PREDICTION OF HOSPITALIZATIONS CAUSED BY RESPIRATORY DISEASES BY USING DATA MINING TECHNIQUES: SOME APPLICATIONS IN CURITIBA, BRAZIL AND THE METROPOLITAN AREA

MATHEUS BITTENCOURT CARDOSO[1] & FABIO TEODORO DE SOUZA[2]
[1]Civil Engineering, Polytechnic School, Pontifical Catholic University of Paraná (PUCPR), Brazil
[2]Postgraduate Program in Urban Management, Pontifical Catholic University of Paraná (PUCPR), Brazil

ABSTRACT

As seen throughout the years, the industrial and technological evolution led to problems that evolved from the constant growth of the globalized economic world, among many other features. One of them is air pollution, which is co-related to human health and welfare. The city of Curitiba and its metropolitan area comprise around 3.5 million people, which are directly affected by over 1 million vehicles in the capital, and the diversified industries that continue to grow. This project consists of combining hospitalization data and other related urban variables such as climate and air pollution. This paper describes the methodology by using a data mining approach to explore data from 2008 to 2016 from Curitiba and six other cities. The aim of this study is to predict future cases of hospitalizations caused by respiratory diseases in Curitiba and the metropolitan area, and to consider atmospheric and air quality. The air pollution parameters used in this study are: suspended particles matter; PM_{10} (particles smaller or equal to 10 micrometers); sulfur dioxide (SO_2); carbon monoxide (CO); ozone (O_3); nitrogen dioxide (NO_2); and the air quality index (AQI). The climate variables are: wind direction; average wind speed; average maximum wind speed; potential and real evaporation; insolation; average cloud cover; total precipitation and days of precipitation; average atmospheric pressure; minimum, maximum and average temperature; and relative humidity. The results of this scientific project may be used by the government for developing actions and interventions aimed at minimizing the adverse effects on the environment and human health. It is also expected that the results could contribute to the planning of environmental, controlling air pollution and improving public policies.

Keywords: data mining, air pollution, prediction, statistics, respiratory diseases.

1 INTRODUCTION

Levels of air pollution increased significantly with the industrial and technological revolutions, alongside some non-reversible exponential population growth in the following years. With larger and more populous cities, unhappy events, such as the great fog that devastated the city of London in England and resulted in more than 4000 deaths, began to bring to the surface the need for a more elaborate study of the air quality.

Despite such catastrophic events, it was only in 1972 that the United Nations decided to organize the first conference in Stockholm, and countries signed the Kyoto Protocol in 1997, which provided targets for the reduction of pollutants between the years of 2008 and 2012 [1]. A study correlated cardiorespiratory hospitalizations with very similar pollutants studied in this paper [2]. Another study concluded that elderly people tended to be hospitalized when exposed to nitrogen dioxide and $PM_{2.5}$ [3]. In the cities of Chicago, Detroit, Houston, Los Angeles, Milwaukee, New York and Philadelphia, heart failures due to daily variation in carbon monoxide, independent of season or temperature, were associated [4]. This paper describes the relationship between atmospheric patterns and hospitalization due to respiratory diseases in the Curitiba Metropolitan Area (CMA).

WIT Transactions on Ecology and The Environment, Vol 211, © 2017 WIT Press
www.witpress.com, ISSN 1746-448X (on-line)
doi:10.2495/AIR170231

2 BACKGROUND

The Environmental Institute of Paraná (IAP) is responsible for measuring air quality data for the cities of Curitiba, Colombo and Araucária, totaling 13 stations. In this region, due to the current implantation of many of those stations, there are still no studies that correlate air pollutant data with respiratory disease cases to predict future cases of hospitalization.

In some pre-existent studies in this research field, Martins et al. [5] investigated the associations between characteristics from the environment, morbidity and habitual diseases. An association between air pollution in São Paulo with pneumonia and flu care for the retirees was identified. Relationships amid childhood respiratory morbidity and air pollutants in Curitiba were also established [6]. Other studies [7], [8] related prejudicial impacts on health and public transportation. The most susceptible population in the US subject to serious health effects from air pollution was associated with those who live very near major regional transportation routes, especially highways [9]. Another study by Brownson et al. [10] reasoned the employer–employee health procedures to prevent future diseases. In extreme cases, Dockery et al. [11] associated air pollution to lethality in six US cities, suggesting that fine particles may lead to the excess of mortality in those cities. This paper focuses on exploring the morbidity dataset in different cities in CMA.

