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

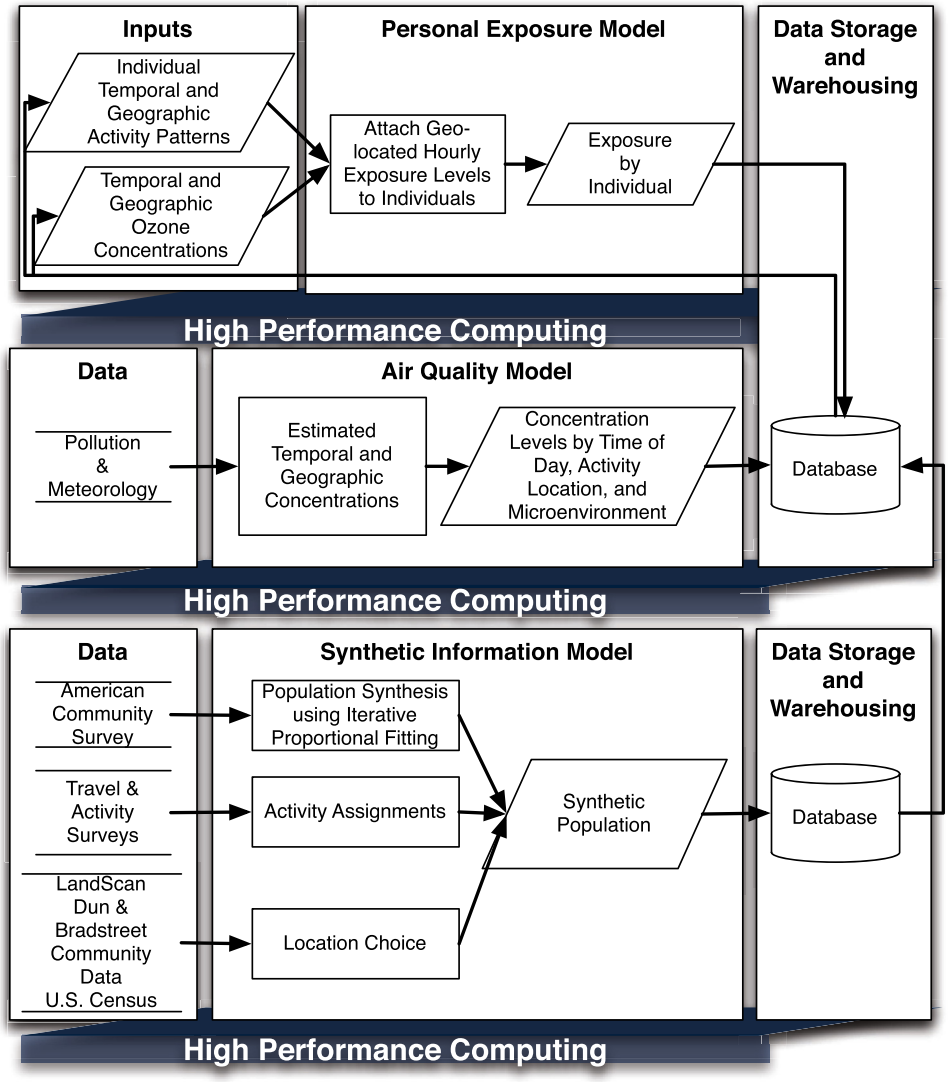

Fig. 1. Conceptual model of the *in silico* analytics platform. The bottom layer represents the synthetic information model, the middle layer represents the air quality estimates, and the top layer represents the personal exposure model.

and demographics for synthetic individuals (Adigaa et al., 2015). The synthetic information model approach has been used to study a wide range of phenomena to support policy decision-making (Barrett et al., 2006; Marathe et al., 2014), including transportation, epidemics (Parikh et al., 2013), and natural and man-made disasters (Barrett et al., 2005; Halloran et al., 2008; Lewis et al., 2013; Marathe and Vullikanti, 2013).

The steps to the generation of the synthetic information model include (1) population synthesis, in which a synthetic representation of each individual and household in a region is created using socioeconomic characteristics from Census data, (2) activity and microenvironment assignments, in which each synthetic individual is assigned a set of activities to perform during the day, along with start and end times based on activity or time-use survey data; and (3) place choice, in which an appropriate location is chosen for each activity for every synthetic individual based on several data sources, including land use patterns and tax data. The main data sources used in the creation of the Houston synthetic information model are described in Table 1.

The American Community Survey (ACS) provides tables of distributions at the Census block group level (a geographical region containing 600 to 3000 individuals) on demographic characteristics, such as age, gender, household income, and household size, which are referred to as marginal distributions. Joint demographic distributions are reconstructed from these marginal distributions using Public Use Microdata Sample (PUMS) data, which provides a 5% representative sample of the ACS data. PUMS information are incorporated into the joint distribution through an iterative proportional fitting technique (Beckman et al., 1996). The process ensures that the synthetic population, including individual demographic characteristics and household composition, match the structure of the true population. Each synthetic household is then located geographically using land-use data and data pertaining to businesses and transportation networks. Realistic activity patterns and their associated activity locations are then added. The activities performed by each synthetic individual on an average day is determined by analyzing the activity patterns in the National Household Travel Survey and linking these patterns to the socio-demographic composition of the households (e.g., income) and individuals within the households (e.g., age, gender). This is done

Table 1
The main data sources used to generate the Houston Metropolitan Area synthetic information model.

Data Source	Description
US Census TIGER (Topologically Integrated Geographic Encoding and Referencing) Data	True population geographic boundaries and demographics to be matched by synthetic population.
US Census American Community Survey (ACS)	Primary source of data used to build the synthetic set of people with the aggregate statistics matching US Census marginals.
Consolidated Human Activity Database (CHAD)	Data on activity sequences by activity location types (e.g., bedroom, office building).
National Household Travel Survey (NHTS)	Data on travel behavior and activity sequences.
Dun & Bradstreet (D&B)	Describes home locations and retail locations used to locate activities.
HERE (formerly NAVTEQ)	Road Network and transportation map.
National Center for Education Statistics (NCES)	Data on school locations.

using a process known as the Fitted-Values Method (Lum et al., 2016). This method preserves within household activity correlations and ensures that activity sequences of synthetic individuals are similar to actual individuals with similar socio-demographic characteristics and household membership.

Activities are further characterized by one or more microenvironments. A microenvironment is defined as a three-dimensional space that is relatively homogeneous with respect to pollutant concentrations for a specified time period (U.S. EPA, 2007). We model 11 microenvironments:

- 6 indoor microenvironments (indoors-residence, indoors-school, indoors-office, indoors-shopping, indoors-bars and restaurants, indoors-other);
- 3 outdoor microenvironments (outdoors-near road, outdoors-public garage or parking, outdoors-other); and
- 2 in-vehicle microenvironments (in-vehicle-cars or trucks, in-vehicle-mass transit).

We adapted the EPA's (U.S. EPA, 2007) mapping of activity location types (e.g., bedroom, park/golf course, office building) in CHAD to microenvironments. Moreover, to link these activity location types to our synthetic information model we created a crosswalk between activities and locations in the synthetic information model to one or more microenvironments. For instance, the activity "home" can be associated with up to two microenvironments – indoors-residence and outdoors-other. Using data from CHAD, we developed empirical distributions to represent the proportion of time spent within each microenvironment while performing a specified activity. We then sampled from these distributions to get time (in seconds) spent within each microenvironment while performing an activity.

The Houston Metropolitan Area synthetic information includes any synthetic individual with either a home or activity location within Harris County, Texas. This resulted in approximately 4.9 million individuals grouped into 1.8 million households. Synthetic individuals can perform up to 6 different types of activities, including staying home, going to work, going to school, shopping, traveling, and other (e.g., socializing at a friend's residence, going to a restaurant); activity types can be performed multiple times by the same individual on the same day. Activities occur in 1.2 million different activity locations (895 thousand housing locations, 166 thousand work locations, 37 thousand shopping locations, 4 thousand schools, and 112 thousand other locations). Fig. 2 shows the household and activity locations across the region, as well as an example set of activities for one household.

