Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
The relationship among air pollution, meteorological factors and COVID-19 in the Brussels Capital Region

Timo Mathys a, Fábio Teodoro de Souza a,b, Demian da Silveira Barcellos b, Ingrid Molderez a,⁎

a Centre for Economics and Corporate Sustainability (CEDON), KU Leuven, Warmoesberg 26, Brussels, Belgium
b Graduate Program in Urban Management (PPGTU), Pontifical Catholic University of Paraná (PUCPR), 1155 Imaculada Conceição St, Curitiba, Parana, Brazil

HIGHLIGHTS

• Mechanisms of how air pollutants and weather can affect COVID-19 are discussed.
• Air quality and weather conditions influence COVID-19 transmission and infection.
• Meteorological and air quality parameters can act as predictors for COVID-19.
• Data mining can provide helpful insights for governmental decision-making in health.
• The ecological fallacy needs to be considered in environmental studies about health.

ABSTRACT

In great metropoles, there is a need for a better understanding of the spread of COVID-19 in an outdoor context with environmental parameters. Many studies on this topic have been carried out worldwide. However, there is conflicting evidence regarding the influence of environmental variables on the transmission, hospitalizations and deaths from COVID-19, even though there are plausible scientific explanations that support this, especially air quality and meteorological factors. Different urban contexts, methodological approaches and even the limitations of ecological studies are some possible explanations for this issue. That is why methodological experimentations in different regions of the world are important so that scientific knowledge can advance in this aspect. This research analyses the relationship between air pollution, meteorological factors and COVID-19 in the Brussels Capital Region. We use a data mining approach that is capable of extracting patterns in large databases with diverse taxonomies. Data on air pollution, meteorological, and epidemiological variables were processed in time series for the multivariate analysis and the classification based on association. The environmental variables associated with COVID-19-related deaths, cases and hospitalization were PM2.5, O3, NO2, black carbon, radiation, air pressure, wind speed, dew point, temperature and precipitation. These environmental variables combined with epidemiological factors were able to predict intervals of hospitalization, cases and deaths from COVID-19. These findings confirm the influence of meteorological and air quality variables in the Brussels region on deaths and cases of COVID-19 and can guide public policies and provide useful insights for high-level governmental decision-making concerning COVID-19. However, it is necessary to consider intrinsic elements of this study that may have influenced our results, such as the use of air quality aggregated data, ecological fallacy, focus on acute effects in the time-series study, the underreporting of COVID-19, and the lack of behavioral factors.

⁎ Corresponding author.
E-mail addresses: mathystimov@outlook.com (T. Mathys), fabio.teodoro@pucpr.br (F.T. Souza), demian.barcellos@gmail.com (D.S. Barcellos), ingrid.molderez@kuleuven.be (I. Molderez).

http://dx.doi.org/10.1016/j.scitotenv.2022.158933
Received 8 June 2022; Received in revised form 6 September 2022; Accepted 18 September 2022
Available online 28 September 2022
0048-9697/© 2022 Elsevier B.V. All rights reserved.
1. Introduction

Air quality is an intensively analyzed topic in scientific research. It has impacts on human health, the environment, and ecosystems. These impacts are mainly observed in big cities, where one finds the primary emission sources of air pollution and higher population densities. Nitrogen dioxide ($NO_2$), carbon dioxide ($CO_2$), and particulate matter (PM) are emitted through combustion, and the most significant emission sources of these pollutants are urban motorized traffic, household emissions, and industrial processes (Wark and Warner, 1976). These pollutants are presumed to contribute to the COVID-19 pandemic, which has brought numerous challenges to the current human population worldwide (Li et al., 2020; Curtis, 2021; Val et al., 2021; Zorän et al., 2022). The people of big cities are especially vulnerable to the risk of getting infected by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for the coronavirus disease 19 (COVID-19) (WHO, 2021). This is primarily due to the higher population density, which facilitates direct transmission by inhaling droplets containing SARS-CoV-2 (Ali et al., 2021; Ortiz and Arkin, 2020).

But research conducted in China, Italy, England, the United States, etc. also suspects that air-borne transmission of COVID-19 is a possible pathway for infections of big city inhabitants and that acute and chronic exposure to air pollution could play a significant role in COVID-19 cases, hospital admissions, ICU admissions, case fatality rates, mortality, prevalence and incidence (Zhu et al., 2020; Conticini et al., 2020; Setti et al., 2020; Travaglio et al., 2021; C. Wu et al., 2020; X. Wu et al., 2020).

