10}(\mu g/m^3)$												
Stations	**Jan 2017**	**Feb 2017**	**Mar 2017**	**Apr 2017**	**Mai 2017**	**Jun 2017**	**Jul 2017**	**Aug 2017**	**Sep 2017**	**…**	**Nov 2019**	**Dec 2019**
S. José Campos	18	20	19	19	22	26	29	31	41	…	19	19
Taubaté	14	18	13	17	20	23	27	29	44	…	15	15
Guaratinguetá	18	21	15	15	18	21	23	24	35	…	17	19
Ribeirão Preto	15	18	21	24	30	38	49	66	65	…	20	16
Campinas-Centro	20	23	20	21	23	24	27	35	43	…	21	18
…	…	…	…	…	…	…	…	…	…	..	…	…

For this study, the sampling intervals were regular, and the time-series was univariate, represented by the pollutant concentration value, collected regularly over time, in a daily sequence. The high dimensionality of the data, whether in the number of objects (stations) or attributes (database columns), can cause problems for the generation of cohesive clusters, since a reduced number of columns in the database form more consistent clusters (MAUGIS et al., 2009; FOGLIATTO, ANZANELLO, 2011). Therefore, in this experiment, the clustering was employed by pollutant, in a specific manner.

In order to find similar time-series a "similarity measure" was applied, based on the distance between the objects. To define whether both must belong to the same cluster, the difference is given by equation 4.1, where *x* and *y* are objects and *m* is the number of attributes. The smaller the distance, the greater the similarity between them (MITSA, 2010; HAN et al., 2011; AGHABOZORGI et al., 2015).

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{4.1}$$

To this end, we used the Euclidean distance, considered adequate to find similar time-series in time and for short periods (in this case, monthly). It is also recommended for comparing numerical quantitative vectors, in addition to being one of the most common methods for measuring similarity in time-series clusters (JAIN, 2010; MITSA, 2010; HAN et al., 2011).

Subsequently, a conventional clustering algorithm is applied to the extracted vectors (Table 4.1). Clustering algorithms can be divided into hierarchical or partitional. In general, hierarchical algorithms have their quality impaired when clustering time-series because they do not allow an adjustment of clusters after division by the divisive or agglomerative method, in addition to having low scalability with large time-series.

In the category of partitional algorithms, the most common are K-means and K-medoids. For this study, we used the K-medoids algorithm, derived from K-means. Indicated for time-series of equal length (granularity), K-medoids uses objects from the base itself as the center of the clusters, called medoids, which makes it less sensitive to outliers, common in time-series. Given the time-series in a cluster, the distance of all-time-series pairs within the group is calculated using a distance measure (AGGARWAL, REDDY, 2013; AUSTIN et al., 2013; JIN, HAN, 2017).

In partitional algorithms, the method creates an initial partition and, after each iteration, tries to improve the partitioning by moving objects from one group to another. The general criterion of good partitioning is that, when dividing the subsets, the elements of a cluster are more similar to one another, and dissimilar among the elements of other clusters (MITSA, 2010; HAN et al., 2011; AGHABOZORGI et al., 2015).

In this study, silhouette coefficient was the method used as a decision criterion to determine the number of clusters and evaluate their quality, which combines the concepts of cohesion and separation (KAUFMAN, ROUSSEEUW, 2005; MITSA, 2010). Equation 4.2 calculates the mean Silhouette value (S_p) which must be between -1 and 1, with positive ideal values close to 1. Being *n* the number of objects in the database and the individual value of the silhouette coefficient of element x_i given by $s(x_i)$, generated by Equation 4.3, the values $a(x_i)$ and $b(x_i)$ correspond, respectively, to the average distance between x_i and all objects in its cluster and the average distance from x_i to another cluster to which x_i does not belong.

$$S_p = \sum_1^n \frac{s(x_i)}{n} \tag{4.2}$$

$$s(x_i) = \frac{\big(b(x_i) - a(x_i)\big)}{max\{a(x_i), b(x_i)\}} \tag{4.3}$$

When viewing the clustering-based time-series dataset, the pollutant clusters were represented by geolocation, through the latitude and longitude parameters of each monitoring station. As São Paulo presents a geographically wide area, it was essential to find spatial patterns that correspond to the levels of pollution (severe, moderate or less polluted) in the different regions of the state.

The K-medoids algorithm was implemented in the Python language, using the Scikit-learn (PEDREGOSA et al., 2011), Pyclustering (NOVIKOV, 2019), Pandas (REBACK et al., 2020), and Geopy (KOSTYA, 2020) open-source libraries, specific for machine learning.

Time-series clustering is an approach widely used as an exploratory technique for other mining algorithms, such as rule discovery, attribute selection, classification, indexing, anomaly detection, among others (MITSA, 2010; HAN et al., 2011). Thus, the cluster analysis result was used as input for the application of association rules, considering the cluster of stations categorized as severely polluted in all pollutants PM_{10}, $PM_{2.5}$, CO, NO_2, SO_2, and O_3. The result of the application of association rules showed relations between the behavior of each pollutant

and the meteorological conditions in its critical periods of pollution, identified in its temporal variation as the month with the highest concentration.

4.2.3 Association Rules

The temporal behavior of pollutants identified by the application of the clustering technique elucidates the variation in concentration of each pollutant over the years, a recurring pattern for the entire period (2017 to 2019), which demonstrates the importance of an investigation of the dominant meteorological factors during periods of high concentration. Proposed by Agrawal et al. (1994), association rule mining is based on finding relations between items in the database. These associations result in association rules, which show elements that occur together in a transaction (AGRAWAL, SRIKANT, 1994).

The Apriori association rules algorithm was applied to run an exhaustive search in the extraction of relevant patterns and to be a reference in this task. To ensure a reduced exponential search space and contribute to processing speed, the algorithm applies the property "if a set of items is frequent, then all its subsets must also be frequent", hence for an infrequent itemset it is unnecessary to generate any subsets of it as a candidate, as they will also be infrequent and discarded (HAN et al., 2011; CASTRO, FERRARI, 2016).

Next, the concepts involved in creating association rules are presented. Be an itemset $I = \{i_1, \ldots, i_n\}$ ordered in a transactional database and $X = \{x_1, \ldots, x_n\}$ a set of transactions, so that each transaction $x_i \in X$ is composed of an *itemset*, such that $x_i \subset I$. Since X and Y are disjoint itemsets, that is, $X \cap Y = \emptyset$, an association rule is expressed in the form $X \rightarrow Y$, where X is called the antecedent of the rule (cause) and Y is called the consequent (consequence), which can be translated as "if X then Y" indicating that when X occurs, Y also tends to occur (AGRAWAL, SRIKANT, 1994; MUELLER, 1995; MITSA, 2010).

The Apriori algorithm works with categorical databases. Thus, the database sample presented in Table 4.2 had its attributes categorized as lower or higher than the average annual value. The result is a transactional base, where there are searches for association rules, which items are related.

Transactional database sample with categorical values, higher or lower than the average annual value. For this example, we considered the pollutant SO_2 and the meteorological conditions TEMP, RH, WS, RI, and GSR in June, 2017, involving all stations.

Table 4. 2 - Transactional database sample with categorical values, higher or lower than the average annual value. For this example, we considered the pollutant SO_2 and the meteorological conditions TEMP, RH, WS, RI, and GSR in June, 2017, involving all stations with pollutant monitoring.

Stations	TEMP	RH	WS	RI	GSR	SO_2
Cubatao_v. parisi	Higher	Lower	Lower	Lower	Higher	Higher
Cubatao_centro	Higher	Lower	Lower	Higher	Higher	Higher
Santos_pontapraia	Higher	Lower	Lower	Lower	Higher	Higher
Paulínia	Higher	Higher	Lower	Higher	Higher	Higher
…	…	…	…	…	…	…

Note: Higher = Higher_Average; Lower = Lower_Average.

The process of searching for association rules was conducted in two phases. First, all possible combinations between the attributes are considered to discover frequent itemsets with a minimum support value (minsup). At this step, several rules are found, but not all of them are interesting and useful. Then, rules that do not meet a minimum confidence limit (minconf) are discarded (HAN et al., 2011; CASTRO, FERRARI, 2016).

Given the rule $X \rightarrow Y$, the support (Equation 4.4) is the coverage of the rule and represents its frequency in the transaction base (CASTRO, FERRARI, 2016), that is, the ratio between the number of R transactions that contain the itemset $X \cup Y$, and the total number of R transactions. Confidence (Equation 4.5) is the accuracy of a rule, defined as the conditional probability $P(Y|X)$. In this case, given the consequent Y of the rule, the antecedent X also happens (MUELLER, 1995).

$$Support\ (X \rightarrow Y) = \frac{|X \cup Y|}{|R|} \tag{4.4}$$

$$Confidence\ (X \rightarrow Y) = \frac{|X \cup Y|}{|X|} \tag{4.5}$$

The adequacy of the rule for evaluating the problem depends on the rule's support and confidence values. In this work, the *Apriori* algorithm was implemented in the Python language, using the "*mlxtend*" (RASCHKA, 2018) and "*Pandas*" (REBACK *et al.*, 2020) libraries. Using association rules, the experiments provided the levels at which pollutants and climatic factors are influenced by each other, in addition to the temporal relationship between these elements. The minimum support (minsup) and minimum confidence (minconf) values considered were higher than 80%.

4.3 Results and discussion

The partitional clustering algorithm K-medoids was used to form the groups, by pollutant, and enabled a better understanding of the regional characteristics of air pollution. The preference for monthly seasonality, generated by the daily averages of pollutant concentrations, avoids the high dimensionality of the data that can influence the formation of groups (MAUGIS et al., 2009; FOGLIATTO, ANZANELLO, 2011).

To define the number of groups for each of the pollutants studied here and to assess the quality of the clusters, the silhouette coefficient was used. The closer the coefficient value is to 1, the more cohesive the group, with values ranging from -1 to 1. After 100 runs of the *k-medoids* algorithm, applied to the average monthly concentration database for each pollutant, the results indicated good adequacy of the groups, indicating the number of groups $k = 2$ with the following silhouette values for pollutants $PM_{2.5}$ (0.72), SO_2 (0.69) and CO (0.44) and the formation of $k = 3$ groups, pollutants NO_2 (0.58), O_3 (0.41), and PM_{10} (0.26). As an example, Figure 4.1 shows that, for pollutant NO_2, we tested 34 different groups (k-1) related to the number of stations. The relation between the silhouette coefficient value corresponds to the number k of groups, with $k = 3$ being the best corresponding value.

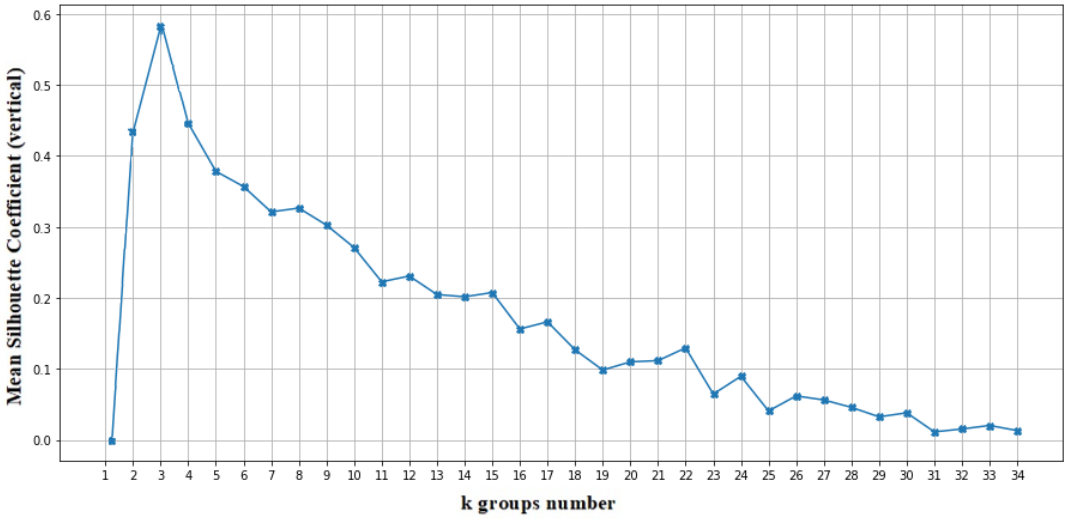


Figure 4. 1 - The silhouette coefficient was the decision criterion in defining the number of k groups, applied to the database of each pollutant. In this example, for pollutant NO_2, the algorithm tested 35 different groups considering the number of monitoring stations in the database (k-1).

Table 4.3 shows a directly proportional relationship between the dimension of the database and the number of groups recommended by the silhouette coefficient. Pollutants PM_{10}, $NO_{2,}$ and O_3 with formation of three groups have more than 50 automatic monitoring stations distributed throughout the state, and pollutants $PM_{2.5}$, CO, and SO_2 with two groups carry out the monitoring with a number of less than 30 stations per pollutant. The stations were separated into pollution levels (severely, moderately, or less polluted), so that Group 1 (G1) is composed of stations with high concentration of the pollutant, Group 2 (G2) with intermediate levels, and Group 3 (G3) of stations with low concentration of the same pollutant. Similar to this proposal, other studies also adopted the cluster analysis method to explore the regional distribution of pollutants, Zhao et al. (2016) classified air pollution levels in 31 Chinese cities, and Huang et al. (2015) analyzed the spatial and temporal distribution of $PM_{2.5}$ pollution in the city of Xi'an, China. Arce et al. (2018) identified correlations and incidents between atmospheric pollutants in the metropolitan cities of Ecuador, which identified distinct groups with large-scale ozone pollution. The results were considered consistent and efficient in identifying patterns of behavior of pollutants.

For the period under analysis, the Annual Arithmetic Average (AAA) of pollutants exceeds the annual limits established by the WHO (WHO, 2019), except for NO_2, which presents a better condition for Groups 2 and 3. The standard deviation maintains similar values between groups of the same pollutant, decreasing from Group 1 to 3, with a coefficient of variation from 15 to 36%, indicating good data homogeneity and the same quality of the clusters. The new Global Air Quality Guidelines (AQG) published by the WHO (WHO, 2021) recommend even lower standards and adjusted the annual concentration limits (PM_{10}: from 20µg/m^3 to 15µg/m^3; $PM_{2.5}$: from 10µg/m^3 to 5µg/m^3, O_3: 60µg/m^3, in the most intense six months; NO_2: from 40µg/m^3 to 10µg/m^3). A comparative assessment pointed to high pollution in all groups by particulate matter (PM_{10} and $PM_{2.5}$) and NO_2 reaching annual averages of 3 to 4 times higher than recommended (G1: PM_{10} - 60.75µg/m^3; $PM_{2.5}$ - 17.41µg/m^3; NO_2 - 41.89µg/m^3). For all the factors exposed, it is clear the need to adopt containment measures to improve the current scenario.

Table 4. 3 - Cluster data by pollutant generated by the K-medoids algorithm, from 2017 to 2019: monitored and selected stations, stations by cluster, comparison of annual mean concentrations, and standard deviation.

Pollutants		**PM_{10}**	**$PM_{2.5}$**	**CO**	**NO_2**	**SO_2**	**O_3**
Monitored/Selected Stations		58/42	31/21	20/18	51/36	18/14	51/46
WHO standard		**20 MAA**	**10 MAA**	**9 ppm(8hs)**	**40 MAA**	**20 (24hs)**	**100 (8hs)**
Group 1	No. Stations	5	15	5	11	4	33
	MAA (µg/m^3)	**60.75**	**17.41**	**0.78 ppm**	**41.89**	**10.27**	**53.01**
	Standard deviation	17.40	4.73	0.13	6.60	2.15	10.12
	Coefficient of variation	28,64	27,16	16,66	15,75	20,93	19,09
Group 2	No. Stations	16	6	13	18	10	6
	MAA (µg/m^3)	**29.23**	**13.24**	**0.47 ppm**	**24.73**	**2.57**	**40.30**
	Standard deviation	8.97	4.83	0.10	6.0	0.55	8.49
	Coefficient of variation	30,68	36,48	21,27	24,26	21,40	21,06
Group 3	No. Stations	21	-	-	7	-	7
	MAA (µg/m^3)	**22.63**	-	-	**15.38**	-	**27.11**
	Standard deviation	6.77	-	-	4.4	-	6.24
	Coefficient of variation	29,91			28,60		23,01

Note: MAA – Mean Annual Average
Selected stations - Monitoring stations with up to 10% of missing data for the entire period.

In terms of spatial distribution, Figure 4.2 lists the stations that constitute the groups and shows the differences between pollutants and the most polluted stations (G1). The cluster highlights the strong presence of high concentrations of $PM_{2.5}$ and O_3 in several areas of the state of São Paulo, also reported by Andrade et al. (2017) and Boian and Andrade (2012). This behavior continued during the days of the truck drivers' strike in the state of São Paulo in 2018 (LEIRIÃO et al., 2020) and in other countries, China (ZHAO et al., 2016; YAO et al., 2020) and Korea (LEE et al., 2021). Therefore, there are more stations with high levels of pollution (G1) for pollutants O_3 (G1:33, G2:6, G3:7) and $PM_{2.5}$ (G1:15, G2:6). For pollutants SO_2, CO, and PM_{10} in a reduced number of stations, PM_{10} (G1:5, G2:16, G3:21), for SO_2 (G1:4, G2:10) and CO (G1:5, G2:13). Research carried out in Brazil by Andrade et al. (2017) and Chiquetto et al. (2022) in six Brazilian states, including São Paulo, state that in metropolitan regions the

primary sources of air pollution are vehicular and industrial emissions and the burning of biomass, which explains at least 40% of $PM_{2.5}$ pollution. Ozone (O_3) precursor pollutants are released into the atmosphere mainly through vehicular and industrial combustion processes. Ozone levels increase in peripheral regions of large urban centers, in the directions in which the wind blows (CANÇADO et al., 2006; CETESB, 2020).

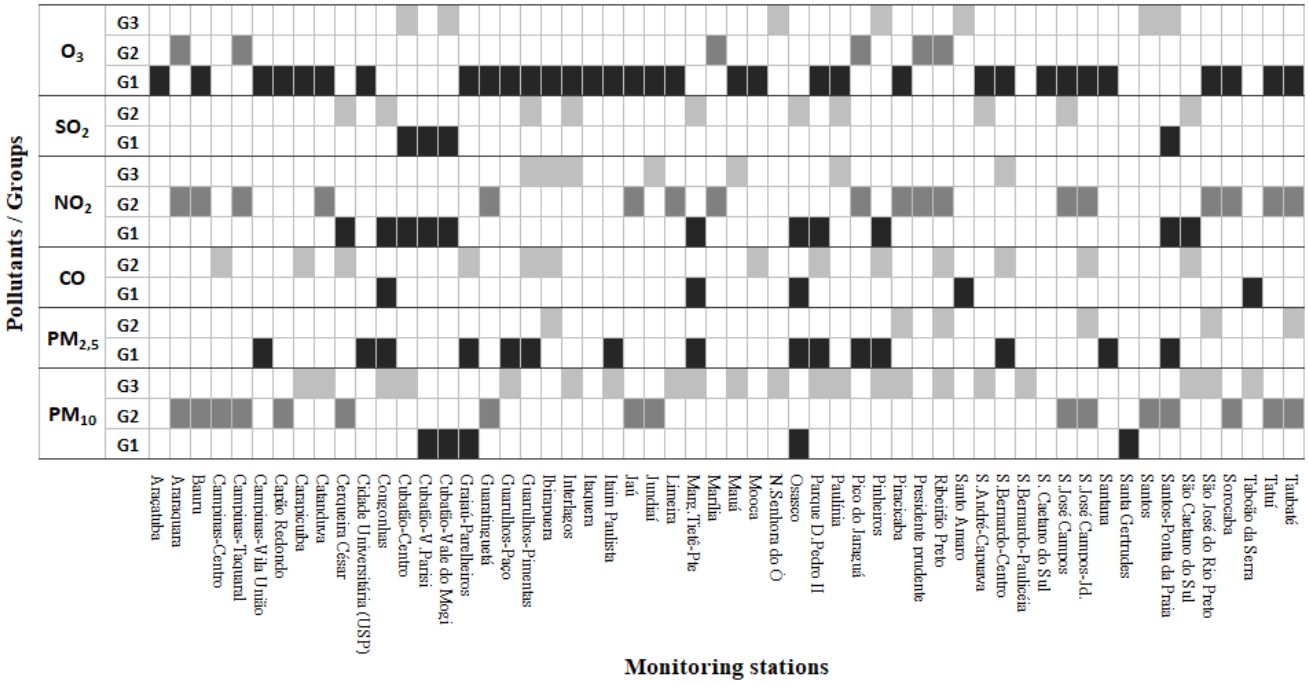


Figure 4. 2 - List of monitoring stations by pollutant and their corresponding groups, generated by the K-medoids algorithm, for the period from 2017 to 2019. In the darkest color, Group 1 (G1) with high concentration of the pollutant, followed by Group 2 (G2) with intermediate levels, and in the lightest color, Group 3 (G3), with low concentration of the pollutant.