2.2. Air quality estimation

According to the Texas Commission on Environmental Quality (TCEQ), the air quality in Houston, Texas is monitored more closely and analyzed with more intensity than perhaps anywhere in the country — if not in the world (Texas Commission on Environmental Quality, 2005). Ozone is measured hourly at 47 sites (US EPA, 2012). One hour ambient meteorological (temperature, relative humidity, and wind speed) data from the monitors are publicly available. For this demonstration, hourly recordings on August 26, 2008, of ambient ozone concentration levels (O_3 ppb-hours) across 39 EPA monitors in the Houston Metropolitan Area is used.[1]

The first step to coupling the environmental pollutant data with the synthetic information model is to assign hourly ambient ozone concentration levels to each of the 1.2 million activity locations. This is done using the inverse weighted distance from each location to the 39 monitors, a standard method for assigning ozone concentration (Hwang and Jaakkola, 2008; Lu and Wong, 2008; Salam et al., 2005). Given that $m_j(t)$ is the concentration measured by monitor $j \in \{1, \ldots, n\}$ at hour h, the pollutant concentration c_l at each activity location $l \in \{1, \ldots, L\}$ at hour $h \in \{1, \ldots, 24\}$ is calculated as follows:

$$c_l(h) = \sum_{j=1}^{n} \frac{\frac{1}{d_{lj}^2} m_j(h)}{\sum_{j=1}^{n} \frac{1}{d_{lj}^2}}, \quad (1)$$

where d_{lj} is the great circle distance between location l and monitor j (Nychka et al., 2015). The next step is to compute concentration levels by microenvironment for each activity location and for each hour of the day (see Section 2.1). We apply one of two methods used for estimating concentration levels within microenvironments: (1) mass balance and (2) factors (U.S. EPA, 2007).

The mass balance method assumes ozone concentration is spatially uniform at any specific time and is used to compute concentration levels for all the indoor microenvironments (e.g., indoors-residence, indoors-schools). It is estimated using the following parameters (U.S. EPA, 2007):

- Air exchange rate R_{ae} is the rate that air flows in and out of a microenvironment.
- Decay rate R_{de} is the removal rate of ozone from a microenvironment due to deposition, filtration, and chemical reaction.

The decay rate $R_{de,l}$ at a given activity location is a function of the ambient concentration level $c_l(h)$. The air exchange rate R_{ae}^{ME} at a given activity location is a function of the ambient concentration level $c_l(h)$, the microenvironment (e.g., indoors-residence, indoors-school), the average outdoor temperature, and the presence of air conditioning (US EPA, 2014b). We sample from lognormal

[1] We selected August 26, 2008 because it represents a warm day in Houston with time varying levels of ozone and is within the same time span of the study performed by Ensor (Ensor et al., 2013). On August 26, 2008, 39 of the total 47 monitors recorded ozone readings.

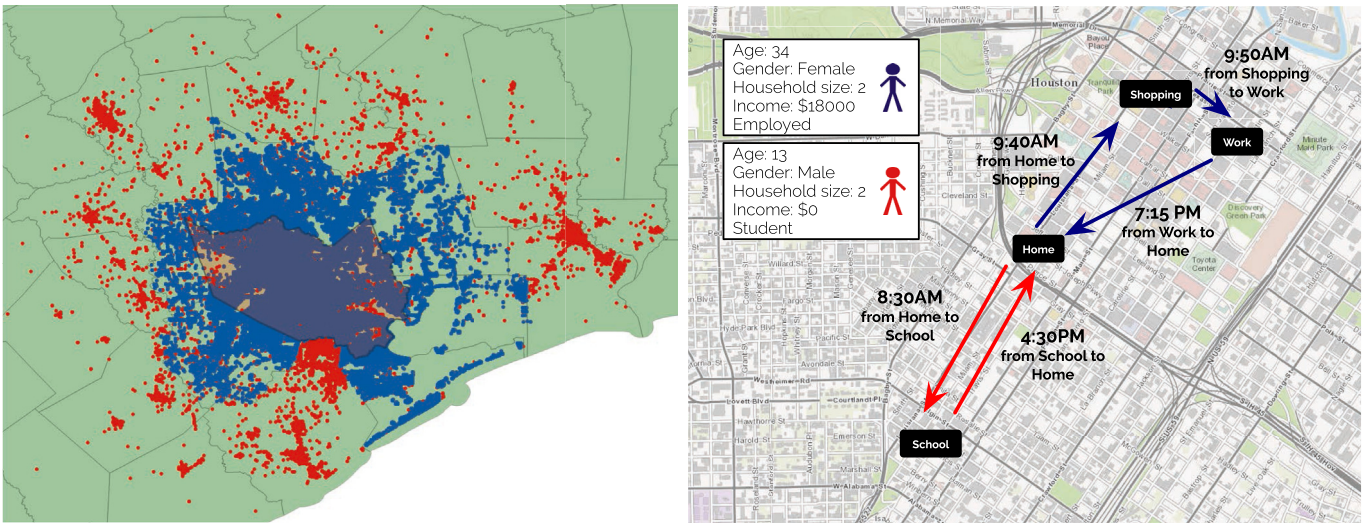

Fig. 2. Activity Locations. (left) The red points represent the 1.2 million activity locations for the synthetic population and the blue points represent the 895 thousand housing locations. The shadow is the boundary of Harris County, Texas. (right) A "close-up" of the activity patterns on an average day for a two-person household. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

distributions for Houston provided by the EPA (US EPA, 2014b) to assign $R_{de,l}$ and $R^{ME}_{ae,l}$ to each of the indoor microenvironments within each activity location. The concentration level $c^{ME}_l(h)$ at an activity location l at time h for microenvironment ME is calculated as follows (U.S. EPA, 2007):

$$c^{ME}_l(h) = \frac{\Delta c^{in}_l(h)}{R_{combined}} + \left(c^{ME}_l(h-1) - \frac{\Delta c^{in}_l(h)}{R_{combined}} \right) \frac{1 - e^{-R_{combined}}}{R_{combined}},$$
$$\text{where } \Delta c^{in}_l = c_l(h) * R^{ME}_{ae,l},$$
$$R_{combined} = R^{ME}_{ae,l} + R_{de,l} \tag{2}$$

The factors method is simpler and is used to estimate concentration levels during in-vehicle travel. The penetration factor f_{pe}, which represents the fraction of ozone entering the microenvironment from the outside via air exchange, is the only parameter used. We assign $f_{pe,l}$ to activity locations by sampling from a normal distribution (US EPA, 2014b). We compute in-vehicle concentration levels using the following equation (U.S. EPA, 2007):

$$c^{ME}_l(h) = c_l(h) * f_{pe,l} \tag{3}$$

Concentration levels for outdoor microenvironments (e.g., outdoors-other, outdoors-near road) are assumed to be equal to the estimated ambient concentration level $c_l(h)$ (U.S. EPA, 2007).

2.3. The personal exposure model

The personal exposure model is developed by coupling the individual activity patterns from the synthetic information model to the air quality estimates. Specifically, we have point-in-time movement of synthetic individuals across their geo-located activity locations and microenvironments. Additionally, we have point-in-time concentration levels by geo-located activity location and microenvironment. We map the point-in-time movement for each synthetic individual to the corresponding ozone concentration level at that time, geo-location, and microenvironment. Note that the assignment of concentration exposure to travel is made by splitting the travel time between the origin and destination activity locations. Therefore, travel exposure is calculated based on the concentration of ozone at the origin and destination activity locations during the time of travel.