2.1 Air quality data in the Curitiba Metropolitan Area (CMA)

The data were obtained with hourly or daily frequency. Therefore, the data were transformed into monthly datasets using the arithmetic average, to match the other climatic and morbidity data (monthly).

Total suspended particulates (TSPs) and inhalable particulate matters (PI/PM_{10}) represent solid and liquid materials suspended in the atmosphere, such as different types of dust, pollen, etc. The particle size is the criteria used to classify these materials. Thicker particles can get trapped in the nose and throat, causing discomfort and irritation, and making it more susceptible for flu-like illness to spread over the body. Thinner dust can cause damage to the respiratory tract and carry other pollutants into the lung alveoli, causing dissatisfactory chronic side-effects related to respiration, cardiac issues, and even cancer. Fine particulates were also associated with higher reporting of bronchitis [12]. People who remain in places highly polluted by inhalable particles are more vulnerable to diseases in general.

The emission of sulfur dioxide (SO_2) is related to the use of fossil fuel in both vehicles and industrial facilities. Because it is a highly soluble gas in the mucous membranes of the upper airway, it can cause irritation and increase in mucus production, discomfort in breathing and aggravation of respiratory and cardiovascular problems. Another effect related to SO_2 refers to the fact that it is one of the precursor's pollutants of acid rain, a global effect of atmospheric pollution, and is responsible for the deterioration of materials, acidification of bodies of water and the destruction of forests.

The emission of carbon monoxide (CO) is directly related to the combustion process of gasoline, alcohol or diesel. It is considered a systemic asphyxiant because it is a substance that impairs the oxygenation of tissues. The effects of human exposure to CO are associated with several problems. Since the affinity of hemoglobin with CO is 210 times greater than with oxygen, the carboxyhemoglobin formed in the blood can have serious consequences, such as mental confusion, impaired reflexes, unconsciousness, the stopping of brain functions and, in extreme cases, death.

Ozone (O_3), when formed near the soil, behaves as a toxic pollutant and can cause eye irritation and reduced lung capacity. It aggravates respiratory diseases, decreases resistance against infections and is responsible for pulmonary dysfunctions, such as asthma. The ozone interferes with photosynthesis and causes damage to works of art and metal structures.

Nitrogen dioxide (NO_2) may cause irritation to the mucous membrane, manifested by rhinitis and severe lung damage, similar to those caused by pulmonary emphysema [13].

A selection of the highest values in a day for both O_3 and NO_2 were taken; those 28, 30 or 31 values became the monthly values throughout the arithmetic mean.

Curitiba, whose geographic coordinates are S25°25’ W49°16’, has five stations. Colombo, S25°17’ W49°13’, has one station. The other seven stations are located in Araucária, S25°35’ W49°24’. The locations of the stations are shown in Fig. 1.

2.2 Morbidity and climatological data

Since morbidity data via DataSUS (National Health Service database) is restricted to monthly values from January 2008 to August 2016, it was decided to work on monthly data in the equivalent period. Thus, it is possible to obtain the climatological data of the city of Curitiba with a monthly frequency via INMET (National Institute of Meteorology), including: wind direction, average wind speed (m/s), average maximum wind speed (m/s), evaporation (mm), insolation (h), total precipitation (mm), pressure (mbar), maximum and minimum average and average temperature (Celsius), relative humidity (%), etc.