For the spatial analysis, we generated the geolocation of the groups for each pollutant, involving all monitored stations in the state of São Paulo, represented by latitude and longitude (Figure 4.3), for the time interval from 2017 to 2019. The groups are indicated on the map as G1, G2, and G3 in red, yellow, and blue, respectively. The group with the highest concentration of PM_{10} involves few, geographically distant cities. In the countryside of the state, the city of Santa Gertrudes, on the São Paulo coast the city of Cubatão (V. Parisi and Vale do Mogi stations), and the RMSP represented by the Grajaú-Parelheiros station. Since PM_{10} comes from combustion processes, resuspended dust, grinding, material crushing, dust-laden wind, secondary aerosol formed in the atmosphere, among others, it is possible to better understand this distribution. In Santa Gertrudes, the main source is the ceramic industry with national

production, known as the Ceramic Hub (CETESB, 2020). The Grajaú-Parelheiros station frequently exceeds the daily emission limits, considered to be severely polluted with particulate matter (PM_{10}, $PM_{2.5}$) and O_3. CETESB reports that since 2013 there has been movement of heavy vehicles on the road near the station, transporting solid waste to landfill, in addition to the industrial profile of the region (CETESB, 2019). In Cubatão, a study by Nardocci et al. (2013) found for each increment of 10μg/m³ of PM_{10} an excess of hospitalizations of 4.25% for total respiratory diseases, 5.74% for respiratory diseases in children aged under 5 years, and 2.29% for cardiovascular diseases in individuals aged over 39 years.

For pollutant $PM_{2.5}$, the analysis points to most stations of the G1 being located in the RMSP, Campinas, and Baixada Santista. Except for the Campinas region, which is also influenced by fires, the main source of pollutants in these regions is the burning of fuels by motor vehicles and more intense industrial emissions (Godoy et al., 2021; Huang et al., 2015). Comparative results obtained by Araújo and Rosário (2020) identified, from satellite data, that the regions most polluted by $PM_{2.5}$ in metropolitan areas of São Paulo, Campinas, and Baixada Santista are the same found by Angelevska et al. (2021) in Macedonia's urban regions, resulting from road transport. Due to a large number of inhabitants, São Paulo is the municipality in Brazil whose highest number of deaths is attributable to exposure to particulate matter (CORÁ et al., 2020). Yu et al. (2022) found associations between prolonged exposure to $PM_{2.5}$ and cancer mortality in Brazilian cities from 2010 to 2018.

Of the 18 stations with CO monitoring, despite the monitoring also taking place in other metropolitan cities in the countryside of the state, such as Campinas, São José dos Campos and Ribeirão Preto, G1 is composed of few stations, all in the RMSP: Santo Amaro, Congonhas, Taboão da Serra, Osasco and Marginal Tietê. According to CETESB (CETESB, 2020), motor vehicles account for about 97% of CO emissions in the RMSP. Despite the increase in the number of motor vehicles over the years, in compliance with the increasingly strict limits of PROCONVE and PROMOT, current concentrations are lower than those observed in the 2000s, but showed a slight increase in 2019 due to road changes in the region. Moisan et al. (2018) observed that 54% of the $PM_{2.5}$ concentration is composed of CO, demonstrating a direct relationship between these pollutants in metropolitan regions.

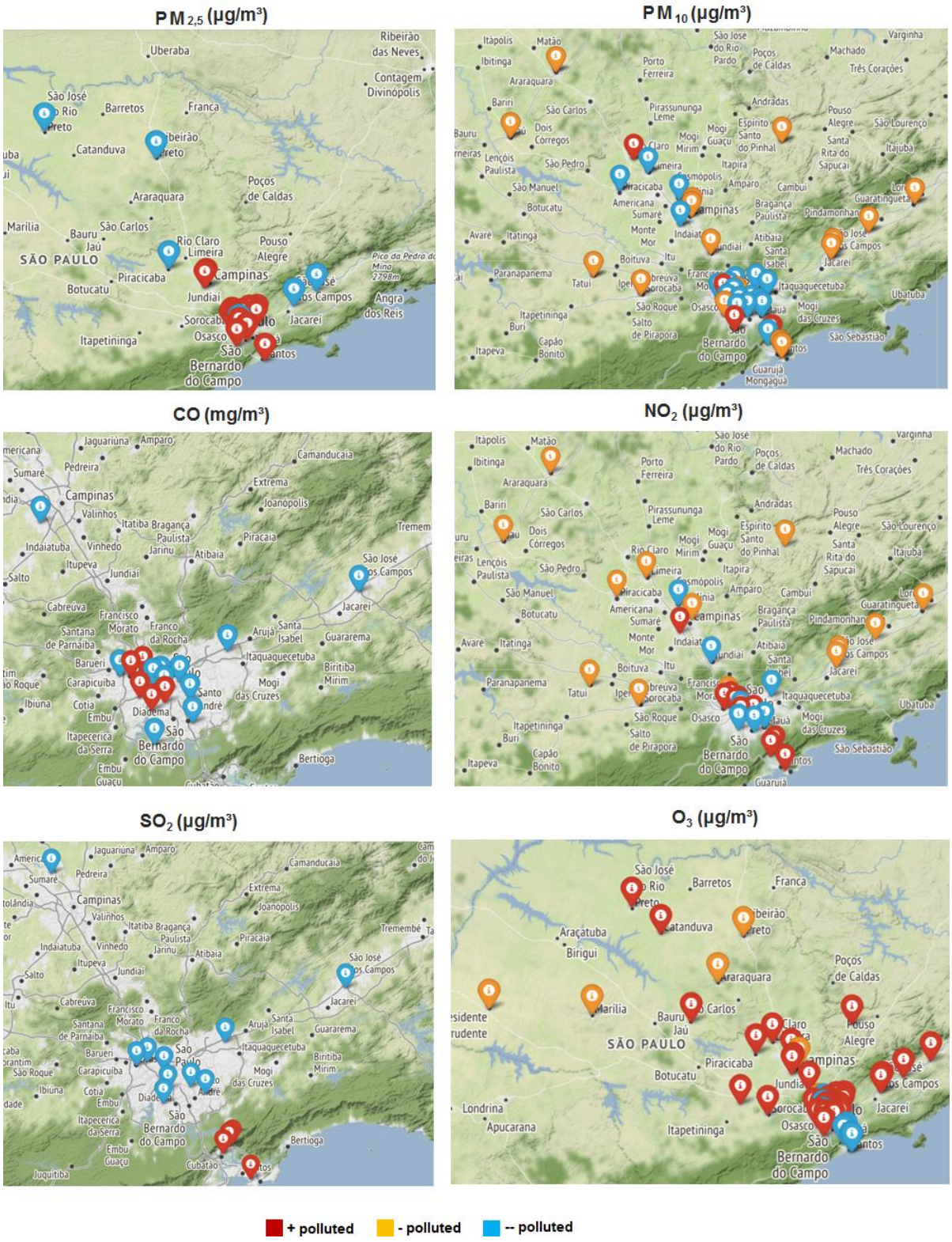


Figure 4. 3 - Spatial visualization of clusters by pollutants, generated by the K-medoids algorithm, for the time interval from 2017 to 2019. The groups were identified as G1 (red), G2 (yellow), and G3 (blue).

The main sources of nitrogen oxide NO_2 are emissions from heavy vehicles and industrial processes (SARRA, MÜLFARTH, 2021). The cluster shows that the high concentration stands out in stations in the RMSP and coast, with the largest number of vehicles, either by industrial or port activity, respectively. Although vehicle emission control plans have managed to reduce the concentration of NO_2 and CO, the results still show that they remain in the RMSP. An urban mobility survey shows that 75% of the most extended, most distant, and motorized trips in the RMSP are due to work and proposes a better spatial distribution of jobs as a potential for reducing effects and equity (CHIQUETTO et al., 2022).

Ozone is one of the most monitored pollutants in the entire state. Of the 51 stations, 33 are part of G1, with similar behavior of concentration among several stations. On the other hand, the stations with lower concentrations are located in the RMSP and coast, that is, 7 stations (G3) whose concentrations are 50% lower than G1 (G1: 53.01 and G3: 27.11). The pollutant is not emitted directly into the atmosphere, being produced photochemically by solar radiation on nitrogen oxides and volatile organic compounds. Therefore, the highest concentrations of ozone are found in stations further away from traffic routes. Nitrogen monoxide emitted in the burning of fuels acts on the consumption of available ozone; thus, O_3 concentrations tend to be low near the traffic ways. High O_3 concentrations can affect agricultural production due to its high oxidative power (CETESB, 2020).

Resulting mainly from activities such as burning diesel, coal, and oil in power plants or copper smelting, SO_2 has a history of high concentrations on the coast (CETESB, 2019), which is consistent with the result of the clustering, with G1 of SO_2 in three stations in Cubatão and in the station in Santos, both on the coast of the state. An epidemiological time-series study found an association between short-term health effects and exposure to air pollutants in Cubatão, and SO_2 showed an association of 3.51% with cardiovascular diseases in people aged over 39 years (NARDOCCI et al., 2013).

The temporal behavior of pollutant concentrations in the G1 group highlights the critical months with high rates for the period from 2017 to 2019, represented in Figure 4.4. The monthly peaks occurred in the same periods for the groups of the same pollutant, recurringly, suggesting a seasonal trend in the three years.

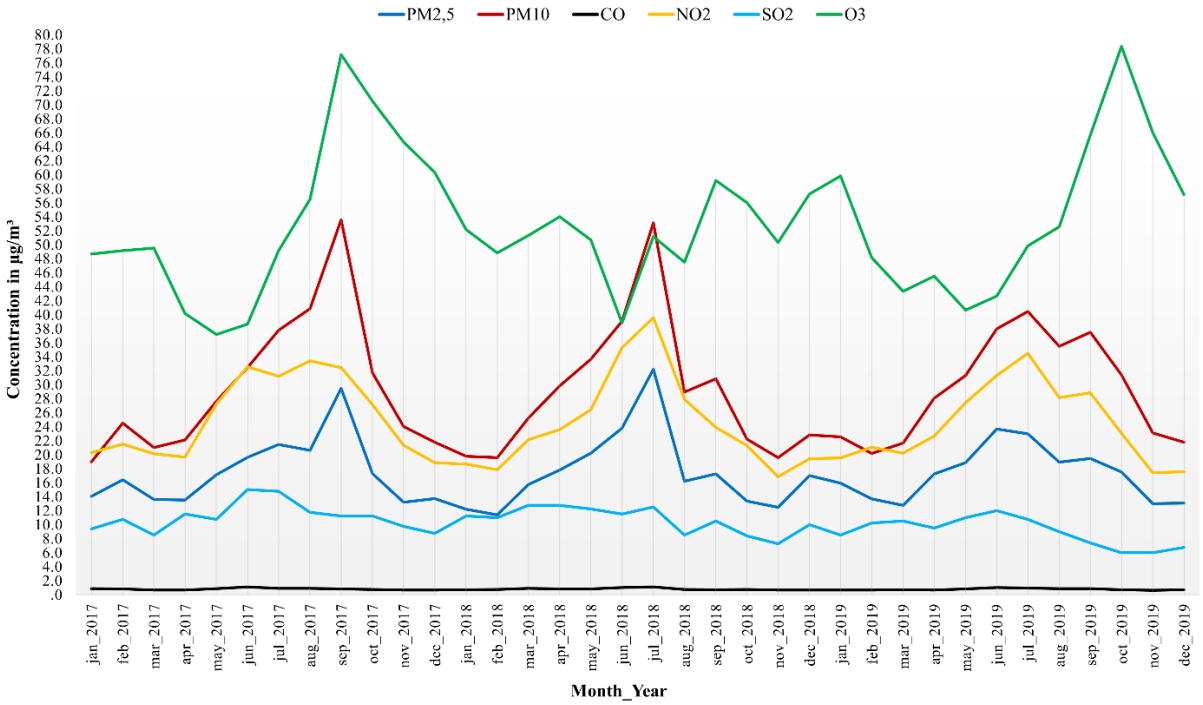


Figure 4. 4 - Temporal variation of the pollutants generated by the K-medoids algorithm and comparison of monthly average concentrations of the G1 group of each pollutant, in 2017–2019.

The concentrations of atmospheric pollutants in the winter months were significantly higher than in the other months, except for O_3, as shown in Figure 4.4. The annual variation of O_3 was opposite to other pollutants, with higher concentrations in summer and lower in winter. This is due to the dependence on increased intensity of solar radiation for the formation of O_3 (ATKINSON, 2000; CETESB, 2020). Meteorological conditions such as less precipitation, weaker winds, and lower temperatures in winter make it difficult for particles to be eliminated by rain or diffusion, increasing the concentration of atmospheric pollutants near the soil surface.

The seasonality of PM_{10} is quite similar to that of $PM_{2.5}$; this is because $PM_{2.5}$ corresponds to about 60% of the concentration of PM_{10} (CETESB, 2020). The study of Wang et al. (2018) found that the ratios of $PM_{2.5}$ to PM_{10} were around 40% in cities in northwest China (ZHAO et al., 2016). Both pollutants were more present in winter, with peaks in September/2017, July/2018, and June/2019. There was a general increase in particulate matter levels in 2017 and 2018, with a reduction in 2019. The same seasonality between the groups (G1, G2, and G3) and strong similarity by concentration between stations of the same group, with great variation between stations of different groups, for the two pollutants PM_{10} and $PM_{2.5}$.

A temporal pattern was also found for CO and NO_2, with significantly higher concentrations in winter, in both groups G1 and G2, with higher concentrations in July/2018

and June/2019, and a small variation in 2017. This characteristic was also observed in northern China, which confirms the relationship of weather conditions being associated with high pollutants concentrations, so that the air is stagnant by light winds and traps pollutants close to the surface, raising their concentrations (ZHAO et al., 2016).

There are records of decreasing SO_2 emission since 2014 due to the change in the sulfur content of diesel in 2013 and of gasoline in 2014, industrial and automotive (CETESB, 2019). However, in 2017 there were higher levels of concentration compared to 2018 and 2019 (Figure 4.5), with a large difference in concentration between the groups (G1: 10.27, G2: 2.57). Peak months are also in winter in June/2017, April/2018, and June/2019 (Figure 4.5). According to CETESB (CETESB, 2020), trucks are responsible for the highest SO_2 emissions due to sulfur in fossil fuels (diesel and gasoline), with a significant effect on the pollutant with lower concentration detected for increased precipitation, lower temperature, and increased humidity level.

Conversely, high concentrations of O_3 occur in summer, in periods of hot and dry weather, with high solar radiation (FEPAM, 2016). For this study, the peak months were September/2017, December/2018, and October/2019 (Figure 4.5), with little difference in concentrations between the groups, being a pattern of the pollutant, regardless of the concentration level (Figure 4.5). Cançado *et al.* (2006) reported the increase in ozone levels between spring and early autumn, in peripheral regions of large urban centers, in the directions in which the winds blow, with concentration peaks in the middle of the morning, due to morning traffic, reaching its maximum mid-afternoon with reduction in the evening.

It is important to note in Table 4.4 the coefficient of variation between the minimum and maximum monthly averages of concentrations during the year, for the same pollutant. For pollutants $PM_{2.5}$, $PM_{10,}$ and O_3, the variation was from 35% to 50%, while for pollutants CO, $NO_{2,}$ and SO_2, the concentrations doubled in the peak period, with a coefficient of variation from 51% to 64%. Therefore, the annual temporal variation of pollutants associated with serious pollution events may be related to meteorological phenomena for the same period (BISHT, SEEJA, 2018; CETESB, 2019; LI et al., 2020).

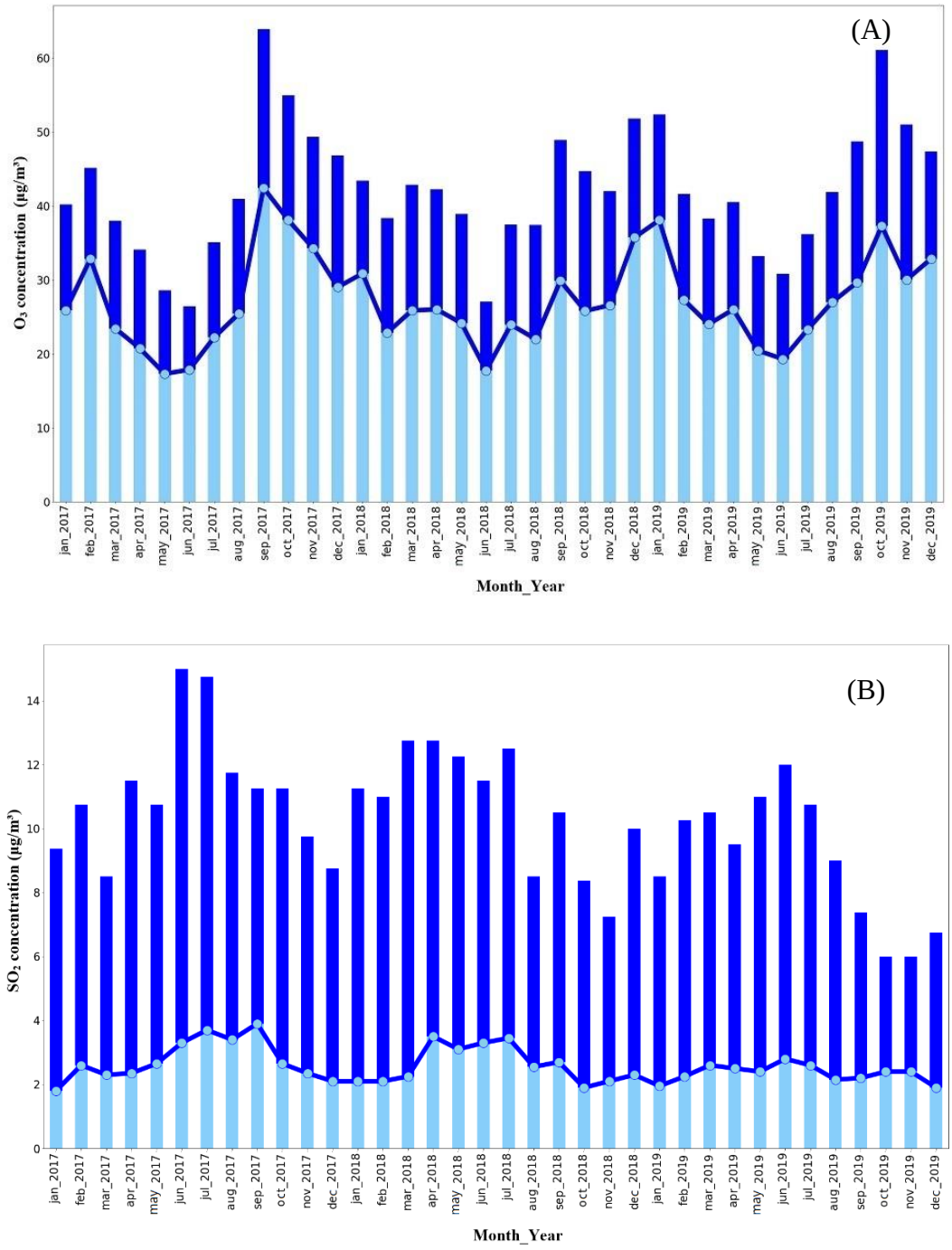


Figure 4. 5 - Monthly temporal variation of pollutants (A) O_3 and (B) SO_2, 2017–2019, generated by the K-medoids algorithm, with the differences in concentrations between the severely polluted groups (G1) in dark blue and the less polluted groups (G3) in light blue.

CETESB data from 2010 to 2019 indicate that May to September was the period in which meteorological conditions were most unfavorable to the dispersion of pollutants in the atmosphere, in 24% of the total days in the year (CETESB, 2020). With the association rules algorithm (Apriori), based on a transactional database, it was possible to identify this relation

and verify meteorological weather conditions: TEMP, WS, RH, RI, and GSR are predominant in the regions where pollution episodes occur, with an increase in the concentration of pollutants PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3 and CO for the period.

Table 4. 4 - Minimum and maximum monthly averages (in μg/m³) of G1, which compose the most polluted stations for all pollutants, 2017–2019. We note the months with the highest and lowest concentration of each pollutant.

Period/ Pollutants	**2017**		**2018**		**2019**	
	Minimum	**Maximum**	**Minimum**	**Maximum**	**Minimum**	**Maximum**
$PM_{2.5}$	November	September	February	July	March	June
	13.20	29.48	11.40	32.20	12.77	23.67
PM_{10}	January	September	January	July	December	June
	34.5	99.00	37.50	105.5	37.00	81.50
CO	December	June	December	July	November	June
	0.64	1.10	0.65	1.10	0.62	1.02
NO_2	April	September	February	July	November	July
	34.09	53.13	33.18	59.22	32.63	53.09
SO_2	December	June	November	April	November	June
	8.5	15.00	7.25	12.75	6.00	12.00
O_3	June	September	June	December	June	October
	26.41	63.92	27.03	51.77	30.83	61.04

Some studies have applied this technique to discover relationships between different attributes in the database involving pollutants. An example is a study by Cagliero et al. (2016) in Milan, Italy, who found correlations between the levels of pollutants, weather factors, and traffic conditions by using association rules. When the temperature is warm, PM_{10} and $PM_{2.5}$, CO, and O_3 pollutant levels tend not to be critical. In addition, the presence of a medium/high number of vehicles is positively correlated with a fairly high concentration of PM_{10}. The use of diesel vehicles is critical for PM_{10} emissions, whereas gasoline engine vehicles do emit a significant amount of CO. The study of Li et al. (2020) proposes the framework of space fusion after time fusion in the process of rules fusion and show a correlation between the pollutants $PM_{2.5}$, NO_2, and O_3 in China and between $PM_{2.5}$ and PM_{10} in the US, considering the distribution of monitoring stations in China and microstations in the US and local climatic characteristics. In Brazil, in Curitiba, Souza and Rabelo (2016) showed different air quality seasonal standards in three regions, which reflected differently in the association of respiratory

problems for the three monitoring stations, generating rules that converge in the worst air condition in winter when the three seasons are inadequate.