Synthetic individuals may come in contact with varying concentration levels depending on their movement across activity locations and between microenvironments within the hour. Therefore, we compute average hourly concentration exposure $e_i(h)$ for synthetic individual i by taking the weighted average of the concentration levels over the course of hour h. The output is the exposure matrix $\mathbf{E} \in \mathbb{R}^{H \times I}$, where H is the number of hours in a day (24) and I is the number of synthetic individuals (approximately 4.9 million).

3. Results

In this section we present the results of our model, particularly focused on investigating the heterogeity of individual, household, and neighborhood-level exposures across the study area (Section 3.1). We then explore results obtained by running several scenarios of the model assuming different levels of spatiotemporal resolution (Section 3.2).

3.1. Exposure heterogeneity

We explore results of the model with particular focus on the spatiotemporal variation of exposures across the study area. The results shown here apply the methodology discussed in the previous section. We therefore assume that individuals move across the environment using their second-by-second activity sequences, ozone concentration levels vary hourly across a 24-h period, and ambient concentration levels are adjusted for indoor, in-vehicle, and outdoor microenvironments.

The 24-h average exposure concentrations by zip code across the entire synthetic population with homes in Harris County are shown in Fig. 3. The lighter shaded areas represent zip codes with relatively low average exposure levels while the darker shaded areas represent zip codes with relatively high exposure levels. We find that the population living in the northeast region of the county generally experiences lower levels of ozone exposure while the southeast and central west regions experience some of the highest exposure levels. Even when averaged at the zip code level, exposure

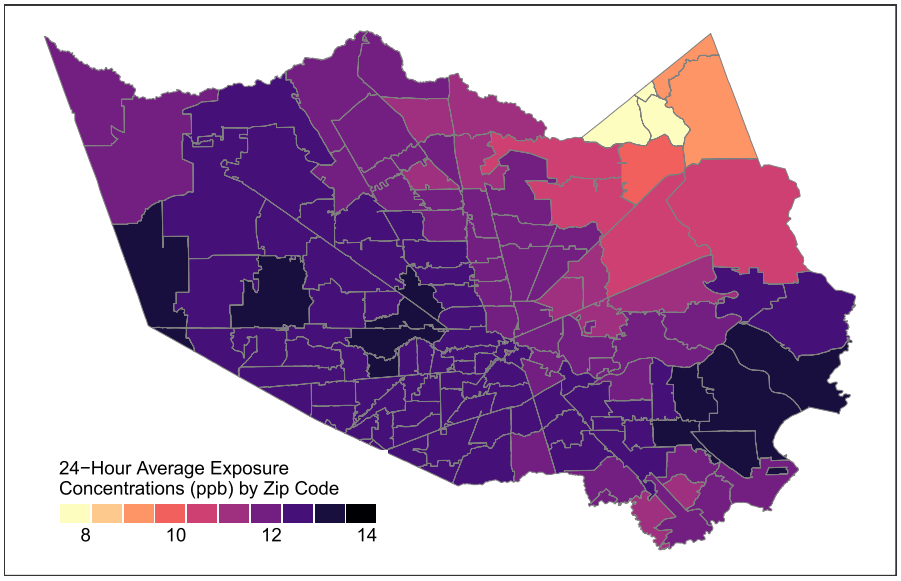

Fig. 3. 24-h average exposure concentrations by zip code on August 26, 2008. The lighter shaded areas represent zip codes with relatively lower average exposure levels. The darker shaded areas represent zip codes with relatively higher average exposure levels.

levels are not homogenous across the county suggesting that where you live is an important consideration in estimating individual exposure levels.

We can drill in further and look at 24-h average exposure levels by household. Fig. 4 illustrates "close-ups" of two neighborhoods. Each square represents a household and the colors indicate the households level of exposure (darker squares represent households with higher average exposure levels, lighter squares represent households with lower average exposure levels). The neighborhood on the right is in the downtown area of Houston and the neighborhood on the left is in the northeast region of the county. As expected, households in the northeast neighborhood have generally lower average exposure levels. However, we can more clearly see here that there is heterogeneity within the small geographic area. There are households in the northeast neighborhood that experience exposure levels just as high as those in downtown Houston, and conversely, there are households in downtown Houston that experience exposure levels as low as many of those in the northeast region. This demonstrates that where you go during the day should be considered when estimating individual levels of exposure.

The platform further allows us to trace individuals over the

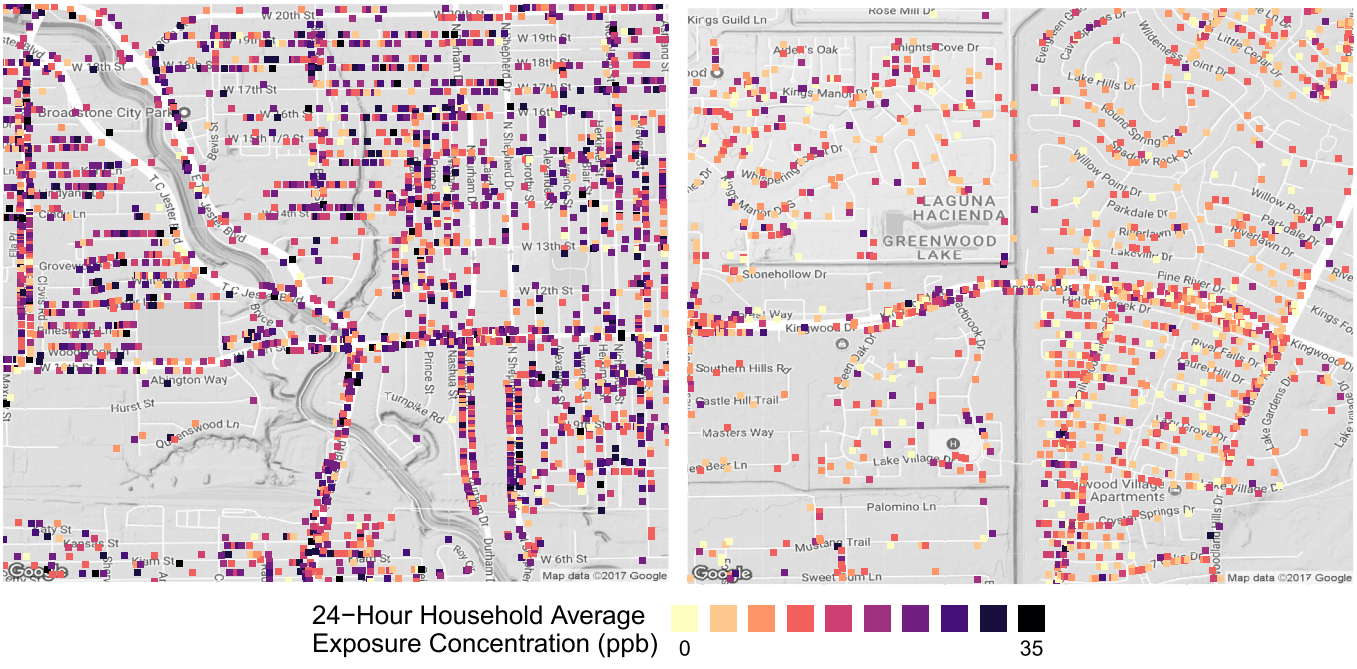

Fig. 4. 24-h average exposure concentrations for individual households in two neighborhoods. (left) A neighborhood in downtown Houston where ozone levels were generally high. (right) A neighborhood northeast of downtown Houston where ozone levels were generally low.