There are three known ways infectious exposure to respiratory fluids carrying SARS-CoV-2 can lead to a new infection: (1) the inhalation of air carrying very fine droplets and aerosol particles (<9 μm) containing SARS-CoV-2, (2) deposition of large droplets (>20 μm) onto the exposed mucous membrane and (3) touching infected inanimate surfaces followed by getting into contact with the mucous membrane (Centers for Disease Control and Prevention – CDC, 2021). Although the second way has been established as the most common for COVID-19 infections, there is growing evidence concerning the first way of exposure. Droplets are released during the exhalation of an infected individual. The larger droplets settle to the ground after an approximate distance of 1.5 m, but the fine and air-borne droplets, which are also released through exhalation, shrink through evaporation, making them small enough to remain suspended in the air for more extended periods (from 15 min to several hours) (CDC, 2021; Van Doremalen et al., 2020). This phenomenon has been scientifically supported by some studies that have proved the aerosolization of bacteria and viruses and their long-range transport across the atmosphere while being carried by soil dust or organic aggregates in sea spray (Yamaguchi et al., 2012; Chen et al., 2010; Aller et al., 2005; Reche et al., 2018). The research of Van Doremalen et al. (2020) indicates that aerosol and fomite transmission is plausible for SARS-CoV-2. Nevertheless, the risk of infection by air-borne transmission is higher in a closed and poorly ventilated environment than outdoors, and the infection risk decreases over time and with the distance the droplets travel. Aerosol droplets mix with the airstream and become diluted within it. In addition, there is a loss of viral viability and infectiousness of SARS-CoV-2 over time due to environmental factors (e.g., temperature, absolute humidity and UV radiation) (Lim et al., 2021). Other studies found detectable SARS-CoV-2 RNA on PM but concluded the concentrations were too low to represent a major transmission vector for SARS-CoV-2 outdoors (Setti et al., 2020; Chirizzi et al., 2021).

There is conflicting evidence concerning the possible correlation between air pollution and the number of COVID-19 cases. Some studies find significant correlations and others do not, and biases are present in both types of studies. For the particulate matter of <2.5 μm of diameter ($PM_{2.5}$), its increasing concentration has a significant positive correlation with COVID-19 cases (Curtis, 2021; Zhu et al., 2020; Liang et al., 2020; Petroni et al., 2020). This positive correlation has also been established for particulate matter of <10 μm of diameter ($PM_{10}$) (Zhu et al., 2020), $NO_2$ (Zhu et al., 2020; Liang et al., 2020; Ogen, 2020), $O_3$ (Zhu et al., 2020) while for $SO_2$ results indicated a significant negative correlation (Zhu et al., 2020). The correlation with $NO_2$ may be a proxy for urbanicity (Liang et al., 2020). Most of these studies show a positive correlation but are biased due to ecological fallacy and possible confounding factors that were not considered. A review article from Villeneuve and Goldberg (2020) critiques how some studies were executed and to which biases they are susceptible. Other studies, for example, Lim et al. (2021), found no significant correlation with any air pollutant (Lim et al., 2021). Studies also found that chronic exposure to atmospheric contaminants may act as a favorable context for the spread and virulence of SARS-CoV-2 in a population that is more likely to develop respiratory and cardiac conditions (Fattorini and Regoli, 2020). Positive correlations between COVID-19 mortality and long-term exposures to $PM_{2.5}$ (Travaglio et al., 2021; C. Wu et al., 2020; X. Wu et al., 2020), $PM_{10}$ (Travaglio et al., 2021), $NO_2$ and $O_3$ (Travaglio et al., 2021) support these findings.

Some authors suspect $PM_{2.5}$ and $PM_{10}$ particles act as carriers for the virus through inertial impaction, interception, Brownian diffusion, or coagulation. However, Belosi et al. (2021) disproved the three first mechanisms but found for ultrafine particles (<0.1 μm), a not yet negligible but small chance virus-laden aerosols could act as a sink for PM through coagulation. Yet, Belosi et al. (2021) also concluded from their research that the likelihood of air-borne transmission of SARS-CoV-2 is very low in outdoor conditions, excluding crowded public areas. This transmission mechanism could be more relevant for indoor community environments, in which further studies are necessary to investigate the potential risks. Although, earlier studies have shown that aerosols could contain biological materials such as viruses and bacteria (Després et al., 2012) and that the infectivity of viruses could be influenced by their interaction with PM (Groulx et al., 2018). In addition, other studies had results supporting that air pollution influences the spread and contagion capacity of some viruses (Cienczwicki and Jaspers, 2007), such as the avian influenza virus (H5N1) (Chen et al., 2010; Horne et al., 2018), the human respiratory syncytial virus (Horne et al., 2018), or the measles virus (Chen et al., 2017; Peng et al., 2020). Next to Belosi et al., other researchers do support that the contagion and spread of SARS-CoV-2 positively correlate with air pollution. Setti et al. (2020) concluded that outdoors, a cluster could be created between PM and SARS-CoV-2 and due to the reduction of the PM’s diffusion coefficient, the persistence of the virus in the atmosphere enhanced.

By looking at the possible explanations for these statistics in the scope of microbiology, one can find that air pollution is an important contributor to pro-inflammatory responses in human cells leading to an overexpression of inflammatory cytokines and chemokines synthesized by alveolar macrophages (Conticini et al., 2020). This means that the air pollutants irritate a person’s respiratory system and cause lung cells to produce more chemicals (i.e., cytokines and chemokines). This phenomenon triggers the immune system to activate through signals. This process has been chiefly observed with exposure to $PM_{2.5}$ and $PM_{10}$, causing the elevated production of interleukin IL-6 and IL-8 and other cytokines in human bronchial cells (Longhin et al., 2018; Yang et al., 2019; Pope et al., 2016). Becker and Soukup (1999) also found results that suggest alterations of the macrophage-mediated inflammatory responses to viral infection by $PM_{10}$ exposure, which means that air pollution also blocks the signals from infected cells to bring white blood cells that can eliminate the infection. This may result in increased spread of pathogens and more viral pneumonia-related hospital admissions (Becker and Soukup, 1999). Exposure to $PM_{2.5}$ can also lead to Chronic Obstructive Pulmonary Disorder (COPD), infections in the upper respiratory system (i.e., the nose, the nasal cavity and the pharynx) and the lower respiratory system (i.e., larynx, the trachea, the bronchi and the lungs) (Xu et al., 2016). Exposures to other pollutants such as $O_3$, $SO_2$ and $NO_2$ also correlate with higher IL-6, IL-8, IL-17 levels and tumor necrosis factor alpha (TNFα) (i.e., cytokines) in patients (Perret et al., 2017; Kurai et al., 2018; Che et al., 2016; Cho et al., 2007). This affects the first line of defense (i.e., the

