Environmental Modelling & Software 103 (2018) 146–157

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Estimating individualized exposure impacts from ambient ozone levels: A synthetic information approach

Bianica Pires [a], *, Gizem Korkmaz [a], Katherine Ensor [b], David Higdon [a], Sallie Keller [a], Bryan Lewis [a], Aaron Schroeder [a]

[a] Biocomplexity Institute of Virginia Tech, Arlington, VA 22203, United States
[b] Department of Statistics, Rice University, Houston, TX 77251, United States

ARTICLE INFO

Article history:
Received 11 July 2017
Received in revised form
19 January 2018
Accepted 11 February 2018

Keywords:
Synthetic populations
Air quality
Ozone
Microenvironment
Personal exposure

ABSTRACT

There is ample evidence that short-term ozone exposure is associated with increased respiratory symptoms. Many studies, however, aggregate the population, activities, or concentration levels of the pollutant across space and/or time, failing to capture critical variations in the exposure levels. We couple spatiotemporal air quality estimates of ozone with a synthetic information model of the Houston Metropolitan Area, allowing us to attach exposure levels to individuals based on exact times, geo-locations, and microenvironments of activities. Several scenarios of the model are run at different levels of resolution. When we maintain the spatiotemporal resolution of the data, the proportion of the population that experiences sharp increases in short-term exposure increases substantially. This can be particularly important if experienced by sensitive populations given the increased risk for adverse health effects. We find that individuals in the same zip code, neighborhood, and even household have varying levels of exposure.

1. Introduction

Exposure is the contact an individual has with a pollutant — a function of both the concentration of the pollutant in the environment and the time an individual is exposed to that pollutant (US EPA, 2011). Localized specific exposures to ozone can dramatically increase health risks for cardiac events and asthma (Ensor et al., 2013, 2014; Raun et al., 2014; Davis and Ensor, 2006). For instance, each 20 parts per billion (ppb) increase in ambient ozone (O_3) concentration in the previous one to three hours is associated with a 4.4% increased risk of having an out-of-hospital cardiac arrest, which is particularly significant given that 90% of these cases result in death (Ensor et al., 2013). Accurate representations of the magnitude, frequency, and duration of localized exposure to a pollutant requires that we account for individual activity patterns across both space and time (Vallero, 2014). Many studies, however, average the activities of individuals in time and space into 12- or 24-h summaries (Klepeis et al., 2001; Leech et al., 2002; Matz et al., 2014). Summarizing the data in this way may miss important variations in ozone exposure across a population.

In this paper, we couple spatiotemporal air quality estimates of ozone (O_3 ppb-hours) across the Houston Metropolitan Area of Texas to a data-informed synthetic information model. The synthetic information model is a simulated population that is representative of the true population of the Houston Metropolitan Area, including the composition of households, the demographics of individuals, and their movement throughout the course of the day (Adigaa et al., 2015; Parikh et al., 2013). We maintain the spatiotemporal resolution of the data such that individual exposure estimates may vary on a second-by-second basis as synthetic individuals move across the geography, entering and leaving geo-located microenvironments each with their own unique ozone concentration estimates. It overcomes the limitations of traditional approaches as it is informed by how people move through their activities during the day, allowing us to attach specific exposure levels to the synthetic individuals based on the exact geo-location of the activity and time of day. Our research has been motivated by the many studies (Ensor et al., 2013, 2014; Raun et al., 2014) that have shown that the resolution of the data is an important consideration as it can impact important health effects that could be translated into lifesaving behavioral and policy changes.

Exposure modeling has been an active research area from a

* Corresponding author.
E-mail address: bpires@vt.edu (B. Pires).

https://doi.org/10.1016/j.envsoft.2018.02.007
1364-8152/

variety of different research approaches (Duan, 1982; Jerrett et al., 2005; Ryan and LeMasters, 2007; Zou et al., 2009). Community-based studies (Rosenthal et al., 2008; Silverman et al., 2010; Wellenius et al., 2012) identify a significant impact on health from air pollution levels but do not directly measure individual exposure. Early research using microenvironment monitoring models, for instance, assigned exposure concentrations based on time spent within different microenvironments (e.g., indoor locations, outdoors) but did not assign these microenvironments to individual-level activity patterns (Fugas, 1975). Extending this work, models incorporated activity diaries allowing for some estimates of the variability and distribution of individual exposure, but did not account for temporal variations in concentration levels (Ott and Flachsbart, 1982). Other models attach air pollution exposure to populations at a group-level, based on the demographics of a subset of individuals, the geographic location of homes and activities, or a set of microenvironments. While some of these studies model representative individuals (Kousa et al., 2002; Burke et al., 2001; Özkaynak et al., 2008), they either stop short in their ability to trace individuals throughout the course of the day or in modeling a representative population of the geographic location in question. A subset of exposure models have focused on calculating personal exposure to emissions and other pollutants while traveling, accounting for factors such as transportation mode, vehicle type, and transportation routes (Hatzopoulou et al., 2011; Hülsmann et al., 2011). While these models seek to accurately reflect exposure during travel, they do not account for the individual's full daily course of activities.