The wind direction was transformed into a geographic coordinate through the table provided by INMET. According to the technical note from the station networks, all wind

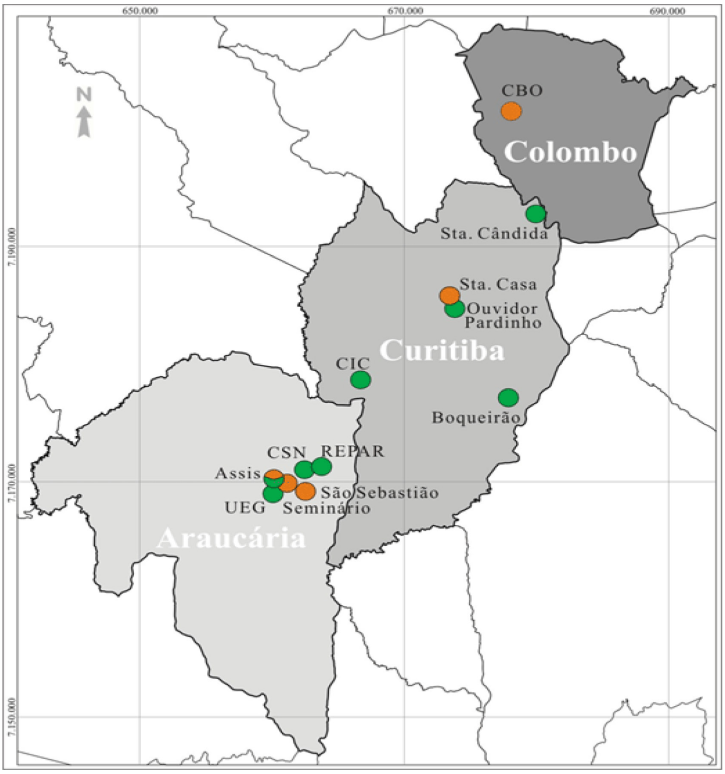


Figure 1: Stations located in Curitiba and nearby cities.

values are measured by Vaisala WT521 transmitters that register wind gusts every 0.25 seconds and calculate averages every 3 seconds for both speed and direction. The stations use these averages of 3 seconds to calculate averages every 10 minutes, thus making a very precise measurement. The pressure is measured every 5 seconds in the Automatic Weather Station. Precipitation is measured every 10 seconds and is cumulative. All measurements previously described are digital.

Analogue measurements contemplate temperature, humidity and radiation. During temperature measurement, an excitation current of 1 mA and voltage measurements are used on a Pt100 element using a 100-ohm resistor. All three are measured every 5 seconds [14].

2.3 Objectives

This research aims to study the relationships between climatological data, air pollution data, total morbidity and respiratory disease morbidity. In this sense, the method proposed focuses on the mathematical and statistical relationships that may predict future hospitalizations due to respiratory problems in the CMA. The major objective, then, is to find relevant patterns and identify the outcome for each municipality.

Therefore, specific objectives include identifying if any chemical components present in the air are more likely to cause hospitalizations; identifying if there is a correlation between climatic and air data also with general hospitalizations; identifying if data from one city influences the hospitalizations of another.

3 METHODS

A data mining project basically consists of three steps: data collection, data preparation and data modelling (multivariate analyses, explainable models and predictive models).

3.1 Data collection and preparation

The work begins with an extensive collection of data from IAP, DATASUS and INMET, which were separated in an inventory table for decision analyses. Those data values are arranged in their most varied form. The air quality information provided by the IAP is the most variable form of data, with different measurement frequency over the years.

According to the National Environment Council Resolution 03/90 (CONAMA), which indicates standards for air quality, the following will be taken into consideration in this study: TSPs, smoke, inhalable particles (PI or PM_{10}), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3) and nitrogen dioxide (NO_2), in addition to the total value of the Air Quality Index. Due to this enormous alternation found between the documents provided by the IAP and the variation in measurement processes over the years, some data needed to be transformed from day–day and hour–hour to monthly records. Also, there are data available from one station that doo not exist on the others due to inconsistency in measuring and storing by those groups. The morbidity data from DATASUS for the selected cities were selected and divided between respiratory morbidity and total morbidity, where it is always possible to draw a parallel and percentages between the two, month to month.

After data collection, it is necessary to prepare them, eliminating constant or sparse variables that do not add up to the work [15]. There are five different ways of dealing with missing data: ignoring them and running the risk of not having enough data to complete a good job or an imputation method. These can be defined in four ways: a) filling-in data manually, which would be extremely extensive in this specific case; b) replacing missing values with a constant, which can greatly distort the study since not all values are equal;

c) use of the average or fashion, which may be satisfactory but does not consider the relationship of the variables; or d) using the most probable value method. In this study, the missing data were not replaced and just ignored.