---

1. Ecological fallacy in epidemiology, this is the failure in reasoning that arises when an inference is made about an individual based on aggregate data from a group (Villeneuve and Goldberg, 2020).

2. Are considered confounding factors: poverty, quality of healthcare, availability of health insurance, population size, vaccination, governmental policies, etc. (Villeneuve and Goldberg, 2020; Katoto et al., 2021).
cilia). The cilia weaken over time due to chronic exposure to air pollution. This facilitates virus invasion in the lower airways, puts the patient at risk for developing more severe viral infections (Gao et al., 2020), and makes patients more suitable for any infective agent (Conticini et al., 2020).

Curtis (2021) analyzed the co-morbidity of COVID-19 with air pollution through bio-monitoring. Curtis (2021) found interesting results for PM$_{2.5}$, CO, active smoking, incense, pesticides, heavy metals, dust/sand, toxic waste sites, volcanic emissions, and wildfires. These variables were associated with pollutant-related oxidative stress in the airways, inflammation in the lungs and other tissues, and the pollutant-driven alteration of the ACE-2 enzymes in respiratory and other cells. It has also been proven that there is co-morbidity between COVID-19 and underlying health conditions that can be developed through chronic exposure to air pollution. These conditions (i.e., hypertension, COPD, non-TH2 asthma, lung cancer, tumors, heart disease, liver disease, kidney disease, obesity, diabetes, immunodeficiencies such as complications in the early type-1 interferon secretion capacity) exacerbate COVID-19 severity (Domingo and Rovira, 2020; Horne et al., 2018; C. Wu et al., 2020; X. Wu et al., 2020; Brandt and Mersh, 2021; Gao et al., 2020; Leung et al., 2020; Zhu et al., 2020). Anastassopoulou et al. (2020) discovered that components of the immune response to SARS-CoV-2 determine different severities of COVID-19 while genes related to the initial stages of infection, i.e., the binding to the cell receptors and entry, determine different susceptibilities to SARS-CoV-2 (Anastassopoulou et al., 2020). Tsai et al. even hypothesize that prolonged exposure to atmospheric pollution “could induce persistent modifications of the immune system” (Tsai et al., 2019).

Meteorology is considered to directly impact air pollution (Pérez et al., 2020). Air pollutants are removed from the atmosphere through wet deposition (e.g., rainout, washout), dry deposition (e.g., sedimentation, diffusion, turbulence), or chemical reactions (e.g., oxidation) (Beckett et al., 1998). Meteorological factors are also presumed to impact COVID-19 mortality. A study in over 120 Chinese cities found an association between ambient temperature and COVID-19 infections (Xie and Zhu, 2020). In addition, Cai et al. (2021) also provide results suggesting that daily average wind velocity (DAWV) was significantly associated with the SARS outbreak. They also concluded that, to some extent, that daily average air pressure (DAAP), daily average relative humidity (DARH), and daily hours of sunshine (DHS) may have influenced the outbreak as well (Cai et al., 2007). They formulated hypotheses to explain the impacts of DAWV, DARH, and DHS. High wind velocity could facilitate the dilution and removal of the air-borne droplets and thus reduce their suspension time in the air. Lower relative humidity favors the survival of enveloped SARS-CoV on inanimate surfaces and in the air. And daily hours of sunshine are directly related to exposure to ultraviolet light, which destroys virus infectivity (Cai et al., 2007; Biasin et al., 2021). For SARS-CoV-2, Bhaganganar & Bhamireddi performed a high-fidelity numerical simulation of COVID-19 case study in real-time using a HYSDSPLIT (Hybrid Single-Particle Lagrangian Integrated Trajectory) model and found in their results that air-borne transmission of SARS-CoV-2 is possible and that the transmission and dilution of the suspended droplets are determined by atmospheric stability regimes (i.e., low wind speed, low level of turbulence, cool moist ground conditions). These factors favor the transmission enabling the virus to survive up to 30 min and travel up to 2 km before diluting (Bhaganganar and Bhamireddy, 2020). Also, several studies found out that COVID-19 cases increase in the winter in combination with dry conditions and hypothesize this is due to the effect of the environment on viral stability, transmission, behavior change of people from outdoor activities to indoor activities, and changes in immunity levels (Mariyama et al., 2020; Brasil et al., 2020). In addition, Barcellos et al. (2021) found predictions of higher COVID-19 mortality with low solar radiation in Manaus and drought periods in São Paulo using the classification based on association rules method.

