A data mining approach to elucidate the relationships between air pollution and respiratory diseases in big cities

Fabio Teodoro de Souz

Pontifical Catholic University of Parana, Brazil

Abstract

The sustainability of large cities is controlled by consumption, disposal, and environmental capacity. The weather patterns have been affected by the quick growing of the cities. These imbalances imply climate changes and negative consequences to the public health. In addition, due to the explosive growth in carbon dioxide emissions from fossil-fuel usage, researchers emphasize the importance in improving the quantitative control of the global carbon cycle as a central element to understand the patterns and projections of climate change. It is also discussed the importance in attributing observed CO_2 variations to human or natural cause. This research focuses on better understanding the relationships between air pollution and respiratory diseases. The methodology consists in applying data mining techniques on hospitalization due to respiratory diseases organized with atmospheric and urban variables. The knowledge acquired from this study - which is still in the early phase of data collection - could be useful for urban management and public health policies. Some qualitative associations between air pollutants in Curitiba and respiratory morbidity of childhood population have been discussed. Curitiba has a metropolitan area with population around 3 million. Some scientists highlighted the necessity of spreading methodological experiences from medium-size cities with relatively stable emissions to the more complex and representative environments of megacities (metropolitan areas with populations greater than 10 million). Moreover, this research should verify if the use of data mining techniques may potentially contribute to explain air pollution associated to the augment of the anthropogenic CO_2 signal in urban environments of megacities. In the last two decades, urbanization has intensified, and in Brazil, about 90% of the population now lives in urban centers. Atmospheric patterns have changed owing to the high growth rate of cities, with negative consequences for public health. This research aims to elucidate the spatial patterns of air pollution and respiratory diseases. A data-based model to aid local urban management to improve public health policies concerning air pollution is described. An example of data preparation and multivariate analysis with inventories from different cities in the Metropolitan Region of Curitiba was studied. A predictive model with outstanding accuracy in prediction of outbreaks was developed. Preliminary results describe relevant relations among morbidity scales, air pollution levels, and atmospheric seasonal patterns. The knowledge gathered here contributes to the debate on social issues and public policies. Moreover, the results of this smaller scale study can be extended to megacities. Background: Data measuring airborne pollutants, public health and environmental factors are increasingly being stored and merged. These big datasets offer great potential, but also challenge traditional epidemiological methods. This has motivated the exploration of alternative methods to make predictions, find patterns and extract information. To this end, data mining and machine learning algorithms are increasingly being applied to air pollution epidemiology. Methods: We conducted a systematic literature review on the application of data mining and machine learning methods in air pollution epidemiology. We carried out our search process in PubMed, the MEDLINE database and Google Scholar. Research articles applying data mining and machine learning methods to air pollution epidemiology were queried and reviewed. Results: Our search queries resulted in 400 research articles.

This work is partly presented at World Conference on Climate Change