3.2 Multivariate analyses

Following the data compilation, the Joining Tree Method, K-Means Method, Principal Component Analysis (PCA) and Autocorrelation Matrix were performed. K-Means represents an iterative process that reclassifies objects in similar classes based on the mean values of each class. The Clustering Tree Method groups data by building a dendrogram and considering the computation of Euclidean distances to create the branches [16]. Fig. 2 illustrates an example of a dendrogram which groups the morbidity variables ('M_' in the graph) from different cities (CWB = Curitiba; CLBO = Colombo; PIRQ = Piraquara; CMPLR = Campo Largo), different lags (T0 or actual month; T1 next month; T2 next two months) and a meteorological variable (average pressure in mbar on the graph). It is possible to note a similarity among air pressure and morbidity in the next two months. The abbreviation TOT represents Total Diseases, and DR, Respiratory Disease (see Table 1 for definition of acronyms).

Table 2 illustrates two clusters grouped by the k-Means Method. It is possible to identify some similarity in grouping as proposed by the Tree Method.

According to Wherry [17], PCA is a way of reducing the number of data using a method of rotating axes, thus defining new loads for the same data. The variable *average pressure* correlated with morbidity variables (M_) can also be seen in Fig. 3.

The autocorrelation matrix, according to Pearson [18], represents the categorization of the variables between perfect negative, perfect positive and no correlation –1, +1 and 0, respectively. Fig. 4 shows some linear coefficients concerning morbidity variables and air quality variables.

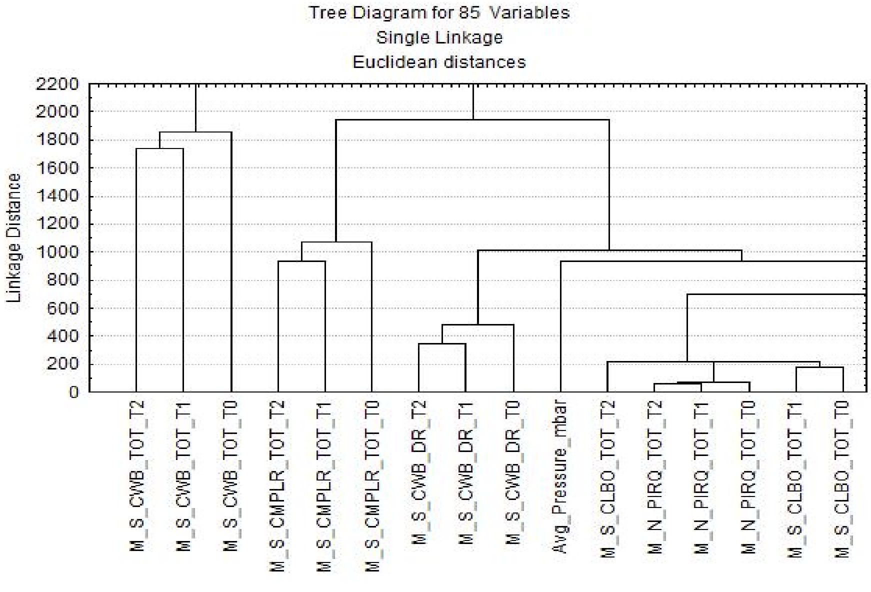


Figure 2: Example of a clustering method via tree analysis.

Table 1: Definition of acronyms.

Initials	*Meaning*
AMTE	Almirante Tamandaré
ARC	Araucária
CMPLG	Campo Largo
CLBO	Colombo
CWB	Curitiba
PINH	Pinhais
PIRQ	Piraquara
DR	Respiratory diseases
TOT	Total diseases
T0	Actual month
T1	Month + 1
T2	Month + 2

Table 2: Clusters from k-Means Method.