Therefore, the scientific literature has shown paradoxical results regarding the influence of environmental variables on COVID-19 cases and deaths, and varied global environmental contexts where research has been carried out play an essential role in these contrasts. Different methodologies to identify associative patterns between environmental and health and even the limitations of ecological studies also can result in these distinctions. Thereby it is essential to test different methodological approaches in different regions. This paper analyzed the associations among concentrations of air pollutants, meteorological variables and the number of COVID-19 cases, hospital admissions and mortality in the Brussels Capital Region (BCR). We use the data mining approach, which can extract valuable knowledge from extensive databases with many different taxonomies. Many studies have already used this approach to explain several phenomena of great complexity (e.g., Barcellos and Souza, 2022; Silva et al., 2022; Duarte et al., 2016). In this research, analyses were conducted to respond to the following research question: Meteorological variables and the concentrations of PM$_{2.5}$, PM$_{10}$, O$_3$, NO$_2$ and black carbon observed in the municipalities of the BCR are associated with the number of confirmed cases and deaths for COVID-19?

2. Data & methods

In order to answer the research question in this article, two methods were used. The first method used was a multivariate analysis on Statistica software (version 14.0). In the multivariate analysis we performed a joining (tree-clustering) algorithm and k-means cluster analysis. The second research is a predictive model called Classification-based association (CBA) with the CBA DMII software (version 2.0) from the National University of Singapore. The CBA generates a set of rules from a dataset with several input variables (e.g., meteorological, air quality, confounders and vaccination) and an output variable. This method required the discretization of the data. These two methods used confirmed cases, daily new hospitalizations and deaths from COVID-19 as outcome variables with seven lag-days for each outcome variable.

2.1. Data acquisition

The data that was collected can be classified in four categories: (1) air quality data, (2) meteorological data, (3) epidemiological data and (4) data of the strictness of the preventive measures taken by the Belgian government in the fight against the COVID-19 pandemic and. Each of these data categories comes from a Belgian institute that provided the data in the form of excel-tables by mail or on internet. In the following paragraphs the sources for the data acquisition are discussed. In Appendix A a table can be found with an overview of the metric units and the abbreviation used for each variable from this point onward.

The data for the air quality variables comes from the Belgian Interregional Environment Agency (IRCELine). They provided the interpolated daily mean concentration of PM$_{2.5}$, PM$_{10}$, O$_3$, NO$_2$ and black carbon for every municipality of the BCR in the period between 25/02/2020 and 01/10/2021. Since the data is not measured with individual monitoring stations for every BCR municipality, they used the 4 × 4 km$^2$ RIO calculation method to provide the interpolated concentrations based on an area-weighted average. That method calculates the area of a municipality that overlaps with every RIO 4 × 4 km$^2$ grid. The data is then collected by calculating the gridcell concentration of every overlapping area (F. Fierens (VMM employee), personal communication, 09/11/2021).

The data for the meteorological variables were obtained from the Belgian Royal Institute for Meteorology (KMI). They provided for the period between 25/02/2020 and 01/10/2021 and for every municipality of the BCR the data for the following variables: air pressure, average temperature, maximum temperature, minimum temperature, relative humidity, absolute humidity, dew point, precipitation, wind speed, global radiation, direct horizontal radiation, sun duration and evapotranspiration (E. Delhaye (KMI employee), personal communication, 16/11/2021).

The data for the epidemiological variables comes from the website of Sciensano (https://epistat.wiv-isp.be/covid/). From the open data available on their website, the following data was retrieved: Confirmed COVID-19 cases, COVID-19 new hospitalizations, COVID-19 deaths and vaccinations against COVID-19 (shots 1, 2 & 3) for every BCR municipality in the period between 25/02/2020 and 01/10/2021.

The data for the strictness of the corona measures was given through a timeline of all the measures in the period from 25/02/2020 to 01/10/2021.
(https://www.hln.be/binnenland/overzicht-van-de-lockdown-op-18-maart-tot-de-recentste-coronamaatregelen-dit-is-het-parcours-dat-belgie-tot-nu-heeft-doorlopen—a470e8f/). Periods of lockdown and periods with more than two strict measures out of lockdown (e.g., curfew, limitation of two guests at home, closing of HORECA, etc.) were categorized as “strict”. Periods with only mild measures (e.g., limits on amount of people in stores, limitation of five guests at home, table limitations on terraces) were categorized as “soft”. Everything in-between was categorized as “mild”.

2.2. Data preparation

For the multivariate analysis, a dataset was created with 32 input variables which are listed in Appendix A in and 571 cases (e.g., every day between 09/03/2020 and 30/09/2021). All the variables in this dataset are related to the entire BCR, the output variables are confirmed cases, new hospitalizations and deaths from COVID-19 in the BCR. Seven lag days were added to each of these output variables to allow prediction generation. Also, for the input variables of precipitation, global radiation and direct horizontal radiation lag days were calculated. The sum of 3, 5, 7, 10 and 14 days for precipitation and the average of 7, 10 and 14 days for global and direct horizontal radiation. This extra action was executed to compare the results of this research with the research of Barcellos et al. (2021) where the variables of precipitation and radiation showed a significant relationship on COVID-19 hospitalization and mortality in the cities of Rio de Janeiro, Sao Paulo and Manaus (Barcellos et al., 2021).

