Check for
updates

Distributions of Human Exposure to Ozone During Commuting Hours in Connecticut Using the Cellular Device Network

Owais GILANI, Simon URBANEK, and Michael J. KANE

Epidemiologic studies have established associations between various air pollutants and adverse health outcomes for adults and children. Due to high costs of monitoring air pollutant concentrations for subjects enrolled in a study, statisticians predict exposure concentrations from spatial models that are developed using concentrations monitored at a few sites. In the absence of detailed information on when and where subjects move during the study window, researchers typically assume that the subjects spend their entire day at home, school, or work. This assumption can potentially lead to large exposure assignment bias. In this study, we aim to determine the distribution of the exposure assignment bias for an air pollutant (ozone) when subjects are assumed to be static as compared to accounting for individual mobility. To achieve this goal, we use cell-phone mobility data on approximately 400,000 users in the state of Connecticut, USA during a week in July 2016, in conjunction with an ozone pollution model, and compare individual ozone exposure assuming static versus mobile scenarios. Our results show that exposure models not taking mobility into account often provide poor estimates of individuals commuting into and out of urban areas: the average 8-h maximum difference between these estimates can exceed 80 parts per billion (ppb). However, for most of the population, the difference in exposure assignment between the two models is small, thereby validating many current epidemiologic studies focusing on exposure to ozone.

Supplementary materials accompanying this paper appear online.

Key Words: Environmental epidemiology; Exposure assignment bias; Human mobility; Pollutant modeling.

1. BACKGROUND AND MOTIVATION

Epidemiologic studies have established associations between various air pollutants and a number of adverse health outcomes for adults and children. Air pollutants, such as ground-level ozone (O_3), particulate matter, carbon monoxide, lead, sulfur dioxide, and nitrogen dioxide, have been shown to worsen health outcomes such as heart rate variability, cardio-

Owais Gilani (✉), Bucknell University, Lewisburg, PA, USA (E-mail: *owais.gilani@bucknell.edu*). Simon Urbanek, AT&T Labs Research, Bedminster, NJ, USA. Michael J. Kane, Yale University, New Haven, CT, USA.

pulmonary mortality, acute myocardial infarction, low birth weight, development and exacerbation of asthma, reduced lung function, and acute respiratory symptoms (Pope III et al. 2002, 2004; Peel et al. 2005; Zanobetti and Schwartz 2005; Chen et al. 2006; Brauer et al. 2007; Delamater et al. 2012; Pedersen et al. 2013; Sacks et al. 2014).

These associations between exposure to air pollutants and adverse health outcomes have been established using various epidemiologic study designs. Some studies have been conducted on an ecological scale, where the unit of analysis is at an aggregated level (such as counties or census tracts). In these studies, an aggregate measure of a health outcome (such as cause-specific mortality rate over time) is correlated with an aggregate measure of pollutant exposure over the study area (Peel et al. 2005; Chen et al. 2006; Delamater et al. 2012; Sacks et al. 2014). While these studies are useful in determining associations on the population level, they are subject to the ecological fallacy so that conducting inference on an individual level is problematic. Other studies have utilized case–control or cohort designs (prospective and retrospective), that have conducted the analysis on an individual scale (Gent et al. 2003; Brauer et al. 2007). In such studies, health data are available on individuals (often in great detail), which are then correlated with data on individual-level pollutant exposure. However, assigning pollutant concentration exposure to individuals is rather challenging. Giving each study participant an air pollution monitor to carry with them all day during the study duration is extremely expensive and impractical. Therefore, epidemiologic studies analyzing the effects of air pollutants on adverse health outcomes at an individual level typically estimate pollutant concentrations at subject residences or work places using various approaches (that vary in their level of sophistication) (Pope III et al. 2002; Gent et al. 2003; Jerrett et al. 2005; Brauer et al. 2007; Pedersen et al. 2013).

A common approach to estimate pollutant concentration at an area of interest (e.g., subject residence) uses observed or monitored air pollutant concentration data at limited sites and time durations to predict concentrations at unsampled sites and time durations. This prediction may be achieved using some very simple approaches such as assigning the same concentration as the closest monitored site or an average from a few closest sites (Rage et al. 2009), or estimating via techniques such as inverse distance weighting (Neidell 2004). More sophisticated approaches include land-use regression models (Turner et al. 2016) and spatial/temporal interpolation techniques using geostatistical models of varying complexity, e.g., universal kriging (Jerrett et al. 2005; Rage et al. 2009).

Another popular approach for assigning pollutant concentrations to subject residences is using some deterministic computer model output that simulates either the underlying pollutant chemistry, e.g., the Community Multiscale Air Quality or CMAQ model (Byun and Schere 2006), or the dispersion process of air pollutants from their source(s), e.g., CALINE4 (Benson 1992). Since these simulations are computationally rather expensive, such models provide predictions on a grid where each grid cell covers a large spatial area (e.g., CMAQ provides predictions on grid cells of size $12 \, \text{km} \times 12 \, \text{km}$). Epidemiologic studies utilizing these models determine the cell within which a subject residence or work place is located, and assign the corresponding pollutant exposure to the subject. Some recent studies have combined these two approaches together to develop data fusion models. In these models, data from both observed concentrations at limited sites and outputs from deterministic computer

models are jointly used to provide predictions at unsampled locations at a fine spatial and temporal resolution (Sacks et al. 2014; Turner et al. 2016).