Cluster 1	*Cluster 3*
M_S_CMPLR_TOT_T0	M_S_CWB_TOT_T0
M_S_CMPLR_TOT_T1	M_S_CWB_TOT_T1
M_S_CMPLR_TOT_T2	M_S_CWB_TOT_T2
M_S_CWB_DR_T0	-
M_S_CWB_DR_T1	-
M_S_CWB_DR_T2	-

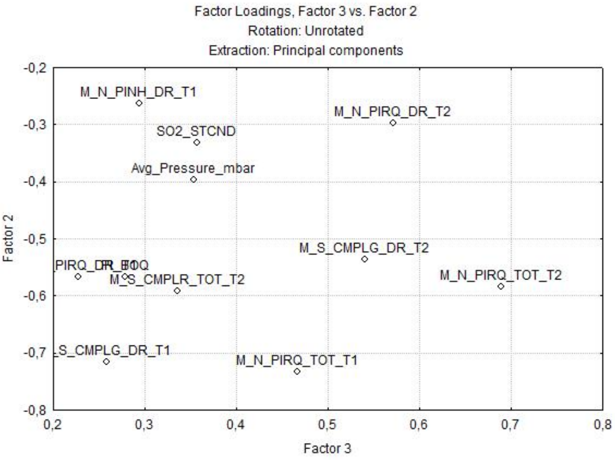
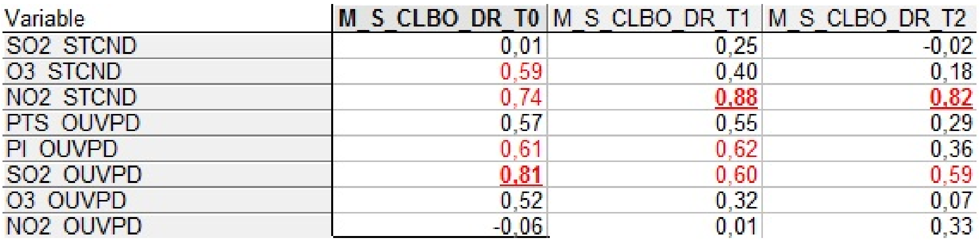


Figure 3: Example of PCA analysis correlating the average pressure with morbidity in different cities as the dendrogram (tree analysis).



Variable	M S CLBO DR T0	M S CLBO DR T1	M S CLBO DR T2
SO2 STCND	0,01	0,25	-0,02
O3 STCND	0,59	0,40	0,18
NO2 STCND	0,74	0,88	0,82
PTS OUVPD	0,57	0,55	0,29
PI OUVPD	0,61	0,62	0,36
SO2 OUVPD	0,81	0,60	0,59
O3 OUVPD	0,52	0,32	0,07
NO2 OUVPD	-0,06	0,01	0,33

Figure 4: Linear relations of morbidity in Colombo with SO_2 and NO_2 in other cities.

With the data properly prepared, it is possible to build the models and create rules that determine cases of respiratory morbidity.

3.3 Hospitalization prediction (classification rules)

The modelling consists of the creation of *classification rules* which are used to search for correlations in the variables within the database. The program (CBA, Singapore University) creates those rules by using advanced algorithms and by making multiple passes over the existing data. The variables must be divided into statistical intervals – quartiles or thirds for example. If a variable such as O_3 has values between 15 and 25, the rule item would be 15_<_O3_<_25. In the first check, the algorithm checks the frequency of each rule item. In subsequent checks, it creates candidate rule items leaving seeds for future passes and to decide whether it is a frequent rule item or not. From those agglomerated rule items, the classifier produces the *classification rules* which are all possible accurate rules [19].

4 RESULTS

The final table that was prepared, which had 104 registers and 149 variables, was divided in thirds when carrying out discretization, with values around 33.3%. Models for predicting three classes of morbidity according to these thirds were constructed. The results of a classifier are usually illustrated in a confusion matrix (as seen in Fig. 5). This table allows the calculation of two important metrics to evaluate the classifier's performance. Those analyses are determined as follows:

