Machine Learning to Forecast Medical Attentions of Pneumonia Cases in Colombian Cities: An implementation with Air Quality, Meteorological and Admission Data

Juan David Gutiérrez (jdgutierrez@udes.edu.co)
Universidad de Santander, Bucaramanga, Santander, Colombia

Research article

Keywords: Air pollution-aerosols, artificial intelligence, meteorological variables, respiratory disease.

Posted Date: August 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-53367/v1

Abstract

Background: Previous authors have evidenced the relationship between air pollution-aerosols and meteorological variables with the occurrence of pneumonia. Forecasting the number of attentions of pneumonia cases may be useful to optimize the allocation of healthcare resources and support public health authorities to implement emergency plans to face an increase in patients. The purpose of this study is to implement four machine-learning methods to forecast the number of attentions of pneumonia cases in the five largest cities of Colombia by using air pollution-aerosols, and meteorological and admission data.

Methods: The number of attentions of pneumonia cases in the five most populated Colombian cities was provided by public health authorities between January 2009 and December 2019. Air pollution-aerosols and meteorological data were obtained from remote sensors. Four machine-learning methods were implemented for each city. We selected the machine-learning methods with the best performance in each city and implemented two techniques to identify the most relevant variables in the forecasting developed by the best-performing machine-learning models.

Results: According to R^2 metric, random forest was the machine-learning method with the best performance for Bogotá, Medellín and Cali; whereas for Barranquilla, the best performance was obtained from the Bayesian adaptive regression trees, and for Cartagena, extreme gradient boosting had the best performance. The most important variables for the forecasting were related to the admission data.

Conclusions: The results obtained from this study suggest that machine learning can be used to efficiently forecast the number of attentions of pneumonia cases, and therefore, it can be a useful decision-making tool for public health authorities.

Introduction

Pneumonia is an acute respiratory infection that affects the lungs. In a person affected by pneumonia, its alveoli are filled with pus and fluid, which makes breathing painful and limits the exchange of gas [1]. Viruses, bacteria, and fungi can cause pneumonia [2]. Depending on the severity, signs and symptoms may include coughing, shortness of breath, fever, sweating and shaking chills, fatigue, chest pain, nausea, vomiting, or diarrhea [3]. The etiologic agents may spread via airborne droplets from a cough or sneeze. Additionally, this disease can be transmitted via blood, especially during and shortly after birth [4].

According to the World Health Organization, 808,694 children aged 5 and younger died to pneumonia in 2017. This disease accounts for 15% of all deaths of children under five [1]. In 2017, the average death rate by pneumonia in Latin America was estimated to be 86.5 per 100,000 inhabitants, but this rate raises to 344.1 per 100,000 among people aged 70+ [5]. In Colombia, in this same year, the average death rate was estimated to be 37.9 per 100,000 inhabitants, whereas for individuals aged 70+, the rate was 140.7 per 100,000 [5].

Previous research showed the positive relationship between air pollutants such as particulate matter of < 2.5 micrometers [6,7], sulfate [8,9] and nitrogen dioxide [10,11] and the incidence of pneumonia. Similarly, meteorological variables have been associated to the disease, mainly temperature, relative humidity, and rainfall [12–14].

Most of these previous works have focused on implementing time series analysis to assess the relationship between air pollutants and meteorological variables with the occurrence of pneumonia. However, forecasting the number of cases using multivariate time series is often limited and not sufficiently accurate because these methods have difficulties to handle the multiple complex nonlinear relationships between environmental variables and the incidence of pneumonia. Machine-learning methods are an alternative to forecasting the number of cases of pneumonia from environmental data [15].

On this paper, we shall implement four methods of machine-learning to forecast the number of medical attentions of pneumonia cases in the five most populated cities of Colombia based on air pollution-aerosols, meteorological and admission variables. Additionally, we shall implement two techniques to identify relevant variables in the forecast developed by the machine-learning methods. Our results are to show the potential of machine-learning methods in forecasting the attentions of pneumonia cases, and in monitoring environmental and admission variables in Colombia.

Materials And Methods

Attentions Data

Daily attentions of pneumonia cases were obtained from the individual reports provided by healthcare provision systems from January 2009 to December 2019. The daily data were grouped by epidemiological week to obtain weekly cumulative attentions. The five most populated cities in Colombia were selected for the study. The cities selected were Bogotá, Medellín, Cali, Barranquilla, and Cartagena.