For the CBA, the same data was used as in the dataset for the multivariate analysis, but a dataset had to be created for each output variable and for each lag period. Also, datasets for one and two lag weeks were created. There were 21 datasets with daily data (e.g., one dataset for each output variable: confirmed cases [t1-t7], new in hospital [t1-t7], deaths [t1-t7]) which each had the same 32 input variables as in the multivariate analysis, one output variable and 564 cases (e.g., from 09/03/2020 until 23/09/2021). There were also 6 datasets with weekly data (e.g., one dataset for each output variable: confirmed cases [w1-w2], new in hospital [w1-w2], deaths [w1-w2]). These datasets each had 30 input variables which are listed in Appendix A, one output variable and 82 cases [2020 week 9–2021 week 40]. The CBA DM II also required the discretization of the data and the elimination of decimals. Discretization is meant that continuous data was transformed into three categories, one for each tercile (high, medium and low). In discretization, some variables were multiplied by 1000. These variables were pm2_5, pm10, o3, no2, bc, global_rad (7d, 10d, 14d), direct_horiz_rad (7d, 10d, 14d) and sun_duration.

2.3. Data modeling

This article comprises two different analyses. First, a multivariate analysis was performed in order to observe which environmental variables could play an influential role in the variance of the epidemiological variables. And secondly, a classification based on association rules was performed in order to extract predictions on the output variables (i.e., epidemiological variables) with a given set of environmental input variables (e.g., air pollution and meteorological variables).

In the multivariate analysis, two analyses were performed. First, a cluster analysis was performed on the database, with dendrogram (joining tree-clustering) and k-means cluster analysis. The dendrogram in Fig. 1 gives a first overview of the variables sorted in different clusters based on their Euclidian distance. This analysis allows the verification of which variables have the same characteristics by classifying the variables in different groups. The formula of the Euclidian distance is represented in Eq. (1). Based on the results of the joining tree-clustering analysis, a k-means cluster analysis was performed with the number of clusters identified in the dendrogram. In this case, the requested number of clusters was set to 10 to see which clusters would separate and which ones would stay intact compared to the dendrogram. Given a set of observations \((x_1, x_2, \ldots, x_n)\) with each observation being a d-dimensional real vector, k-means clustering aims to distribute the \(n\) observations into \(k\) \((\leq n)\) sets \(S = \{S_1, S_2, \ldots, S_k\}\).
in order to minimize the within-cluster sum of squares or variance. The formula for k-means clustering can be found in Eq. (2).

\[
d(p, q) = \sqrt{(p - q)^2}
\]

where, \(d(p, q)\) is the distance between the two points \(p\) and \(q\); and \(\sqrt{(p - q)^2}\) is absolute value of the numerical difference of the coordinates of \(p\) and \(q\).

\[
\text{arg}_k \min \sum_{i \in S_i} |x - \mu_i|^2 = \text{arg}_k \min \sum_{i=1}^d |S_i| \ Var \ S_i
\]

where, \(n\) is the number of observations, \(k\) is the number of sets; \(S_i\) is the set of observations; \(\mu_i\), mean of points in \(S_i\); and \(\text{Var}_S\) is (1/m)th unobserved stochastic error term with mean zero and finite variance.

The second method used was the CBA, this method generates a classifier (a predictive model) based on association rules (patterns of cause-effect in the database). These rules are generated in the configuration where if \(x\) (i.e., the antecedent value) is true, then \(y\) (i.e., the consequent value) is also true. Each rule has a support (\(S\%\)) and a confidence (\(C\%\)), these are statistical parameters that filter the classification rules. The support of a rule \(x \rightarrow y\) in a dataset \(D\) represents the percentage of records, \(x\) and \(y\), which were classified correctly in relation to all records in the dataset. The confidence of a rule \(x \rightarrow y\) in dataset \(D\) represents the percentage of records that the rule or classifier was correct. In other words, a rule \(x \rightarrow y\) only holds in \(D\) with confidence \(C\%\) and support \(S\%\) of the cases in \(D\) that contain \(x\) and are labeled with class \(y\) and if \(5\%\) of the cases in \(D\) contain \(x\) and are labeled with class \(y\) (Liu et al., 1998). The minimal support for this analysis was set to 5 \% and the minimal confidence to 80 \%. The settings used to generate the models in the CBA were With Rule Pruning in Pruning Control3 and Use Small Rule in Rule Usage Control. Each classifier also generates a confusion matrix. This is a table that presents the classification frequencies for each class of the model (true positive, false positive, true negative, false negative). The correct classifiers (true positive and true negative cases) are represented on the main diagonal. The accuracy of a classifier is the sum of the matrix's main diagonal, divided by the total number of values and multiplied by hundred. The error rate is the difference between a hundred and the accuracy (Barcellos et al., 2021). Error rates above 10 \% were excluded from this analysis.

3. Results

In the following section, the results of the two analyses are interpreted.

3.1. Results from multivariate analysis

The first analysis performed was the multivariate analysis with the two cluster analyses (e.g., joining tree clustering and k-means clustering).