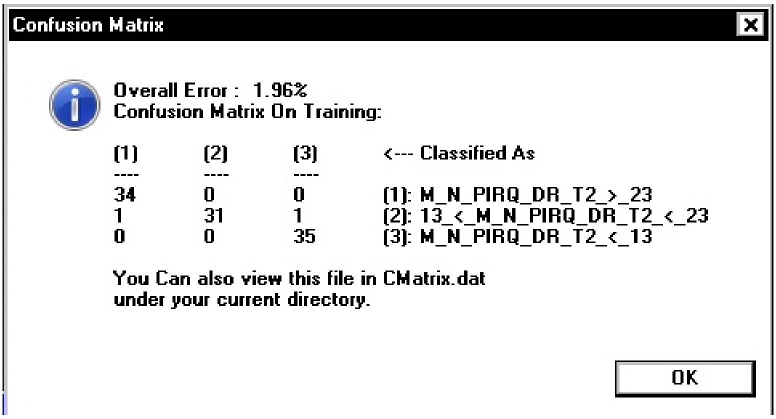
A. Accuracy:

$$\text{Accuracy} = \left\{\left(\frac{t\_pos}{pos}\right)\cdot\left[\frac{pos}{(pos+neg)}\right]\right\} + \left\{\left(\frac{t\_neg}{neg}\right)\cdot\left[\frac{neg}{(pos+neg)}\right]\right\},$$

where:
t_pos = positive samples classified as positives;
pos = positive samples;
t_neg = negative samples classified as negatives;
neg = negative samples.
(t stands for true)

B. Precision:

$$\text{Precision} = \left(\frac{t\_pos}{t\_pos+f\_pos}\right)\ or\ \left(\frac{t\_neg}{t\_neg+f\_neg}\right),$$



```
Overall Error :  1.96%
Confusion Matrix On Training:

[1]       [2]       [3]       <--- Classified As
----      ----      ----
34        0         0         [1]: M_N_PIRQ_DR_T2_>_23
1         31        1         [2]: 13_<_M_N_PIRQ_DR_T2_<_23
0         0         35        [3]: M_N_PIRQ_DR_T2_<_13

You Can also view this file in CMatrix.dat
under your current directory.
```

Figure 5: An example of 98.04% accuracy.

Precision = $\left(\frac{t\_pos}{t\_pos+f\_pos}\right)$ or $\left(\frac{t_neg}{t_neg+f_neg}\right)$where:
f_pos = negative samples rated as positives
f_neg = positive samples rated as negatives
(f stands for false)

4.1 Morbidity prediction

Some models were built to predict hospitalizations in seven cities of the metropolitan area of Curitiba using CBA as described in section 3.3. Those cities were previously selected by analyses of wind direction in the years 2005–2007. Two extra variables were created for each city, T1 and T2, which represent the actual month plus one and two. In addition to the respiratory morbidity, the total morbidity, which can be any disease that led to hospitalization, was taken into account when creating the database. In this sense, there were four future morbidity variables for each city, totalizing 28 models for predicting morbidity. Table 3 shows the results of the morbidity predictive classifiers. The computed metrics, accuracy and precision were obtained through cross validation.

As it can be seen in Table 3, there is a very good performance on the validation of the models for all classifiers. Therefore, these models could be used to precisely advance in two months the level of morbidity in all those cities. This advanced information helps the public health authorities in suitably managing outbreaks.

4.2 Classification rules

The rules built are then divided into two parts: an ‘IF’ that represents a condition and a consequence, and a ‘THEN’ represented by ->. There are two parameters that are matters of importance when creating a rule: support (cover) and confidence, both stated in percentages by the program.

Example rule 1:
IF CO_UEG_>_0_52 and M_S_CMPLG_DR_T0_<_150
THEN -> M_S_AMTE_DR_T1_<_40
(21.277% 100.000% 10 10 21.277%)

Table 3: Accuracy and precision for respiratory diseases (DR) and total diseases (TOT).