Aerosols Data

Data on air pollution-aerosols corresponded to aerosol optical depth measured by the Moderate Resolution Imaging Spectroradiometer, a space-borne instrument [16]. Daily data on air pollution-aerosols are available on the NASA product: Modern-Era Retrospective Analysis for Research and Applications, Version 2 [17]. The air pollution-aerosols included in this study were Black Carbon Surface Mass Concentration (BCSMASS), Dimethylsulphide Surface Mass Concentration (DMSSMASS), Dust Surface Mass Concentration of 2.5 μm in diameter (DUSMASS25), SO_4 Surface Mass Concentration (SO4SMASS), and Sea Salt Surface Mass Concentration of 2.5 μm in diameter (SSSMASS25). All data were converted to $\mu g/m^3$. Daily data were grouped by epidemiological week to obtain weekly average

data. Spatial matching between the values of air pollution-aerosols and the cities included in the study was performed using the raster package of R [18].

Meteorological Data

Daily data on rainfall and temperature were obtained from Modern-Era Retrospective analysis for Research and Applications, Version 2 [19]. The daily data were grouped by epidemiological week to obtain weekly cumulative rainfall and weekly average temperature data. Spatial matching among the weekly values of rainfall and temperature and the five cities being evaluated was performed using the raster package [18] of R.

We included lags of up to 4 weeks for air pollution-aerosols and meteorological variables, which we considered sufficient to capture the necessary time for the period of incubation of the disease and the time to visit a healthcare facility, and the report of a new pneumonia case.

Admission Data

In the models, we included the following as admission data: The year (2009 to 2019), the epidemiological week (Epiweek) with values from 1 to 52 or 53 for each year, and the week consecutive (Consweek), with values ranging from 1 to 573 for the entire study period.

Machine-learning Methods

Four machine-learning methods were implemented to forecast the number of attentions of pneumonia cases in each city. The methods implemented were Extreme Gradient Boosting (XGBoost), Random Forest (RF), Support Vector Machines (SVM), and Bayesian Adaptive Regression Trees (BART).

XGBoost is used to implement gradient boosted decision trees. The method is an approach where new models that predict the residuals or errors of prior models are created and then added together to make the final prediction [20]. RF combines several randomized decision trees and aggregates to do their predictions by averaging [21]. The objective of SVM is to find a hyperplane in an N-dimensional space that distinctly classifies the data points [22]. BART is a nonparametric Bayesian regression approach which uses dimensionally adaptive random basis elements [23].

In each city, the response variable was the number of attentions of pneumonia cases per week. We used, as predictor variables in the machine-learning methods, the air pollution-aerosols and meteorological variables with lags of up to 4 weeks, as well as the year, the Epiweek and the Consweek.

Each machine-learning method was trained and tested on a partitioned 70/30 percentage split of the dataset by stratified random sampling for each city. The method of 10-fold cross-validation was used for

training the dataset. Additional file 1 shows the parameters of each machine-learning method implemented. The performance of the forecasting was evaluated with the R^2 metric. We used the package caret [24] of R to implement the machine-learning methods.

Machine-learning Model Interpretation

We implemented the techniques of permutation feature importance and feature interaction to provide explanations and to analyze the behavior and forecasting of the best-performing machine-learning model in each city.

The permutation feature importance is an approach that classifies the contribution of each variable based on its precision. This means that a variable can be significantly important if changing its values (permutation) increases the model error, which means that the model needs this variable to perform more accurate forecasting. On the other hand, if the model error shows no change when varying the values, the variable does not contribute or influence the model when making the forecast [25]. The permutation feature importance was estimated with 500 iterations.

Feature interaction explains the interaction between variables. The technique states that the effect that a variable can have on the forecast is probably influenced by other variables. Therefore, this method recognizes that variables can be interconnected and that not only does a variable by itself have an influence on the machine-learning model, but that the interaction between variables can also have an effect on how the model is making its forecast [26].

The package iml of R [25] was used to implement the techniques of permutation feature importance and feature interaction.

Results

Attentions of Pneumonia Cases

Between 2009 and 2019, a total of 1,199,890 attentions of pneumonia cases were reported in the five cities selected for the study. Bogotá was the city with the highest number of attentions with 457,343 attentions reported, whereas Cartagena, with 88,164 attentions, was the city with the lowest number of attentions. Every city, except for Bogotá, showed an evident incremental tendency in the number of attentions of pneumonia cases (Fig. 1), with a reduction in the number of attentions in 2015 (Fig. 1b, 1c, 1d and 1e).

Forecast Performance

The R^2 metric was estimated using the test dataset to evaluate the performance of the forecast developed by the machine-learning methods. In average, the RF method had the best performance for the five cities (R^2 = 0.80 sd = 0.03) and the worst average performance was observed in SVM (R^2 = 0.50 sd = 0.11). As for Bogotá, Medellín, and Cali, the best performance was achieved through RF, for Barranquilla, BART had the best performance, and for Cartagena, the best performance corresponded to XGBoost (Tab. 1). In average, the best performance, according the R^2 metric, was observed in Cali and Barranquilla. Bogotá was the city where the performance was the lowest for the four machine-learning methods implemented.