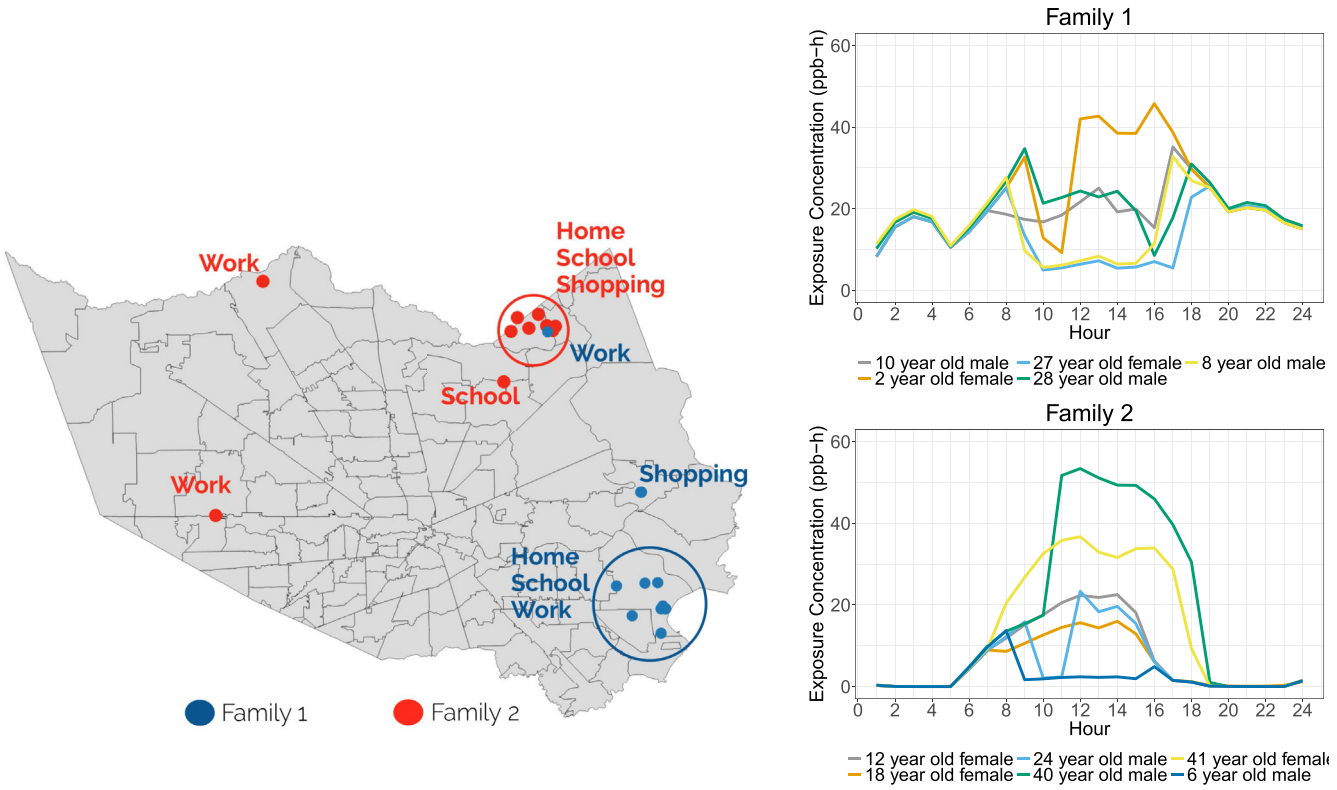

Fig. 5. Hourly average ozone exposure concentrations for two synthetic families located in different Houston neighborhoods on August 26, 2008. The 24-h average exposure level for Families 1 and 2 is 18.6 ppb and 8.8 ppb, respectively.

course of the day. Hourly average exposure concentrations for two synthetic families are given in Fig. 5. These are two demographically similar families located in different Houston neighborhoods (Family 1 is located in the southeast and Family 2 is located in the northeast). This demonstrates that ozone exposure is heterogeneous even between members within a household, illustrating the significance of maintaining the resolution of the data at the individual level when deriving exposure estimates.

Finally, we explore results by demographic subpopulations. For illustrative purposes, we look at individuals by age group across two zip codes: zip code 77339, where concentration levels were relatively low (median concentration levels were 8.9 ppb-h), and zip code 77505, where concentration levels were relatively high (median concentration levels were 30.3 ppb-h). Fig. 6 shows the 1-h average exposure concentrations across different age groups for the synthetic residents of the low concentration (LC) zip code (blue) and the high concentration (HC) zip code (pink). The lines represent median hourly exposure levels, the dark shaded areas represent the 25th and 75th quartiles, and the light shaded areas represent the remaining individuals. As expected, across all age groups the majority of synthetic individuals in LC are exposed to lower ozone levels than those in HC. However, while median exposures are lower in zip code LC, some individuals in this zip code reach exposure levels just as extreme as in zip code HC, which has some the highest levels of exposure in the region. Within each zip code, we also find some trends. In LC, adults between the ages of 18 and 64 have the highest exposure levels. This makes sense since they are more likely to travel outside of their home zip code to go to work, where concentration levels may be much higher. On the other hand, we see no meaningful differences between average exposures across the ages in HC.

3.2. Scenario evaluation

In this section, we run three scenarios of the model to explore the potential impact that modeling at different levels of spatio-temporal resolution has on individual-level exposures across a population. Scenario 1 assumes individuals stay home all day and calculates exposure by using geo-located hourly concentration levels. Scenario 2 moves the individuals through their time activity sequences for the day and uses geo-located 24-h average concentrations. Scenario 3 (as described in Sections 2 and 3.1) moves the individuals through their time activity sequences for the day and uses geo-located hourly concentration levels.

Daily (24-h) average mean and peak exposure concentrations across the entire synthetic population for the three scenarios are shown in Fig. 7. Lighter colors represent zip codes with lower average exposure levels while darker colors represent zip codes with higher average exposures. Within each scenario and across both average and peak exposures we see similar spatial trends – the northeast region shows relatively lower average exposure levels and the southeast and central west areas show relatively high average exposure levels. Examining average exposures (shown on top), we find that when individuals are assumed to stay home (Scenario 1) estimated average exposure concentrations are lower than when individuals are allowed to move spatially according to their activity sequences. As a consequence, average exposure levels may be underestimated when making this simplifying assumption. On the other, examining average peak exposures (shown on the bottom) we find that assuming daily average concentration levels (Scenario 2) results in comparably lower estimates of average peak exposure. This makes sense since this scenario eliminates any temporal variations (such as spikes and dips) in ozone levels. Scenario 2 is likely limited in its ability to identify cases with significant short-term peak exposures.