To run the Apriori algorithm, a new dataset per pollutant was generated, as mentioned in Table 4.2, with data related to the months identified as maximum (month with the highest concentration of the pollutant during the year), highlighted in Table 4.4. The algorithm analyzes all the possibilities between the items and applies minimum support and confidence values equal to or greater than 80% for the selection of "strong rules". Repeated rules with the same itemsets were not considered. The main rules of this period are presented in Table 4.5.

Table 4. 5 - Rules obtained by the Apriori algorithm for the annual averages of each variable, pollutant, and meteorological conditions. In the antecedent of the rules, the meteorological conditions for the consequent "concentration higher than annual average" of each pollutant and its respective parameters of support and confidence.

Year	[**Antecedent**]	→ [**Consequent**]	**Rules**	**Support**	**Confidence**
2017	[RI lower than average, RH lower than average, GSR higher than average]	→ [$PM_{2.5}$ higher than average]	9	87.5%	100%
2018	[GSR lower than average, RI lower than average, TEMP lower than average, WS lower than average]		44	100%	100%
2019	[RH lower than average, RI lower than average, TEMP lower than average]		9	83.33%	100%
2017	[RI lower than average, TEMP higher than average]	→ [PM_{10} higher than average]	11	100%	100%
2018	[TEMP higher than average, RI lower than average, WS lower than average]		49	90%	100%
2019	[TEMP higher than average, WS lower than average]		49	98%	100%
2017	[RH higher than average, TEMP higher than average, GSR higher than average, WS lower than average]	→ [CO higher than average]	179	100%	100%
2018	[TEMP higher than average, WS lower than average, GSR higher than average]		179	100%	100%
2019	[WS lower than average, GSR higher than average, RH lower than average]		601	100%	100%

2017	[WS lower than average, RI lower than average]	→ [NO_2 higher than average]	11	83.33%	100%
2018	[WS lower than average, GSR lower than average, TEMP lower than average]		179	87.5%	100%
2019	[TEMP lower than average, GSR lower than average, WS lower than average, RI lower than average]		179	100%	100%
2017	[RI higher than average, GSR higher than average, RH higher than average, TEMP higher than average, WS lower than average]	→ [SO_2 higher than average]	601	100%	100%
2018	[TEMP higher than average, RH higher than average, WS lower than average]		179	100%	100%
2019	[GSR higher than average, TEMP higher than average, WS lower than average, RH higher than average]		44	100%	100%
2017	[TEMP higher than average, RI lower than average, GSR higher than average, RH lower than average]	→ [O_3 higher than average]	179	88.88%	100%
2018	[TEMP higher than average, GSR higher than average, RI higher than average, RH lower than average]		49	88.88%	100%
2019	[TEMP higher than average, RI lower than average, WS higher than average, GSR higher than average, RH lower than average]		601	88.88%	100%

It is observed that the rules presented have support above 83.33% and confidence of the 100% in total. In this case, the support indicates the meteorological factors (antecedent) that occur together in the base with the frequency of 83.33% or more, when the pollutant is higher than average. Confidence indicates that when the pollutant is higher than average, there is 100% certainty of the meteorological factors (antecedent) that will also occur. Therefore, in a temporal summary, for the months with a higher than average concentration of pollutants, we can highlight:

- $PM_{2.5}$: Lower than average TEMP and RH, milder WS, and low RI, the lower or higher than average GSR rate did not change this condition.

- PM_{10}: Lower than average WS and RI and higher than average TEMP always present at high concentration.

- CO: Higher than average TEMP and GSR and milder WS are related to high concentration, but high or low RH is also frequent.

- NO_2: condition of present meteorological factors lower than average (TEMP, GSR, WS, and RI).

- SO_2: Higher than average RI, GSR, RH, and TEMP, but milder WS.

- O_3: Higher than average TEMP, GSR, WS, and lower than average RH, with higher or lower than average RI index not changing this condition.

Once the seasonality of pollutants characterized by the effects of associated meteorological conditions has been identified, the relevance is the possibility of exploring in a more specific way which meteorological factors are present during the high concentration of each pollutant. Table 4.5 shows that there is a frequent relationship (2017 to 2019) between weather conditions associated with the increase in the concentration of each pollutant. For example, the high concentration of $PM_{2.5}$ is related to periods when there is little rain, and this behavior is repeated in three consecutive years. Still, the high pollutant concentration is associated to other meteorological conditions, such as humidity, temperature, and wind lower than annual average. Likewise, the increase in PM_{10} is related to high temperatures, the CO with hard solar radiation, the NO_2 with mild wind speed, SO_2 with higher than average temperature and humidity and low wind, and O_3 related to high temperatures and hard solar radiation but low humidity.

In a meteorological analysis, relative humidity and rainfall are significantly associated with pollutants under the following conditions: $PM_{2.5}$ (RI and RH lower than average), PM_{10} (RI lower than average), NO_2 (RI lower than average), SO_2 (RH and RI higher than average), O_3 (RH lower than average and RI lower or higher than average) and CO (RH lower or higher than average), as found in other studies (ZHAO et al., 2016; SHRESTHA, 2022). We can therefore consider that the increase in humidity or more rains can decrease the levels of PM_{10}, $PM_{2.5}$, NO_2, and O_3 pollution. The reduction of humidity and little rain combined with other associated meteorological factors (Table 4.5) can contribute to the control of SO_2 and CO. Given the relationship between the pluviometric index and the relative humidity of the air, it is essential to understand that the humidity of the air is related to the amount of water vapor present in the atmosphere, which determines the dry or humid condition and varies daily,

however the high humidity in the atmosphere that favors the occurrence of rains (FEISTEL, HELLMUTH, 2021).

The light wind was a unanimous behavior for all pollutants, affecting the concentration. Only one rule for the pollutant O_3 with 88.88% support and 100% confidence presented strong wind in the background but associated with high temperature, hard solar radiation, little rain, and low humidity. Strong wind disperses contaminants but favors carrying them over long distances (SHRESTHA, 2022). The temperature lower or higher annual average was an antecedent in all pollutants' rules. The same happened with the global solar radiation, only absent for the PM_{10}. The conditions associated with each pollutant were: CO, NO_2, and O_3 (TEMP and GSR higher than average), MP_{10} (TEMP higher than average), $MP_{2.5}$, and NO_2 (TEMP lower than average and GSR lower than average). This leads us to conclude that temperature and global solar radiation impact pollutants, but always with other meteorological factors. Generally, high temperatures and hard solar radiation are accompanied by little rain, low humidity, and little wind, resulting in a high concentration of pollutants. Temperature is associated with radiation as it results in heating the atmosphere. The air absorbs little solar radiation, mainly heated by the energy released from the Earth's surface (ZHAO et al., 2016, LIU et al. 2020).

Considering the periods with highest concentrations of pollutants and which meteorological factors influence their increase, the rules presented in Table 4.5 contribute as warning indicators that they present a greater risk to the population. The results indicate that meteorological conditions contribute to critical pollution episodes for all pollutants. Other studies prove that meteorological factors impair the dispersion of pollutants. Moisan et al. (2018) and Liu et al. (2021) reported that, in the winter months, the behavior of pollutants is influenced by the meteorological conditions of the season, which is consistent with the rules obtained. CETESB (2020) reported that during the winter of 2019 there was a predominance of hot and dry air mass throughout the state of São Paulo, with low ventilation and absence of rain, hindering the dispersion of pollutants. Similarly, frequent, and high exposure to pollutants is related to increased hospital visits and hospitalizations, such as $PM_{2.5}$particles, which can remain in the atmosphere for several weeks (NODARI, SALDANHA, 2016; MACHIN, NASCIMENTO, 2018).

4.4 Conclusions

In order to understand the current air pollution situation in São Paulo, this work investigated multi-polluting spatial and temporal patterns (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3, and CO) using data clustering techniques involving 56 air quality monitoring stations covering urban, coastal, and countryside areas from 2017 to 2019. The analysis sought reveals disparities in the spatial distribution of pollution concentrations throughout São Paulo state, as well as the seasons most impacted by each pollutant and their temporal variations. The meteorological parameters (air temperature, relative humidity, wind speed, rainfall, and global solar radiation) associated with critical periods of concentration in different regions were generated by the association rule mining technique and involved the choice of periods of change in regular seasons per pollutant, for groups of severely polluted stations. In the spatial analysis by cluster, the stations belonging to one cluster have characteristic and similar behavior patterns over time, categorized into pollution levels, with quality assured by the silhouette coefficient, resulting in the formation of two to three groups per pollutant. The temporal profile of the cluster allowed the identification of the seasonality of pollutants between groups, demonstrating the importance of investigating the unfavorable meteorological factors prevailing in periods of high concentration. With the analysis by association rules, it was possible to explore, in a more specific way, which meteorological parameters were present during the months of the peak concentration of each pollutant, considering the rules with more significant support and confidence (above 83.33%), which guaranteed a high frequency of the same pattern in the data. Although easily accessible, the limitations found for this study were the unavailability of some stations for data from the entire time series (2017-2019), mainly meteorological. This forced the discard of some stations, which does not affect the analysis because it deals with a large volume of data, but would make it even more comprehensive, including other cities. The data mining techniques were adequate for the proposed study and guarantee replicability for scenarios in which the spatial and temporal patterns of air pollution and probable associations are essential, justifying future work, which may include new parameters aimed at socioeconomic factors, health, or transport. In conclusion, the results demonstrated potential benefits since the problems related to the emission of pollutants were originated regionally and should preferably be faced from these same perspectives. Therefore, the importance of specific studies at the state scale, based on the behavior of pollutants in different regions and related climate changes, provides subsidies for the generation of effective policy measures to avoid critical periods of exposure of the population and to mitigate the emission of pollutants.

5. Short-term relation between air pollutants and hospitalizations for respiratory diseases: analysis by temporal association rules

5.1 Introduction

The use of strategies to reduce the concentration of air pollutants, in the short and long term, can prevent the development of new diseases and moderate the occurrence of existing ones (BELLINGER et al., 2017; BERGMANN et al., 2020; NADALI et al., 2022; HU et al., 2022; ZHU et al., 2022). Studies associated with the effects of pollution on health have shown that short-term exposures that can be harmful to health are those measured in days or minutes. They still constitute acute air pollution episodes related to lung diseases, especially bronchitis, and asthma, triggering respiratory infections and cardiovascular disease (ischemia, arrhythmia, and heart failure), in addition to the increase in hospitalizations and visits to emergency units (LAM et al., 2016; NODARI & SALDANHA, 2016; BERGMANN et al., 2020; MERCAN et al., 2020; ZHU et al., 2022). However, the effects of long-term exposure, generated over a lifetime or for years, are the result of prolonged coexistence with pollutants and are associated with increased mortality and exposure to carcinogens, in addition to the development of chronic respiratory diseases and worsening of heart problems (YANAGI et al., 2012; SOUZA & RABELO, 2016; LIU & PENG, 2018; SALAMEH et al., 2018; LEIRIÃO et al., 2020; SOMPORNRATTANAPHAN et al., 2020; BAI et al., 2022).

The patterns established as indicators of air quality are generated due to the frequency of acute episodes of concentration of a group of pollutants: inhalable particulate matter (PM_{10}), fine particulate matter ($PM_{2.5}$), ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2) and sulfur dioxide (SO_2), which are universally monitored and act differently on human health; however, only a limited number of cities have air quality monitoring networks (CANÇADO et al., 2006; POLEZER et al., 2022). It is evident that the most vulnerable groups are at high risk, even at low levels of exposure; children due to an immature immune system, older adults due to less efficient immunity combined with age comorbidities, and people with pre-existing chronic diseases (MACHIN & NASCIMENTO, 2018; MIRANDA et al., 2021; SHUDAN et al., 2022).

According to the World Health Organization (WHO), the new Global Air Quality Guidelines (AQG) are more restrictive about the levels considered safe, as they provide clear evidence that air pollution affects different aspects of health in even lower concentrations than previously thought (WHO, 2021; POLEZER et al., 2022). The Air Quality Life Index (AQLI), published by the Energy Policy Institute of the United States, quantifies the long-term relationship between air pollution and life expectancy. The report published in 2021 indicated that 79% of the population lives in areas where the concentration of particulate matter has exceeded the recommended limit, affecting health more than any other condition and resulting in a lost life expectancy per person of approximately two years (LEE & GREENSTONE, 2021; IEMA, 2022). The concern is even more significant in low and middle-income countries, where large populations suffer daily from severe air pollution due to large-scale urbanization, industrial expansion, and economic development based on inefficient public policies (ANDRADE et al., 2017; MIRAGLIA & GOUVEIA, 2014).

The WHO also states that emissions generally have specific, predictable temporal patterns and that seasonal variations are determinant and can enhance their effects (MORAES et al., 2019; WHO, 2021). Therefore, to ensure acceptable limits and ideal conditions for people and the environment, a multi-pollutant analysis that considers the temporality and duration of critical events of high concentrations, capable of revealing, through time intervals, the behavior of the pollutants with the most significant influence on health is essential. Although there are already several studies, the update of the WHO guidelines is based on studies carried out in Europe, North America, Asia, and Oceania, with few in South America. In developing countries, even with much data available on health systems, little knowledge of the relation between health and long or short-term exposure to high levels of air pollution is generated. Much of the data is analyzed by statistical methods with limited capacity for temporal analysis of multiple variables simultaneously (AMEER et al., 2019; BERGMANN et al., 2020; REPRESA et al., 2019; RYBARCZYK & ZALAKEVICIUTE, 2018).

Degraded air quality, especially in urban centers, has been associated with worsening diseases (LI et al., 2017; CHIQUETTO et al., 2022; LEIRIÃO et al., 2022; LIU et al., 2022). In addition to many monitoring stations being located in urban areas and few in rural areas, studies have focused mainly on metropolitan areas, which often exceed air quality patterns. It is relevant to assess whether the behavior is similar in surrounding cities and other areas, such as the countryside and coastline. In the China, most monitoring stations are located in urban areas with high pollution levels. Especially for rural regions, spatial representativeness

decreases, which impairs the assessment of air pollution and human health (BAI et al., 2022). A study found that the concentration of PM2.5 shifted over time, demonstrating an internal regional and inter-regional transport with positive spatial autocorrelation. The most polluted cities, characterized by industry, are distributed further to the north and the least polluted ones to the south. The reduction of high concentrations of $PM_{2.5}$ in the most polluted cities contributed to the reduction of $PM_{2.5}$ pollution in the south (LIU et al., 2022).

Research involving air pollution uses association rule extraction algorithms to assess associations between pollutants, emission sources, and their effects on air quality. As proposed by Agrawal et al. (1993), association rule mining is premised on finding relations between items in the database by generating rules, which, in turn, reveal associations between elements that occur together. The possibility of events repeating themselves from time to time is substantial. Still, there are limitations in the literature regarding the temporal order in the association of data and their relationships. Classical algorithms are not applicable for extracting multiple rules from temporal data and treat time as a simple numerical attribute, usually extracted from sequential data, ordered linearly in time (KAM & FU, 2000; LAXMAN & SASTRY, 2006; WINARKO & RODDICK, 2007; MITSA, 2010). The incorporation of the temporal aspect to association rules is a factor of great importance for data mining. Temporal association rules look for exciting relations or patterns in large data sets when the temporal attribute and its relationships are considered at intervals. For example, “event A occurs after event B occurs”, or “both events A and B occur before event C occurs” (MITSA, 2010; PAYUS et al., 2013; CAGLIERO et al., 2016; LI et. al., 2019; LI et al., 2020). Due to each domain's specificities, temporal data are studied in variable granularities and frequencies; although practical, the analysis of time points offers a different level of detail than data mining with temporal or even frequent temporal intervals, being more advantageous and relevant.

Related works such as by Nguyen et al. (2018) found patterns in the form of temporal association rules. The rules is a particular type of cancer treatment and its consequent set of co-occurring toxicities. Rules are built by temporal decision trees from data sequencing. The algorithm only determines the temporal directions of the associations, as reported by the author, however, it is relatively insensitive to temporal interval constraints between cancer treatments and diagnostics, therefore, ineffective in revealing temporal progression. In the capital of Malaysia, an urban and industrialized city, Payus et al. (2013) evaluated by association rules that exposure to high concentrations of PM_{10}, CO, and high temperatures is strongly associated with an increase in the number of hospitalizations of patients for respiratory problems, with

more than 90% confidence. The Apriori algorithm was applied to the dataset, but did not contain a temporal dimension, as the timestamps were discarded to generate the association rules. Gomes et al. (2013) showed a significant association between the increase in the number of hospitalizations for asthma in children in 27 cities in Greater São Paulo due to exposure to PM concentrations and meteorological factors. To detect the association, the Poisson probability distribution was used, whose exponential represents the relative risk, given by the ratio between the probabilities of the explanatory variable. In the United States, Moskovitch and Shahar (2009) found, in a set of records of diabetic patients, complex temporal patterns of clinical records, organized in a tree of significant temporal patterns with different frequencies.

In this way, this study uses the time factor to discover temporal association rules to analyze hospitalizations for respiratory diseases when the WHO concentration limits are exceeded for the pollutants: PM_{10}, $PM_{2.5}$, NO_2, CO, O_3, and SO_2, from January 2017 to October 2021. Three cities in São Paulo, Brazil, with different characteristics (in the metropolitan, coastal, and countryside regions) were considered here. The main contributions of this work are 1) Mining multi-pollutant temporal rules (temporal implications between variables) and not just identifying patterns in a time series; 2) Investigate the frequent daily temporal behavior of the relation between pollutants and the increase in the number of hospitalizations for respiratory diseases; 3) Present which daily time intervals (before and after) the relation occurs; 4) Identify which pollutants and their combinations are most harmful to health.

5.2 Data and methods

5.2.1 Study area

According to the Institute of Energy and Environment of São Paulo (Santana et al., 2012; IEMA, 2022), Brazil still has high levels of concentration compared to other countries and remains above that indicated by the WHO over the last 22 years, even in the 2020 and 2021 pandemic years. The state of São Paulo has the largest industrialized area in Brazil, located in the country's southeastern region. According to the Brazilian Institute of Geography and Statistics (IBGE), it is the most populous state with 46.6 million inhabitants, the equivalent to 21.9% of the Brazilian population, and it has the country's largest GDP with 31.6% (IBGE, 2021).

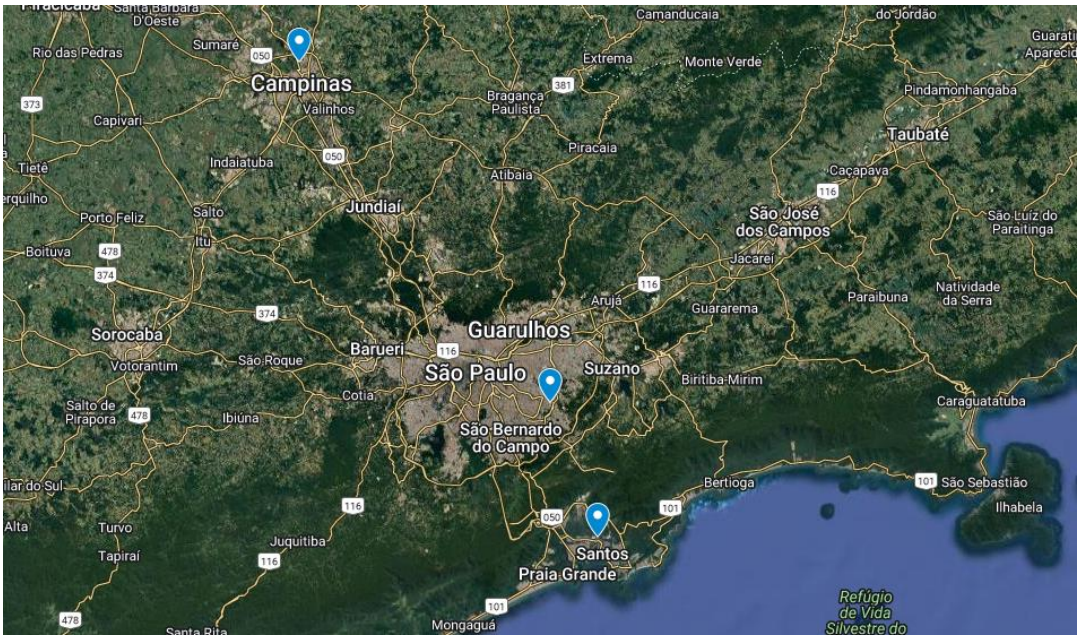


Figure 5.1 - Location of the cities considered in the study, in São Paulo (metropolitan), Santos (coastal), and Campinas (countryside).

This study considers three cities in the state of São Paulo (Figure 5.1), which, from January 2017 to October 2021, had high levels of air pollution and the highest number of pollutants monitored by CETESB (QUALAR 2021), but with different topographic characteristics: São Paulo, located in the Metropolitan Area of São Paulo (RMSP), the coastal city of Santos, and the city of Campinas, in the countryside of the state.

5.2.2 Dataset

Data collection involved two public databases: air pollution data were obtained from CETESB's QUALAR (QUALAR, 2021) online platform, which automatically monitors air quality with 62 stations in 36 municipalities, 29 of them being in the RMSP and the other 33 stations in the countryside and coastal regions. Daily data on all admissions in the Brazilian Unified Health System (SUS) were provided by TABNET/TabWin of the Department of Informatics of SUS (DataSUS, 2021), which comprises only public health data.