Table 1
Performance values using R^2 metric and testing dataset

	XGBoost	RF	SVM	BART
Bogotá	0.74	0.79	0.36	0.76
Medellín	0.75	0.76	0.51	0.72
Cali	0.82	0.83	0.67	0.79
Barranquilla	0.83	0.83	0.52	0.84
Cartagena	0.80	0.79	0.45	0.69
Average (sd)	0.79 (0.04)	0.80 (0.03)	0.50 (0.11)	0.76 (0.06)

sd = standard deviation. Note that a value of R^2 =1.00 corresponds to a perfect performance, and a value of R^2 =0.00 corresponds to a worst performance.

Machine-learning Method Interpretation

We implemented two techniques to analyze the behavior of the machine-learning model with the best performance in each city. The techniques of permutation feature importance and feature interaction showed that the Consweek and Epiweek variables were the most important variables in the machine-learning methods with the best performance for each city (Fig. 2). According to the technique of permutation feature importance, the rest of variables had low contribution to the forecast (Fig. 2a, 2c, 2e, 2g, 2i), with values of mean squared error after 500 permutation reaching < 2.5.

Most variables interacted with each other, since few interactions had a zero value of interaction strength (Fig. 2b, 2d, 2f, 2h and 2j). The Consweek variable had the highest degree of interaction with the other variables in the five cities, and therefore, this is the interaction with the largest influence in the forecast developed by the best-performing methods in each city (RF, BART and XGBoost).

The interaction between Consweek and the rest of variables influenced, in every city, the forecast obtained by > 0.10, reaching values of > 0.35 in Cartagena (Fig. 2j). Other variables such as Epiweek (Fig. 2b, 2f

and 2j), SO4MASS with lag = 3 (Fig. 2d), year (Fig. 2f and 2h), Temperature with lag = 1 (Fig. 2h), DUSMASS25 with lag = 2 (Fig. 2h), and Rainfall (Fig. 2j) had interaction values relatively high in some cities.

Discussion

Four machine-learning methods were implemented to forecast the weekly attentions of pneumonia cases in the five largest cities of Colombia, and as result, the method with the best performance was identified for each city. To the best of our knowledge, no previous studies have implemented machine-learning methods in the forecasting of the number of attentions of pneumonia cases. This is the first work to implement and compare various machine-learning methods with the purpose of forecasting the number of attentions of pneumonia cases using air pollution-aerosols, meteorological and admission data.

Our results showed that, in Bogotá, the average performance for the four machine-learning methods implemented is lower than in the other cities. We believe that this fact may be related to the seasonal pattern of the time series of attentions of pneumonia cases in this city. The time series of attentions in Bogotá showed a seasonal pattern with peaks in the months of April and May, which coincide with the first rainfall season in the region [27]. This seasonal pattern suggests the necessity of using other methods to forecast the number of attentions of pneumonia cases in this city. Some possible methods in the time series approach include multivariate vector autoregression models [28] or Bayesian structural time series models [29], and machine-learning methods such as recurrent neural networks [30].

Comparing the most important predictor variables in our results to most previous works [31–34] shows no coincidences, and that circumstance is related to the pattern of the time series and the approach implemented in this paper. Firstly, the week consecutive was the most important variable in the machine-learning methods with the best performance in each city, which seems to be associated to the incremental tendency observed in the time series of attentions, including Bogotá in particular since 2016 [Additional file 2]. Similarly, the epidemiological week was the most important variable in Bogotá according to the technique of permutation feature importance, which may be explained by the seasonal pattern of the time series of attentions in this city, along with the study period.

On the other hand, while our approach consisted of implementing several machine-learning models to forecast the number of attentions of pneumonia cases based on air pollution-aerosols, meteorological and admission data, previous authors focused on exploring the relationship between air quality and meteorological data and the occurrence of pneumonia by using, in most cases, statistical methods such as generalized linear models [31,35–37], generalized additive models [33,38–42], or autoregressive integrated moving average [34,43–45]. This difference in approaches illustrates the dissimilarities between predictive models, which generally provide high precision but low explicability, and explanatory models that generally provide high explicability but low precision [46–48].

Every city included in the study showed a decrease in the number of attentions of pneumonia cases in 2015. This fact was observed even in Bogotá in spite of the seasonal pattern in the time series of

attentions in this city [Additional file 2]. This reduction in the number of attentions in 2015 coincided with a decrease in the gross domestic product of Colombia and it is consistent with the relationship observed by other authors between pneumonia and economic growth measured as gross domestic product [49,50]. In the case of Colombia, the country experienced a contraction of its economy in 2015 as result of the reduction in oil prices [51], the country's main export.

