Environmental Pollution 226 (2017) 21–29

Contents lists available at ScienceDirect

Environmental Pollution

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

National-scale exposure prediction for long-term concentrations of particulate matter and nitrogen dioxide in South Korea☆

Sun-Young Kim [a,*], Insang Song [b]

[a] *Institute of Health and Environment, Seoul National University, Seoul, South Korea*
[b] *Department of Geography, Seoul National University, Seoul, South Korea*

ARTICLE INFO

Article history:
Received 29 November 2016
Received in revised form
16 March 2017
Accepted 19 March 2017

Keywords:
Exposure prediction
Korea
National model
Nitrogen dioxide
Particulate matter

ABSTRACT

The limited spatial coverage of the air pollution data available from regulatory air quality monitoring networks hampers national-scale epidemiological studies of air pollution. The present study aimed to develop a national-scale exposure prediction model for estimating annual average concentrations of PM_{10} and NO_2 at residences in South Korea using regulatory monitoring data for 2010. Using hourly measurements of PM_{10} and NO_2 at 277 regulatory monitoring sites, we calculated the annual average concentrations at each site. We also computed 322 geographic variables in order to represent plausible local and regional pollution sources. Using these data, we developed universal kriging models, including three summary predictors estimated by partial least squares (PLS). The model performance was evaluated with fivefold cross-validation. In sensitivity analyses, we compared our approach with two alternative approaches, which added regional interactions and replaced the PLS predictors with up to ten selected variables. Finally, we predicted the annual average concentrations of PM_{10} and NO_2 at 83,463 centroids of residential census output areas in South Korea to investigate the population exposure to these pollutants and to compare the exposure levels between monitored and unmonitored areas. The means of the annual average concentrations of PM_{10} and NO_2 for 2010, across regulatory monitoring sites in South Korea, were 51.63 μg/m3 (SD = 8.58) and 25.64 ppb (11.05), respectively. The universal kriging exposure prediction models yielded cross-validated R^2s of 0.45 and 0.82 for PM_{10} and NO_2, respectively. Compared to our model, the two alternative approaches gave consistent or worse performances. Population exposure levels in unmonitored areas were lower than in monitored areas. This is the first study that focused on developing a national-scale point wise exposure prediction approach in South Korea, which will allow national exposure assessments and epidemiological research to answer policy-related questions and to draw comparisons among different countries.

1. Introduction

There has been increasing evidence supporting the association between long-term exposure to air pollution and human health (Pope and Dockery, 2006; Hoek et al., 2013). Most cohort studies of long-term air pollution and health have particularly focused on traffic-related air pollutants such as particulate matter less than or equal to 10 or 2.5 μm in diameter (PM_{10} or $PM_{2.5}$), nitric oxide (NO), and nitrogen dioxide (NO_2) (Beelen et al., 2014; Kaufman et al., 2016; Laden et al., 2006; Pope et al., 2004).

One of the most important challenges faced by these studies has been the unavailability of individual-level air pollution measurements. Measurements are only available at limited numbers of monitoring sites. To overcome this limitation, many cohort studies have adopted exposure prediction models based on statistical approaches and regulatory air quality monitoring data, and estimated air pollution concentrations at participants' homes. The most common statistical modeling approaches employed have been regression methods, namely, land use regression and geostatistical techniques such as kriging (Hoek et al., 2008; Jerrett et al., 2005). Land use regression explains the spatial variability of air pollution using predictors, and kriging additionally incorporates spatial correlation structures. These models often considered hundreds of geographic variables, computed in the Geographic Information System (GIS), as predictors representing local or regional pollution

☆ This paper has been recommended for acceptance by Charles Wong.

* Corresponding author. Institute of Health and Environment, Seoul National University, Seoul, 08826, South Korea.
E-mail address: puha0@snu.ac.kr (S.-Y. Kim).

http://dx.doi.org/10.1016/j.envpol.2017.03.056
0269-7491/

sources. Either a limited number of selected variables or a few summary predictors estimated by dimension reduction were included (Eeftens et al., 2012; Sampson et al., 2011). In addition to geographic variables, other studies improved the prediction models by including satellite imagery data or air quality model outputs as predictors in the models (Kloog et al., 2011; Lindstrom et al., 2013; van Donkelaar et al., 2015; Young et al., 2016).

Although these exposure prediction models have mostly been developed for city areas (Eeftens et al., 2012; Hoek et al., 2008; Keller et al., 2015), some studies have expanded the modeling domains to a national scale while retaining the high spatial resolution (Hystad et al., 2011; Knibbs et al., 2014; Novotny et al., 2011; Sampson et al., 2013; Vienneau et al., 2010). There are large areas without regulatory air quality monitoring sites in the vicinity, which hamper the assessment of air pollution concentrations that the people residing in such areas are exposed to. This limitation also prevents resulting health analyses using the existing national-scale health data from large cohorts or government-generated databases. Developing national exposure prediction models would support exposure assessments and health analyses on a national scale to provide important policy-related information, such as national patterns of air pollution and health risks and areas that need to be targeted for air pollution reduction and that contain vulnerable populations.