- QUALAR database: the maximum daily concentrations of each pollutant obtained for each city. As made available, for the pollutants, PM_{10}, $PM_{2.5}$, and SO_2 were the maxima in 24h. For NO_2, the average maximum of 1h per day, and for O_3 and CO, the maximum average of 8h per day. As it involves a larger area, the average of the daily maxima of three

stations was generated for the city of São Paulo: Congonhas, Parque D. Pedro, and Marginal Tietê. In Campinas, missing data from the Vila União station were completed by the nearest station - Centro (PM_{10}: 4.8%, PM_{10}: 9.8%, O_3: 10.7%, NO_2: 15.2%). For Santos, only one station (Ponta da Praia).

- DataSUS Database: Daily hospital admissions for respiratory diseases CID-10 (J00 to J099) were collected according to the International Classification of Diseases – CID. Data per patient were: MUNIC_RES – municipality of residence; DI_INTER – date of admission; DIAG_PRINC – main diagnostic code. We emphasize that the patient's city of residence was considered and identified by the IBGE code (São Paulo – 355030, Santos – 354850, and Campinas – 350950).

Adding up the daily hospitalizations for respiratory diseases in each municipality (Table 5.1), the annual total of hospitalizations was 223,433 for São Paulo, 5,661 for Santos, and 26,339 for Campinas. It is important to highlight that the data provided does not include hospitalizations from the private sector or visits that did not generate hospitalizations, only SUS Health establishments. According to the São Paulo State Health Department data, more than 25 million people are exclusive SUS users, representing about 58% of the state's 43 million inhabitants (MENDES, 2018).

Table 5. 1 - Total hospitalizations for respiratory problems CID-10 (J00 to J099) in the Unified Health System (SUS) from 2017 to 2021.

City	Stations	Profile	2017	2018	2019	2020	2021*	Total
São Paulo	Congonhas	Metropolitan	52,493	52,336	50,619	38,072	29,913	223,433
	Parque D. Pedro							
	Marginal Tietê							
Campinas	Vila União	Countryside	6,402	5,713	5,735	4,604	3,885	26,339
	Centro							
Santos	Santos	Coast	1,688	1,629	1,168	658	518	5,661

* Note: October 2021.

The last and next technique was adopted to fill in missing values in the databases, which replaced the missing value with the average between the previous and the next day (Plaia and Bondi, 2006). The annual average was considered in consecutive intervals of days with missing values.

The database was organized in a multivariate daily format (composed of daily values for each variable) from January 2017 to October 2021. Temporality was expressed through a

temporal attribute date, and the generated temporal intervals represent the values assumed by the other attributes in such intervals.

Table 5.2 is a sample of the database of this study. Each table row includes the record of maximum daily values of concentration of pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3, and CO), the daily average of hospital admissions for acute respiratory problems, and the date of occurrence of each set of values. The dates were used to identify the days when each pollutant exceeded the daily concentration limits defined by the WHO (CETESB, 2019); the maximum daily concentrations collected, independently for each city, were compared and categorized as lower or higher. To categorize the values of the number of daily hospitalizations, an annual average of hospitalizations was generated for each city, and values lower or higher than the annual average of admissions were defined.

Table 5. 2 - Sample of a city's database for application of the ARMADA algorithm.

Date	**PM_{10}**	**$PM_{2.5}$**	**CO**	**SO_2**	**NO_2**	**O_3**	**Admissions**
2019-01-01	lower	higher	lower	lower	lower	higher	higher
2019-01-02	higher	higher	higher	lower	lower	higher	higher
2019-01-03	lower	higher	lower	lower	lower	lower	lower
2021-10-29	lower	lower	higher	higher	lower	lower	lower
2021-10-30	higher	lower	lower	lower	lower	higher	higher
2021-10-31	higher	higher	lower	higher	higher	higher	higher

5.2.3 Temporal Association Rules

Time data can be represented by: (i) events – an occurrence in time, (ii) intervals – start and end time, or (iii) time series – events based on regular intervals (WANG et al., 2006; WANG et al., 2018). When dealing with facts that last for a while, instead of analyzing the data as an instantaneous occurrence in chronological order, we consider the temporality of the data in intervals that can have different durations and frequencies, which allows us to find more significant patterns and a better understanding of the relations between intervals (WINARKO & RODDICK, 2007). However, most studies in the literature opt for data linked to univariate time series. Multivariate time series, which has a set of n variables, each with a time series, requires computational complexity to estimate the relevance of a pattern in time (MITSA, 2010; KAM & FU, 2000; RAJ et al., 2022).

The ARMADA algorithm (An algorithm for discovering Richer relative temporal Association rules from interval-based data) proposed by Winarko and Roddick (2007) seeks to find temporal patterns of data based on time intervals. It extends the MEMISP (MEMory Indexing for Sequential Pattern Mining) algorithm by Lin and Lee (2002), which explores frequent temporal patterns. Since the MEMISP originated from the combination of the Apriori (AGRAWAL et al., 1993) and GSP (AGRAWAL & SRIKANT, 1994) algorithms, based on the candidate generation and testing methodology and also on the FP-Grow (HAN & KAMBER, 2006) and PrefixSpan algorithms (HAN, KAMBER & PEI, 2011) based on the growth of patterns, the ARMADA algorithm is the result of best practices already incorporated with MEMISP, influenced by algorithms from both methodologies. ARMADA has already been tested on synthetic datasets. According to Winarko and Roddick (2007) and Mitsa (2010), it proved superior to the algorithms from which it was generated and more efficient for finding frequent temporal patterns in large databases.

For this study, the ARMADA algorithm was used for two specific purposes: (1) to find all frequent temporal patterns and (2) to generate association rules from the frequent patterns, which correlate the events in the database, considering the association, and the duration intervals of such events.

In the 1st step, the algorithm searches for frequent patterns recursively, with the help of an index table, which associates each pattern with a start and end time and considers the state of each variable in the pattern identification (WINARKO & RODDICK, 2007; MITSA, 2010).

Formally, given a temporal database $D = \{t1. . . tn\}$, each record ti has an id and a temporal attribute with a start and end time, where *start_time<end_time*, indicating the relatively short time interval, compared to the total period under analysis. Where S is the set of all possible states for the data in *D*, when a state $s \in S$ holds for a while, called the state interval, it is represented according to the tuple *<start_time, s, end_time>*. If s is a unique state in *S*, we can say that s is a temporal pattern *<s>*. For example, in a hypothetical database that records the concentration of an atmospheric pollutant ($PM_{2.5}$) from a monitoring station (Santos) for a given period, the pollutant concentration value is considered the state *<s>*, represented by *<01/01/2019, 37µg/ m3, 01/02/2019>*. The algorithm's first task is to search for frequent states so that, in the end, the database is transformed into a collection of sequences of states.

In this study, the database is composed of continuous daily attributes of maximum concentrations of pollutants and average hospitalizations for respiratory diseases, organized by the temporal attribute date (Table 5.2). To find the patterns, the algorithm identifies time points

in which the continuous attributes (pollutants or hospitalizations) present a state or behavior of interest, defined as a pollutant concentration above the limit established by the WHO. The union of immediately neighboring time points allows the construction of time intervals.

The time interval between the intervals of states *i* and *j*, for $i < j$, being represented by $gap\ (i, j) = start\_time_{sj} - end\_time_{si}$, is the difference between the start point of *j* and the end point of *i*. A time constraint *(maximum_gap)* can be considered to remove less frequent patterns and only consider relations between state intervals that have a *gap* that meets the predefined *maximum_gap*.

The time interval between the intervals of states *i* and *j*, for $i < j$, being represented by $gap\ (i, j) = start\_time_{sj} - end\_time_{si}$, is the difference between the start point of *j* and the end point of *i*. A time constraint *(maximum_gap)* can be considered to remove less frequent patterns and only consider relations between state intervals that have a *gap* that meets the predefined *maximum_gap*.

In the 2nd step, the execution of the algorithm, the association rules are generated from the frequent patterns. A temporal association rule is an X→Y expression where X and Y are regular temporal patterns. X is the antecedent pattern of the rule, and Y is the consequent pattern, tending to occur when X occurs. In cases where the rule is associated with a time point, it is represented as Ri[t]: X→Y, and when it holds for a time interval, it is presented as: Ri[−t, +t]: X→Y, where Ri is the identification of the rule and [−t, +t] the time interval in which it was identified, so that −t is the smallest time point and +t is the largest one (AGRAWAL & SRIKANT, 1994; MITSA, 2010; HAN et al., 2011).

The relevance and validation of rules depend on the support and trust values. Given the X→Y rule, the support (Equation 5.1) is the rule's coverage and represents the frequency of pattern occurs in the database. Therefore, it is obtained by the ratio of the number of transactions that have the pattern through the total amount of transactions (CASTRO & FERRARI, 2016). The confidence (Equation 5.2) of a temporal association rule X→Y is the ratio between the Y and X standard deviation (MUELLER, 1995).

$$support\ (R_i) = \frac{|X \cup Y|}{|R|}. \tag{5.1}$$

$$confidence\ (R_i) = \frac{\sigma(Y)}{\sigma(X)}. \tag{5.2}$$

After finding frequent patterns with *support>minsup*, the algorithm allows the generation of temporal association rules from these patterns, with *confidence>minconf*,

resulting in pollutants associated with increased hospitalizations, considering the time interval in which this relation occurs. The problem with exploiting “richer” temporal association rules is generating all rules with confidence greater than or equal to the minimum confidence (*minconf*) specified by the user. This work aims to identify the best combination of support and confidence, the most representative of the databases.

In this study, the ARMADA algorithm was implemented in Python, using open-source libraries specific for machine learning: scikit-learn, mlxtend, pandas, seaborn, matplotlib, apriori, association_rules and numpy (PEDREGOSA et al., 2011; RASCHKA, 2018; REBACK et al., 2020).

5.3 Results and discussion

5.3.1 Descriptive data analysis

In this experiment, three databases were generated for the cities of São Paulo (RMSP), Campinas (countryside), and Santos (coast). Each base with a total of 1,765 records from January 1, 2017 to October 31, 2021, was composed of seven categorical attributes of concentration of monitored pollutants (PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, and O_3), and the total daily hospitalizations were composed of respiratory diseases in the respective locations. Table 5.3 presents the summarized statistics of each variable (maximum, minimum, mean, and standard deviation). Despite the differences between the cities regarding the concentration of pollutants and a number of hospitalizations, with few exceptions, the daily maximum of pollutants is above the recommended by the WHO in the three regions, with variability related to local meteorology and emissions, also found in similar studies, according to the characteristics of each region (AMATO et al., 2020; KACHBA et al., 2020; LEIRIÃO et al., 2022). With the new Global Air Quality Guidelines (AQG) published by the WHO, the recommended daily patterns have been reduced from 50μg/m^3 to 45μg/m^3 (PM_{10}) and from 25μg/m^3 to 15μg/m^3 ($PM_{2.5}$) and the annual standard of NO_2 decreased from 40μg/m^3 to 10μg/m^3 (WHO, 2021).

Table 5. 3 - Descriptive statistics of pollutants and hospitalizations from each database: São Paulo (RMSP), Campinas (countryside), Santos (coast), from 2017 to 2021.

Cities	**Pollutants**	**PM_{10}**	**$PM_{2.5}$**	**CO**	**SO_2**	**NO_2**	**O_3**	**Admission**
	WHO pattern(μg/m^3)	45	15	10	40	200	100	(daily)
São Paulo (RMSP)	Maximum	122	71	3	8	223	177	299
	Minimum	9	6	0	1	22	5	16
	Average	35	21	1	3	84	65	127
	Standard deviation	17.1	10.3	0.5	1.2	30.5	27.4	41.4
Campinas (countryside)	Maximum	89	60	-	-	136	162	40
	Minimum	7	5	-	-	0	12	1
	Average	27	19	-	-	41	70	15
	Standard deviation	10	9.1	-	-	24.8	22.1	5.6
Santos (coast)	Maximum	106	48	-	44	140	127	16
	Minimum	6	3	-	0	0	1	1
	Average	29	16	-	11	55	45	4
	Standard deviation	13.6	6.7	-	9	16.8	14.6	2

Note:

WHO pattern until 2021, as it involves data from 2017 to 2021 (WHO, 2021).
Daily parameters considered: MP_{10}, $MP_{2.5}$, and SO_2 (Daily Maximum), NO_2 (1-hour Average Maximum), and O_3 and CO (8-hour Average Maximum).
The stations in Campinas do not monitor SO_2 and CO pollutants, and in Santos there is no monitoring of CO.

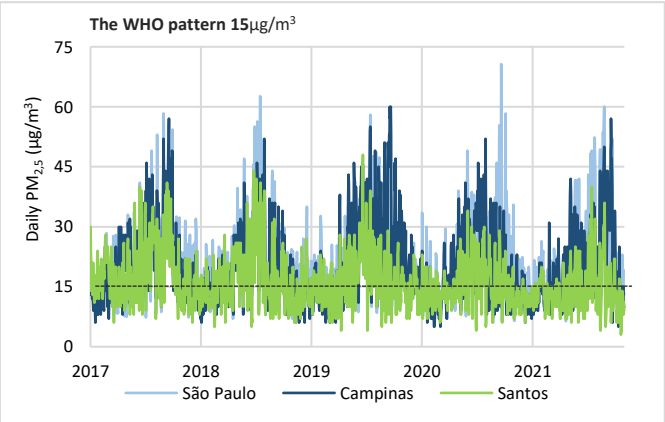
According to Table 5.3, PM_{10}, $PM_{2.5}$, and O_3 recorded high values in the three cities, SO_2 on the coast and NO_2 in the RMSP. The behavior of PM_{10} was similar to that observed for $PM_{2.5}$ in all cities. Andrade et al. (2017) state that there has been an increase in secondary pollutants $PM_{2.5}$ and O_3 in the last 30 years in the state of São Paulo. The studies by Tadano et al. (2021) and Araújo and Rosário (2020) highlight the same cities (São Paulo, Campinas, and Baixada Santista) as the most polluted regions by $PM_{2.5}$ in the state of São Paulo as found by Godoy and Silva (2022). In the study by Miranda et al. (2021) in the period from 2011 to 2017, high concentrations of $PM_{2.5}$ (28-63 μg/m^3), PM_{10} (52-110 μg/m^3) and O_3 (49-135 μg/m^3) were also found in São Paulo. When considering the $PM_{2.5}$ limit of 15 μg/m^3 for 24 hours, the daily average is higher than recommended, and the maximum values recorded were practically four times higher (São Paulo – 71 μg/m^3, Campinas – 60 μg/m^3, and Santos 48 μg/m^3). The same occurred for PM_{10}, with high maximum concentrations, above the limit of 45μg/m^3 (São Paulo – 122μg/m^3, Santos – 106μg/m^3, and Campinas – 89μg/m^3).

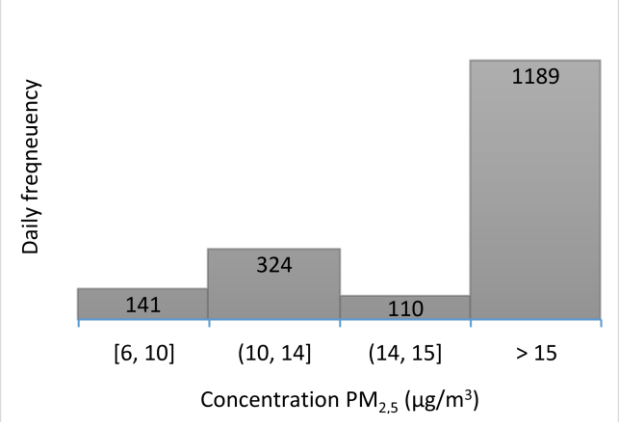
Maximum O_3 was 177μg/m^3 in the RMSP, 162μg/m^3, countryside, and 127μg/m^3, on the coast. Very present in hot seasons, the recommendation is a maximum daily average of 8 hours of 100μg/m^3 and a restriction of three to four days of exceedance per year. Studies have

shown that long-term exposure to O_3 is related to mortality from respiratory problems, but there is already evidence that this association persists even at deficient levels of exposure (CANÇADO et al., 2006; ALVIM et al., 2017).

SO_2 presented concentrations above the recommended (40µg/m^3) and was found in high concentrations along the coast of São Paulo. In Santos, the maximum concentration was 44µg/m^3 with an average of 11µg/m^3, with a much lower rate in the RMSP (maximum of 8µg/m^3 and average of 3µg/m^3). According to CETESB (2019), this is due to the large port and industrial complex in the region. A study by Wang and Corbett (2007) presented the costs and benefits of SO_2 reduction found in high concentrations on the west coast of the USA. Regarding the health impacts of the pollutant, Nardocci et al. (2013) found an association between SO^2 exposure and short-term health effects in cardiovascular disease on the coast of São Paulo. Finally, in São Paulo, NO_2 exceeded the recommended 200µg/m^3 with maximum concentrations of 223µg/m^3, mainly due to the emission of pollutants from motor vehicles and large industrial enterprises, in addition to the logistics sector and the local airport. A fundamental relationship for improving urban mobility in the megacity of São Paulo was highlighted by Chiquetto et al. (2022), who defend the need for urban planning integrated with transport to reduce the commute distance between jobs and housing, the main factor for emission of NO_2 currently in the region.

For a better understanding of the data, temporal aspects were considered in the analysis, initially only with the seasonality of pollutants between regions, for a general demonstration of the frequency of peak concentrations of pollutants and an increase in the number of hospitalizations (Figure 5.2). Then, the search for temporal patterns in each location and the multi-polluting relationship with hospitalizations by temporal association rules – the main contribution of this study – will be presented with data analysis.





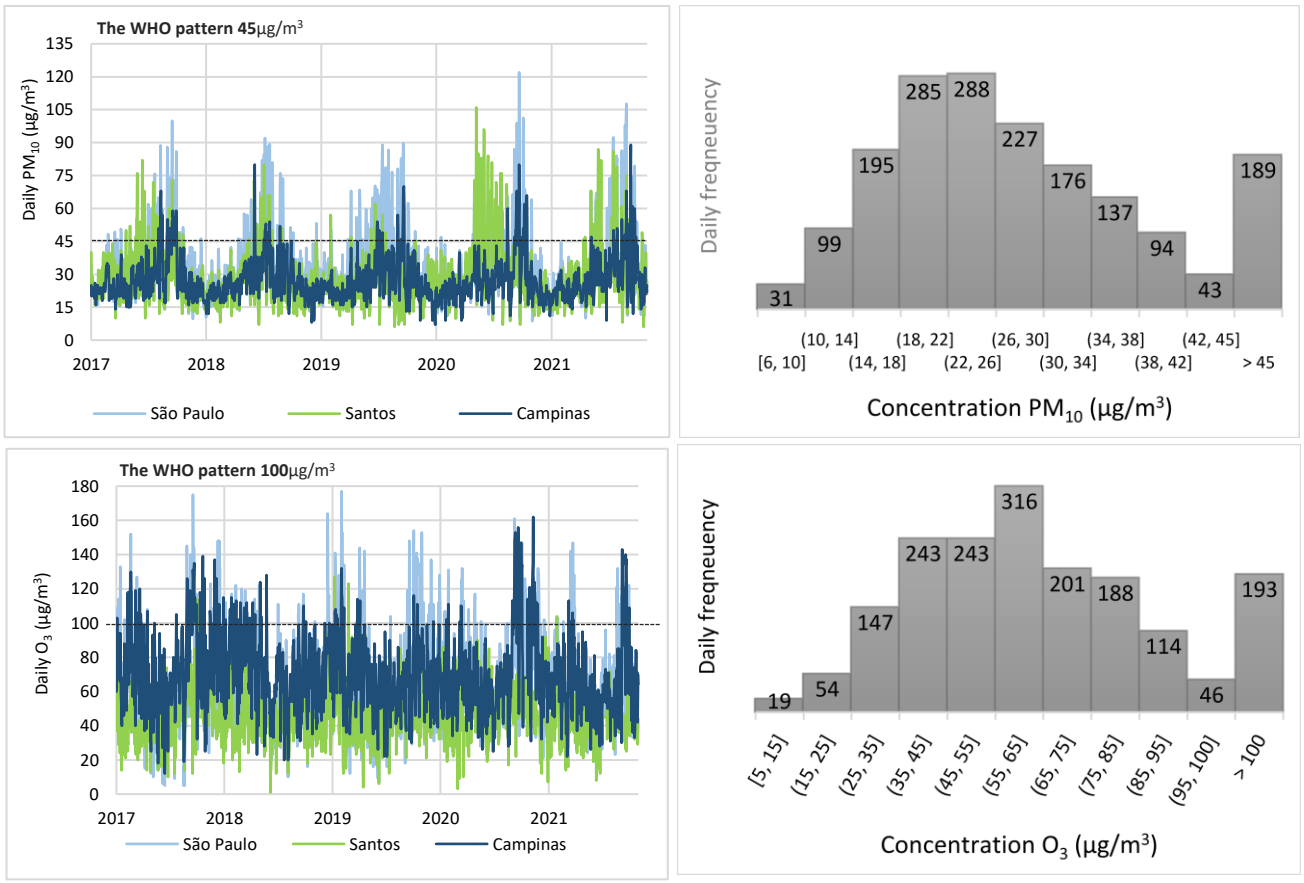


Figure 5.2 - Seasonality of inhalable and fine particulate matter pollutants (PM_{10}, $PM_{2.5}$) and Ozone (O_3) which recorded high values in the three cities. Due to the change in the values recommended by the WHO in 2021, the reference values were maintained from 2017 to 2021.