Our study has some limitations that are worth mentioning. Firstly, we assumed that all attentions of pneumonia cases reported to the individual reports from the healthcare provision system corresponded to community-acquired pneumonia and it was not possible to exclude hospital-acquired pneumonia or other forms of pneumonia from the database. Secondly, our data on air pollution-aerosols corresponded to remote sensor data of aerosol optical depth captured by the Moderate Resolution Imaging Spectroradiometer and not in-situ measurements for each city. This approach seems to be proper, considering the source of data (NASA's satellites) and that, in Colombia, there is not a reliable network to monitor air quality in main cities. However, the DMSSMASS variable showed atypical data for Bogotá (a unique value of 0.00) and Medellín (two unique values of 0.00 and 0.01). The cause of these atypical values is not fully clear, but it might suggest that there are limits in the use of aerosol optical depth method for this air pollutant, particularly in tropical cities with an altitude of over 1,500 meters above sea level.

Conclusions

In this study, we implemented four machine-learning methods to forecast the number of weekly attentions of pneumonia cases in five Colombian cities based on air quality, meteorological and admission data. The results obtained show that RF, XGBoost and BART can accurately forecast the number of attentions of pneumonia cases.

The results show that the percentage of variance in the weekly attentions of pneumonia cases (dependent variable) can be explained by over 76% (R^2 metric) by the machine-learning methods implemented when the method with the best performance is selected for each city. These findings show that machine-learning methods have potential in forecasting the number of attentions of pneumonia cases in Colombia.

Abbreviations

XGBoost = Extreme gradient boosting; RF = Random Forest; SVM = Support Vector Machines; BART = Bayesian Adaptive Regression Trees; Epiweek = Epidemiological week; Consweek = Week consecutive; Temperature_1 = Temperature with lag = 1; Temperature_2 = Temperature with lag = 2; Temperature_3 = Temperature with lag = 3; Temperature_4 = Temperature with lag = 4; Rainfall_1 = Rainfall with lag = 1; Rainfall_2 = Rainfall with lag = 2; Rainfall_3 = Rainfall with lag = 3; Rainfall_4 = Rainfall with lag = 4; BCSMASS = Black Carbon Surface Mass Concentration; BCSMASS_1 = Black Carbon Surface Mass Concentration with lag = 1; BCSMASS_2 = Black Carbon Surface Mass Concentration with lag = 2;

BCSMASS_3 = Black Carbon Surface Mass Concentration with lag = 3; BCSMASS_4 = Black Carbon Surface Mass Concentration with lag = 4; DMSSMASS = Dimethylsulphide Surface Mass Concentration; DMSSMASS_1 = Dimethylsulphide Surface Mass Concentration whit lag = 1; DMSSMASS_2 = Dimethylsulphide Surface Mass Concentration whit lag = 2; DMSSMASS_3 = Dimethylsulphide Surface Mass Concentration whit lag = 3;DMSSMASS_4 = Dimethylsulphide Surface Mass Concentration whit lag = 4; DUSMASS25 = Dust Surface Mass Concentration of 2.5 μm in diameter; DUSMASS25_1 = Dust Surface Mass Concentration of 2.5 μm in diameter with lag = 1; DUSMASS25_2 = Dust Surface Mass Concentration of 2.5 μm in diameter with lag = 2; DUSMASS25_3 = Dust Surface Mass Concentration of 2.5 μm in diameter with lag = 3; DUSMASS25_4 = Dust Surface Mass Concentration of 2.5 μm in diameter with lag = 4; SO4SMASS = SO_4 Surface Mass Concentration; SO4SMASS_1 = SO_4 Surface Mass Concentration with lag = 1; SO4SMASS_2 = SO_4 Surface Mass Concentration with lag = 2; SO4SMASS_3 = SO_4 Surface Mass Concentration with lag = 3; SO4SMASS_4 = SO_4 Surface Mass Concentration with lag = 4; SSSMASS25 = Sea Salt Surface Mass Concentration of 2.5 μm in diameter; SSSMASS25_1 = Sea Salt Surface Mass Concentration of 2.5 μm in diameter with lag = 1; SSSMASS25_2 = Sea Salt Surface Mass Concentration of 2.5 μm in diameter with lag = 2; SSSMASS25_3 = Sea Salt Surface Mass Concentration of 2.5 μm in diameter with lag = 3; SSSMASS25_4 = Sea Salt Surface Mass Concentration of 2.5 μm in diameter with lag = 4.

Declarations

Ethics approval and consent to participate

The research did not pose any risk, since we used anonymized information of public access at a municipal level, and biological samples were not collected. Therefore, an environmental impact was not produced. In that sense, ethical approval was not required and declaration of interest is not declared.

Consent for publication

Not applicable. The research does not include details relating to an individual person.