The dendrogram from the cluster analysis represented in Fig. 1 indicates that the variables can be assembled in seven clusters based on the Euclidean distances between the variables. The first cluster consisted of air pressure. This observation was also made by Barcellos et al. (2021). The second cluster consisted of the BCR confirmed [t0-7] variables. The third cluster consisted of BCR_sum_precipitation 3d, 5d, 7d, 10d and 14d. The fourth cluster consisted of BCR_precipitation, BCR_rel_hum, BCR_ET0 and BCR_pm2.5. The fifth cluster consisted of all the BCR deaths [t0; t7] and BCR_no2, BCR_pm10, BCR_max_temp, BCR_avg_temp and BCR_pm2.5. This cluster is interesting and requires further analysis because it holds both epidemiological and environmental variables. The last cluster consisted of BCR_abs_hum, BCR_dew_pt, BCR_min_temp, BCR_sum_duration, BCR_cum_days_w/o_rain, BCR_global_rad, BCR_avg_global_rad (7d, 10d, 14d), BCR_direct_horiz_rad, BCR_avg_direct_horiz_rad (7d, 10d, 14d), BCR_wind_speed, strictness measures, BCR_days_w/o_rain and BCR_bc. BCR_avg_global_rad (7d, 10d, 14d), BCR_direct_horiz_rad, BCR_avg_direct_horiz_rad, BCR_avg_temp, BCR_pm2_5, the maximum temperature and the average temperature of BCR could act as potential inputs for the prediction model for deaths from COVID-19 in the BCR. No other cluster showed promising results.

3.2. Results from classification based on association rules

The second analysis that was performed was a Classification based on association rules with the CBA-DMII software. Fig. 2 represents an overview of the predictive model's accuracies, and Appendix B shows the models' support and confidence.

For the confirmed cases of COVID-19, the CBA-DMII generated 896 rules from 9 datasets (e.g. BCR_confirmed_cases [t1; t7] and [W1; W2]). These rules were then filtered based on their support (e.g., minimum 8 \%), confidence (e.g., minimum 90 \%), error rate (e.g., maximum 10 \%) and the coherence between environmental and epidemiological variables in their relationships. Two of the datasets generated interesting rules with the CBA DMII program, these were the BCR_confirmed_W1 and the BCR_confirmed_W2 datasets. Both classifiers had an error rate of 0.00 \%. After filtering, twelve rules were kept. An overview of the interesting predictive models for confirmed cases is represented in Table 2 and in Appendix B. The datasets for the other outcome variables for confirmed cases in the BCR and the other rules from the same sets either had an error rate that was too high, support or confidence that were too low or the rules were uninteresting for this research.

For the new hospitalizations related to COVID-19 in the BCR, the CBA-DMII generated 1103 rules from 9 datasets (e.g. BCR_new_in_hosp [t1; t7] and [W1; W2]). After filtering the rules, eighteen rules from six datasets (e.g. BCR_new_in_hosp_t2, BCR_new_in_hosp_t3, BCR_new_in_hosp_t4, BCR_new_in_hosp_t7, BCR_new_in_hosp_W1 and BCR_new_in_hosp_W2) were kept for interpretation. The same filtering properties were used for the confirmed cases of COVID-19. An overview of the interesting rules for confirmed cases is represented in Table 3 and in Appendix B. The datasets for the other outcome variables for new hospitalizations in the BCR and the other rules from the same sets either had an error rate that was too

### Table 1

| Cluster | Variables |
|---------|-----------|
| Cluster 1 | BCR_pressure |
| Cluster 2 | BCR_new_in_hosp_t1; BCR_new_in_hosp_t2; BCR_new_in_hosp_t3; BCR_new_in_hosp_t4; BCR_new_in_hosp_t5; BCR_new_in_hosp_t6; BCR_new_in_hosp_t7; |
| Cluster 3 | BCR_sum_precipitation, 14d |
| Cluster 4 | BCR_sum_precipitation, 10d |
| Cluster 5 | BCR_pm2_5; BCR_pm10; BCR_no2; BCR_avg_temp; BCR_max_temp; BCR_percentage; BCR_days_w/o_rain; BCR_bc |
| Cluster 6 | BCR_avg_global_rad (7d, 10d, 14d) |
| Cluster 7 | BCR_avg_global_rad (7d, 10d, 14d) |
| Cluster 8 | BCR_avg_global_rad (7d, 10d, 14d) |
| Cluster 9 | BCR_direct_horiz_rad; BCR_avg_direct_horiz_rad, BCR_avg_temp, BCR_pm2_5, the maximum temperature and the average temperature of BCR could act as potential inputs for the prediction model for deaths from COVID-19 in the BCR. No other cluster showed promising results. |
high, support or confidence that was too low or the rules were uninteresting for this research.

For the deaths related to COVID-19 in the BCR, the CBA-DMII generated 1085 rules from 9 datasets (e.g., BCR_deaths [t1; t7] and [W1; W2]). After filtering the rules, fourteen rules from two datasets (e.g., BCR_deaths_W1 and BCR_deaths_W2) were kept for interpretation. The same filtering properties were used for the confirmed cases and new hospitalizations of COVID-19. An overview of the interesting rules for confirmed cases is represented in Table 4 and in Appendix B. The datasets for the other outcome variables for deaths in the BCR and the other rules from the same sets either had an error rate that was too high, support or confidence that was too low or the rules were uninteresting for this research.

4. Discussion

From the multivariate analysis, the following input variables were deemed interesting for further investigation as they showed to have a significant relation with the epidemiological output variables of deaths related to COVID-19.