In general, concentrations were significantly higher in winter, except for O_3 in summer. A total of 255,433 cases of hospitalizations for respiratory diseases were included in this study (Table 5.1), with a high level of hospitalization observed in winter, a common condition found in related studies (LAM et al., 2016; LIU, 2019; ESCOBAR, 2020). In São Paulo, hospitalizations showed a marked dispersion, ranging from 16 to 299 daily hospitalizations, a daily average of 127 occurrences, and a standard deviation of 41.4 (Table 5.3).

Despite the average number of hospitals which stays in São Paulo being higher than in Campinas (minimum: 1, maximum: 40, average: 15, standard deviation: 5.6) and Santos (minimum: 1, maximum: 16, average: 4, standard deviation: 2), the patterns between the three cities were close. Considering the proportion of population size with average daily hospitalizations, Campinas had a slightly higher index (São Paulo = 0.0010%, Campinas = 0.0012% and Santos = 0.0009%).

Respiratory diseases showed a similar seasonality between cities. In a sample from 2019 (Figure 5.3), it is possible to notice that hospitalizations for respiratory diseases peak in early autumn (late March), which remains to a lesser extent during winter. This behavior is recurrent and demonstrates a more intense effect in the transition from summer to colder temperatures on the incidence of respiratory diseases, this is the period of temporal analysis in the following stages of this research. Lam et al. (2016) report that higher levels of humidity and ozone in the warm season and low humidity in the cold season were associated with more hospitalizations for respiratory diseases. In Canada, Parajuli et al. (2021) found positive seasonal correlations for $PM_{2.5}$ and O_3 in hot periods and $PM_{2.5}$ and NO_2 in cold ones from 2001 to 2012.

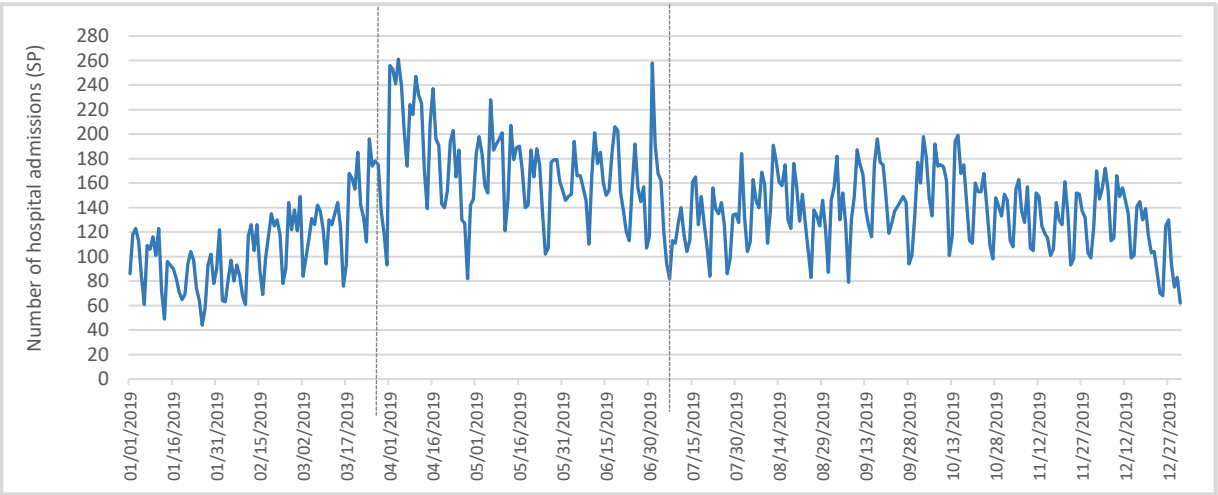


Figure 5.3 - Seasonality of hospitalizations for respiratory diseases represented by a sample of data, in São Paulo, in 2019.

Although these results show a possible association, more than seasonal effects are needed to assess the short-term risks and effects of multi-pollutant exposure to hospitalizations for respiratory problems, especially with such a large volume of data. The results can be ratified with a specific algorithm to mine temporal data, and the temporal patterns with rules that associate hospitalizations and pollutants concentrations over time can be analyzed. In the analysis by temporal association rules below, these data will be explored to identify which pollutants were related before, during, and after hospital discharge.

5.3.2 Analysis by temporal association rules

After the pre-processing step of the database, composed of the temporal attribute (date), categorized attributes: concentration of pollutants PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, O_3, and the number of hospitalizations, the ARMADA algorithm was applied to the mining of temporal patterns and rules. Unlike time series studies, which adopt implicit temporality, which, in turn, requires the chaining and organization of records to indicate the chronological order of events, this study adopted the explicit format of temporal data, with daily intervals as descriptors. Before starting to read the database, some initial parameters were defined:

- States of interest: the variables had their values categorized to emphasize differences in behavior. For each pollutant, daily concentration values were classified as higher or lower than the limits established by the WHO. When the database presents other uncategorized or non-temporal attributes, does not affect the results.
- Consequent target: association rules have the format antecedent → consequent, which indicates the relationship between them. The variable "hospitalizations" represented by the total number of daily hospital admissions for respiratory diseases was the consequent target, when it was above the annual average. As background, we have the states of interest of pollutants on days of high hospitalizations for the entire research period (2017 to 2021). According to the literature, studies seek to explain variations in hospitalizations for respiratory problems by counting hospitalizations over time (GOUVEIA et al., 2017; SANTOS et al., 2021).

The experiments were generated in two steps, according to Winarko and Roddick (2007): 1) Search for patterns associated with temporal intervals to identify the most frequent patterns and (2) Represent the daily temporal relations by association rules from the patterns found. Starting with the first step, the database was read recursively, looking for transactions in which the consequent target condition and states of interest were met. With the help of an index table, a new ID was assigned to each transaction located, with the registration of the date and state of the pollutants in this transaction, generating at the end a list of points of interest that make up a new database (Figure 5.4).

Still, in this first step, the time intervals are constructed with the time points associated with the transactions. To generate the intervals, the normalization of temporal patterns by Höppner (2001), called lifespan by Ale and Rossi (2000), was considered, and proposed to

identify the pattern duration in a database. Therefore, it was possible to obtain on which days there was an increase in the number of hospitalizations and for how long they were sustained (time interval of interest).

	Date	PM25	PM10	CO	SO2	NO2	O3	hospitalizations
0	02/15/2017	lower	lower	lower	lower	lower	higher	higher
1	02/16/2017	lower	lower	lower	lower	lower	higher	higher
2	02/17/2017	lower	lower	lower	lower	lower	higher	higher
3	02/18/2017	lower	lower	lower	lower	lower	higher	higher
4	02/21/2017	lower	lower	lower	lower	lower	higher	higher
...	...	...	...	...	...	...	...	...
298	09/21/2021	higher	higher	lower	lower	lower	lower	higher
299	09/22/2021	higher	higher	lower	lower	lower	lower	higher
300	09/27/2021	higher	higher	lower	lower	lower	higher	higher
301	09/28/2021	higher	higher	lower	lower	lower	higher	higher
302	09/29/2021	higher	higher	lower	lower	lower	higher	higher

Figure 5. 4 - Demonstration of a portion of the Table of Indexes and respective ID of the time points of interest, with pollutant states and consequent target (São Paulo database).

In Figure 5.5, we observed that the state *(s)* of each variable composes a pattern, represented by the triple *(b, s, f)*, which describes the condition of each variable in a time interval, defined by the start *(b)* and end *(f)* time points. As in São Paulo, from 02/15 to 02/18/2017 (Figures 5.4 and 5.5), a pattern can be recurrent with different durations. In this case, the high concentration of O_3 (state of interest) may have been reflected in the hospital discharge (consequent target). Therefore, the duration of each pattern is determined by its permanence time in consecutive days, joined by the 1st occurrence of the pattern *(b)* and the last *(f)*, similar to the temporalization process presented by Akhlagh et al. (2012).

For this, the intervals were ordered according to their starting time points *(b1, s1, f1), (b2, s2, f2), ..., (bn, sn, fn), ...* where $bi \leq bi+1$ and $bi < fi$. Therefore, the sequential intervals are replaced by a single interval for the pattern, resulting from the union of both *(min (bi, bj), s, max (fi, fj))*. The algorithm checks each time point in the index table for the proximity of existing intervals. A new interval is constructed when there is no interval on the day before or after the pattern (state) under analysis. The entire base period (2017 to 2021) was considered in this step.

With a small sample from January to May 2017 from the São Paulo database (Figure 5.5), it is possible to notice that the patterns are repeated at different intervals, as is the case of

(O_3: higher) from 02/15 to 02/18, 02/21, 03/10 and 04/15/2017, the pattern ($PM_{2.5}$: higher) on 02/23, 03/17, 03/25, 03 to 05/05 and 05/09 and finally the pattern ($PM_{2.5}$: higher; PM_{10}: higher) on 04/05 and 05/25/2017.

	b	f	PM 25	PM 10	CO	SO2	NO2	O3	hospitalizations	duration (days)
0	17/02/15	17/02/18	lower	lower	lower	lower	lower	higher	higher	4
1	17/02/21	17/02/21	lower	lower	lower	lower	lower	higher	higher	1
2	17/02/23	17/02/23	higher	lower	lower	lower	lower	lower	higher	1
3	17/03/10	17/03/10	lower	lower	lower	lower	lower	higher	higher	1
4	17/03/17	17/03/17	higher	lower	lower	lower	lower	lower	higher	1
5	17/03/25	17/03/25	higher	lower	lower	lower	lower	lower	higher	1
6	17/04/05	17/04/05	higher	higher	lower	lower	lower	lower	higher	1
7	17/04/15	17/04/15	lower	lower	lower	lower	lower	higher	higher	1
8	17/05/03	17/05/05	higher	lower	lower	lower	lower	lower	higher	3
9	17/05/09	17/05/09	higher	lower	lower	lower	lower	lower	higher	1
10	17/05/25	17/05/25	higher	lower	lower	lower	lower	lower	higher	1

Figure 5. 5 - Time intervals and duration when a pattern is maintained for one or more days. Sample for São Paulo database, from January to May 2017.

Observing Table 5.4, it is possible to found that approximately 36% (34% to 38%) of the total days with hospitalizations was above the annual average, one or more pollutants had concentrations above those recommended by the WHO. Therefore, high concentrations of pollutants may reflect on the same day for increased hospitalizations. The values that indicate this relationship were: in São Paulo, of the 831 days with hospitalization above the annual average, 303 days were a high concentration of one or more pollutants, likewise for Campinas (749/283) and Santos (568/192). Still, it reinforces that continuous exposure to pollutants occurring day(s) before hospital discharge can contribute to worsening health. Previous studies report effects for urgent and acute hospitalization (LAM et al., 2016; MATOS et al., 2019) and others significantly associated with different days of delay depending on the patient's condition (KHOSRAVI et al., 2020; PARAJULI et al., 2021; MOURA, 2021).

Thus, it was possible to obtain how many time intervals we had in each city and their respective durations (São Paulo – 208, Campinas – 171, and Santos – 144), shown in Table 5.4. Campinas had more consecutive days of hospital admissions, with a maximum of 11 days from 06/11 to 06/21/2019 and 07/13 to 07/23/2020. We observed that the consecutive days of high hospitalizations rarely exceed the frequency of one day (65% to 77%), two days (13% to 20%), and three days (3% to 10%), being a typical behavior of the cities, with slight variations. Therefore, more significant air pollution generated more days with high hospitalizations,

presented in Table 5.3 (descriptive statistics of the database). The literature indicates that the growth of 10μg/m^3 in the concentration of atmospheric pollutants increases the risk of emergency treatment for diseases respiratory (MATOS et al., 2019; CLAY et al., 2021).

Table 5. 4 - List of the number of transactions in the databases of São Paulo, Campinas, and Santos (2017 to 2021) in which the records with consequent targets and states of interest were located with their respective time intervals.

Number of records for	**São Paulo**	**Campinas**	**Santos**
Consequent target (in days): hospitalization for respiratory problems rate higher than the annual average	831 (47%)	749 (42%)	568 (32%)
States of interest (in days): pollutants with a concentration higher than the WHO pattern	303 (36%)	283 (38%)	192 (34%)
Consecutive days of high hospitalizations	208 intervals: 5 days – 1% 4 days – 3% 3 days – 9% 2 days – 13% 1 day – 74%	171 intervals: 11 days – 1% 6 days – 1% 5 days – 1% 4 days – 2% 3 days – 10% 2 days – 20% 1 day – 65%	144 intervals: 5 days – 1% 4 days – 3% 3 days – 3% 2 days – 16% 1 day – 77%

As expected, the first step was the search for frequent patterns and their time intervals. In the second step, temporal association rules were generated. This process begins with the reading of the time intervals database. When identifying a new pattern, the counter is started *(n=1)* and incremented *(n=n+1)* when the same pattern is found in the other time intervals until it registers all occurrences of the same pattern. The process is repeated for all the most frequent patterns, forming the temporal association rules. Support and confidence values of all generated rules are calculated to select the most frequent (highest support) and strongest (highest confidence) rules. Table 5.5 lists the main rules for daily temporal relations by city. The antecedent "high concentration of one or more pollutants" points to the consequent "high number of hospitalizations" from 2017 to 2021.

As a result, the rules show that the pollutants $PM_{2.5}$, PM_{10}, and O_3 are associated with increased hospitalizations in the RMSP (São Paulo) and countryside (Campinas), with particulate matter predominating in the most supportive rules (Table 5.5). In São Paulo, $PM_{2.5}$ and PM_{10} combined occurred with a frequency of 38.5%; in Campinas, only $PM_{2.5}$ alone had a frequency of 66.1%. Exposure to O_3 presented a lower rate compared to particulates, with

maximum support of 17.5%. In Santos, the rule was slightly different and showed that SO_2 contamination was directly related to hospitalizations, with 43.85% support.

The results were considered consistent in identifying temporal patterns by association rules, mainly because they correspond to the most present pollutants in regions and periods of high concentrations consistent with data reported by the CETESB (2019) and CETESB (2020). As SO_2 emissions are directly proportional to the sulfur content of fuels used by marine diesel engines, this leads to a high concentration of the pollutant on the coast, also observed in the USA, Canada, and Mexico (WANG & CORBETT, 2007), as well as in the coastline of Portugal (PEREIRA et al., 2007). Epidemiological studies found an association between SO_2 and the incidence of acute respiratory diseases in children on the day of exposure (NASCIMENTO et al., 2020) and cardiovascular diseases, two to three days after exposure (NARDOCCI et al., 2013).

For the entire study period, the association rules show that pollutants $PM_{2.5}$, PM_{10}, and O_3 at high levels, above the pattern recommended by the WHO, are present in the rule's antecedent when there is occurrence of hospital discharge due to respiratory problems (consequent). The effect of fine particles is considered more harmful to health, and this direct relation to hospitalizations can be observed in the three regions. The harmful effects of $PM_{2.5}$ occur both in the short term with acute effects and in the long term with chronic results when it accelerates the decline of lung function throughout life (NASCIMENTO et al., 2020). In India (RAJAK AND CHATTOPADHYAY, 2020), in Brazil (GONÇALVES et al., 2022), and Spain (AMATO et al., 2014) found strong associations between respiratory disease and higher rates of hospitalization or hospital visit to short-term exposures to particulate matter, as in this study, the main risk factor for respiratory health problems. Li et al. (2020) also found a correlation between $PM_{2.5}$, NO_2, and O_3 pollutants in China and between $PM_{2.5}$ and PM_{10} in the USA.

High ozone concentrations currently represent the main air pollution problem in São Paulo (ALVIM et al., 2017). Previous studies found that combined O_3 and $PM_{2.5}$ pollutants have consistent associations with respiratory hospitalizations and should be avoided at high levels, mainly if associated with low humidity and low temperature (LAM et al., 2016; PARAJULI et al., 2021). However, high O_3 concentration was not considered a determining factor for more hospitalizations for asthma in the study by Nadali et al. (2022). The CO pollutants monitored in São Paulo and NO_2 in São Paulo and Santos were not associated with the increase in hospitalizations.

In the RMSP, support indicates that high $PM_{2.5}$ and PM_{10} (antecedents) occur together at the base with a frequency equal to or greater than 38.5% when hospitalization is higher than the annual average. Confidence indicates that when hospitalization is above average, there is 77% certainty that $PM_{2.5}$ and PM_{10} pollutants (antecedents) are in high concentration. We can also conclude that peaks in $PM_{2.5}$ concentration increase hospitalizations with 29.3% support and O_3 with a lower frequency of 15.9%. SO_2 and CO did not exceed WHO patterns for this region in the study period. The background of the rules also leads us to conclude that simultaneous exposure to high levels of several pollutants is less common. The study found little risk for multi-pollutants due to low co-occurrences between air pollutants. Still, multi-contaminant air pollution is common in cities and puts human health at risk in a cumulative way (NASCIMENTO et al., 2020).

Table 5.5 - Temporal Association Rules obtained by the ARMADA algorithm. In the antecedents, the pollutants in the "higher" state than the WHO pattern for the consequent "hospitalization above the annual average" and their respective support and confidence parameters.

[**Antecedent**]	**Support**	**Confidence**
São Paulo		
[**$PM_{2.5}$: higher, PM_{10}: higher**, CO: lower, SO_2: lower, NO_2: lower, O_3: lower]	38.5%	77%
[**$PM_{2.5}$: higher**, PM_{10}: lower, CO: lower, SO_2: lower, NO_2: lower, O_3: lower]	29.3%	58.6%
[$PM_{2.5}$: lower, PM_{10}: lower, CO: lower, SO_2: lower, NO_2: lower, **O_3: higher**]	15.9%	31.8%
[**$PM_{2.5}$: higher, PM_{10}: higher**, CO: lower, SO_2: lower, NO_2: lower, **O_3: higher**]	10.1%	20%
[**$PM_{2.5}$: higher**, PM_{10}: lower, CO: lower, SO_2: lower, NO_2: lower, **O_3: higher**]	4.8%	15%
[$PM_{2.5}$: lower, **PM_{10}: higher**, CO: lower, SO_2: lower, NO_2: lower, O_3: lower]	1.4%	30%
Campinas		
[**$PM_{2.5}$: higher,** PM_{10}: lower, NO_2: lower, O_3: lower]	66.10%	94%
[$PM_{2.5}$: lower, PM_{10}: lower, NO_2: lower, **O_3: higher**]	17.5%	40%
[**$PM_{2.5}$: higher, PM_{10}: higher**, NO_2: lower, O_3: lower]	5.3%	10%
[**$PM_{2.5}$: higher**, PM_{10}: lower, NO_2: lower, **O_3: higher**]	4.7%	8%
[**$PM_{2.5}$: higher, PM_{10}: higher**, NO_2: lower, **O_3: higher**]	3.5%	7%
[$PM_{2.5}$: lower, **PM_{10}: higher**, NO_2: lower, O_3: lower]	2.3%	15%
[$PM_{2.5}$: lower, **PM_{10}: higher**, NO_2: lower, **O_3: higher**]	1.2%	23%
Santos		
[$PM_{2.5}$: lower, PM_{10}: lower, **SO_2: higher**, NO_2: lower, O_3: lower]	43.85%	80%
[**$PM_{2.5}$: higher**, PM_{10}: lower, SO_2: lower, NO_2: lower, O_3: lower]	14.65%	46%
[$PM_{2.5}$: lower, **PM_{10}: higher**, SO_2: lower, NO_2: lower, O_3: lower]	13.9%	28%
[**$PM_{2.5}$: higher**, PM_{10}: lower, **SO_2: higher**, NO_2: lower, O_3: lower]	11.8%	35%
[**$PM_{2.5}$: higher**, **PM_{10}: higher**, **SO_2: higher**, NO_2: lower, O_3: lower]	8.3%	4%
[**$PM_{2.5}$: higher**, **PM_{10}: higher**, SO_2: lower, NO_2: lower, O_3: lower]	6.2%	70%
[$PM_{2.5}$: lower, **PM_{10}: higher**, **SO_2: higher**, NO_2: lower, O_3: lower]	1.4%	51%

In Campinas, for the highest incidence of registered respiratory diseases, $PM_{2.5}$ also showed a high concentration, with 66.1% of occurrences predominant in frequent patterns.

Next, O_3 is associated with 17.5% frequency. The combined pollutants ($PM_{2.5}$, PM_{10}, and O_3) showed a low frequency, the same behavior as the RMSP. There was also no record of hospitalizations associated with NO_2. In Santos, the rules expressed 43.85% of hospitalizations attributed to SO_2 exposure, $PM_{2.5}$ with 14.65%, and PM_{10} with 13.9%, showing a lower association with respiratory problems. The CO pollutants monitored in São Paulo and NO_2 in São Paulo and Santos were not associated with increased hospitalizations.

Although the temporal rules are sufficient to state which pollutants are associated with high hospitalization in different regions, a short-term analysis is still needed to visualize on which days and intervals each pattern occurred and the temporal relations between them. To describe the temporal relations, Allen's Interval Algebra (AIA) was applied using the BEFORE and AFTER relations. The graphical representation (Figure 5.6) contributes to the interpretation and identification of relations, the horizontal axis represents the interval in days, and the vertical axis contains the rules.