Availability of data and materials

The dataset and script in R are available at https://github.com/juandavidgutier/Forecasting-attentions-of-pneumonia

Competing interests

The author declare that they have no competing interests.

Funding

This research had not economic support

Authors' Contributions

Conceptualization: Juan David Gutiérrez. Data curation: Juan David Gutiérrez. Methodology: Juan David Gutiérrez. Formal analysis: Juan David Gutiérrez. Software: Juan David Gutiérrez. Writing: Juan David Gutiérrez.

Acknowledgements

The author would like to thank the Colombian Public Health Surveillance System for providing the dataset. To José Daniel Salazar and Santiago Carvajal from the National University of Colombia for their contributions in the development of scripts in R.

References

1. WHO. Pneumonia [Internet]. 2018 [cited 2020 Jul 18]. Available from: https://www.who.int/news-room/fact-sheets/detail/pneumonia
2. National Center for Immunization and Respiratory Diseases (NCIRD). Causes of Pneumonia | CDC [Internet]. 2020 [cited 2020 Jul 18]. Available from: https://www.cdc.gov/ncird/
3. WHO. Pneumonia: Symptoms [Internet]. 2018 [cited 2020 Jul 18]. Available from: https://www.who.int/westernpacific/health-topics/pneumonia
4. Nissen MD. Congenital and neonatal pneumonia. Paediatr Respir Rev. 2007 Sep 1;8(3):195–203.
5. Global Burden of Disease Collaborative Network. GBD Results Tool | GHDx [Internet]. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017 (GBD 2017) Results. 2018 [cited 2020 Jul 18]. Available from: http://ghdx.healthdata.org/gbd-results-tool
6. Zhang Z, Hong Y, Liu N. Association of ambient Particulate matter 2.5 with intensive care unit admission due to pneumonia: a distributed lag non-linear model. Sci Rep. 2017 Aug 17;7(1):8679.
7. Duan Z, Han X, Bai Z, Yuan Y. Fine particulate air pollution and hospitalization for pneumonia: a case-crossover study in Shijiazhuang, China. Air Qual Atmosphere Health. 2016 Nov 1;9(7):723–33.
8. Shi W, Liu C, Norback D, Deng Q, Huang C, Qian H, et al. Effects of fine particulate matter and its constituents on childhood pneumonia: a cross-sectional study in six Chinese cities. The Lancet. 2018 Oct 1;392:S79.
9. Xiao Q, Liu Y, Mulholland JA, Russell AG, Darrow LA, Tolbert PE, et al. Pediatric emergency department visits and ambient Air pollution in the U.S. State of Georgia: a case-crossover study. Environ Health. 2016 Nov 25;15(1):115.

10. Cheng F-J, Lee K-H, Lee C-W, Hsu P-C. Association between Particulate Matter Air Pollution and Hospital Emergency Room Visits for Pneumonia with Septicemia: A Retrospective Analysis. Aerosol Air Qual Res. 2019;19(2):345–54.
11. Pirozzi CS, Jones BE, VanDerslice JA, Zhang Y, Paine R, Dean NC. Short-Term Air Pollution and Incident Pneumonia. A Case–Crossover Study. Ann Am Thorac Soc. 2017 Dec 28;15(4):449–59.
12. Qiu H, Sun S, Tang R, Chan K-P, Tian L. Pneumonia Hospitalization Risk in the Elderly Attributable to Cold and Hot Temperatures in Hong Kong, China. Am J Epidemiol. 2016 Oct 15;184(8):570–8.
13. Wiemken T, Mattingly W, Furmanek S, Guinn B, English C, Carrico R, et al. Impact of Temperature Relative Humidity and Absolute Humidity on the Incidence of Hospitalizations for Lower Respiratory Tract Infections Due to Influenza, Rhinovirus, and Respiratory Syncytial Virus: Results from Community-Acquired Pneumonia Organization (CAPO) International Cohort Study. Univ Louisville J Respir Infect [Internet]. 2017 May 22;1(3). Available from: https://ir.library.louisville.edu/jri/vol1/iss3/7
14. Tian D, Jiang R, Chen X, Ye Q. Meteorological factors on the incidence of MP and RSV pneumonia in children. PLOS ONE. 2017 Mar 10;12(3):e0173409.
15. Jhuo S-L, Hsieh M-T, Weng T-C, Chen M-J, Yang C-M, Yeh C-H. Trend Prediction of Influenza and the Associated Pneumonia in Taiwan Using Machine Learning. In: 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). 2019. p. 1–2.
16. National Aeronautics and Space Administration - NASA. Moderate-resolution Imaging Spectroradiometer. MODIS [Internet]. 2017 [cited 2020 Jul 19]. Available from: https://modis.gsfc.nasa.gov/about/media/modis_brochure.pdf
17. Bosilovich MGL. MERRA-2: File Specification [Internet]. 2015 Sep [cited 2020 May 9]. Available from: https://ntrs.nasa.gov/search.jsp?R=20150019760
18. Rdocumentation. raster package | R Documentation [Internet]. raster v3.0-12. 2020 [cited 2020 May 9]. Available from: https://www.rdocumentation.org/packages/raster/versions/3.0-12
19. National Aeronautics and Space Administration-NASA. Modern-Era Retrospective analysis for Research and Applications, Version 2 MERRA-2 [Internet]. 2017 [cited 2019 Oct 15]. Available from: https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/data_access/
20. Brownlee J. A Gentle Introduction to XGBoost for Applied Machine Learning [Internet]. Machine Learning Mastery. 2016 [cited 2020 Jun 18]. Available from: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
21. Biau G, Scornet E. A random forest guided tour. TEST. 2016 Jun 1;25(2):197–227.
22. Schölkopf B, Smola AJ, Bach F, Scholkopf MD of the MPI for BC in TGPB. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press; 2002. 658 p.
23. Berk RA. Statistical Learning from a Regression Perspective. Springer; 2016. 366 p.
24. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2005;28(5):1–26.

