Atmospheric Environment 151 (2017) 1–11

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches

Cole Brokamp [a, b, *], Roman Jandarov [b], M.B. Rao [b], Grace LeMasters [b, c], Patrick Ryan [a, b]

[a] Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[b] Department of Environmental Health, University of Cincinnati, Cincinnati, OH, USA
[c] Division of Asthma Research, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

HIGHLIGHTS

- Land use models based on regression (LUR) and random forest (LURF) were created for elemental PM2.5
- LURF models were more accurate and precise than LUR models for most elements.
- Random forest may be used in future land use models for more accurate exposure assessment.

ARTICLE INFO

Article history:
Received 30 August 2016
Received in revised form
28 November 2016
Accepted 29 November 2016
Available online 1 December 2016

Keywords:
Elemental PM2.5
Land use regression
Random forest

ABSTRACT

Exposure assessment for elemental components of particulate matter (PM) using land use modeling is a complex problem due to the high spatial and temporal variations in pollutant concentrations at the local scale. Land use regression (LUR) models may fail to capture complex interactions and non-linear relationships between pollutant concentrations and land use variables. The increasing availability of big spatial data and machine learning methods present an opportunity for improvement in PM exposure assessment models. In this manuscript, our objective was to develop a novel land use random forest (LURF) model and compare its accuracy and precision to a LUR model for elemental components of PM in the urban city of Cincinnati, Ohio. PM smaller than 2.5 μm (PM2.5) and eleven elemental components were measured at 24 sampling stations from the Cincinnati Childhood Allergy and Air Pollution Study (CCAAPS). Over 50 different predictors associated with transportation, physical features, community socioeconomic characteristics, greenspace, land cover, and emission point sources were used to construct LUR and LURF models. Cross validation was used to quantify and compare model performance. LURF and LUR models were created for aluminum (Al), copper (Cu), iron (Fe), potassium (K), manganese (Mn), nickel (Ni), lead (Pb), sulfur (S), silicon (Si), vanadium (V), zinc (Zn), and total PM2.5 in the CCAAPS study area. LURF utilized a more diverse and greater number of predictors than LUR and LURF models for Al, K, Mn, Pb, Si, Zn, TRAP, and PM2.5 all showed a decrease in fractional predictive error of at least 5% compared to their LUR models. LURF models for Al, Cu, Fe, K, Mn, Pb, Si, Zn, TRAP, and PM2.5 all had a cross validated fractional predictive error less than 30%. Furthermore, LUR models showed a differential exposure assessment bias and had a higher prediction error variance. Random forest and other machine learning methods may provide more accurate exposure assessment.

1. Introduction

1.1. Land use regression models

Many air pollution exposure assessment methods assume that the spatial distribution of air pollutant concentrations are directly related to the use of the surrounding land. Physical features like

[*] Corresponding author. Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.
E-mail address: cole.brokamp@cchmc.org (C. Brokamp).

http://dx.doi.org/10.1016/j.atmosenv.2016.11.066
1352-2310/© 2016 Elsevier Ltd. All rights reserved.

elevation as well as the location and intensity of known pollutant sources including industrial emitters and traffic have been found to correlate well with pollutant concentrations (Briggs, 2005; Kolovos et al., 2010). Specifically, land use regression (LUR) uses predictors within a regression framework and has been the main focus of many land use models, becoming a popular tool for exposure assessment in air pollution research (Ryan et al., 2007; Henderson et al., 2007; Kashima et al., 2009; Ross et al., 2006). However, land use modeling is a complex problem due to the high spatial and temporal variations in pollutant concentrations on the local scale (Briggs et al., 1997; Beelen et al., 2010). LUR models have provided valuable insights and while more complex approaches have been applied to variable selection, the methodology has not included current predictive machine learning techniques. Therefore, there is an opportunity to improve the accuracy and precision of land use models, resulting in better exposure assessment for air pollution related epidemiological studies.

1.2. *Using random forest in land use models*

Land use models inherently use a high number of features that are highly correlated, for example, the length of highways within 100, 200, 300, and 400 m. Selection of which features to use in the final model is the outstanding challenge in land use model building and several approaches have been implemented (see Ryan and LeMasters 2007 for a review), most of which revolve around stepwise variable selection in a regression framework. Inclusion of correlated predictors generate problems for regression, often leading to unstable model estimates and variance inflation (Hastie et al., 2005). Although methods like variance inflation tests and influence statistics exist to combat this problem, they work by removing variables from the model that might otherwise be useful for prediction. Another challenge rising from regression-based land use models is the difficulty in capturing non-linear relationships and complex interactions. Because of the usually small sample size ($n = 20$ to 40) and very large number of possible predictors ($p = 50$ to over 500), it is often not feasible to evaluate all possible regression models.