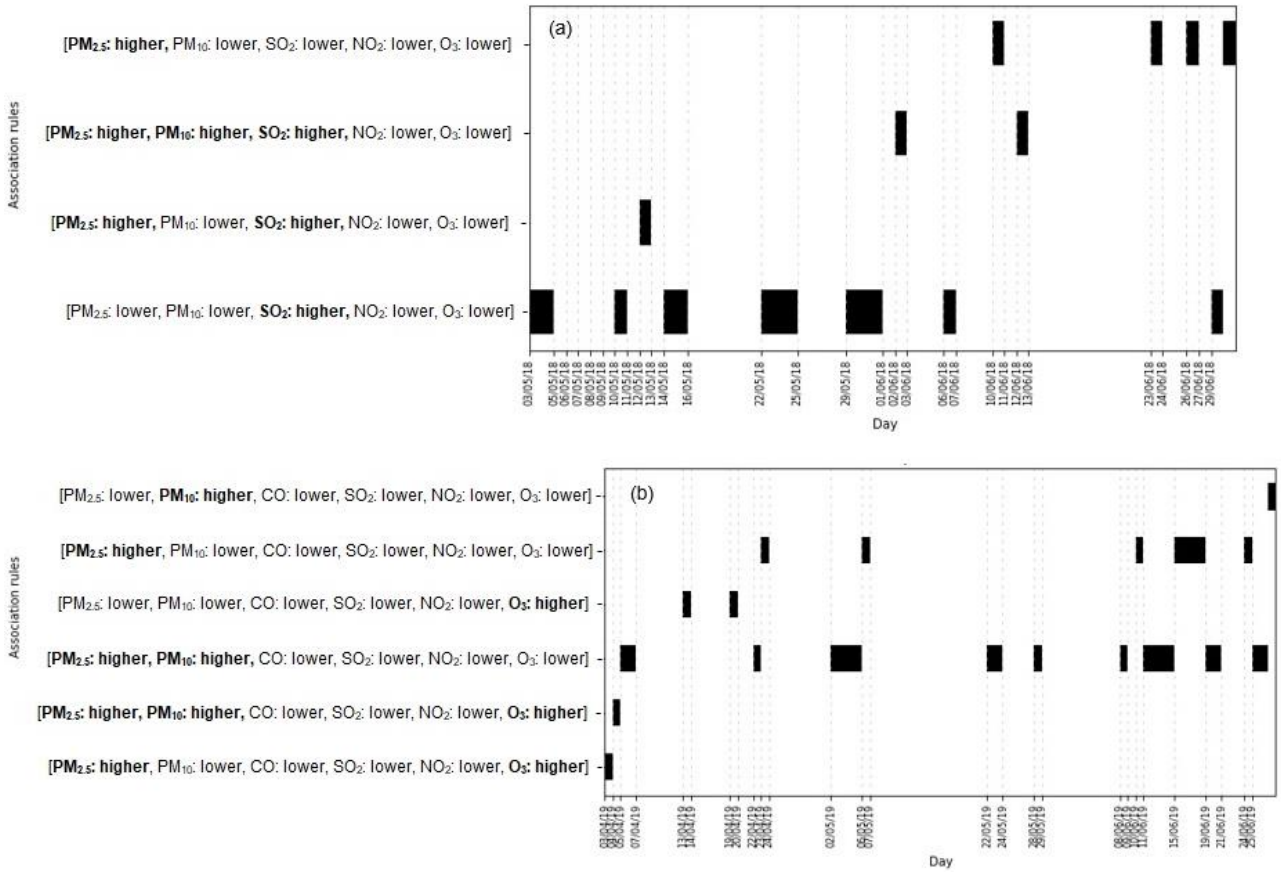


Figure 5.6 - Temporal relations between the association rules for the consequent "hospitalizations above the annual average" during the peak of hospitalizations: a) Santos from Maio to June 2018 and b) São Paulo from April to June 2019.

There was a peak of hospitalizations in 2018 in Santos city between May and June, which can be better understood in Figure 5.6 (a). Note that the most frequent rule was [SO_2: higher] from 05/03 to 06/07, one day (05/12) with [$PM_{2.5}$: superior, SO_2: superior] and twice (02 and 06/13) with [$PM_{2.5}$: superior, PM_{10}: superior, SO_2: superior]. As of 06/10/2018, the rule [$PM_{2.5}$: superior] is more frequent. Therefore, visualization allows simple and intuitive interpretations, which reveal, through temporal analysis, the relations between pollutants and hospitalizations for respiratory problems. Such a contribution was not found in the literature and exposed the short-term relation between temporal patterns.

During the time series analysis for São Paulo in 2019 (Figure 5.3), we identified an increase in hospitalizations from April to June 2019. With the graph of temporal relations in Figure 5.6 (b), we can see which pollutants were associated. The month of April started with the rules [$PM_{2.5}$: higher, PM_{10}: higher, O_3: higher] followed by intense periods of high concentration of [$PM_{2.5}$: higher, PM_{10}: higher] interspersed with short intervals of [O_3: higher] and [$PM_{2.5}$: higher]. The time interval highlighted are from 02 to 05/06/2019 and 11 to 06/15/2019 for the rule [$PM_{2.5}$: higher, PM_{10}: higher], which affect the high number of hospitalizations for five consecutive days. The same occurs from 15 to 06/19 for the rule [$PM_{2.5}$: higher]. On the other days, the attributes presented values within the expected behavior.

This first analysis considers the effects of pollution on the same day when there was a high concentration of one or more pollutants, which showed a direct association between pollutants and local respiratory admissions. However, studies report a delay in the effects of air pollution during the exposure period for individuals. That is, hospitalizations may be associated with pollution levels from previous days (YIN et al., 2017; MATOS et al., 2019; KHOSRAVI et al., 2020; NADALI et al., 2022). Therefore, it is essential to investigate the temporal association between hospitalizations and pollution levels in the earlier days and the behavior of pollutants and hospitalizations in the following days. The algorithm predicts a time constraint (*maximum_gap*) applied from selecting one of the temporal association rules, which considers only relationships between state intervals that meet the predefined *gap*. In this study, the *maximum_gap* = 4, four days was adopted before (*gap* = -1, *gap* = -2, *gap* = -3, *gap* = -4) and four days after (*gap* = 1, *gap* = 2, *gap* = 3, *gap* = 4) the intervals in which there were hospitalizations above the annual average. The choice of the gap is based on studies that evaluated the impact of pollution on health, ranging from 2 to 5 days earlier (MATOS et al., 2019; KHOSRAVI et al., 2020; NADALI et al., 2022.

To show the behavior of pollutants and hospitalizations for the São Paulo database from 2017 to 2021, the most frequent temporal association rule for the entire period was investigated: [$PM_{2.5}$: higher, PM_{10}: higher, CO: lower, SO_2: lower, NO_2: lower, O_3: lower] → [hospitalization: higher]. In Table 5.6, it is possible to observe the result with the most frequent rules (greater support and trust). The algorithm checks all the dates on which the rule occurred (from 2017 to 2021) and records the most frequent behavior of each pollutant and hospitalizations four days before and four days after, successively, for each gap.

Table 5. 6 - Temporal Association Rules and their support and trust. The algorithm evaluates the delay effect of gap = -1 to gap = -4 and subsequent gap = 1 to gap = 4 at all time points where the most frequent rule occurs in São Paulo, from 2017 to 2021.

GAP	[Antecedent] → [Consequent]	Support	Confidence
-4	[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: higher]	17.5%	56.00%
-3	[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: higher]	12.5%	43.47%
	[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: lower]	11.25%	47.36%
-2	[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: higher]	22.5%	72.00%
-1	[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: lower]	20.00%	57.00%
	[$PM_{2.5}$: higher] → [admissions: higher]	15.00%	48.00%
0	**[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: higher]**	**38.50%**	**77.00%**
1	[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: lower]	36.25%	70.73%
	[all pollutants: lower] → [admissions: higher]	22.50%	46.15%
2	[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: higher]	22.50%	40.90%
	[all pollutants: lower] → [admissions: higher]	18.75%	34.09%
3	[$PM_{2.5}$: higher, PM_{10}: higher] → [admissions: lower]	21.25%	41.46%
	[all pollutants: lower] → [admissions: higher]	21.25%	43.58%
4	[all pollutants: lower] → [admissions: lower]	22.25%	51.42%
	[all pollutants: lower] → [admissions: higher]	21.25%	37.77%

On the day before (gap = -1) the peak admission of all dates in which the rule occurred, pollutants $PM_{2.5}$ and PM_{10} had concentrations above the WHO limit (support: 20% and confidence: 57%), but hospitalizations were lower. Exposure to $PM_{2.5}$ and PM_{10} pollutants on previous days (gap = -2 and gap = -4) generate hospitalizations on the same day and new hospitalizations on the following days (gap = -1 and gap = -3) after exposure, even with low concentrations. The range of support was from 17.5% to 22.5% and confidence was from 56% to 72%. Another pattern that preceded the peak of hospitalizations was the increase in the concentration of $PM_{2.5}$ (support: 15% and confidence: 48%).

In the studies that sought the relationship between exposure to pollutants and their effects on health in China, Yin et al. (2017) found that PM_{10} exposure was significantly associated with daily mortality with a delay of up to 2 days. In Canada, the most consistent effects of O_3 exposure appeared one day after exposure (PARAJULI et al., 2021). In the say way in Iran, from 2014 to 2019, the most significant adverse effects found in hospitalization for asthma were observed the next day (NADALI et al., 2022). In Brazil, exposure to $PM_{2.5}$ it showed an increase of $10\mu g/m^3$ in the pollutant concentration in medium-sized cities and on the fifth day was significant the hospitalizations (MANTOVANI et al., 2016).

Although it is important to understand the behavior of pollutants in the period that precedes hospital discharges, it is equally important to analyze what happens after this period, if the pollutants remain high in the following days, for how long, and the effect on hospitalizations, almost not addressed by the studies that deal with this short-term relationship in the works. The study by Matos et al. (2019) found by Poisson regression, with daily hospitalization of respiratory diseases, daily concentrations of air pollutants, temperature, humidity, and rainfall, for PM_{10}, an increase of 2.43%, 2.73%, and 3.29% in the accumulated of 5, 6, and 7 days, respectively.

In Table 5.6 and Figure 5.7, an analysis of the most frequent rules in São Paulo from 2017 to 2021 (gap = 0), the most frequent and strong rules indicate that the pollutants that caused the increase in hospitalizations stayed for three days (gap = 1 to gap = 3) above the WHO limits, oscillating in smaller hospitalizations (gap = 1) and again larger (gap = 2 and gap = 4), in a decreasing trend in pollutant concentration levels over time. In parallel, another rule was reducing all pollutants on the other days (gap = 1 to gap = 4), but still maintaining high hospitalizations. Days after high hospitalizations, the risk of hospitalizations due to respiratory diseases was associated with lower concentrations of all pollutants.

Gap = 1

	PM25	PM10	CO	SO2	NO2	O3	hospitalization	size	support	confidence
0	higher	higher	--	--	--	--	lower	29	0.3625	0.707317
1	--	--	--	--	--	--	higher	18	0.2250	0.461538
2	higher	--	--	--	--	--	higher	9	0.1125	0.230769
3	--	--	--	--	--	--	lower	8	0.1000	0.195122
4	higher	higher	--	--	--	higher	lower	8	0.1000	0.205128
5	--	higher	--	--	--	--	higher	3	0.0375	0.076923
6	higher	--	--	--	--	--	lower	3	0.0375	0.073171
7	--	--	--	--	--	higher	higher	1	0.0125	0.025641
8	higher	higher	--	--	--	higher	lower	1	0.0125	0.02439

Gap = 2

	PM25	PM10	CO	SO2	NO2	O3	hospitalization	size	support	confidence
0	--	--	--	--	--	--	lower	24	0.3000	0.666667
1	higher	higher	--	--	--	--	higher	18	0.2250	0.409091
2	--	--	--	--	--	--	higher	15	0.1875	0.340909
3	higher	--	--	--	--	--	lower	7	0.0875	0.194444
4	higher	--	--	--	--	--	higher	6	0.0750	0.136364
5	higher	higher	--	--	--	higher	higher	4	0.0500	0.090909
6	higher	higher	--	--	--	higher	lower	3	0.0375	0.083333
7	higher	higher	--	--	--	--	lower	2	0.0250	0.055556
8	--	--	--	--	--	higher	higher	1	0.0125	0.022727

Gap = 3

	PM25	PM10	CO	SO2	NO2	O3	hospitalization	size	support	confidence
0	--	--	--	--	--	--	higher	17	0.2125	0.435897
1	higher	higher	--	--	--	--	lower	17	0.2125	0.414634
2	--	--	--	--	--	--	lower	11	0.1375	0.268293
3	higher	higher	--	--	--	--	higher	10	0.1250	0.25641
4	higher	--	--	--	--	--	higher	8	0.1000	0.205128
5	higher	--	--	--	--	--	lower	6	0.0750	0.146341
6	higher	--	--	--	--	higher	lower	4	0.0500	0.097561
7	higher	higher	--	--	--	higher	higher	3	0.0375	0.076923
8	--	--	--	--	--	higher	lower	2	0.0250	0.04878
9	--	higher	--	--	--	--	higher	1	0.0125	0.025641
10	higher	higher	--	--	--	higher	lower	1	0.0125	0.02439

Gap = 4

	PM25	PM10	CO	SO2	NO2	O3	hospitalization	size	support	confidence
0	--	--	--	--	--	--	lower	18	0.2250	0.514286
1	--	--	--	--	--	--	higher	17	0.2125	0.377778
2	higher	higher	--	--	--	--	higher	14	0.1750	0.311111
3	higher	--	--	--	--	--	higher	9	0.1125	0.2
4	higher	--	--	--	--	--	lower	7	0.0875	0.2
5	higher	higher	--	--	--	--	lower	7	0.0875	0.2
6	higher	higher	--	--	--	higher	higher	4	0.0500	0.088889
7	higher	higher	--	--	--	higher	lower	2	0.0250	0.057143
8	--	--	--	--	--	higher	higher	1	0.0125	0.022222
9	higher	--	--	--	--	higher	lower	1	0.0125	0.028571

Figure 5.7 - Sample of all Temporal Association Rules and support and confidence parameters, for gap = 1 to gap = 4. An analysis of the most frequent rule in São Paulo from 2017 to 2021.

It is important to recognize that differences in the composition of pollutants and people's vulnerability contribute to worsening pollution's effects on the population. The study by Machin and Nascimento (2018) states that a 5μg/m^3 increase in $PM_{2.5}$ concentrations increases the risk of respiratory hospitalizations by 20% and 38% (89 admissions) in children from Cuiabá, state of the Amazon Region in Brazil. In a large study of older adults, there was convincing evidence that ozone exposure did not find cardiovascular problems in this population but did confirm effects on the lung, even at low levels of exposure (FRAMPTON et. al., 2017). High pollution levels can increase the mortality risk associated with pollution by up to 15%, often due to slight variations (MANTOVANI et al., 2016; Nascimento et al., 2020). Therefore, the results show that adopting the new WHO guidelines (WHO, 2021) for concentration limits considered stricter and safer is urgent.

5.4 Conclusion

The State of São Paulo has the largest industrialized area and vehicular sources of emissions in Brazil, and many municipalities have been presenting high levels of air pollution. This study investigated the relationship between exposure to critical multi-polluting air pollution events (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3, and CO) with hospitalizations for respiratory diseases in the metropolitan area (São Paulo city), countryside (Campinas city), and coast (Santos city), from 2017 to 2021. Time series studies are insufficient to assess the risks and effects of short-term exposure to large-scale pollution. Differently, analysis by temporal association rules showed the potential to identify implications between variables over time and extracted patterns from multidimensional data when WHO concentration limits are exceeded

for pollutants. The application of the technique was able to show more frequent patterns associated with time intervals and presented the daily relations by association rules.

With the duration of each pattern, it was possible to obtain in which day intervals there was an increase in the number of hospitalizations and for how long they were sustained. We found that in approximately 35% of the total number of days with hospitalization above the annual average, one or more pollutants had concentrations above those recommended by the WHO on the same day. From the frequent patterns for the entire study period, the temporal association rules show that $PM_{2.5}$, PM_{10}, and O_3 pollutants are strongly associated with increased hospitalizations. In the RMSP ($PM_{2.5}$ and PM_{10} with 38.5% support and 77% confidence), in Campinas ($PM_{2.5}$ with 66.1% support and 94% confidence), and the pollutant O_3 with maximum support of 17.5%. On the coast (Santos), the rule showed that SO_2 high concentration is directly related to hospitalizations, with 43.85% of support and 80% of confidence. The CO pollutants monitored in São Paulo and NO_2 in São Paulo and Santos were not associated with the increase in hospitalizations.

The short-term analysis between pollutants and their health effects indicates on the previous day to the peak hospitalization the pollutants had concentrations above the WHO limit. The exposure to $PM_{2.5}$ and PM_{10} pollutants in the previous days (gap = -2 and gap = -4) generated hospitalization on the same day and new admissions in the following after exposure. The most frequent and robust rules indicate that the pollutants that caused the increase in hospitalizations remain for three days above limits, oscillating in smaller hospitalizations on the 1st day and again higher on the 2nd and 3rd days. In parallel, another rule shows that lower concentrations of all pollutants on other days (gap = 1 to gap = 4) maintain high hospitalizations. Therefore, lower concentrations are also related to higher hospitalizations, even within the limits established by WHO.

These results are important due to the scarcity of this approach in the literature and were considered consistent in identifying temporal patterns by association rules, mainly because they correspond to the most present pollutants in the respective regions and to the periods of high concentrations reported by CETESB. Another contribution is the intuitive graphical interface for investigating the daily temporal behavior between patterns and their relations by periods of interest. The limitations were the unavailability of pollutant data for the entire time series, solved with data from the nearest station of the same characteristic which was mentioned in the methodology, which guaranteed good results. Future works may include new meteorological variables to investigate their influence on this study. In conclusion, high pollutant exposure was

significantly associated with daily respiratory hospitalization, even with a delay. We identified which pollutants and combinations are most harmful to health in each region. Likewise, even with concentrations lower than the limits defined by the WHO, pollution impacts the number of hospitalizations.

6. Conclusion

As previously evidenced, the main objective of this work was to explore data mining techniques to find spatial and temporal patterns of air pollutants in cities in the State of São Paulo and to investigate meteorological conditions associated with critical periods of concentration of such pollutants, in addition, to analyze the relationship and short-term evolution in the increase of hospitalizations due to respiratory diseases. Since the temporality associated with the data is an essential contribution to the analysis and extraction of knowledge in the study of air quality and other areas, we consider this to be the most significant contribution, in addition to being multidisciplinary research generated from real data, which aims to seek in practice. These techniques are more adequate to the demands that the context lacks.

One of the themes that emerged from this research involved spatial aspects considered fundamental for a comprehensive study to understand the current situation of air pollution across the State and reveals disparities in the spatial distribution of pollution exposure data across the territory. This work investigated the multipollutant spatial and temporal patterns (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3, and CO) using the data grouping technique, involving stations with air quality monitoring, covering urban, coastal, and inland areas (GODOY et al. 2021; GODOY, SILVA, 2022a; GODOY, SILVA, 2022b).

In the spatial analysis by partitional clustering by pollutant, which used the K-medoids algorithm, the stations belonging to the formed groups showed characteristic and similar behavior patterns over time, categorized into pollution levels, with the quality of the groups validated by the silhouette coefficient, resulting in the formation of two to three groups per pollutant, which led to different levels of pollution. The temporal profile of the clustering allowed identifying the seasonality of the pollutants, which demonstrated the importance of investigating the unfavorable meteorological factors predominant in periods of high concentration.

By association rules, it was possible to explore, more specifically, which meteorological parameters (air temperature, relative humidity, wind speed, rainfall, and global solar radiation) were associated with the critical periods of concentration of each pollutant in different regions. The technique was applied based on the choice of periods of regular seasonal changes per pollutant, for groups of severely polluted stations, considering the rules with greater support

and confidence (above 83.33%), which guarantee a high frequency of the same pattern in the data.

The rules can help generate warning signs for possible increases in the concentration of pollutants, since the results confirm a relationship between high-concentration episodes and certain atmospheric conditions occurrence (GODOY et al. 2021; GODOY, SILVA, 2022a; GODOY, SILVA, 2022b). With this, we found the importance of specific studies based on the behavior of pollutants in different regions and related climate changes, which provide subsidies for the generation of effective political measures to avoid critical periods of exposure to the population and the mitigation of pollutant emissions.

The temporal analysis detected trends and periods of variation in the concentration of pollutants. It made it possible to investigate the time window between the occurrences of hospitalizations for respiratory diseases associated with exposure to critical events of multipollutant air pollution (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3 and CO) in the metropolitan region (São Paulo), countryside (Campinas) and coast (Santos), from 2017 to 2021. Time series studies proved insufficient in assessing the risks and effects of short-term exposure to large-scale pollution. Through an algorithm extension for mining temporal association rules, the technique searched for more frequent patterns associated with temporal intervals and presented daily relationships by association rules, which extracted patterns from multivariate (multipollutant) data when the thresholds of WHO concentrations are exceeded for pollutants.

With the duration of each pattern, it was possible to obtain in which intervals of days there was an increase in the number of hospitalizations and for how long they were sustained, with lags of up to 4 days in hospitalization after exposure, which allowed a more specific analysis, something relevant due to the scarcity of this approach in the literature by temporal association rules. Another contribution was the intuitive graphical interface for investigating the daily temporal behavior between patterns and their relationships by periods of interest.

In conclusion, the results demonstrate potential benefits since the problems related to the emission of pollutants originate regionally and should preferably be faced from these same perspectives. The results also showed that high exposure to pollutants was significantly associated with daily respiratory hospitalization, even with delay and identifying which pollutants and combinations are most harmful to health in each region. Finally, the findings from our experiments showed that the techniques could produce consistent scores compared to previous research and provide possible subsidies for air quality management in São Paulo.