25. Molnar C, Bischl B, Casalicchio G. iml: An R package for Interpretable Machine Learning. J Open Source Softw. 2018;3(26):786.
26. Molnar C. Interpretable Machine Learning [Internet]. Interpretable Machine Learning. 2020 [cited 2020 Jun 30]. Available from: https://christophm.github.io/interpretable-ml-book/
27. Vargas-Luna A, Santos AC, Cardenas E, Obregon N. Analysis of distribution and spatial interpolation of rainfall in Bogota, Colombia. Dyna. 2011;78(167):151–9.
28. Wei WWS. Multivariate Time Series Analysis and Applications. John Wiley & Sons; 2018. 540 p.
29. Harvey AC. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press; 1990. 547 p.
30. Bianchi FM, Maiorino E, Kampffmeyer MC, Rizzi A, Jenssen R. Recurrent Neural Networks for Short-Term Load Forecasting: An Overview and Comparative Analysis. Springer; 2017. 74 p.
31. Glick AF, Tomopoulos S, Fierman AH, Elixhauser A, Trasande L. Association Between Outdoor Air Pollution Levels and Inpatient Outcomes in Pediatric Pneumonia Hospitalizations, 2007 to 2008. Acad Pediatr. 2019 May 1;19(4):414–20.
32. Nhung NTT, Amini H, Schindler C, Kutlar Joss M, Dien TM, Probst-Hensch N, et al. Short-term association between ambient air pollution and pneumonia in children: A systematic review and meta-analysis of time-series and case-crossover studies. Environ Pollut. 2017 Nov 1;230:1000–8.
33. Li D, Wang J, Zhang Z, Shen P, Zheng P, Jin M, et al. Effects of air pollution on hospital visits for pneumonia in children: a two-year analysis from China. Environ Sci Pollut Res. 2018 Apr 1;25(10):10049–57.
34. Ruchiraset A, Tantrakarnapa K. Time series modeling of pneumonia admissions and its association with air pollution and climate variables in Chiang Mai Province, Thailand. Environ Sci Pollut Res. 2018 Nov 1;25(33):33277–85.
35. Kongchouy N, Choonpradub C, Kuning M. Methods for Modeling Incidence Rates with Application to Pneumonia Among Children in Surat Thani Province, Thailand. Chiang Mai J Sci. 2010;37(1):29–38.
36. Marrie TJ, Durant H, Yates L. Community-Acquired Pneumonia Requiring Hospitalization: 5-Year Prospective Study. Rev Infect Dis. 1989 Jul 1;11(4):586–99.
37. Negrisoli J, Nascimento LFC, Negrisoli J, Nascimento LFC. Atmospheric pollutants and hospital admissions due to pneumonia in children. Rev Paul Pediatr. 2013 Dec;31(4):501–6.
38. Huh K, Hong J, Jung J. Association of meteorological factors and atmospheric particulate matter with the incidence of pneumonia: an ecological study. Clin Microbiol Infect [Internet]. 2020 Mar 14 [cited 2020 Jul 29]; Available from: http://www.sciencedirect.com/science/article/pii/S1198743X20301488
39. Liu Y, Kan H, Xu J, Rogers D, Peng L, Ye X, et al. Temporal relationship between hospital admissions for pneumonia and weather conditions in Shanghai, China: a time-series analysis. BMJ Open. 2014 Jul 1;4(7):e004961.

