Environmental Research 159 (2017) 427–434

Contents lists available at ScienceDirect

Environmental Research

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

A spatio-temporal prediction model based on support vector machine regression: Ambient Black Carbon in three New England States

Yara Abu Awad[a,*], Petros Koutrakis[a], Brent A. Coull[a,c], Joel Schwartz[a,b]

[a] Department of Environmental Health, Harvard T.H. Chan School of Public Health, 401 Park Drive, Boston, MA 02215, USA
[b] Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02215, USA
[c] Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02215, USA

ARTICLE INFO

Keywords:
Black Carbon
Air pollution
Prediction
Support Vector Regression
Machine learning

ABSTRACT

Fine ambient particulate matter has been widely associated with multiple health effects. Mitigation hinges on understanding which sources are contributing to its toxicity. Black Carbon (BC), an indicator of particles generated from traffic sources, has been associated with a number of health effects however due to its high spatial variability, its concentration is difficult to estimate. We previously fit a model estimating BC concentrations in the greater Boston area; however this model was built using limited monitoring data and could not capture the complex spatio-temporal patterns of ambient BC. In order to improve our predictive ability, we obtained more data for a total of 24,301 measurements from 368 monitors over a 12 year period in Massachusetts, Rhode Island and New Hampshire. We also used Nu-Support Vector Regression (nu-SVR) – a machine learning technique which incorporates nonlinear terms and higher order interactions, with appropriate regularization of parameter estimates. We then used a generalized additive model to refit the residuals from the nu-SVR and added the residual predictions to our earlier estimates. Both spatial and temporal predictors were included in the model which allowed us to capture the change in spatial patterns of BC over time. The 10 fold cross validated (CV) R^2 of the model was good in both cold (10-fold CV R^2 = 0.87) and warm seasons (CV R^2 = 0.79). We have successfully built a model that can be used to estimate short and long-term exposures to BC and will be useful for studies looking at various health outcomes in MA, RI and Southern NH.

1. Introduction

Fine ambient particulate matter (with an aerodynamic diameter of 2.5 μm – $PM_{2.5}$) has been widely associated with multiple health effects including cardiovascular and lung-cancer mortality following both chronic (Dockery et al., 1993; Krewski et al., 2009; Lepeule et al., 2012; Pope et al., 1995) and acute (Analitis et al., 2006; Schwartz and Marcus, 1990) exposures. As it is composed of a mixture of heterogeneous substances, efficiently mitigating $PM_{2.5}$ hinges on understanding the health effects of components arising from different sources. In particular, particles arising from fuel combustion have been independently associated with mortality (Laden et al., 2000; Ostro et al., 2007).

Black Carbon (BC) has been identified primarily as a marker of diesel traffic, followed by general traffic with minor contributions from biomass combustion in the U.S. (Sasser, 2012). Time-series studies, which rely on daily exposures at one or several central monitors have shown associations with respiratory (Bremner et al., 1999), cardiovascular (CVD) and total mortality (Maynard et al., 2007). While time-series studies look at the effect of acute exposures, other studies have also shown an association between chronic BC exposure and each of increased blood pressure (Schwartz et al., 2012), faster rates of lung function decline (Lepeule et al., 2014), impaired cognitive function (Power et al., 2010), and all-cause, cardiovascular, lung cancer and cardiopulmonary mortality (Beelen et al., 2008; Filleul et al., 2005; Smith et al., 2009).

However, relying on central monitors rather than individual level exposures leads to high exposure misclassification as there is spatial variability in concentrations (Clougherty et al., 2008; Künzli et al., 2005). While individual exposures may be determined through the use of personal monitoring devices, this limits sample size and duration of follow-up. Using personal monitors, Jansen et al. (2005) collected exposure data of 16 participants for 2 weeks and found an association between BC exposure and increased airway inflammation among asthmatics.

Another approach to estimate individual exposures is Land Use Regression (LUR) (Hoek et al., 2008; Ryan and LeMasters, 2007). LUR models, which account for spatial variability by using data on spatial predictors of emissions to predict exposure, are able to capture

* Correspondence to: HSPH Landmark Center, Box 15677, 401 Park Drive West, Boston, MA 02215, USA.
E-mail address: yara.abuawad@mail.harvard.edu (Y. Abu Awad).

http://dx.doi.org/10.1016/j.envres.2017.08.039
Received 11 July 2016; Received in revised form 18 August 2017; Accepted 21 August 2017
Available online 18 September 2017
0013-9351/

variations in exposure among study participants residing in different locations.

Several previous land use regression models were based on short duration intensive monitoring campaigns which could result in insufficient temporal resolution.

If pollution controls such as Diesel fuel composition and retrofit of particle filters on buses reduce exposure in areas heavily impacted by Diesel buses, but not elsewhere, the spatial pattern can change over time, and an LUR model will typically fail to pick up such spatio-temporal changes. Changes in traffic patterns and density over time can similarly produce changes in the spatial distribution of BC. Moreover, if the year with the intensive monitoring campaign had atypical weather, such as an unusual number of inversions, or more or less transported BC than usual due to differences in the tracking of weather fronts and prevailing winds, the estimated spatial distribution over the entire study period may not adequately reflect the spatial distribution at any given time.