City	*Accuracy T1 %*	*Precision T1 %*	*Accuracy T2 %*	*Precision T2 %*
AMTE_DR	95.74	95.83	97.83	97.62
ARC_DR	93.20	93.08	95.10	95.16
CMPLG_DR	95.15	94.98	95.10	94.95
CLBO_DR	96.12	94.44	96.08	94.44
CWB_DR	97.09	96.93	95.10	94.93
PINH_DR	92.23	92.29	93.14	93.13
PIRQ_DR	97.09	97.05	98.04	97.98
AMTE_TOT	100	100	93.48	92.85
ARC_TOT	98.06	98.03	99.02	99.04
CMPLG_TOT	97.09	97.10	96.08	95.95
CLBO_TOT	97.09	97.10	99.02	99.12
CWB_TOT	95.15	94.04	91.18	89.64
PINH_TOT	98.06	98.09	96.08	96.01
PIRQ_TOT	97.09	97.05	96.08	95.90

So, IF CO at UEG station is higher than 0.52 and morbidity by respiratory diseases in Campo Largo city is lower than 150 people, THEN the total hospitalizations by respiratory diseases in Almirante Tamandaré City is going to be lower than 40 people. This rule has 100% confidence and its support is 21.27%.

Example rule 2:
IF M_S_CLBO_DR_T0_<_0 and O3_RPR_>_28
THEN -> M_S_CLBO_DR_T1_<_0
(20.388% 100.000% 21 21 20.388%)

So, IF there are no cases of hospitalization in the actual month in Colombo, and O_3 in Repar station is higher than 28, THEN there will be no hospitalizations caused by respiratory diseases in Colombo on the following month. This rule has 100% confidence and the support is 20.38%. As shown in Fig. 1, Colombo is far from Repar's station.

Table 4 illustrates a rules comparison among three levels of Inhalable Particles (PI) from the Ouvidor Pardinho Station (OUVPD), located in Curitiba, and three levels of morbidity caused by respiratory disease in Curitiba City for the next month or one month in advance. It is evident that the higher the PI concentration, the higher the morbidity is in Curitiba in the next month. These three rules acknowledge 33% (sum of the three supports 13.592% + 8.738% + 10.680%) of the phenomenon involving PI and morbidity. If combined with more rules concerning other variables (for example NO_2, SO_2 and others), such combination could bring even higher support and confidence for predictions. These rules could also be used with advantage from the urban management and health policy.

5 CONCLUSION AND NEXT STEPS

Outbreaks of respiratory disease have a strong economic impact on affected countries, especially in developing countries that have a poor response capacity of health services. In this sense, forecasting outbreaks two months in advance may help health institutions prepare for dealing with these extreme situations. The prediction models described in this paper

Table 4: Conditions for hospitalizations from respiratory diseases in Curitiba in T1.

IF	THEN (CWB_DR_T1)
PI_OUVPD_>_18 Vel_V_Med_m_s_<_2_1 (13.592% 100.000%)	M_S_CWB_DR_T1_>_1 250
12_<_PI_OUVPD_<_18 PTS_OUVPD_<_19 (8.738% 100.000%)	1100_<_M_S_CWB_DR _T1_<_1250
PI_OUVPD_<_12 Vel_V_Med_m_s_>_2_35 (10.680% 100.000%)	M_S_CWB_DR_T1_<_1 100

presented high accuracy and precision in the cross-validation process and could, therefore, be used by the authorities in mitigating the effects caused by poor air quality [20].

At the level of urban management, critical levels of air pollution measured at stations could suggest guidelines for minimizing emissions in these problem regions.

At the public health level, predictive models with high accuracy could alert campaigns to also mitigate the effects of pollution and prepare teams for the management of disease outbreaks caused by air pollution.

Further studies may be necessary to characterize other factors affecting morbidity, such as land use and natural phenomena in the region. It should be noted that the methodology described here is based on the acquisition of knowledge from monitoring data; however, the knowledge of specialists could usefully be incorporated to improve the quality of the forecasts. A hybrid model combining more than one data-mining technique could also be developed [21].

ACKNOWLEDGEMENTS

Special thanks to the CNPq for their financial support of the research and to the following institutes that provided the data: Environmental Institute of Paraná (IAP), State Health Secretary (SESA), Institute of Urban Planning of Curitiba (IPPUC) and Mary’s Protection Center of Children and Teenagers (CEDIN).

REFERENCES

[1] Costa, R., Jr., *Aula de Poluição do Ar*, Departamento de Engenharia Ambiental, Centro Tecnológico. https://goo.gl/ppzDAp, UFES, 2005. Accessed on: 15 Dec. 2016.