The limitations found were the unavailability of some stations for the meteorological data of the entire time series (2017-2019), which led to the discarding of some stations, which did not affect the analysis because it is a large volume of data. In some cases, the unavailability of pollutant data was solved with data from the nearest station, with the same characteristics mentioned in the methodology of the articles.

The data mining techniques were suitable for the proposed study and ensured replicability. Therefore, as possible future research, other scenarios in which spatial and temporal patterns and their association matter can be carried out, including new parameters aimed at socioeconomic factors, transport, or other diseases associated with air quality. We also highlight that for the evolution of the methodology discussed here into a framework, we need to automate the data collection and pre-processing process with public databases.

References

ABE, K. & MIRAGLIA, S. (2018). 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), 47, 61-73. https://doi.org/10.5327/Z2176-947820180310

AGGARWAL, C. & REDDY, C. (2013). Data Clustering: algorithms and applications. CRC Press, (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series).

AGHABOZORGI, S., SHIRKHORSHIDI, A. S., WAH, T. Y., SOLTANIAN, H., & HERAWAN, T. (2015) Spatial and Temporal Clustering of Air Pollution in Malaysia: A Review. International Conference on Agriculture, Environment and Biological Sciences. Antalya, Turkey, 213 – 219.

AGRAWAL, R. & SRIKANT, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20. VLDB, Santiago de Chile, 12-15, 487-499.

AGRAWAL R, IMIELINSKI T, SWAMI A (1993) Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD CONFERENCE ON MANAGEMENT OF DATA, p. 207 – 216, Washington, DC, USA. ACM Press - New York, NY, USA. DOI: 10.1109/ACCESS.2019.2930004

AKHLAGH M M, TAN S C, KHAK F (2012) Temporal data classification and rule extraction using a probabilistic decision tree. In: International Conference on Computer Information Science (ICCIS), 2., 2012, Kuala Lumpur. Proceedings... [S.l.]: IEEE, p. 346–351.

ALE J M, ROSSI G H (2000) An approach to discovering temporal association rules. In: ACM Symposium on Applied Computing, Como. Proceedings... Nova Iorque: ACM, 2000. p. 294–300.

ALVIM D S, GATTI L V, CORRÊA SM, et al. (2017) Main ozone-forming VOCs in the city of Sao Paulo: observations, modeling, and impacts. Air Qual Atmos Health 10, 421–435. https://doi.org/10.1007/s11869-016-0429-9

AMATO F, ALASTUEY A, ROSA J, CASTANEDO Y G, CAMPA A M S, PANDOLFI M, LOZANO A, GONZÁLEZ J C, QUEROL X (2014). Trends of road dust emissions contributions on ambient air particulate levels at rural, urban and industrial sites in southern Spain. Atmos. Chem. Phys., 14, 3533–3544, https://doi.org/10.5194/acp-14-3533-2014, 2014

AMATO, F.; LAIB, M.; GUIGNARD, F.; KANEVSKI, M. (2020). Analysis of air pollution time-series using complexity-invariant distance and information measures. Physical A: Statistical Mechanics and Its Applications, 547. https://doi.org/10.1016/j.physa.2020.124391

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

ANDRADE M F, KUMAR P, FREITAS E D, YNOUE R Y, MARTINS J, MARTINS L, NOGUEIRA T, MARTINEZ P P, MIRANDA R M, ALBUQUERQUE T, GONÇALVES F L T, OYAMA B, ZHANG Y (2017) Air quality in the megacity of São Paulo: Evolution over the last 30 years and future perspectives. Atmospheric Environment, 159, 66–82. https://doi.org/10.1016/j.atmosenv.2017.03.051

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

ANGELEVSKA, B., ATANASOVA, V., & ANDREEVSKI, I. (2021) Urban air quality guidance based on measures categorization in road transport. Civil Engineering Journal, 7(2), 253–267. https://doi.org/10.28991/cej-2021-03091651

ARAÚJO, J.; ROSÁRIO, N. (2020). 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. https://doi.org/10.5327/Z2176-947820200552

ARCE, D., LIMA, F.; ORELLANA, M.; ORTEGA, J.; SELLERS, C. (2018). Discovering behavioral patterns among air pollutants: A data mining approach. Enfoque UTE, 9(4), 168 - 179. https://doi.org/10.29019/enfoqueute.v9n4.411

ATKINSON, R. (2000). Atmospheric chemistry of VOCs and NOx. Atmospheric Environment, 34, 2063–2101. http://dx.doi.org/10.1016/S1352-2310(99)00460-4

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

BAI H, YAN R, GAO W, WEI J, SEONG M (2022) Representatividade espacial das estações de monitoramento de PM 2.5 e sua implicação para a avaliação em saúde. Air Qual Atmos Health 15, 1571-1581 (2022). https://doi.org/10.1007/s11869-022-01202-2

BATISTA, A. F. M.; CHIAVEGATTO, A. D. P. Machine Learning aplicado à Saúde (2019). Workshop: Machine Learning. *In*: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADO À SAÚDE, 19. *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.

BELLINGER, C.; MOHOMED JABBAR, M.; ZAIANE, O; VARGAS, A.O. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health, 17(1), 1–19. https://doi.org/10.1186/s12889-017-4914-3

BERGMANN S, LI B, PILOT E, CHEN R, WANG B, YANG J (2020) Effect modification of the short-term effects of air pollution on morbidity by season: A systematic review and meta-analysis. Science of the Total Environment, 716. https://doi.org/10.1016/j.scitotenv.2020.136985

BISHT, M. & SEEJA, K.R. (2018). Air Pollution Prediction Using Extreme Learning Machine: A Case Study on Delhi (India). 10.1007/978-981-10-5828-8_18.

BOIAN, C. & ANDRADE, M. F. (2012). Characterization of ozone transport among metropolitan regions. Revista Brasileira de Meteorologia, 27 (2). https://doi.org/10.1590/S0102-77862012000200009

BONT, J., JAGANATHAN, S., DAHLQUIST, M., PERSSON, A, STAFOGGIA, M, LJUNGMAN, P. (2022). Ambient air pollution and cardiovascular diseases: an umbrella review of systematic reviews and meta-analyses. J Intern Med. 2022; 291: 779– 800. https://doi.org/10.1111/joim.13467

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

CAGLIERO, L., CERQUITELLI, T., CHIUSANO, S., GARZA, P., RICUPERO, G., & XIAO, X. (2016). Modeling Correlations among Air Pollution-Related Data through Generalized Association Rules. IEEE International Conference on Smart Computing, SMARTCOMP. https://doi.org/10.1109/SMARTCOMP.2016.75 01707

CANÇADO, J. E. D., BRAGA, A, PEREIRA, L. A. A., ARBEX, M. A., SALDIVA, P. H. N., & SANTOS, U. P. (2006). Repercursões clínicas da exposição à poluição atmosférica. Jornal Brasileiro de Pneumologia, 32 (Supl. 1): S5 – S11. Available at: http://www.scielo.br/pdf/jbpneu/v32s2/a02v32s2.pdf. Accessed on: Feb. 09, 2020.

CARBON BRIEF (2021). Analysis: Which countries are historically responsible for climate change? Available at: https://www.carbonbrief.org/analysis-which-countries-are-historically-responsible-for-climate-change/ Accessed on: Jul. 10, 2022.

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

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

CEMADEN (2020). CENTRO NACIONAL DE MONITORAMENTO E ALERTAS DE DESASTRES NATURAIS. Available at: http://www.cemaden.gov.br. Accessed on: Mai. 18, 2020.

CÉSAR, A. C. G.; NASCIMENTO, L. F. C.; MANTOVANI, K. C. C.; VIEIRA, L. C. P. (2016). 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. https://doi.org/10.1016/j.rppede.2015.12.005

CETESB (2019). Companhia Ambiental do Estado de São Paulo. 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. Available at:

https://cetesb.sp.gov.br/ar/wp-content/uploads/sites/28/2019/05/Relat%C3%B3rio-de-Qualidade-do-Ar-2017.pdf. Accessed on: Mai. 08, 2019.

CETESB (2020). COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO. Poluentes. Available at: https://cetesb.sp.gov.br/ar/poluentes/. Accessed on: Feb. 24, 2020.

CETESB (2020b). COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO. Relatório Operação Inverno. 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.

CLAY K, MULLER N Z, WANG X (2021). Recent increases in air pollution: evidence and implications for mortality. Review of Environmental Economics and Policy, v. 15, n. 1, p. 154-162 https://doi.org/10.1086/712983

CHEN, J., HOEK, G. (2020). Long-term exposure to PM and all-cause and cause-specific mortality: A systematic review and meta-analysis. Environment International, volume 143, 105974, ISSN 0160-4120, https://doi.org/10.1016/j.envint.2020.105974.

CHIQUETTO, J.B., LEICHSENRING, A. R., RIBEIRO, F.N.D., & RIBEIRO, W. C. (2022). Work, housing, and urban mobility in the megacity of São Paulo, Brazil, Socio-Economic Planning Sciences, Volume 81. https://doi.org/10.1016/j.seps.2021.101184.

CHUNG, C.Y., YANG, J., HE, J., YANG, X., HUBBARD, R., & JI, D. (2021). An investigation into the impact of variations of ambient air pollution and meteorological factors on lung cancer mortality in Yangtze River Delta, Science of The Total Environment, Volume 779. https://doi.org/10.1016/j.scitotenv.2021.146427

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (2019). 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. Available at: https://cetesb.sp.gov.br/ar/wp-content/uploads/sites/28/2019/05/Relat%C3%B3rio-de-Qualidade -do-Ar-2017.pdf. Accessed on: Mai. 08, 2019.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (2020). Relatório Operação Inverno. 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.

CONAMA (2018). CONSELHO NACIONAL DO MEIO AMBIENTE. Resolução CONAMA no 491, de 19 de novembro de 2018. Dispõe sobre qualidade do ar. Diário Oficial da República Federativa do Brasil, Poder Executivo, Brasília, DF, 21 nov. 2018. Seção 1, 155-156. Available at: http://www2.mma.gov.br/port/conama/legiabre.cfm?codlegi=740 Accessed on: Dec. 10, 2019.

CORÁ, B., LEIRIÃO, L., & MIRAGLIA, S. (2020). Impacto da poluição do ar na saúde pública em municípios com elevada industrialização no estado de São Paulo. Revista Brasileira de Ciências Ambientais, 1-12. https://doi.org/10.5327/Z2176-947820200671

DataSUS (2021). Informatics Department of the Unified Health System (in Portuguese: Departamento de Informática do Sistema Único de Saúde). Sistema de informação sobre

mortalidade e morbidade. Available at: http://datasus.saude.gov.br/informacoes-de-saude/tabnet. Accessed on: 20 August 2021.

DIMITRIOU, K. (2016). 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. https://doi.org/10.4209/aaqr.2016.05.0214

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

ESCOBAR, H. (2020). Dados comprovam aumento de eventos climáticos extremos em São Paulo. Jornal da USP. Available at: https://jornal.usp.br/ciencias/ciencias--ambientais/dados-comprovam--aumento-de-eventos-climati-cos-extremos-em-sao-paulo/. Accessed on: Mai. 05, 2020.

FRAMPTON, MW, BALMES, JR, BROMBERG, PA, STARK, P, ARJOMANDI, M, HAZUCHA, MJ et al. (2017). Multicenter Ozone Study in older Subjects (MOSES): part 1. Effects of exposure to low concentrations of ozone on respiratory and cardiovascular outcomes. Boston (MA): Health Effects Institute (Research Report 192 Part 1; https://www.healtheffects.org/ publication/multicenter-ozone-study-older-subjects-moses-part-1-effects-exposurelow-concentrations. Accessed on: Feb. 21, 2021.

FEISTEL, R., & HELLMUTH, O (2021). Relative humidity: A control valve of the steam engine climate. Journal of Human, Earth, and Future, 2(2), 140–182. https://doi.org/10.28991/HEF-2021-02-02-06

FOGLIATTO, F.S. & ANZANELLO, M.J. (2011). Selecting the Best clustering variables for grouping mass-customized products involving workers' learning. International Journal of Production Economics, Elsevier, 130(2), 268-276. http://dx.doi.org/10.1016/j.ijpe.2011.01.009

GALVÃO JR., P. A., ROVEDA, S. R. M. M., & VIEIRA, H. E. (2022). Hybrid Models Applied to Create a Classification Index of Fire Risk Levels in Brazil. Brazilian Journal of Environmental Sciences (Online), 1–11. https://doi.org/10.5327/Z2176-94781286

GODOY, A.R.L., SILVA, A.E.A. (2022a). Spatial patterns and temporal variations of pollutants at 56 air quality monitoring stations in the state of São Paulo, Brazil. Environmental Monitoring and Assessment 194, 910. DOI: 10.1007/s10661-022-10600-z

GODOY, A.R.L.; SILVA, A.E.A. (2022b). Análise de Material Particulado Fino no Interior do Estado de São Paulo por Agrupamento de Dados. In: XIX Congresso Nacional de Meio Ambiente de Poços de Caldas, Poços de Caldas - Minas Gerais. Anais do 19° Congresso Nacional do Meio Ambiente de Poços de Caldas, 2022. p. 59-68.

GODOY, A. R. L., SILVA, A. E. A., BUENO, M. C., POZZA, S. A., COELHO, G. P. (2021). Application of machine learning algorithms to PM2,5 concentration analysis in the state of São Paulo, Brazil. Revista Brasileira de Ciências Ambientais, 56, 152-165. https://doi.org/10.5327/Z21769478782

GOMES, A. C. D. S., LUCIO, P. S., & SPYRIDES, M. H. C. (2013). Influence of Pollution from Particulate Matter at the Hospitalizations of Asthmatic Children in area of Great São Paulo. Revista Brasileira de Geografia Física, 6(4), 749. <https://doi.org/10.26848/rbgf.v6i4.233066>

GONÇALVES, F. L. T.; CARVALHO, L. M. V.; CONDE, F. C.; LATORRE, M. R. D. O.; SALDIVA, P. H. N.; BRAGA, A. L. F. (2005). 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. https://doi.org/10.1016/j.envint.2004.08.004

GONÇALVES P. B., NOGAROTTO D. C., CANTERAS F. B., POZZA S. A. (2022). The relationship between the number of COVID-19 cases, meteorological variables, and particulate matter concentration in a medium-sized Brazilian city. Brazilian Journal of Environmental Sciences (Online), 57(2), 167–178. https://doi.org/10.5327/Z217694781300

GOUVEIA, N., CORRALLO, F. P., LEON, A. C. P., JUNGER, W., FREITAS, C. U. (2017). Air pollution and hospitalizations in the largest Brazilian metropolis. Revista de Saúde Pública, v. 51. https://doi.org/10.11606/S1518-8787.2017051000223.

GOVENDER, P. & SIVAKUMAR, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollution Research, 11(1), 40 - 56. https://doi.org/10.1016/j.apr.2019.09.009

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

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

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

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

HELLAN, S P, LUCAS, C G, & GODDARD, N H (2022). Bayesian Optimisation for Active Monitoring of Air Pollution. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), 11908-11916. https://doi.org/10.1609/aaai.v36i11.21448

HÖPPNER F (2001) Learning temporal rules from state sequences (2001). In: Workshop on Learning from temporal and Spatial Data, Seattle, USA. Proceedings…IJCAI'01, 2001, Seatle: p. 25-31.

HU Y, JI J S, ZHAO B (2022) Restrictions on indoor and outdoor NO2 emissions to reduce disease burden for pediatric asthma in China: A modeling study, The Lancet Regional Health - Western Pacific, 24, 100463. https://doi.org/10.1016/j.lanwpc.2022.100463

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

IAP (2020). Instituto Ambiental do Paraná. Fontes de Poluição Atmosférica. Available at: http://www.iap.pr.gov.br/pagina-1415.html Accessed on: Feb. 05, 2022.

IBGE (2021). Instituto Brasileiro de Geografia e Estatística. Available at: https://cidades.ibge.gov.br/brasil/sp/panorama. Accessed on: Jul. 05, 2022.

IBGE (2022). Instituto Brasileiro de Geografia e Estatística. Cidades e Estados. Available at: https://www.ibge.gov.br/cidades-e-estados. Accessed on: Mai. 05, 2022.

IEMA (2022). Instituto de Energia e Meio Ambiente. Qualidade do Ar. Available at: http://energiaeambiente.org.br/wp-content/uploads/2022/06/RelatorioAnual_IEMA_20 22.pdf. Accessed on: Mai. 27, 2022.

INPE (2019). Instituto Nacional de Pesquisas Espaciais. 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.

JAIN, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

JIANG, X.Q., MEI, X.D., FENG, D. (2016). Air pollution and chronic airway diseases: what should people know and do? J Thorac Dis. 2016 Jan;8(1): E31-40. https://doi.org/10.3978/j.issn.2072-1439.2015.11.50

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

JOÃO, R. S. (2020). Mineração de regras de associação temporais envolvendo dados quantitativos contínuos. Thesis (Doctoring) – Universidade Federal de São Carlos, São Carlos. Available at: < https://repositorio.ufscar.br/handle/ufscar/13877>. Accessed on: May 1, 2020.

KACHBA Y, CHIROLI D M G, BELOTTI J, ALVES T A, TADANO Y S, SIQUEIRA H (2020) Artificial Neural Networks to Estimate the Influence of Vehicular Emission Variables on Morbidity and Mortality in the Largest Metropolis in South America. Sustainability, 12, 2621 https://doi.org/10.3390/su12072621

KAM OS, FU AWC (2000). Discovering temporal patterns for interval-based events. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1874, 317 - 326. https://doi.org/10.1007/3-540-44466-1_32

KARASAWA, Eliane; SOUSA, Elaine P. M.. (2022). TRUMiner: Mineração de Regras Temporais em Bases de Séries Multivariadas e Heterogêneas. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 37., Búzios. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação. p. 403-408. ISSN 2763-8979. DOI: https://doi.org/10.5753/sbbd.2022.226199.

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

KHOSRAVI T, HADEI M, HOPKE P K, et al. (2020) Association of short-term exposure to air pollution with mortality in a middle eastern tourist city. Air Qual Atmos Health 13, 1223–1234 https://doi.org/10.1007/s11869-020-00875-x

KOSTYA, E. (2020). GEOPY. Available at: <https://pypi.org/project/geopy/>. Accessed on: Jan. 26, 2020.

KAWASHIMA, AB, MARTINS, LD, RAFEE, SAA et al. Desenvolvimento de um inventário espacializado de emissões atmosféricas para as principais fontes industriais do Brasil. Environ Sci Pollut Res 27, 35941–35951 (2020). https://doi.org/10.1007/s11356-020-08281-7

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

LAM, H. C., LI, A. M., CHAN, E. Y., & GOGGINS, W. B. (2016). The short-term association between asthma hospitalizations, ambient temperature, other meteorological factors, and air pollutants in Hong Kong: a time-series study. Thorax, 71, 1097-109. http://dx.doi.org/10.1136/thoraxjnl-2015-208054

LAXMAN, S. &, SASTRY, P.S. (2006). A survey of temporal data mining. Sadhana, 31, 173–198. https://doi.org/10.1007/BF02719780

LEE, Y., CHOI, Y., AN, H., PARK, J., & GHIM, Y. S. (2021). Cluster analysis of atmospheric particle number size distributions at a rural site downwind of Seoul, Korea. Atmospheric Pollution Research, 12(6). https://doi.org/10.1016/j.apr.2021.101086

LEE K., GREENSTONE M. (2021). Air quality life index annual update. Energy Policy Institute, University of Chicago. Available at: https://aqli.epic.uchicago.edu/wp-content/uploads/2021/08/AQLI_2021-Report.EnglishGlobal.pdf Accessed in July 2022.

LEIRIÃO, L. F. L., DEBONE, D., PAULIQUEVIS T., ROSÁRIO, N. M. E., MIRAGLIA, S. G.E.K. (2020) Environmental and public health effects of vehicle emissions in a large metropolis: Case study of a truck driver strike in São Paulo, Brazil. Atmospheric Pollution Research, 11(6), 24–31, 2020. https://doi.org/10.1016/j.apr.2020.02.020

LEIRIÃO L F L, DEBONE D, MIRAGLIA S G E K (2022). Does air pollution explain COVID-19 fatality and mortality rates? A multi-city study in São Paulo state, Brazil. Environmental Monitoring and Assessment 194, 275 (2022). https://doi.org/10.1007/s10661-022-09924-7

LI, D.; LIU, J.; ZHANG, J.; GUI, H.; DU, P.; YU, T.; CHENG, Y. (2017). Identification of long-range transport pathways and potential sources of PM2.5 and PM10 in Beijing from 2014 to 2015. Journal of Environmental Sciences, 56, 214–229. https://doi.org/10.1016/j.jes.2016.06.035

LI T, LI Y, AN D, HAN Y, XU S, LU Z, CRITTENDEN J (2019). Mining of the association rules between industrialization level and air quality to inform high-quality development in

China. Journal of Environmental Management, 246, 564–574. https://doi.org/10.1016/j.jenvman.2019.06.022

LI Z, ZHOU W, LIU X, QUIAN Y, WANG C, XIE Z, MA H (2020). Research on Association Rules Mining of Atmospheric Environment Monitoring Data. Technology-Inspired Smart Learning for Future Education. Singapore: Springer. https://doi.org/10.1007/978-981-15-5390-5_8

LIN MY, LEE SY (2002). Fast discovery of sequential patterns by memory indexing. In: Data Warehousing and Knowledge Discovery: 4th International Conference, DaWaK 2002 Aix-en-Provence, France, September 4–6, 2002 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 150–160. ISBN 978-3-540-46145-6.