Random forests are resistant to these problems. A key advantage of random forest is its ability to capture complex and non-linear relationships between predictors and the outcome with small sizes of training data. Random forests may be more accurate predictors of pollutant concentrations if they can indeed capture more patterns based on land use data. A random forest has been empirically shown to estimate concentrations of nitrogen dioxide based on land use data in the urban area of Geneva with a lower error when compared to regression (Champendal et al., 2014), although the authors did not compare the model's cross validated performance with a traditional land use regression model. We hypothesize that land use random forest (LURF) models, as compared to LUR models, will result in more accurate and precise estimates of PM2.5 elemental component concentrations.

1.3. *Random forests*

Random forests (James et al., 2013; Liaw and Wiener, 2002) are often implemented in prediction analyses because of their increased accuracy and resistance to multi-collinearity and complex interaction problems as compared to linear regression (Hastie et al., 2005). The technique itself is an ensemble learning method that builds on bagging – specifically the bootstrapped aggregation of several regression trees – to predict an outcome. Bagging is most often used to reduce the variance of an estimated prediction function and is most useful for models which are unbiased but have a high variance, like regression trees (Hastie et al., 2005). Random forests, first proposed by Breiman (Breiman et al., 1984), modify the bagging technique by ensuring that the individual trees are decorrelated by using a bootstrap sample for each tree and also randomly selecting a subset of predictors for testing at each split point in each tree. The random forest comes with the advantages of tree-based methods, namely the ability to capture complex interactions and maintain low bias, while at the same time alleviating the problem of high variance of predictions usually associated with tree-based methods by growing the individual trees to a very deep level (usually one observation per terminal node) and averaging their predictions.

1.4. *Land use models for elemental PM2.5 components*

Particulate matter (PM) is a complex mixture of chemical and elemental constituents and epidemiological studies have shown that these components and their sources are associated with adverse cardiovascular and respiratory health outcomes in adults (Zanobetti et al., 2009; Simkhovich et al., 2008; Dockery, 2009). Further studies suggest that certain components of PM2.5 are responsible for adverse health effects and characterizing these health effects of PM components has been identified as a research priority by the National Research Council for the National Academies (N. R. C. U. C, 2004). Recently, successful LUR models have been developed for PM components in twenty areas in Europe as a part of the ESCAPE study (de Hoogh et al., 2013) and for an urban area in Canada (Zhang et al., 2015). These land use models have allowed for assessment of exposure to individual components of PM and the study of their association with health outcomes (Beelen, 2015; Eeftens et al., 2014; Hampel et al., 2015). Although some models have been developed, limited information on PM components has impeded progress in identifying their health effects (Bell et al., 2007).

2. Methods

2.1. *Elemental PM2.5 measurements*

Measurements were collected at 24 sites across Cincinnati, Ohio as a part of CCAAPS, with full details available elsewhere (Ryan et al., 2007). Briefly, sites were selected based on the location of the CCAAPS cohort as well as wind direction, and proximity to pollution sources. Nine of the total sites were located within 400 m of a major roadway, while the rest of the sites were all located at least 1500 m away from a major roadway. Fig. 1 shows the location of the CCAAPS sampling sites and the birth addresses for the CCAAPS cohort. Between 2001 and 2005, PM2.5 samples were collected on 37-mm Teflon membrane filters and 37-mm quartz filters with Harvard-type Impactors. The increase in weight of the Teflon filters after sampling was used to determine the total PM2.5 mass (Hu et al., 2006) and X-ray fluorescence was used with the quartz filters to determine elemental concentrations for a total of 38 elements. Traffic related air pollution (TRAP) was calculated as the fraction of elemental carbon that was attributable to traffic by using a multivariate receptor model (Henry, 2000, 2003), UNMIX, to identify source signatures. One of the signatures was identified as TRAP because it was similar to comparison measurements conducted for cluster sources of trucks and buses (Hu et al., 2006) in Cincinnati, Ohio. Mean elemental concentrations for each site were calculated as averages and were considered missing if at least 75% of their measurements were classified as below the threshold of measurement certainty. For implementation of the land use models, in addition to total PM2.5 and TRAP, we restricted our building of models to the following eleven elements, which were selected for their previous association with health effects and a