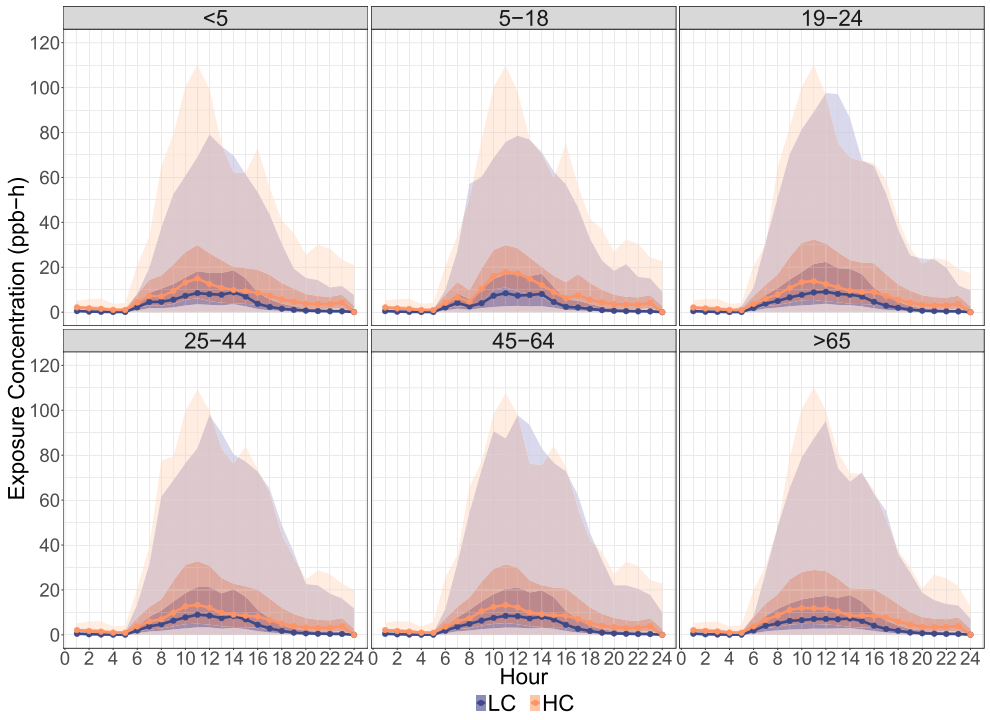

Fig. 6. Hourly average exposure concentrations across different age groups for two zip codes: zip code 77339, where ozone concentration levels were generally low (LC) and zip code 77505, where ozone concentration levels were generally high (HC). The lines represent the median hourly exposure levels, the dark shaded areas represent the 25th and 75th quartiles, and the light shaded areas represent the 10th and 90th quartiles.

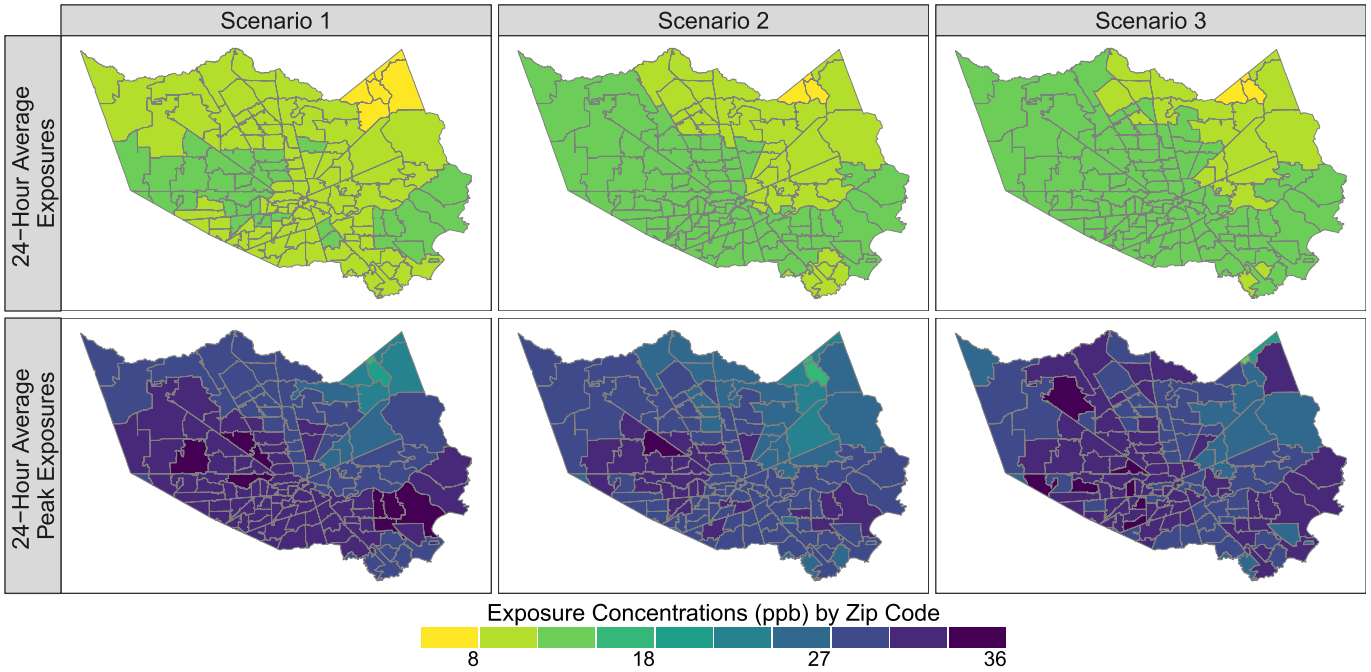

Fig. 7. The 24-h average exposure concentration of residents on August 26, 2008 across the three scenarios. Scenario 1 allows geo-located concentrations to vary hourly and assumes individuals stay home. Scenario 2 uses 24-h average geo-located concentrations and assumes individuals move through their time-sequenced activities. Scenario 3 allows geo-located concentrations to vary hourly and assumes individuals move through their time-sequenced activities. (top) Average exposure concentrations by zip code in Harris County. (bottom) Average peak exposure concentrations by zip code in Harris County.