Most national exposure prediction models have been developed in North America and Western Europe (Hystad et al., 2011; Novotny et al., 2011; Sampson et al., 2013; Vienneau et al., 2010). National prediction models that operate in different areas can provide insights into the differences and similarities between countries. The aim of this study was to develop a national-scale exposure prediction approach with fine-scale spatial variability using universal kriging and dimension-reduced predictors for estimating annual average concentrations of PM_{10} and NO_2 during 2010 in South Korea. We selected the year 2010 for our study, as there were sufficient numbers of monitoring sites available and a quinquennial population and housing census, a data source of a large portion of geographic variables, was conducted this year (Yi et al., 2016). To gain insights into the performance of our prediction models, we compared our primary approach with two alternative exposure prediction approaches. Using our exposure prediction models, we also investigated the distributions of the predicted PM_{10} and NO_2 concentrations across residential census output areas to assess population exposure levels.

2. Materials and methods

2.1. Data

2.1.1. Regulatory air quality monitoring data

We obtained hourly PM_{10} and NO_2 concentrations measured at the 294 regulatory air quality monitoring sites operating in 2010 in South Korea (48 million people in 100,033 km^2) from the National Institute of Environmental Research (Fig. 1). The South Korean regulatory air quality monitoring network consists of four types of monitoring site: urban background, urban roadside, regional background, and national background (MOE, 2011). Urban background sites were established for monitoring air pollution exposure level of the population and are located mostly in municipality buildings in highly populated areas with no major pollution sources in the adjacent areas. Urban roadside sites, located next to large and busy roads, focus on traffic-related air pollution. Regional background sites are located in rural areas, measuring the background level air pollution, while national background sites are located along the coastlines, where there are no domestic pollution sources, and therefore they are used for detecting air pollution that has been affected by neighboring countries. Because we aimed to develop prediction models based on pollution sources within the country, which were represented by geographic variables, the six national background sites were excluded.

Given hourly measurements of PM_{10} and NO_2, we calculated daily averages for the days when hourly measurements were recorded for more than 75% (18 h) of the day. Then, the annual averages were calculated at the sites meeting the following minimum inclusion criteria. For a site to be selected, it was required to have more than 75% (274 days) of daily data, at least one daily measurement in each of the 10 months, and no more than 45 consecutive days without daily measurements. After excluding the 11 sites that did not meet the inclusion criteria, the remaining 277 sites were used for our model development. The annual average concentrations of PM_{10} and NO_2 were natural log-transformed to approximate normality.

2.1.2. Geographic variables

To obtain geographic characteristics contributing to the spatial variability of PM_{10} and NO_2, we computed 322 geographic variables at each location using ArcGIS 10.2. The conceptual background of these variables, data sources, and detailed procedures for data processing and variable computation are provided in Eum et al. (2015). There were two types of variables (proximity and buffer) in eight categories representing plausible pollution sources. The list of geographic variables is shown in Table S1. The eight categories were traffic, demographic characteristics, land use, physical geography, transportation facilities, emissions, vegetation, and altitude. The source data, including road networks, census, and vegetation index were mostly collected or generated in 2010. As an exception, the land use data were derived from more than 800 maps completed in 2007 and updated in 2009 for certain urban areas (Eum et al., 2015). The proximity variables were computed as the distance to the nearest feature (e.g., major road/airport), while the buffer variables were aggregates or summaries (e.g., mean, median, or percentage) of a characteristic feature (e.g., population/road length) within a circular buffer area. We excluded variables with little variability and recoded some geographic variables to effectively reflect the relationships with air pollution. Variables were excluded when more than nine-tenths of data points were the most common values. Furthermore, land use variables, calculated as percentage values, were eliminated when their maximum values were less than 10%. The number of geographic variables after exclusion decreased from 322 to 284. All the distance variables were log transformed.

2.1.3. Census output area

To assess population exposure levels, we used census output areas which are the smallest territorial units for which census data are available in South Korea (Kang et al., 2007). In 2010, there were 83,463 census output areas, with a median area of 0.02 km^2 (mean of 1.21 km^2) and an average of 572 residents. Previous studies predicted air pollution concentrations at the centroids of census tracts or mesh areas to represent population exposure (Hystad et al., 2011; Knibbs et al., 2014). However, these centroids may fall in areas where few people live. To ensure that the residential locations are well represented, we selected the largest residential area in each census output area and used the centroid of that area. The boundaries of the census output areas were obtained from the Statistical Geographic Information System of Statistics Korea (http://sgis.kostat.go.kr). Land cover maps indicating residential areas were obtained from the Environmental Geographic Information Service (http://egis.me.go.kr) operated by the South Korean Ministry of Environment. The residential areas were identified by using image-classification algorithms on converted land surface