Moving away from aggregate-level exposure calculations and accounting for an individual's daily activities, more recent exposure models have developed synthetic populations to represent each individual in a geographic location. For instance, the Environmental Protection Agency (EPA)'s Air Pollutants Exposure (APEX) model develops a synthetic population that characterizes the study area and utilizes the Consolidated Human Activity Database (CHAD), a repository of harmonized human activity data, to assign activity patterns to synthetic individuals (US EPA, 2012). In this model, a person's exposure is obtained by mapping the activities reported in the survey into several microenvironments (e.g., indoors-residence), each with an estimated exposure rate. Geography is measured at the level of Census tracts (a geographic region with a population size between 1,200 and 8,000 people). Activities are assigned in hourly chuncks and individuals may move from their assigned home tract only for work. Other synthetic population models are limited in their use for estimating environmental exposure to contaminants that vary over both space and time. These methods aggregate activities into percent time (e.g., percent of day) and allocate the aggregated time to an activity location (Lenormand and Deffuant, 2013; Namazi-Rad et al., 2014; Wheaton et al., 2009; US EPA, 2014a). As an example, a study of Sydney, Australia uses traditional synthetic population models with single daily exposure values, and couples the percent time spent at various locations to these daily average exposure levels (Newth, 2012).

Other studies focus on direct measurement of human exposure through personal monitors or home-based centers but cost, measurement accuracy, and logistics limit the use of this approach on a scale large enough to provide continuous community-wide understanding of exposure (Weisel et al., 2005; US EPA, 2015). More recent studies have used personal monitoring systems as a way to estimate regional concentration levels of ozone and other air pollutants (Xu et al., 2017). For instance, Nikzad et al. (2012) used sensors and a mobile phone application to collect data from 16 participants in San Diego, California over the course of a month. Regional estimates derived from this data were then compared to levels measured by EPA monitors across the county. While results are promising for generating air quality estimates, these studies do not account for the individual activity sequences of the population.

Our contribution is the development of an *in silico* analytics platform that uses the synthetic information model to understand the impact of air quality over physical space and time on individual exposure levels. We estimate hourly ozone levels for the geo-coordinates (latitude, longitude) associated with the activity locations in our synthetic information model and allow synthetic individuals to move across this geography based on their second-by-second time sequenced activities. We combine individual activity patterns with the microenvironment modeling approach, assigning each geo-located activity location hourly exposure rates over a set of microenvironments. Few studies consider human activity patterns in detail at the population level due to availability of data to accurately generate representative populations and their movement and the computational burden associated with working with such vasts amounts of data. Coupling this information to pollutant levels that vary spatiotemporally and by microenvironment adds another level of difficulty. We overcome such computational challenges by applying high performance computing and database management techniques and expand on current exposure models by integrating data sources across many publicly and commercially available datasets. This allows us to trace ozone exposures of each synthetic individual as they move throughout the course of the day. The spatiotemporal resolution of the data handled by this platform provides additional flexibility for comparing results across different sets of assumptions.

2. Methodology

In this section we describe the *in silico* analytics platform we developed, which couples a synthetic information representation of the residents in the Houston Metropolitan Area to spatiotemporal air quality estimates. We begin by describing the synthetic information model developed at Virginia Tech (Marathe et al., 2014) (Section 2.1). It includes socio-demographically relevant activity sequences and the movement of each individual in the population through their sequences second-by-second during the day. This allows aggregation of time intervals to match the environmental quality data (e.g., hourly intervals) (Section 2.2). We then discuss the methodology used to determine individual-level exposures to ozone (Section 2.3). The output of the platform is the exposure profiles for the roughly 4.9 million synthetic individuals in the Houston Metropolitan Area. Fig. 1 illustrates the conceptual model of the *in silico* analytics platform. This platform provides an integrated database that can be used and reused for the analysis of various studies related to the synthetic information model and air quality. Moreover, given the fidelity of the data the platform can process, we have the flexibility to explore results under various sets of assumptions.

2.1. Synthetic information model

The first step in creating the *in silico* analytics platform is to generate the synthetic information model for the Houston Metropolitan Area. The synthetic information model is a set of synthetic people, each associated with demographic variables, located geographically at specific points in time and place, such as homes and schools, each associated with specific geo-locations. It is created by integrating a variety of databases from commercial and public sources, including statistical surveys, administrative data, and data on the built environment (e.g., buildings, roads, and land use), through a process that preserves the confidentiality of the individuals in the original data sets, yet produces realistic attributes