This year to year variability in meteorology can be accounted for by obtaining multiple years of daily BC measurements, and including interaction terms between land use terms (that are surrogates for BC emissions) and mixing height and wind speed. Such terms model how concentrations vary for a given amount of emissions. Resulting exposure estimates are valuable for examining shorter term effects of BC on acute events (blood pressure etc.) as well as reflect change in the spatial distribution of BC in that year resulting from meteorology. Moreover, if BC measurements are available for many years, LUR models can capture the impact of changes in fuel and pollution controls, which will improve predictions by better capturing spatio-temporal variation in BC levels.

We previously fit such a model predicting black carbon in for the years 1999–2004 in the greater Boston area (Gryparis et al., 2007). Subsequently, resulting exposure predictions have been used to show an association with a variety of health outcomes in Boston-area cohorts, including: increased blood pressure (Alexeeff et al., 2011), atherosclerosis (Wilker et al., 2013) and decline in cognitive function (Power et al., 2010; Suglia et al., 2008).

Despite evidence generated by studies using existing BC predictions, they have some limitations. First, due to the moderate size of the data set spanning over four years, the data did not support very large models that allowed for complex spatio-temporal patterns likely present in the true pollution fields. Second, again due to the modest number and location of the available BC monitors over a decade ago, the range over time and space over which the model can produce reliable estimates is limited to the greater Boston area of the model. Therefore, now that more data is available, more advanced models incorporating nonlinear terms and higher order interactions, with appropriate regularization of parameter estimates, would likely improve prediction. In addition, New England is home to a large number of cohort studies that offer the opportunity to examine novel outcomes and biomarkers that provide evidence about the biological pathways responsible for observed health effects.

Hence, we expanded the geographic area of the original model by adding data from three states: Massachusetts, New Hampshire and Rhode Island, updated monitoring data up to and including 2011, and applied machine learning techniques that allow for complex and nonlinear relations between predictors and BC.

Unlike prediction using regression modeling, machine learning does not require that we make any assumptions regarding the functional form of the relationship between predictors and the variable of interest. Instead, machine learning uses the data provided and, within the limiting parameters specified, builds the prediction model. The model we chose to use; nu-Support Vector Regression has been demonstrated to predict ambient air pollutants with good generalizability (Hajek, 2015; Lu and Wang, 2005; Sotomayor-Olmedo et al., 2013) and is discussed in further detail below.

One limitation of the nu-SVR approach is that while it captures all possible interactions, it only partially captures nonlinearities. Generalized additive models (GAM) on the other hand can capture complex nonlinear relationships through smooth terms. This approach can be thought of as simple gradient boosting. Gradient boosting has been used in machine learning techniques like random forests to find the best fitting model by iteratively fitting a new model to the error from previous models (Natekin and Knoll, 2013). Fitting multiple relatively weak models can yield a better fit than one strong model. Here we implemented a simple form of this technique and boosted the power of our nu-SVR model using a second weaker model.

Our goal is to employ this approach to generate exposure predictions with greater predictive power than in previous models for Black Carbon levels.

2. Data and methods

2.1. Monitoring data

In order to capture the spatial and temporal variability of Ambient Black Carbon (BC) in the study area of interest, we added measurements from the greater Boston area, Cape Cod, Western and Central Massachusetts in addition to Rhode Island and New Hampshire from 2000 to 2011. To improve our ability to separate spatial, temporal, and spatio-temporal variability in BC levels, we focused on adding locations with a large number of repeated measurements. In total, monitoring data was obtained from five sources described below. A total of 24,301 observations were included from 368 unique monitoring locations (see Fig. 1 for a map of monitors). The mean BC concentration was 0.59 μg/m^3 with a standard deviation of 0.42 μg/m^3 and a maximum of 3.25 μg/m^3.

2.2. Sources of measurements

As part of a **National Institute of Environmental Health Sciences** (NIEHS) funded study, we carried out measurements at 53 sites between 2006 and 2008 around the Boston area, which were selected based on gaps in previous spatial measurements. We obtained 2798 24-h measurements using an Aethalometer® (Model AE-16 by Magee Scientific Corp.).

The Northeast States for Coordinated Air Use Management (NESCAUM) conducted a monitoring study to look at the spatial variability of pollution generated from traffic sources in the Boston area between 1999 and 2003 (Allen, 2014). BC was measured using an Aethalometer at 12 sites. This study provided 4767 24-h observations.

The Interagency Monitoring of Protected Visual Environments ("IMPROVE") is a network of monitor sites in national parks and wilderness areas, that measures Elemental Carbon (EC) via thermal/optical reflectance (IMPROVE, 2016). We obtained 2478 24-h measurements from the Quabbin Summit and Cape Cod locations from 2001 to 2011.

The U.S. **Environmental Protection Agency** (EPA) requires states to monitor PM2.5. Teflon® filters used to collect 24-h $PM_{2.5}$ ambient measurements throughout MA and Southern NH were obtained from the state environmental agencies, and we analyzed them for BC using a smokestain reflectometer (EEL Model M34D by Diffusion Systems Ltd). Reflectance was transformed to absorption coefficients according to ISO 9835. We obtained 6073 measurements from 23 sites in MA and 591 measurements from 7 sites in Southern NH for the years 2000–2011. We also obtained a total of 7285 Aethalometer BC observations from the RI Department of Environmental Management, from 7 sites between 2005 and 2011.

The **Normative Aging study** (NAS) is a longitudinal study of aging established by the Veterans Administration in 1961. We conducted indoor exposure monitoring between 2006 and 2010 at the homes of study participants. Measurements were taken in the main activity room of participants' homes over the period of one week using a Teflon® filter