40. Kim J, Kim J-H, Cheong H-K, Kim H, Honda Y, Ha M, et al. Effect of Climate Factors on the Childhood Pneumonia in Papua New Guinea: A Time-Series Analysis. Int J Environ Res Public Health. 2016 Feb;13(2):213.
41. Ebi KL, Exuzides KA, Lau E, Kelsh M, Barnston A. Association of Normal Weather Periods and El Niño Events With Hospitalization for Viral Pneumonia in Females: California, 1983–1998. Am J Public Health. 2001 Aug 1;91(8):1200–8.
42. Sohn S, Cho W, Kim JA, Altaluoni A, Hong K, Chun BC. Pneumonia Weather': Short-term Effects of Meteorological Factors on Emergency Room Visits Due to Pneumonia in Seoul, Korea. J Prev Med Pub Health. 2019 Mar;52(2):82–91.
43. Rodrigues E, Machado A, Silva S, Nunes B. Excess pneumonia and influenza hospitalizations associated with influenza epidemics in Portugal from season 1998/1999 to 2014/2015. Influenza Other Respir Viruses. 2018;12(1):153–60.
44. Cahyati W, Sari M. Forecasting of Childhood Pneumonia in Semarang City. In Atlantis Press; 2020 [cited 2020 Jul 29]. p. 244–9. Available from: https://www.atlantis-press.com/proceedings/icracos-19/125931374
45. Lim C, Chen M, Chen M. Forecasting Emergency Department Admissions for Pneumonia in Tropical Singapore. Online J Public Health Inform [Internet]. 2018 May 22 [cited 2020 Jul 29];10(1). Available from: https://ojphi.org/ojs/index.php/ojphi/article/view/8327
46. Meyer J. Differences in Model Building Between Explanatory and Predictive Models [Internet]. The Analysis Factor. 2018 [cited 2020 Jul 25]. Available from: https://www.theanalysisfactor.com/differences-in-model-building-explanatory-and-predictive-models/
47. Sainani KL. Explanatory Versus Predictive Modeling. PM&R. 2014;6(9):841–4.
48. Koppius O. Prediction vs. explanation in statistical model building [Internet]. Center for Prevention Implementation Methodology for Drug Abuse and HIV. 2017 [cited 2020 Jul 25]. Available from: http://cepim.northwestern.edu/calendar-events/2017/2/28/otto-koppius-prediction-vs-explanation-in-statistical-model-building
49. Tian Y, Wu Y, Liu H, Si Y, Wu Y, Wang X, et al. The impact of ambient ozone pollution on pneumonia: A nationwide time-series analysis. Environ Int. 2020 Mar 1;136:105498.
50. Norbäck D, Lu C, Zhang Y, Li B, Zhao Z, Huang C, et al. Lifetime-ever pneumonia among pre-school children across China – Associations with pre-natal and post-natal early life environmental factors. Environ Res. 2018 Nov 1;167:418–27.
51. Toro-Córdoba JH, Garavito-Acosta AL, López-Valenzuela DC, Montes-Uribe E. El choque petrolero y sus implicaciones en la economía colombiana [Internet]. Bogotá, Colombia: Banco de la República; 2015 Oct [cited 2020 Jul 25]. Report No.: 906. Available from: https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/6217/be_906.pdf