Table 2
Summary table of significant CBA predictors for confirmed cases of COVID-19.

| Input         | Taxonomies                  | Meteorology        | Air quality | Control | Immunology | Epidemiology        | Output                          |
|---------------|-----------------------------|--------------------|-------------|---------|------------|---------------------|--------------------------------|
| 3rd tercile   | BCR_days_w/o_rain           | BCR_pressure       | BCR_pm2_5   | BCR_confirmed | BCR_confirmed   | Confirmed cases of COVID-19 (tercile 3) |
| 2nd tercile   | BCR_max_temp                | BCR_o3             | BCR_vaccination | BCR_confirmed | BCR_confirmed   |                                    |
| 1st tercile   | BCR_dew_pt                  |                    | Strictness meaures | BCR_confirmed | BCR_confirmed   |                                    |

Table 3
Summary table of significant CBA predictors for new hospitalizations related to COVID-19.

| Input         | Taxonomies                  | Meteorology        | Air quality | Control | Immunology | Epidemiology        | Output                          |
|---------------|-----------------------------|--------------------|-------------|---------|------------|---------------------|--------------------------------|
| 3rd tercile   | BCR_global_rad              |                    | Strictness measures | BCR_new_in_hosp | BCR_confirmed   | New hospitalizations from COVID-19 (tercile 3) |
| 2nd tercile   | BCR_abs_hum                 | BCR_bc             | BCR_vaccination | BCR_confirmed | BCR_confirmed   |                                    |
| 1st tercile   | BCR_avg_direct_horiz_rad_10d| BCR_avg_direct_horiz_rad_14d | BCR_vaccination | BCR_confirmed | BCR_confirmed   |                                    |
COVID-19 in the BCR: the concentration of NO2, PM$_{2.5}$, in the BCR for whom the relation is positive and the average temperature for whom the relation is negative. The relation between COVID-19 mortality and the air pollutants PM$_{2.5}$ and NO$_2$ is also supported by the studies of Conticini et al. (2020), Setti et al. (2020), Travaglio et al. (2021), C. Wu et al. (2020) and X. Wu et al. (2020). As for the average temperature, Moriyama et al. (2020) and Brassey et al. (2020) observed that in the winter in combination with dry conditions there were more cases of COVID-19. A negative relation was also observed in the multivariate analysis between deaths from COVID-19 and average temperature. This means theoretically, in weather conditions with a lower average temperature, an increase in COVID-19 mortality can be expected. Moriyama et al. (2020) hypothesized this trend could be because of the environment on viral stability, transmission and behavioral change of people between seasons.

The classification based on association rules observed several relations between the input and output variables. A high rate of confirmed cases of COVID-19 in the BCR had a rather strong relation with a moderate vaccination rate (i.e., first shot), strict corona measures, a high PM$_{2.5}$ concentration, a moderate NO$_2$ concentration, fewer days without rain per week, low and high precipitations, a moderate sum of precipitation over a period of two weeks, a high average direct horizontal radiation during a period of two weeks, a high direct horizontal radiation, a high average global horizontal radiation over a period of two weeks, moderate average and minimum temperatures and high air pressure. The results for PM$_{2.5}$ confirm their suspected influence on COVID-19 mortality (Travaglio et al., 2021). For the meteorological variables, the results could not be supported by the literature but the impact of these variables on COVID-19 cases and hospitalizations can act as predictors for their association with COVID-19 deaths observed in this analysis.

This research has brought to light some interesting results. The use of data mining techniques can provide useful insights for high-level government decision-making concerning COVID-19 and other epidemic diseases. Barcellos et al. (2021) reached the same conclusion by studying the influence of weather factors and air quality on deaths and cases of COVID-19, by data mining, in three Brazilian metropolises. The greatest strength of this study is the use of these advanced data analysis and modeling techniques. Studies like this and many others (e.g., Barcellos and Souza, 2022; Silva et al., 2022; Duarte et al., 2016) demonstrate the potential of these techniques to support public policies in urban management. However, some limitations need to be considered for the interpretation of the results of this study.

First, for the air quality variables, data is obtained through RIO 4 × 4, this is a calculation method that aggregates data based on measuring points into data for municipalities. This means that the data per municipality cannot be considered 100% valid. However, this method is recognized within IRCEL as a valid and reliable technique for providing federal data.

Second, the time series studies performed in this research (e.g., multivariate analysis and CBA) focus on the acute effects of air pollution on the COVID-19 spread, hospitalization and mortality. This brings forward several biases. For one, the variance of the air pollution variables may have underlying causes that could have a greater influence on the spread and mortality of COVID-19 than air pollution on its own. For example, in periods of lockdown, the air quality in big cities improved a lot while COVID-19 cases, ICU admissions and deaths dropped after a while but in what ratio the reason for the drop the COVID-19 cases, ICU admissions and deaths can be attributed to the decrease in air pollution is still unknown.

Third, the research is victim of ecological fallacy. In epidemiology, this is the failure in reasoning that arises when an inference is made about an individual based on aggregate data from a group (Villeneuve and Goldberg, 2020). In this research, the analyzed data was retrieved and processed in community-level and in BCR-level configurations. This does not necessarily mean that all the individual cases that formed the cases for the output data (epidemiological) were exposed to the same input data parameters (air quality, weather conditions, confounders).

A fourth limitation is the inevitable underreporting of COVID-19 confirmed cases due to the evolution in testing effectiveness and the attribution
of COVID-19-related deaths to other causes. Beyond, some people did not go to the hospital and preferred to take self-medication. This can affect results and also lead to underreporting.