[2] Burnett, R.T., Smith-Doiron, M., Stieb, D., Cakmak, S. & Brook, J.R., Effects of particulate and gaseous air pollution on cardiorespiratory hospitalizations. *Arch. Environ. Health.*, pp. 130–139, 1999.

[3] Neupane, B., Jerrett, M., Burnett, R.T., Marrie, T., Arain, A. & Loeb, M., Long-term exposure to ambient air pollution and risk of hospitalization with community-acquired pneumonia in older adults. *Am. J. Respir. Crit. Care Med.*, **181**, pp. 47–53, 2010.

[4] Morris, R.D., Naumova, E.N., Munasinghe, R.L., Ambient air pollution and hospitalization for congestive heart failure among elderly people in seven large US cities. *Am. J. Public Health.*, **85**(10), pp. 1361–1365, 1995.

[5] Martins, L.C., et al., Poluição atmosférica e atendimentos por pneumonia e gripe em São Paulo, Brasil. *Revista de Saúde Pública*, **36**(1), pp. 88–94, 2002.
Bakonyi, S.M.C., et al., Poluição atmosférica e doenças respiratórias em crianças na cidade de Curitiba, PR. *Rev Saúde Pública*, **38**(5), pp. 695–700, 2004. www.scielo.br/pdf/rsp/v38n5/21758.pdf. Accessed on: 21 Feb. 2017.

[6] Hino, A.A.F., et al., Built environment and physical activity for transportation in adults from Curitiba, Brazil. *Journal of Urban Health*, **17**, pp. 1–17, 2013.

[7] Mosquera, J., et al., Transport and health: A look at three Latin American cities. *Cadernos de Saúde Pública (ENSP. Impresso)*, **29**, pp. 654–666, 2013.

[8] Brugge, D., Durant, J.L. & Rioux, C., Near-highway pollutants in motor vehicle exhaust: a review of epidemiologic evidence of cardiac and pulmonary health risks. *Environmental Health*, **6**(1), p. 23, 2007.

[9] Brownson, R.C., et al., Understanding administrative evidence-based practices. *American Journal of Preventive Medicine*, **46**, pp. 49–57, 2014.

[10] Dockery, D.W., et al., An association between air pollution and mortality in six U.S. cities. *N. Engl. J. Med.*, **329**(24), pp. 1753–1759, 1993.

[11] Dockery, D.W., Cunningham, J., Damokosh, A.I., Neas, L.M., Spengler, J.D. & Koutrakis, P., et al. Health effects of acid aerosols on North American children respiratory symptoms. *Environmental Health Perspectives*, **104**, pp. 500–505, 1996.

[12] Instituto Ambiental do Paraná (IAP), *Indicadores da Qualidade do Ar*. http://www.iap.pr.gov.br/modules/conteudo/conteudo.php?conteudo=59. Accessed on: 19 Nov. 2016.

[13] Ministério da Agricultura, Pecuária e Abastecimento. Instituto Nacional de Meteorologia. Rede de Estações Meteorológicas Automáticas do INMET: Nota Técnica No.001/2011/SEGER/LAIME/CSC/INMET. http://www.inmet.gov.br/portal/css/ content/topo_iframe/pdf/Nota_Tecnica-Rede_estacoes_INMET.pdf. Accessed on: 19 Nov. 2016.

[14] Pyle, D., *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, Inc.: San Francisco, CA, pp. 9–43,1999.

[15] Han, J. & Kamber, H., *Data Mining – Concepts and Techniques*, chapters 6–8, 2001.

[16] Wherry, R.J., *Contributions to Correlational Analysis*, Academic Press: New York, 1984.

[17] Pearson, K., Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London*, *Ser. A*, **187**, pp. 253–318, 1896.

[18] Liu, B., Hsu, W. & Ma, Y., Integrating classification and association rule mining. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998.

[19] Souza, F.T. & Rabelo, W.S., A data mining approach to study the air pollution induced by urban phenomena and the association with respiratory diseases. *11th International Conference on Natural Computation (ICNC)*, IEEE, pp. 1045–1050, 2015.

[20] Souza, F.T., A data-based model to locate mass movements triggered by seismic events in Sichuan, China. *Environ Monit Assess*, **186**(1), pp. 575–587, 2014.