Figures

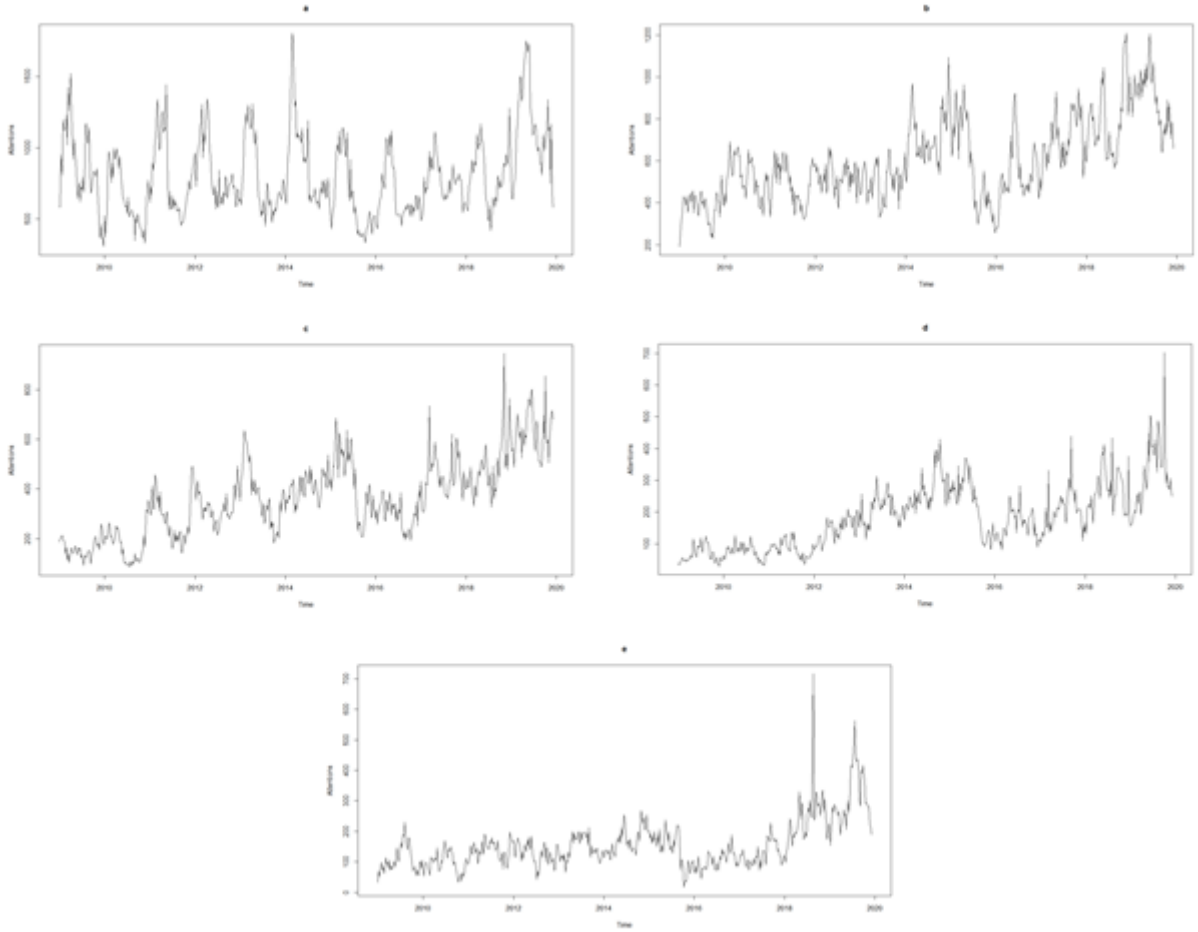

Figure 1

Time series of attentions of pneumonia cases in the top five largest cities of Colombia. a = Bogotá, b = Medellín, c = Cali, d = Barranquilla, and e = Cartagena

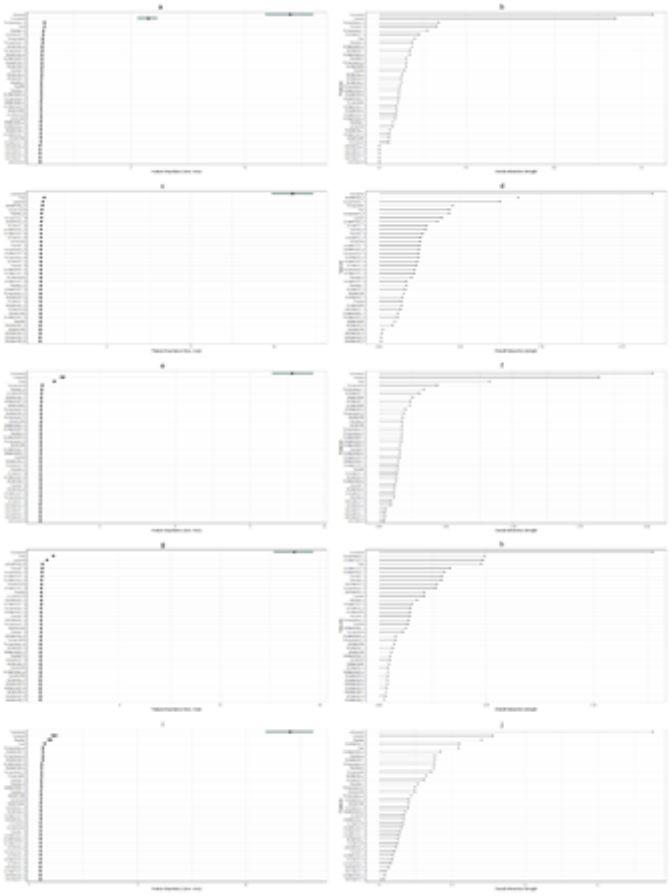

Figure 2

Permutation feature importance (Left) and Feature interaction (Right) ranking for Bogotá (a, b), Medellín (c, d), Cali (e, f), Barranquilla (g, h) and Cartagena (i, j). The error is expressed as a loss in the precision of the forecast, mse = mean squared error. Note that permutation feature importance and feature interaction were calculated for the best-performing machine-learning method for each city (Bogotá, Medellín and Cali = RF, Barranquilla = BART, and Cartagena = XGBoost).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile.docx