A fifth limitation is the exclusion of behavioral factors from the individual cases within the population. Occupation, people lacking health insurance, underlying health conditions, lifestyle, and viral load that people were exposed to were not taken into account while these factors are known to play an important role in the evolution of the spread and contagion of the coronavirus.

The last limitation concerns the positive associations. Even though these have been established, it is not possible to know in what proportion the associations may be attributed to the possible carrying effect of air pollutants or to the aerosolization of the viruses due to the weather conditions or to morbidity between the variables.

5. Conclusion

The environmental variables associated with COVID-19-related deaths were NO₂, PM2.5, temperature, precipitation, air pressure and radiation. For confirmed cases of COVID-19, the variables with associative patterns were O₃, PM₂.₅, temperature, dew point, radiation, precipitation, air pressure and wind speed. Environmental variables associated with COVID-19 hospitalization were black carbon, radiation, air pressure, wind speed, temperature, and precipitation. Combined, these environmental variables with epidemiological factors, can predict intervals of hospitalization, cases and deaths from COVID-19. These findings confirm the influence of meteorological and air quality variables in the Brussels region on deaths and cases of COVID-19. The data mining approach was able to identify with successfully these environmental variables to predict deaths, new hospitalizations and new cases of COVID-19 in the Brussels region. This approach can guide public policies and provide useful insights for high-level government decision-making concerning COVID-19.

However, this research was prone to limitations such as the use of aggregated data, ecological fallacy, focus on acute effects in a time-series study, underreporting of COVID-19 and exclusion of behavioral factors. Especially the ecological fallacy needs to be more considered in environmental studies about public health.

Important gaps for further research are to analyze the effect of air quality and meteorological variables on the spread of COVID-19 and on the biologic effects of their interaction in patients separately. Also, the inclusion of behavioral factors can bring interesting results.

CRediT authorship contribution statement

Timo Mathys: Formal analysis, Methodology, Investigation, Software, Visualization, Writing – original draft, Writing – review & editing. Fábio Teodoro de Souza: Project administration, Resources, Visualization, Conceptualization, Methodology, Software, Validation, Supervision, Writing – review & editing. Demian da Silveira Barcellos: Resources, Visualization, Conceptualization, Validation, Writing – review & editing. Ingrid Molderez: Project administration, Conceptualization, Validation, Supervision, Writing – review & editing.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.158933.
Fattorini, D., Rogelli, F., 2020. Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. Environ. Pollut. 264, 114732. https://doi.org/10.1016/j.envpol.2020.114732.

Gao, Y., Ding, M., Dong, X., Zhang, J., Kursat Azkur, A., Azkur, D., Gan, H., Sun, Y., Fu, W., Li, W., Li, Y., Cao, Y., Yan, Q., Cao, G., Hou, X., Deng, X., Ren, S., Wang, L., Zou, J., Ding, H., Zhang, X., Song, Y., 2020. Exposure to fine particulate air pollution is associated with endothelial injury and systemic inflammation. Circ. Res. 121 (11), 1204–1214. https://doi.org/10.1161/CIRCRESAHA.116.309279.

Reche, J., D’Orta, G., Mladenov, N., Winget, D.M., Suttle, D.A., 2018. Deposition rates of viruses and bacteria above the atmospheric boundary layer. ISME J. 12 (4), 1154–1162. https://doi.org/10.1038/s41396-017-0044-2.

Setti, L., Passarino, F., De Gennaro, G., Barbieri, P., Perrone, M.G., Borelli, M., Palmasini, J., Di Gilio, A., Torsoli, V., Fontana, F., Clemente, L., Pallavicini, A., Ruscio, M., Pisciotti, P., Miani, A., 2020. SARS-cov-2RNA found on particulate matter of Bergamo in northern Italy first evidence. Environ. Res. 188, 109754. https://doi.org/10.1016/j.envres.2020.109754.

Silva, L.F., Fonseca, M.N., Moura, E., Souza, F.T., 2022. Ecosystems services and green infrastructure for respiratory health protection: a data science approach for Paraná, Brazil. Sustainability 18, 1835. https://doi.org/10.3390/su18068135.

Travaglio, M., Yu, Y., Popovic, R., Selley, L., Leal, N.S., Martins, L.M., 2021. Links between air pollution and COVID-19 in England. Environ. Pollut. 268, 115859. https://doi.org/10.1016/j.envpol.2021.115859.

Tui, D.H., Riediker, M., Bercht, A., Paccaud, F., Waehrer, G., Vollenweider, P., Bochud, M., 2019. Effects of short- and long-term exposures to particulate matter on inflammatory marker levels in the general population. Environ. Sci. Pol. 26 (19), 14697–14704. https://doi.org/10.1016/j.envsci.2019.05.014.

Valli, M., Hassanzadeh, J., Mirahmadizadeh, A., Hoseini, M., Maleki, Z., Ghareh, H., 2021. Effect of meteorological factors and air quality index on the COVID-19 epidemiological characteristics: an ecological study among 210 countries. Environ. Pollut. 281, 117855. https://doi.org/10.1016/j.envpol.2021.117855.

Petroni, M., Hill, D., Younes, L., Barkman, L., Howard, S., Howell, L.B., Mirovsky, J., Collins, M.B., 2020. Hazardous air pollutant exposure as a contributing factor to COVID-19 mortality in the United States. Environ. Res. Lett. 15 (9), 094009. https://doi.org/10.1088/1748-9326/abaf86.