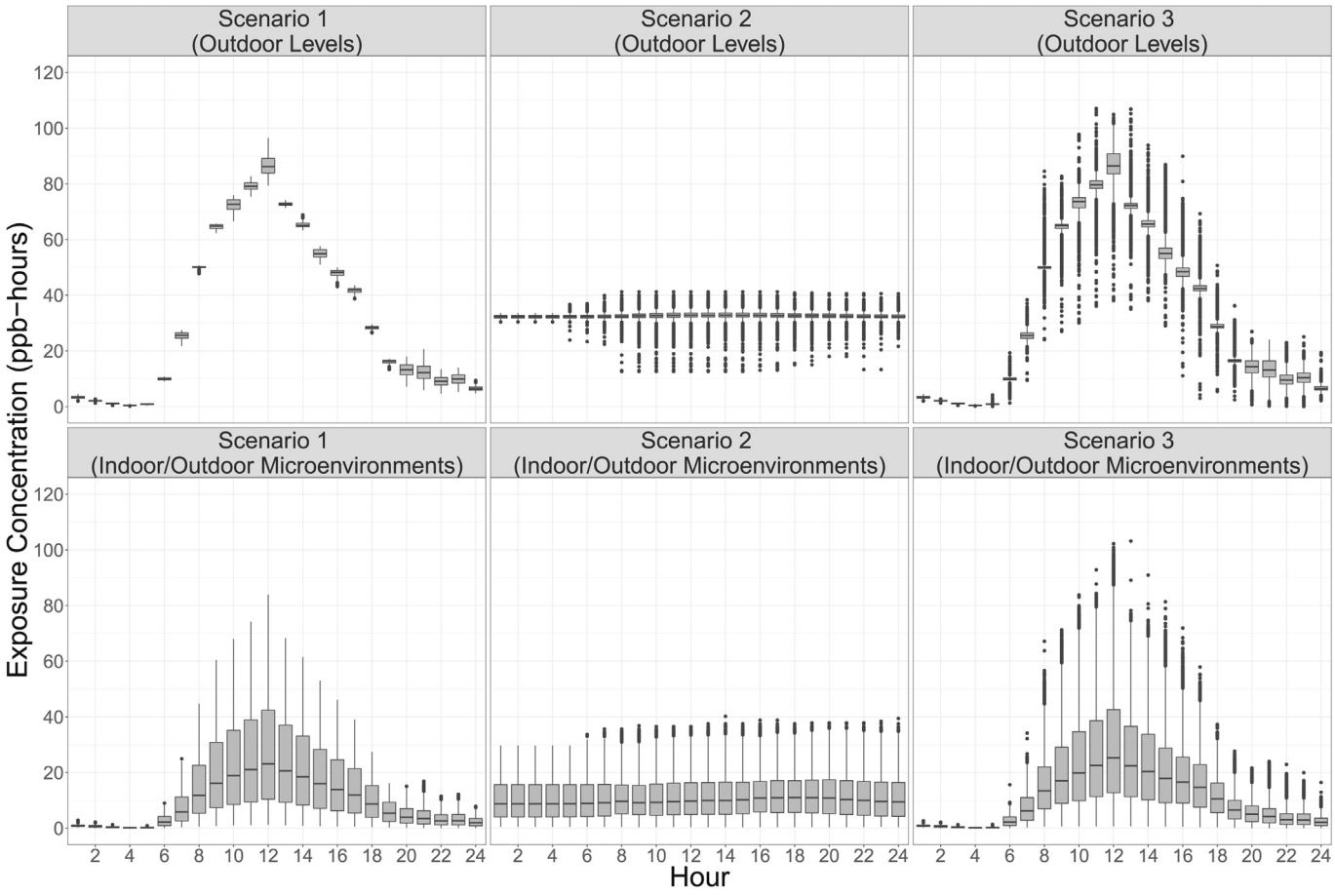

Fig. 8. Hourly average exposure concentrations for the synthetic population in zip code 77026 in Houston calculated according to three scenarios, using ozone levels on August 26, 2008. Median ozone exposure is represented by the line in the middle of the box. The box represents the inter-quartile range (the middle 50% of individual ozone exposure levels). The vertical lines above and below the box represent the upper (75%) and lower quartiles (25%) of exposure levels. The whiskers represent the remaining values. Scenario 1 allows geo-located concentrations to vary hourly and assumes individuals stay home. Scenario 2 uses 24-h average geo-located concentrations and assumes individuals move through their time-sequenced activities. Scenario 3 allows geo-located concentrations to vary hourly and assumes individuals move through their time-sequenced activities. (top) Ambient ozone concentration levels are used as a proxy for ozone exposure. (bottom) Ozone concentration levels are adjusted for indoor, in-vehicle, and outdoor microenvironments.

We can also explore temporal variations in ozone exposure across a population. Fig. 8 gives one-hour average exposure traces for the synthetic population in a Houston neighborhood (zip code 77026) for one day under the three scenarios. The top three charts illustrate results when we use ambient ozone concentration levels as a proxy for personal exposure and the bottom three charts show results when ozone concentration is adjusted for indoor, in-vehicle, and outdoor microenvironments. Hourly median ozone exposure is represented by the line in the middle of each box. The box represents the inter-quartile range (the middle 50% of individual ozone exposure levels), the vertical lines above and below the box represent the upper (75%) and lower quartiles (25%) of exposure levels, and the whiskers represent the remaining values (or extremes). We find that Scenario 1 does not capture the extreme values, which can be particularly important if the highest levels are experienced by already vulnerable populations. Scenario 2, on the other hand, does not capture temporal variations in ozone concentration throughout the course of the day, demonstrating that time sensitive exposures may not be captured. Note that while median exposure levels are generally lower when we account for indoor and outdoor behaviors, the variability of exposure levels across the population is much higher, signifying that some individuals will experience exposure levels as high as in runs where only outdoor levels are modeled. For instance, exposure under Scenario 1 when using outdoor levels looks quite homogeneous across the population. This makes sense since individuals shown live in the same neighborhood and are assumed to stay home all day. When we account for their indoor and outdoor behaviors, however, we find that at peak hours there is nearly a 85 ppb-h difference between individuals experiencing the lowest levels of exposure to ones experiencing the highest. When synthetic individuals are allowed to move (Scenario 3), this range increases to 100 ppb-h.

As discussed in Section 1, previous studies have found that each 20 ppb increase in ozone over a period of one to three hours is associated with a 4.4% increased risk of having an out-of-hospital cardiac arrest (Ensor et al., 2013). We examined whether modeling at finer spatiotemoral resolutions could better capture these potential risks. We computed the one-hour lagged change in ozone exposure for all individuals in zip code 77026 (the same population as shown in Fig. 8). Fig. 9 shows the highest change in ozone exposure experienced by each individual across the three scenarios. When we use outdoor levels only, we find that nearly all individuals under Scenarios 1 and 3 experienced at least once an increase of 20 ppb-h (or higher) in ozone exposure over the course of the day. On the other hand, few synthetic individuals

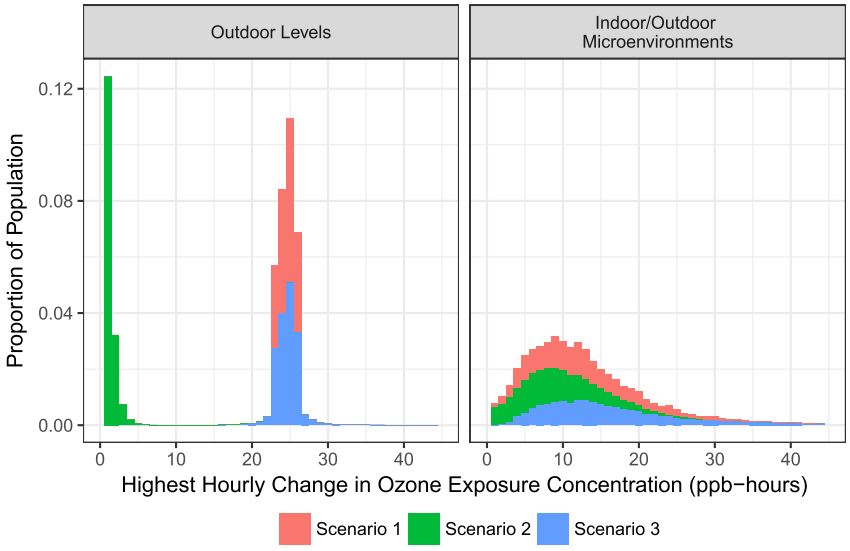

Fig. 9. The highest change in hourly ozone exposure experienced by each individual across the three scenarios, using ozone levels on August 26, 2008. Scenario 1 allows geo-located concentrations to vary hourly and assumes individuals stay home. Scenario 2 uses 24-h average geo-located concentrations and assumes individuals move through their time-sequenced activities. Scenario 3 allows geo-located concentrations to vary hourly and assumes individuals move through their time-sequenced activities. (left) Ambient ozone concentration levels are used as a proxy for ozone exposure. (right) Ozone concentration levels are adjusted for indoor, in-vehicle, and outdoor microenvironments.

experienced any short-term peak exposures and only one was over 20 ppb-h under Scenario 2. When we adjust for indoor, in-vehicle, and outdoor microenvironments we find that almost half of the population experiences one-hour exposure changes of less than 10 ppb under Scenarios 1 and 2. The proportion of the population that experienced a 20 ppb-h (or higher) increase in ozone exposure is 15% in Scenario 1 and 5% in Scenario 2. The distribution is much flatter and has a longer tail under Scenario 3, signifying that more individuals are experiencing more extreme short-term peak exposures. Nearly 30% of the population in Scenario 3 experienced a 20 ppb-h or higher ozone change. Note that for purposes of readability we cut the x-axis at 45 ppb-h. Values can reach up to 82 ppb-h under Scenario 3.

4. Discussion and conclusions

In this paper, we coupled air quality spatiotemporal estimates to the synthetic information model of the Houston Metropolitan Area that captures the detailed individual-level activities throughout the day. One of the major challenges associated with creating high fidelity models is the limitation in computational resources and performance. Developing the synthetic information model, creating air quality estimates, and then coupling them is computationally intensive, requiring a system that can handle processing and joining tables, each with millions of rows of data. While this type of computational challenge may have limited our ability to model at this level of resolution in the past, the availability of high performance computing and database management techniques (e.g., parallel processing, b-tree clustered indexing) allowed for the effective processing, summarization, and exploration of these data. This allowed us to attach specific exposure levels to the synthetic individuals based on the exact time of day, the geo-location of the activity, and the condition of the physical environment. This is crucial to better understand exposure pattern heterogeneity. Moreover, it provided the flexibility to run various scenarios across different levels of spatiotemporal resolution allowing us to compare results when certain assumptions are relaxed. Models that do not account for finer levels of resolution in space and time may be missing important differences in the distribution of exposure in the population.

