Applying Transfer Learning in Air Pollution Exposure Assessment

Jianzhao Bi[1], Shrey Gupta[1], Avani Wildani[1], and Yang Liu[1]

[1]*Emory University*

We apply transfer learning algorithms in estimating ambient air pollution exposure in which sparse air quality measurements cannot support reliable exposure prediction models (e.g., in developing countries). This project will enable accurate air pollution health analysis in developing countries in Asia, South America, and Africa.

Air pollution is a major threat to human health [1]. Numerous epidemiological studies have shown associations of exposure to air pollutants (fine particles, oxides of nitrogen, sulfur dioxide, ozone, etc.) with risk for adverse effects on morbidity and mortality [2]. An accurate estimation of the association between air pollutant exposure and a health outcome heavily relies on the accuracy of pollution exposure data which provide spatiotemporal concentrations of the pollutant in the environment the target population resides. Large-scale, high-resolution air pollution exposure data can be generated by statistical models with ground air quality measurements and associated predictors. Recently, non-parametric machine learning models have been increasingly adopted to improve exposure assessment.

The performance of an exposure prediction model highly depends on the amount and quality of ground air quality measurements. Due to high installation and maintenance cost, routine air quality stations have only sparsely deployed in developing countries. The sparsity of the measurements in space and time significantly hinders the qualities of exposure data and the downstream health analysis. Transfer learning (TL), a machine learning technique, is a promising way to enable accurate exposure assessment in data-poor regions/periods. TL can be used in storing relationships between air pollution exposure and its predictors while modeling in a data-rich region (period) and applying it to a data-poor region (period). With the transferred relationships, only few additional air quality measurements are needed to generate reliable exposure data. Table 1 shows the preliminary results of our transfer learning model to improve historical exposure assessment in the state of California.

Table 1: Transferring the $PM_{2.5}$ exposure model built for the year of 2018 to 2015 in California based on Random Forest.

Model	N	CV R^2	RMSE
Source (2018)	41886	–	–
Target (2015)	1000	0.43	5.65 $\mu g/m^3$
Transfer	41886+1000	0.52	5.17 $\mu g/m^3$

Project Goals and Phases

- *Phase 1*: Due in Spring 2020 – Exploring effective TL algorithms for the task of exposure transferring and developing a preliminary TL model. The preliminary TL model is examined with abundant air quality data in the United States.
- *Phase 2*: Due in Winter 2020 – Expanding the preliminary TL models to the developing countries with less abundant air quality data, e.g., China and India.
- *Phase 3*: Due in Spring 2021 – Developing a "big data" air quality exposure database with as many as high-quality air quality data in the world and automating the TL process to provide on-demand service for air pollution health analysis in developing countries.

Experimentation Requirements: Google Compute Engine

Our experiments deal with processing and training models with large volumes of data, therefore requiring powerful computation resources such as the Google Compute Engines, which can handle CPU and GPU intensive jobs with relatively low demand of memory and disk I/O.

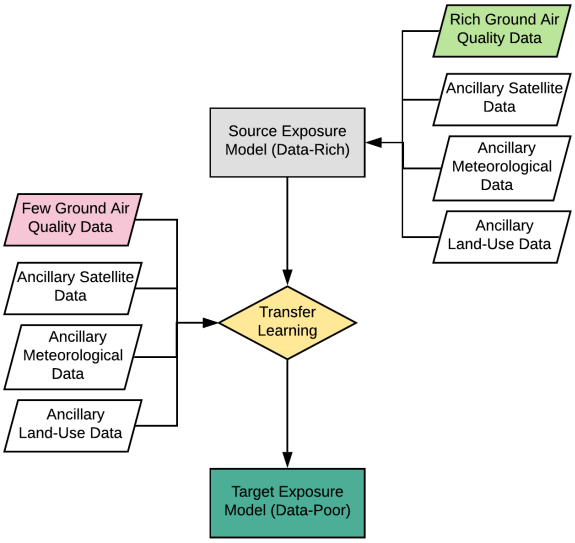

Figure 1: Architecture for the air pollution transfer learning

Our project will be the first to apply the transfer learning algorithms in air pollution exposure assessment in the regions with sparse air quality data. Figure 1 describes our modeling architecture. The results will enable accurate air pollution health analysis in developing countries, which will support the improvement of air quality standards and benefit local residents suffering from severe air pollution issues.

[1] Cohen, A. J. et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet 2017, 389, 1907–1918.

[2] Brook, R. D. et al. American Heart Association Council on Epidemiology and Prevention, Council on the Kidney in Cardiovascular Disease, and Council on Nutrition, Physical Activity and Metabolism. Metabolism, Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. Circulation 2010, 121, 2331–2378.