MAPINS: An intra-city $PM_{2.5}$ modeling web application using a scalable data management and analysis system integrating public multi-source data

Xin Yu, Yuanbin Cheng, Yijun Lin, Yao-Yi Chiang,

Dimitrios Stripelis, Jose Luis Ambite

ABSTRACT: Air quality modeling is of great significance for studying the impact of air pollutants on human health and the urban built environment. Existing works mainly focus on applying various machine learning algorithms or physical simulations to generate models for air quality prediction based on different sources of data including geographic data, traffic emission data, remote sensing data, etc. Third-party software/tools are widely developed for air quality modeling as well as visualization. Many studies rely on third-party software for each of the processes. For example, the IBM Statistical Package for the Social Sciences (IBM SPSS) is widely used in the land use regression (LUR) model to handle the stepwise regression [1][2]. CALINE, a software package developed by the California Department of Transportation (Caltrans), is often used to provide a simulated result of the air pollution caused by road traffic (which most studies take as an explanatory variable [3][4]). The use of third-party software maybe plausible, for example, CALINE is a powerful tool when the study is on a large spatial and temporal scale (e.g., study the population impact on air quality). However, our research aims at developing a complete air quality prediction system that serves as easy access for people to be aware of the air quality within their neighborhoods. Thus, the dependency on third-party software/tools can cause "knowledge silos", that means developers should carry the result of third-party software manually to the next step analysis. This is the case that we do not want it happens during an entire automatic system.

Existing work of evaluating fine scale air quality typically replies on area-specific and expert-selected features (e.g., geographic features) for building an air quality model, which may omit potential important factors and manually reduce the variety of useful spatial data [5]. In this paper, we present an expert-free air quality prediction system, MAPINS. The system applies a data mining approach [6] that utilizes public multi-source data, OpenStreetMap (OSM) and air pollutant data from EPA web service to automatically figure out important geographic features and generate an expert-free method to predict $PM_{2.5}$ concentrations at a fine spatial resolution. The fine-scale system can inform people about surrounding air quality so that people can take preventive actions in advance. Thus, our system can serve as a platform that users can easily query the air pollution statue nearby.

In addition, the required data for building an air quality model are often coming from a variety of data sources in heterogeneous formats. The data can be huge and with a high update frequency (i.e., streaming data), which requires a particular big data infrastructure for storage, access, and analytics. As a result, it remains a challenge to integrate all data processing components that start from **data acquisition**, **pre-processing**, **modeling**, **prediction**, and **visualization** as well as **result dissemination**. To handle large spatial datasets efficiently, we use Apache Spark to build a scalable data management and analysis system to achieve the high performance for air quality prediction. Our system is based on a service-oriented architecture (SOA) to cooperate all the data processing components together to realize an automatic workflow from gathering raw data to the final display of predicting results.

KEYWORDS: Air quality, end-to-end system, SOA, fine scale

References

Briggs, D. Collins, S. Elliott, P. Fischer, P. Kingham, S. Lebret, E. Pryl, K. Reeuwijk, H. V. Smallbone, K. and Veen, A. V. E. (1997). Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science. 11, 7 (1997), 699-718.

Sahsuvaroglu, T. Arain, A Kanaroglou, P. Finkelstein, N. Newbold, B. Jerrett, M. Beckerman, B. Brook, J. Finkelstein, M. and Gilbert, N. L. (2006). A Land Use Regression Model for Predicting Ambient Concentrations of Nitrogen Dioxide in Hamilton, Ontario, Canada. Journal of the Air & Waste Management Association. 56, 8 (2006), 1059-1069.

Li, L. Lurmann, F. Habre, R., Urman R., Rappaport, E. Ritz, B. Chen, J.-C Filliland, F. D. and Wu, J. (2017). Constrained Mixed-Effect Models with Ensemble Learning for Prediction of Nitrogen Oxides Concentrations at High Spatiotemporal Resolution. Environmental science & technology, 51(17), 9920-9929.

Wilton, D. Szpiro, A. Gould, T. and Larson, T. (2010). Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles, CA and Seattle, WA. Science of the total environment, 408(5), 1120-1130.

Liu, C. Henderson, B. H. Wang, D. Yang, X. and Peng, Z. R. (2016). A land use regression application into assessing spatial variation of intra-urban fine particulate matter ($PM_{2.5}$) and nitrogen dioxide (NO_2) concentrations in City of Shanghai, China. Science of The Total Environment, 565, 607-615.

Lin, Y. Chiang, Y.-Y. Pan, F. Stripelis, D. Ambite, J. L. Eckel, S. P. and Habre, R. (2017). Mining public datasets for modeling intra-city $PM_{2.5}$ concentrations at a fine spatial resolution. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (p. 25).

Xin Yu, Master student, Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089

Yuanbin Cheng, Master student, Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089

Yijun Lin, Research Programmer, Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089

Yao-Yi Chiang, Associate Professor (Research), Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089

Dimitrios Stripelis, Ph.D. student, Information Sciences Institute, University of Southern California, Los Angeles, CA 90089

Jose Luis Ambite, Research Associate Professor, Information Sciences Institute, University of Southern California, Los Angeles, CA 90089