Several scenarios of the model were run at different levels of resolution. Scenario 1 and Scenario 2 relax assumptions around the geospatial movement of synthetic individuals and the temporal variations in ambient ozone concentrations, respectively. The first scenario is similar to studies that assumed individuals stayed home or close to home (e.g., within the same Census tract) throughout the course of the day (Burke et al., 2001; Özkaynak et al., 2008; US EPA, 2014a) but modeled the temporal variations of a pollutant. The second scenario most closely resembles studies that used similar 12- or 24-h averages of pollutant levels (Newth, 2012). Scenario 3, on the other hand, maintains the spatiotemporal resolution of the data to compute estimates that account for spatiotemporal variations in ozone concentration levels and the movement of individuals across the geography and different microenvironments. When we assume individuals remain within small geographic areas as in Scenario 1, we may be underestimating exposure of those living in the relatively low concentration areas of Houston, particularly for populations over 18 that are more likely to travel outside of their home neighborhood for work and other activities such as to attend postsecondary schools or go shopping. We saw evidence of this both spatially (lighter regions in Scenario 1 became darker under Scenarios 2 and 3) and temporally (populations aged over 18 were more likely to experience higher exposure to ozone during the day). By not accounting for the geospatial movement of individuals, we could potentially neglect vulnerable subpopulations that live in these types of neighborhoods. Scenario 2 is inadequate for detecting short-term peak exposures which is important for capturing populations sensitive to health effects such as cardiac arrest and asthma attacks. On the other hand, one may need to consider the trade-off between computational cost and data fidelity when studying the effects of certain pollutants to long-term health outcomes (e.g., lung cancer), where understanding long-term exposures to pollutants is important (Vallero, 2014).

Studies have also used ambient concentration levels as a proxy for personal exposure (Newth, 2012). We find that using this assumption likely overestimates individual ozone exposure levels

and may fail to capture populations most at risk for adverse health effects associated with short-term peak exposures. As has been shown in previous studies, increases in ozone levels over the course of a few hours can have major health implications, supporting the need for exposure calculations that account for the detailed spatiotemporal resolution of pollutant levels and population movement (Ensor et al., 2013). Our results found at-risk populations followed a bi-modal distribution when we assume outdoor levels only – nearly all individuals in Scenarios 1 and 3 experienced, at least once, a change in hourly ozone levels above 20 ppb while in Scenario 2 nearly no individuals experienced this. On the other hand, adjusting for indoor, in-vehicle, and outdoor microenvironments resulted in a smoother distribution with greater variability across the population. When considering at-risk populations, Scenario 3 in this case is particularly concerning. Nearly 30% of the population experienced hourly peak exposures above 20 ppb with some as high as 80 ppb – an increase that quadruples the risk of an out-of-hospital cardiac arrest. Keeping in mind the limitations and advantageous across each of the three scenarios, one could cross-reference the at-risk populations under each scenario, potentially allowing policy makers and healthcare professionals to be better informed of the subpopulations at risk.

Data collected from personal monitoring systems allows one to develop individual exposure traces (Nikzad et al., 2012). Extending this to estimate personal exposures across a population, however, remains a challenge. In this study, we traced individual exposures capturing the variability of personal exposure levels across the population in our study area. Because of this, we were able to detect the heterogeneity of exposures, finding that an individual's exposure level throughout the day can be quite different across Houston, within neighborhoods, and within households. This type of analysis can potentially be used to inform resource allocation strategies that better target vulnerable neighborhoods, subpopulations, and individuals.

In further work, we will couple the synthetic information model to an improved spatiotemporal ozone model developed by Ensor (Ensor and Raun, 2015). We will also generate activity patterns that reflect important variations with respect to exposure levels, such as type of day (e.g., weekday vs. weekend) and time of year (e.g., winter vs. summer). The coupling of spatiotemporal air quality estimates with the synthetic information model of the Houston Metropolitan Area in this study resulted in an initial implementation of the *in silico* analytics platform to evaluate disparities in exposure to air pollution at a level of detail not possible with other models. This allowed us to attach specific exposure levels to the synthetic individuals. Furthermore, the heterogeneous exposure levels of the population across time are more accurately reflected, allowing for increased sensitivity to detecting the variation of exposure across the population and within zip codes, neighborhoods, and even households.

References

Adigaa, A., Agashea, A., Arifuzzamana, S., Barretta, C.L., Beckmana, R., Bisseta, K., Chena, J., Chungbaeka, Y., Eubanka, S., Guptaa, E., et al., 2015. Generating a Synthetic Population of the United States. Tech. Rep. NDSSL 15-009. Network Dynamics and Simulation Science Laboratory.

Barrett, C.L., Eubank, S.G., Smith, J.P., 2005. If smallpox strikes Portland…. Sci. Am. 292 (3), 54–61.

Barrett, C., Eubank, S., Marathe, M., 2006. Modeling and simulation of large biological, information and socio-technical systems: an interaction based approach. In: Interactive Computation. Springer, pp. 353–392.

Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic base-line populations. Transportation Research A – Policy and Practice 30, 415–429.

Burke, J.M., Zufall, M.J., Özkaynak, H., 2001. A population exposure model for particulate matter: case study results for PM2. 5 in Philadelphia, PA. J. Expo. Anal. Environ. Epidemiol. 11 (6), 470–489.

Davis, G., Ensor, K., 2006. Outlier detection in environmental monitoring network data: an application to ambient ozone measurements for Houston, Texas. J. Stat. Comput. Simulat. 76 (5), 407–422.

Duan, N., 1982. Models for human exposure to air pollution. Environ. Int. 8 (1–6), 305–309.

Ensor, K., Raun, L., 2015. Spatio-temporal Pollution Estimation from Monitoring Network Data. Tech. rep., Rice University.

Ensor, K.B., Raun, L.H., Persse, D., 2013. A case-crossover analysis of out-of-hospital cardiac arrest and air pollution. Circulation 127 (11), 1192–1199.

Ensor, K.B., Ray, B.K., Charlton, S.J., 2014. Point source influence on observed extreme pollution levels in a monitoring network. Atmos. Environ. 92, 191–198.

Fugas, M., 1975. Assessment of total exposure to an air pollutant. In: Proceedings of the International Conference on Environmental Sensing and Assessment, vol. 2. IEEE, Las Vegas. Nevada, pp. 38–45.

Halloran, M.E., Ferguson, N.M., Eubank, S., Longini, J., Cummings, I.M.D.A., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T.C., Wagener, D., Beckman, R., Kadau, K., Barrett, C., Macken, C.A., Burke, D.S., Cooley, P., 2008. Modeling targeted layered containment of an influenza pandemic in the United States. Proc. Natl. Acad. Sci. U.S.A. 105 (12), 4639–4644. https://doi.org/10.1073/pnas.0706849105. http://www.ncbi.nlm.nih.gov/pubmed/18332436.

Hatzopoulou, M., Hao, J.Y., Miller, E.J., 2011. Simulating the impacts of household travel on greenhouse gas emissions, urban air quality, and population exposure. Transportation 38 (6), 871–887.

Hülsmann, F., Gerike, R., Kickhöfer, B., Nagel, K., Luz, R., 2011. Towards a multi-agent based modeling approach for air pollutants in urban regions. In: Proceedings of the Conference on "Luftqualität an Straßen, pp. 144–166.

Hwang, B., Jaakkola, J., 2008. Ozone and other air pollutants and the risk of oral clefts. Environ. Health Perspect. 116, 1411–1415.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C., 2005. A review and evaluation of intraurban air pollution exposure models. J. Expo. Sci. Environ. Epidemiol. 15 (2), 185–204.

Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M., Switzer, P., Behar, J.V., Hern, S.C., Engelmann, W.H., 2001. The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants. J. Expo. Anal. Environ. Epidemiol. 11 (3), 231–252.