LIU J C, PENG R D (2018). Health effect of mixtures of ozone, nitrogen dioxide, and fine particulates in 85 US counties. Air Quality, Atmosphere & Health 11, 311–324. https://doi.org/10.1007/s11869-017-0544-2

LIU, Y, DONG, J & ZHAI, G (2022). Association between air pollution and hospital admissions for hypertension in Lanzhou, China. Environ Sci Pollut Res 29, 11976–11989, https://doi.org/10.1007/s11356-021-16577-5

LIU L, ZHANG X, ZHONG J, WANG J, YANG Y (2019). The 'two-way feedback mechanism' between unfavorable meteorological conditions and cumulative PM2.5 mass existing in polluted areas south of Beijing. Atmospheric environment, 208, 1-9, https://doi.org/10.1016/j.atmosenv.2019.02.050

LIU X, ZHAO C, SHEN X., et al. (2022) Spatiotemporal variations and sources of PM2.5 in the Central Plains Urban Agglomeration, China. Air Qual Atmos Health 15, 1507–1521 (2022). https://doi.org/10.1007/s11869-022-01178-z

LIU, Y., ZHOU, Y., & LU, J. (2020). Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. Sci Rep 10(1):14518. https://doi.org/10.1038/s41598-020-71338-7

LIU, Y., ZHAO, H., MA, Y., YANG, H., WANG, Y., WANG, H., ZHANG, Y., ZOU, X., WANG, H., WEN, R., WEN, Z., & QUAN, W. (2021). Characteristics of particulate matter and meteorological conditions of a typical air-pollution episode in Shenyang, northeastern China, in winter 2017. Atmospheric Pollution Research, Volume 12, Issue 1. https://doi.org/10.1016/j.apr.2020.09.007.

MACHIN, A. B. & NASCIMENTO, L. F. C. (2018). 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. https://doi.org/10.1590/0102-311X00006617

MANTOVANI K C C, NASCIMENTO L F C, MOREIRA D S, VIEIRA L C P F S, VARGAS N P (2016). Poluentes do ar e internações devido a doenças cardiovasculares em São José do Rio Preto, Brasil. Ciência & Saúde Coletiva [online]., v. 21, n.2 https://doi.org/10.1590/1413-81232015212.16102014

MARTINS, E.H., NOGAROTTO, D. C., MORTATTI, J., & POZZA, S. A. (2019). Chemical composition of rainwater in an urban area of the southeast of Brazil, Atmospheric Pollution Research, 10, 2, 520-530. https://doi.org/10.1016/j.apr.2018.10.003

MATOS E P, REISEN V A, SERPA F S, PREZOTTI FILHO P R, LEITE M F S (2019). Análise espaço-temporal do efeito da poluição do ar na saúde de crianças. Cadernos de Saúde Pública, v. 35, n. 10 https://doi.org/10.1590/0102-311X00145418

MAUGIS, C., CELEUX, G., & MARTIN-MAGNIETTE, M. (2009). Variable selection for clustering with Gaussian mixture models. Biometrics 65 (3), 701-709. 10.1111/j.1541-0420.2008.01160.x

MENDES, E V (2018). Entrevista: A abordagem das condições crônicas pelo Sistema Único de Saúde. Ciência & Saúde Coletiva, 23, 431-436. https://doi.org/10.1590/1413-81232018232.16152017

MIRAGLIA SGK, GOUVEIA N. (2014). Custos da poluição atmosférica nas regiões metropolitanas brasileiras. Cien. Saude Colet 2014; 19(10):4141-4147. https://doi.org/10.1590/1413-812320141910.09232014

MIRANDA A C, SANTANA J C C, YAMAMURA C L K, (2021). Application of neural network to simulate the behavior of hospitalizations and their costs under the effects of various polluting gases in the city of São Paulo. Air Qual Atmos Health 14, 2091–2099. https://doi.org/10.1007/s11869-021-01077-9

MITSA, T. (2010). Temporal data mining. New York: Chapman and Hall, 46-48. https://doi.org/10.1201/9781420089776

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

MORAES, S. L., ALMENDRA, R., SANTANA, P., & GALVANI, E. (2019). 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, 35, 7, 1-16. https://doi.org/10.1590/0102-311x00101418

MOSKOVITCH R, SHAHAR Y (2009). Medical temporal-knowledge discovery via temporal abstraction. AMIA Annu Symp Proc. 2009; 2009: 452-456.

MOURA M N, VITORINO M I, SILVA G G C, ANDRADE V S (2021). Relationship between respiratory diseases and environmental conditions: a time-series analysis in Eastern Amazon. Brazilian Journal of Environmental Sciences (Online), 56(3), 398–412. https://doi.org/10.5327/Z217694781020

MUELLER, A. (1995). Fast sequential and parallel algorithms for association rule mining: a comparison. (2nd edition), Technical report, Faculty of the Graduate School of The University of Maryland.

NABIZADEH, R., YOUSEFIAN, F., MOGHADAM, V.K. et al. (2019). Characteristics of cohort studies of long-term exposure to PM2.5: a systematic review. Environ Sci Pollut Res 26, 30755–30771. https://doi.org/10.1007/s11356-019-06382-6

NADALI A, LEILI M, KARAMI M, ABDOLRAHMAN B, ABBAS A (2022). The short-term association between air pollution and asthma hospitalization: a time-series analysis. Air Qual Atmos Health 15, 1153–1167 https://doi.org/10.1007/s11869-021-01111-w

NARDOCCI, A. C., FREITAS, C. U., LEON, A. C. M. P., JUNGER, W. L., & GOUVEIA, N. D. C. (2013). Poluição do ar e doenças respiratórias e cardiovasculares: Estudo de séries temporais em Cubatão, São Paulo, Brasil. Cadernos de Saúde Pública, 29(9), 1867–1876. https://doi.org/10.1590/0102-311X00150012

NASCIMENTO A P, SANTOS J M, MILL J G, ALBUQUERQUE T T A, REIS Jr N C, REISEN V A, PAGEL E C (2020). Association between the incidence of acute respiratory diseases in children and ambient concentrations of SO2, PM10 and chemical elements in fine particles, Environmental Research, Volume 188 https://doi.org/10.1016/j.envres. 2020.109619.

NGUYEN D, LUO W, PHUNG D, VENKATESH (2018). S. Knowledge-Base d Systems LTARM: A novel temporal association rule mining method to understand toxicities in a routine cancer treatment. Knowledge-Based Systems, 161, 313–328. https://doi.org/10.1016/j.knosys.2018.07.031

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

NODARI, A. S. & SALDANHA, C. B. (2016). Episódios críticos de Poluição Atmosférica no município de Porto Alegre/RS. In: INTERNATIONAL SYMPOSIUM ON ENVIRONMENTAL QUALITY, 10. Available at: http://www.abes-rs.uni5.net/centraldeeventos/_arqTrabalhos/trab_201 609101137020 00000650.pdf. Accessed on: Feb. 20, 2019.

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

NOVIKOV, A. (2019). PyClustering: Data Mining Library. Journal of Open Source Software, 4 (36), 1230. http://dx.doi.org/10.21105/joss.01230

PAHO (2018). ORGANIZAÇÃO PAN-AMERICANA DE SAÚDE. Não polua o meu futuro! O impacto do ambiente na saúde das crianças. Available at: https://iris.paho.org/handle/10665.2/49123. Accessed on: Apr. 08, 2019.

PARAJULI R P, SHIN H H, MAQUILING A, SMITH-DOIRON M (2021). Multi-pollutant urban study on acute respiratory hospitalization and mortality attributable to ambient air pollution in Canada for 2001–2012, Atmospheric Pollution Research, Volume 12. https://doi.org/10.1016/j.apr.2021.101234

PAYUS C, SULAIMAN N, SHAHANI M, BAKAR A (2013). Association rules of data mining application for respiratory illness by air pollution database. International Journal of Basic & Applied Sciences, 13, 3, 11–16.

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. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12, 85, 2825 - 2830. Available at: http://www.jmlr.org/papers/v12/pedregosa11a.html. Accessed on: Mar. 05, 2020.

PEREIRA M C, SANTOS R C, ALVIM-FERRAZ M C M (2007). Air Quality Improvements Using European Environment Policies: A Case Study of SO2 in a Coastal Region in Portugal, Journal of Toxicology and Environmental Health, Part A, 70:3-4, 347-351, DOI: 10.1080/15287390600884990

PINTO, W.P., REISEN, V. A., & MONTE, E. Z. (2018). Previsão da concentração de material particulado inalável, na Região da Grande Vitória, ES, Brasil, utilizando o modelo SARIMAX. Engenharia Sanitária e Ambiental, 23, 2, 307-318. http://dx.doi.org/10.1590/S1413-41522018168758

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

POLEZER G, POTGIETER-VERMAAK S, OLIVEIRA A., et al. (2022). The new WHO air quality guidelines for PM2.5: predicament for small/medium cities. Environ Geochem Health. https://doi.org/10.1007/s10653-022-01307-8

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. (2018). Assessing the impact of PM2.5 on respiratory disease using artificial neural networks. Environmental Pollution, 235, 394-403. https://doi.org/10.1016/j.envpol.2017.12.111

PRANATA, R, VANIA, R, TONDAS, AE, SETIANTO, B, SANTOSO, A. (2020). A time-to-event analysis on air pollutants with the risk of cardiovascular disease and mortality: A systematic review and meta-analysis of 84 cohort studies. J Evid Based Med. 2020; 13: 102–115. https://doi.org/10.1111/jebm.12380

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

QUALAR (2021). Qualidade do Ar. Dados de poluentes. CETESB. Available at: https://cetesb.sp.gov.br/ar/qualar. Accessed on: Mai. 08, 2021.

RAJAK R, CHATTOPADHYAY A (2020). Short- and Long-Term Exposure to Ambient Air Pollution and Impact on Health in India: A Systematic Review, International Journal of Environmental Health Research, 30:6, 593-617, DOI: 10.1080/09603123.2019.1612042

RAJ S, PRASAD M V N K, BALAKRISHNAN R (2022). Spatio-temporal association rule-based deep annotation-free clustering (STAR-DAC) for unsupervised person re-identification, Pattern Recognit., 122, p. 108287, DOI: 10.1016/j.patcog.2021.108287s

RASCHKA, S. (2018). Providing machine learning and data science utilities and extensions to Python's scientific computing stack. The Journal of Open Source Software, 3 (24). DOI:10.21105/joss.00638

REBACK, J., MCKINNEY, W., VAN DEN BOSSCHE, J., AUGSPURGER, T., CLOUD, P., HAWKINS, S., ... & SEABOLD, S. (2020). Pandas-dev/pandas: Pandas 1.2.0, Zenodo. https://doi.org/10.5281/zenodo.3509134.

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

REQUIA, W. J., VICEDO-CABRERA, A. M., AMINI, H., da Silva, G. L., SCHWARTZ, J. D., & KOUTRAKIS, P. (2023). Short-term air pollution exposure and hospital admissions for cardiorespiratory diseases in Brazil: A nationwide time-series study between 2008 and 2018. Environmental esearch, 217, 114794.

REPRESA, N.S., FERNÁNDEZ-SARRÍA, A., PORTA, A., & PALOMAR-VÁZQUEZ, J. (2019). Data Mining Paradigm in the Study of Air Quality. Environmental Processes, 2019. https://doi.org/10.1007/s40710-019-00407-5

ROESSLER, L. H. (2016). Fundação Estadual de Proteção Ambiental (FEPAM). Available at: http://www.fepam.rs.gov.br/biblioteca/geo/bases_geo.asp. Accessed on: Apr. 28, 2020.

RYBARCZYK, Y, ZALAKEVICIUTE, R (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. Applied Sciences, 8(12). https://doi.org/10.3390/app8122570

SADAT, Y. K.; KARIMIPOUR, F.; SADAT, A. K. (2014). 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. (2011). 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. https://doi.org/10.1016/j.atmosenv.2011.02.001

SANTANA, E., BORGES, K. C., FERREIRA, A. L., & ZAMBONI, A. (2021). Padrões de qualidade do ar: uma experiência comparada Brasil, EUA e União Europeia. São Paulo: Instituto de Energia e Meio Ambiente. Available at: <https://iema-site-staging.s3.amazonaws.com/padroes-final01.pdf>. Accessed on: Dec. 15, 2021.

SANTOS, F. S., PINTO, J. A., MACIEL, F. M., HORTA, F. S., ALBUQUERQUE, T. T. A., ANDRADE, M. F. (2019). Avaliação da influência das condições meteorológicas na

concentração de material particulado fino (MP2,5) em Belo Horizonte, MG. Engenharia Sanitária e Ambiental, 24, 2, 371-381. https://doi.org/10.1590/s1413-41522019174045

SANTOS, T. C., CARVALHO, V. S. B, & REBOITA, M. S. (2016). 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, 21, 2, 307-313, https://doi.org/10.1590/s1413-41522016139269

SANTOS, U P, ARBEX, M A, BRAGA, A L F, MIZUTANI, R F, CANÇADO, J E D, TERRA-FILHO, M, & CHATKIN, J M. (2021). Environmental Air Pollution: respiratory effects. Jornal Brasileiro De Pneumologia, 47(1), e20200267. https://doi.org/10.36416/1806-3756/e20200267

SÃO PAULO (2013). Decreto nº 59.113. Estabelece novos padrões de qualidade do ar e dá providências correlatas. Com retificações posteriores. São Paulo. Available at: https://www.al.sp.gov.br/repositorio/legislacao/decreto/2013/decreto-59113-23.04.2013.html. Accessed on: Dec. 02, 2019.

SARRA, S. R. & MÜLFARTH, R. C. K. (2021). The impacts of new Coronavirus Epidemic on the levels of pollutants in the city of São Paulo. Brazilian Journal of Development, 7(4), 40415–40438, 2021. https://doi.org/10.34117/bjdv7n4-482

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

SILVEIRA, C. R., CECATTO, J. R., SANTOS, M. T. P., & RIBEIRO, M. X. (2018). Thematic Spatiotemporal Association Rules to Track the Evolving of Visual Features and Their Meaning in Satellite Image Time Series. In Information Technology-New Generations: 15th International Conference on Information Technology (pp. 317-323). Springer International Publishing.

SHAHEEN, M., SHAHBAZ, M., GUERGACHI, A. (2013). Context Based Positive and Negative Spatio-Temporal Association Rule Mining, Knowledge-Based Systems, volume 37, pages 261-273, ISSN 0950-7051. https://doi.org/10.1016/j.knosys.2012.08.010.

SHRESTHA, S.L. (2022). Quantifying effects of meteorological parameters on air pollution in Kathmandu valley through regression models. Environ Monit Assess 194, 684. https://doi.org/10.1007/s10661-022-10347-7

SHUDAN L, YI Z, RUNMEI M, XIAOFEI L, JINGYUAN L, HONGBO L, PENG S, JINGYI Z, PING L, XUN T, TIANTIAN L, PEI G (2022). Long-term exposure to ozone and cardiovascular mortality in a large Chinese cohort, Environment International, Volume 165. https://doi.org/10.1016/j.envint.2022.107280

SOBRINHO, O.M., RUDKE, A.P., MORAIS, M.V.B., MARTINS L.D. (2023). Efeitos meteorológicos da infraestrutura verde em uma cidade latino-americana de porte médio em desenvolvimento: uma avaliação de modelagem numérica. Sustentabilidade. 2023; 15(2):1429. https://doi.org/10.3390/su15021429

SOMPORNRATTANAPHAN, M., THONGNGARM, T., RATANAWATKUL, P., WONGSA, C., & SWIGRIS, J. J. (2020). The contribution of particulate matter to respiratory

allergy. Asian Pacific Journal of Allergy and Immunology, 38(1), 19 - 28. https://doi.org/10.12932/AP-100619-0579

SOUZA, F. T. & RABELO, W. S. (2016). 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… 1045-1050. https://doi.org/10.1109/ICNC.2015.7378136

SOUZA, W. J. V., SCUR, G., & HILSDORF, W. C. (2018). Eco-innovation practices in the brazilian ceramic tile industry: The case of the Santa Gertrudes and Criciúma clusters, Journal of Cleaner Production, Volume 199. https://doi.org/10.1016/j.jclepro.2018.06.098.

SQUIZZATO, R, NOGUEIRA, T, MARTINS, LD et al. Além das megacidades: rastreando a poluição do ar em áreas urbanas e a queima de biomassa no Brasil. Clim. Atmos. Sci. 4, 17 (2021). https://doi.org/10.1038/s41612-021-00173-y

STAEHLE, C., MAYER, M., & KIRCHSTEIGER, B. (2022). Quantifying changes in ambient NOx, O3 and PM10 concentrations in Austria during the COVID-19 related lockdown in spring 2020. Air Qual Atmos Health. https://doi.org/10.1007/s11869-022-01232-w

TADANO Y S, BACALHAU E T, CASACIO L, PUCHTA E, PEREIRA T S, ALVES T A, UGAYA C M L, SIQUEIRA H V (2021). Unorganized Machines to Estimate the Number of Hospital Admissions Due to Respiratory Diseases Caused by PM10 Concentration. Atmosphere, 12, 1345. https://doi.org/10.3390/atmos12101345

THANGJAI, W., NIWITPONG, S. A., & NIWITPONG, S. (2021). Bayesian confidence interval for ratio of the coefficients of variation of normal distributions: A practical approach in civil engineering. Civil Engineering Journal, 7, 135–147. https://doi.org/10.28991/cej-2021-03091651

WANG, C., CORBETT, J. J. (2007). The costs and benefits of reducing SO2 emissions from ships in the US West Coastal waters, Transportation Research Part D: Transport and Environment, Volume 12, Issue 8. https://doi.org/10.1016/j.trd.2007.08.003.

WANG, L., MENG, J., XU, P., & PENG, K. (2018). Mining temporal association rules with frequent itemsets tree. Applied Soft Computing Journal, 62, 817–829. https://doi.org/10.1016/j.asoc.2017.09.013

WANG, X., SMITH, K., HYNDMAN, R. (2006). Characteristic-Based Clustering for Time Series Data. Data Mining and Knowledge Discovery, 13, 335-364. https://doi.org/10.1007/s10618-005-0039-x

WINARKO, E. & RODDICK, J. F. (2007). ARMADA - An algorithm for discovering richer relative temporal association rules from interval-based data, Data & Knowledge Engineering, 63, 1, 76-90. https://doi.org/10.1016/j.datak.2006.10.009

WHO (2019). World Health Organization. Nine out of ten people worldwide breathe polluted air, but more countries are taking action. Available at: 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: Mai. 08, 2019.

WHO (2021). World Health Organization. Global Air Quality Guidelines. Available at: https://apps.who.int/iris/bitstream/handle/10665/345334/9789240034433-eng.pdf Accessed on: Jul. 18, 2022.

WHO (2021). World Health Organization.Global Air Quality Guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide, and carbon monoxide. Available at: https://apps.who.int/iris/bitstream/handle/10665/345329/9789240034228-eng.pdf. Accessed on: Jul. 11, 2022.

WRI Brasil (2021). O Estado da Qualidade do Ar no Brasil. Available at: https://www.wribrasil.org.br/sites/default/files/wri-o-estado-da-_qualidade-do-ar-no-brasil.pdf Accessed on: Jul. 10, 2022.

XIAO, C., CHANG, M., GUO, P., YUAN, M., XU, C., SONG, X., XIONG, X., LI, Y., & LI, Z. (2020). 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, 11, 1, 11-26. https://doi.org/10.1016/j.apr.2019.08.008

YAO, X., GE, B., YANG, W., LI, J., XU, D., WANG, W., & WANG, Z. (2020). Affinity zone identification approach for joint control of PM2.5 pollution over China. Environmental Pollution, 265. https://doi.org/10.1016/j.envpol.2020.115086

YANAGI, Y.; ASSUNÇÃO, J. V.; BARROZO, L. V. (2012). The impact of atmospheric particulate matter on cancer incidence and mortality in the city of São Paulo, Brazil. Cadernos de Saúde Pública, 28, 9, 1737-1748. https://doi.org/10.1590/S0102-311X2012000900012

YIN P, HE G, FAN M, CHIU KY, FAN M, LIU C, XUE A, LIU T, PAN Y, MU Q, ZHOU M (2017). Poluição atmosférica por partículas e mortalidade em 38 das maiores cidades da China: análise de séries temporais. BMJ (Clinical research ed) 356: j667–j667. https://doi.org/10.1136/bmj.j667

YU, P., XU, R., LI, S., COELHO, M.S.Z.S, MALCOLM, P., ABRAMSON, M. J., & GUO, Y. (2022). Associations between long-term exposure to PM2.5 and site-specific cancer mortality: A nationwide study in Brazil between 2010 and 2018, Environmental Pollution, Volume 302. https://doi.org/10.1016/j.envpol.2022.119070

ZHAO, S., YU, Y., YIN, D., HE, J., LIU, N., QU, J., & XIAO, J. (2016). 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. Environment International, 86, 92–106. https://doi.org/10.1016/j.envint.2015.11.003

ZHU Y, PENG L, LI H, PAN J, KAN H, WANG W (2022) Temporal variations of short-term associations between PM10 and NO2 concentrations and emergency department visits in Shanghai, China 2008–2019. Ecotoxicology and Environmental Safety, 229, Article 113087. https://doi.org/10.1016/j.ecoenv.2021.113087

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