Kousa, A., Kukkonen, J., Karppinen, A., Aarnio, P., Koskentalo, T., 2002. A model for evaluating the population exposure to ambient air pollution in an urban area. Atmos. Environ. 36 (13), 2109–2119.

Leech, J.A., Nelson, W.C., Burnett, R.T., Aaron, S., Raizenne, M.E., 2002. It's about time: a comparison of Canadian and American time-activity patterns. J. Expo. Anal. Environ. Epidemiol. 12 (6), 427–432.

Lenormand, M., Deffuant, G., 2013. Generating a synthetic population of individuals in households: sample-free vs sample-based methods. J. Artif. Soc. Soc. Simulat. 16 (4), 12.

Lewis, B., Swarup, S., Bisset, K., Eubank, S., Marathe, M., Barrett, C., 2013. A simulation environment for the dynamic evaluation of disaster preparedness policies and interventions. J. Publ. Health Manag. Pract. 19 (Suppl. 2), S42–S48. https://doi.org/10.1097/PHH.0b013e31829398eb. http://www.ncbi.nlm.nih.gov/pubmed/23903394.

Lu, G., Wong, D., 2008. An adaptive inverse-distance weighting spatial interpolation technique. Comput. Geosci. 34, 1044–1055.

Lum, K., Chungbaek, Y., Eubank, S., Marathe, M., 2016. A two-stage, fitted values approach to activity matching. International Journal of Transportation 4 (1), 41–56.

Marathe, M., Vullikanti, A.K.S., 2013. Computational epidemiology. Commun. ACM 56 (7), 88–96. https://doi.org/10.1145/2483852.2483871.

Marathe, M., Mortveit, H., Parikh, N., Swarup, S., 2014. Prescriptive analytics using synthetic information. Emerging Trends in Predictive Analytics: Risk Management and Decision Making 1–19.

Matz, C.J., Stieb, D.M., Davis, K., Egyed, M., Rose, A., Chou, B., Brion, O., 2014. Effects of age, season, gender and urban-rural status on time-activity: canadian human activity pattern survey 2 (chaps 2). Int. J. Environ. Res. Publ. Health 11 (2), 2108–2124.

Namazi-Rad, M.-R., Mokhtarian, P., Perez, P., 2014. Generating a dynamic synthetic population–using an age-structured two-sex model for household dynamics. PLoS One 9 (4), 1.

Newth, D., 2012. An integrated agent-based framework for assessing air pollution impacts. J. Environ. Protect. 03 (29), 1135–1146. https://doi.org/10.4236/jep.2012.329132.

Nikzad, N., Verma, N., Ziftci, C., Bales, E., Quick, N., Zappi, P., Patrick, K., Dasgupta, S., Krueger, I., Rosing, T.S., Griswold, W.G., 2012. Citisense: improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system. In: Wireless Health, AC.

Nychka, Douglas, Furrer, Reinhard, Paige, John, Sain, Stephan, 2015. fields: tools for spatial data, R package version 9.0. www.image.ucar.edu/fields.

Ott, W., Flachsbart, P., 1982. Measurement of carbon monoxide concentrations in indoor and outdoor locations using personal exposure monitors. Environ. Int. 8 (1–6), 295–304.

Özkaynak, H., Palma, T., Touma, J.S., Thurman, J., 2008. Modeling population exposures to outdoor sources of hazardous air pollutants. J. Expo. Sci. Environ. Epidemiol. 18 (1), 45–58.

Parikh, N., Youssef, M., Swarup, S., Eubank, S., 2013. Modeling the effect of transient populations on epidemics in Washington DC. Sci. Rep. 3, 3152.

Raun, L.H., Ensor, K.B., Persse, D., 2014. Using community level strategies to reduce asthma attacks triggered by outdoor air pollution: a case crossover analysis. Environ. Health 13. https://doi.org/10.1186/1476-069X-13-58, 58–58.

Rosenthal, F.S., Carney, J.P., Olinger, M.L., 2008. Out-of-hospital cardiac arrest and airborne fine particulate matter: a case-crossover analysis of emergency medical services data in Indianapolis, Indiana. Environ. Health Perspect. 116 (5), 631.

Ryan, P.H., LeMasters, G.K., 2007. A review of land-use regression models for characterizing intraurban air pollution exposure. Inhal. Toxicol. 19 (Suppl. 1), 127–133.

Salam, M., Millstein, J., Li, Y., FW, L., Margolis, H., Gilliland, F., 2005. Birth outcomes and prenatal exposure to ozone, carbon monoxide, and particulate matter: results from the children's health study. Environ. Health Perspect. 113, 1638–1644.

Silverman, R.A., Ito, K., Freese, J., Kaufman, B.J., De Claro, D., Braun, J., Prezant, D.J., 2010. Association of ambient fine particles with out-of-hospital cardiac arrests in New York City. Am. J. Epidemiol. 217.

Texas Commission on Environmental Quality (TCEQ), 2005. The Houston Air Quality Change. Rapid Economic and Population Growth Create a Potent Blend for the Region's Environment. Natural Outlook Spring 2005. Available from. https://www.tceq.texas.gov/assets/public/comm_exec/pubs/pd/020/05-02/houston-x.pdf.

U.S. EPA, 2007. Ozone Population Exposure Analysis for Selected Urban Areas. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Health and Environmental Impacts Division, Research Triangle Park, North Carolina.

US EPA, 2011. Exposure Factors Handbook, 2011 edition. U.S. Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment, Washington Office, Washington, DC.

US EPA, 2012. Human exposure modeling - air pollutants exposure model. available from. https://www.epa.gov/fera/human-exposure-modeling-air-pollutants-exposure-model.

US EPA, 2014. EPA's stochastic human exposure and dose simulation (sheds) model. available from. https://www.epa.gov/chemical-research/stochastic-human-exposure-and-dose-simulation-sheds-estimate-human-exposure.

US EPA, 2014. Health Risk and Exposure Assessment for Ozone. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Health and Environmental Impacts Division, Research Triangle Park, North Carolina.

US EPA, 2015. Relationship between indoor, outdoor, and personal air. research project publication details, available from. http://cfpub.epa.gov.

Vallero, D., 2014. Fundamentals of Air Pollution, fifth ed. Elsevier, Amsterdam, Boston.

Weisel, C., Zhang, J., Turnip, B., Morandi, M., Colome, S., Stock, T., Spektor, D., Korn, L., Winer, A., Alimokhtari, S., Kwon, J., Mohan, K., Harrington, R., Giovanetti, R., Cui, W., Afshar, M., Maberti, S., Shendell, D., 2005. Relationship of indoor, outdoor, and personal air (riopa) study: study design, methods and quality assurance/control results. J. Expo. Anal. Environ. Epidemiol. 15, 123–127.

Wellenius, G.A., Burger, M.R., Coull, B.A., Schwartz, J., Suh, H.H., Koutrakis, P., Schlaug, G., Gold, D.R., Mittleman, M.A., 2012. Ambient air pollution and the risk of acute ischemic stroke. Arch. Intern. Med. 172 (3), 229–234.

Wheaton, W.D., Cajka, J.C., Chasteen, B.M., Wagener, D.K., Cooley, P.C., Ganapathi, L., Roberts, D.J., Allpress, J.L., 2009. Synthesized population databases: a US geospatial database for agent-based models. Tech. rep., RTI International. https://doi.org/10.3768/rtipress.2009.mr.0010.0905 available from. http://www.ncbi.nlm.nih.gov/pubmed/20505787.

Xu, J., Jiang, H., Zhao, H., Stephens, B., 2017. Mobile monitoring of personal nox exposures during scripted daily activities in Chicago, IL. Aerosol and Air Quality Research 17, 1999–2009.

Zou, B., Wilson, J.G., Zhan, F.B., Zeng, Y., 2009. Air pollution exposure assessment methods utilized in epidemiological studies. J. Environ. Monit. 11 (3), 475–490.