Regardless of the sophistication of the method used to assign pollutant exposure to subjects using their residential or work locations, a fundamental challenge that is not addressed is the fact that subjects do not spend all of their time at home or at work. Humans move around during the day, and the patterns of movement likely depend on a number of factors such as time of day, day of week, season, and employment status. By not taking mobility into account, current exposure models may suffer from significant bias in their estimates.

The goal of this study is to determine the distribution of exposure bias to an air pollutant (O_3) when mobility is not taken into account, and to identify groups of individuals for which this bias is large. To achieve this goal, we require information on patterns of human mobility. Mobility is well captured by cell-phone data; however, most available cell-phone-based mobility studies either require users to install an "app" to capture their location (Bayir et al. 2009) or rely on other, more opportunistic approaches (Calabrese et al. 2011). These methods suffer from not sufficiently sampling at the population level and are likely not robust enough to generalize.

Analyses have been performed using cell-phone infrastructure based on Call Detail Records (CDRs), which are generated by applications or phone calls on cellular devices (Becker et al. 2013; Lu et al. 2017; Thuillier et al. 2018; Marques-Neto et al. 2018). These records are generated primarily for billing purposes and include information including location (triangulated via towers), data usage, etc. These records give precise location and duration information about the device while it is in use; however, they do not provide information at times when the device is not in use and therefore have substantial time windows with no information. Analyzing mobility using such data likely leads to biased results, as the data are not representative of mobility for individuals themselves and for the entire population.

The number of studies that explicitly account for human mobility when assigning exposure to air pollutants is fairly small. Nyhan et al. (2018) used CDR data to incorporate a work and home location into pollution exposure assignment, while (Özkaynak et al. 2009) incorporated human movement indirectly through the Environmental Protection Agency's (EPA) Consolidated Human Activity Database. Warren et al. (2017) incorporated longer-term movement, such as relocation to a different state, to analyze susceptibility to environmental pollution during pregnancy.

This study uses Advanced Wireless Service (AWS) data, which provides the most passive and complete data on population-level mobility. The data consist of cell devices that are turned on, that are connected to towers, and that are serviced by at least two towers (discussed in Sect. 2). Very few studies that quantify population-level human mobility at the census tract-level use cell data at the infrastructure level (Becker et al. 2013), and to our knowledge, this is the first study that integrates pollution information with human mobility for an accurate exposure study at this scale. Our approach uses cell-tower-level data to determine patterns of human mobility. We assign exposure to O_3 concentration to mobile devices, which are used as proxies for individuals, in the US state of Connecticut (CT) during a week in summer 2016. Based on individuals' mobility behavior and the tower area they are connected to most between the hours of 8:00 PM and 6:30 AM (which we term the *nighttime local area*), we are able to assess the exposure assignment bias.

Knowledge of the distribution of O_3 exposure assignment bias will be extremely beneficial for environmental epidemiologists studying the effects of O_3 exposure on adverse health outcomes. Minimal exposure assignment bias due to a static pollutant concentration assignment would validate the results of past studies, while a significant exposure assignment bias would provide evidence for the need for human mobility to be explicitly accounted for in future epidemiologic studies.

This paper proceeds as follows: In Sect. 2, we provide details of the cell-phone telemetry data. In Sect. 3, we provide details on the data and model used to provide predictions of O_3 concentrations at the cell-tower sites. In Sect. 4, we provide details on the distribution of O_3 exposure assignment bias, with concluding remarks in Sect. 5.

2. QUANTIFYING HUMAN MOBILITY WITH CELL-TOWER HAND-OFFS

2.1. CELL-TOWER DENSITY IN THE STATE OF CONNECTICUT

Cellular phone towers coordinate data movement between cell phones and other cell-connected devices providing both phone calls and Internet connectivity. Interconnected towers form a network capable of routing data traffic between cell devices or to Internet-connected devices through a base station. US cell-phone networks facilitate the communication of hundreds of millions of devices, transferring vast amounts of data on any given day. This study is restricted to the state of Connecticut (CT). Figure 1 shows the locations of approximately 10,000 cell towers in the state, along with the locations of the primary and secondary roads in the state, taken from the OpenCelliD Web site (https://opencellid.org). The data traffic capacity of an individual tower is limited; to compensate for spatial areas with higher capacity demands, more towers may be erected in those areas. As a result, the spatial distribution of the towers matches the population distribution. This is seen in Fig. 1, with most people residing in the southwest corner, the shoreline, and in metropolitan areas including Hartford, Danbury, Middletown, and Waterbury.

Figure 2 shows the distribution of the distance to the closest tower from each tower. The distribution is roughly exponential. The lower quartile, median, and upper quartile are 2.7, 68.2, and 306 m, respectively. Many towers are within meters of other towers because multiple towers are sometimes built at a single site. This is especially true in highly populated areas, like southwestern Connecticut, where there are more constraints placed on where a tower can be placed as well as the fact that more towers are needed to compensate for higher traffic demands. Eastern Connecticut is generally less populated than the central and southwestern portions of the state, and the towers in these regions may be several kilometers away from the next closest one.

2.2. HAND-OFF TRAJECTORIES IN THE CELL-TOWER NETWORK

A single cellular device connects to the network through a tower. Devices generally connect to the closest tower; however, this can vary due to geographic features, which can occlude communication, and the amount of traffic being routed through the towers.
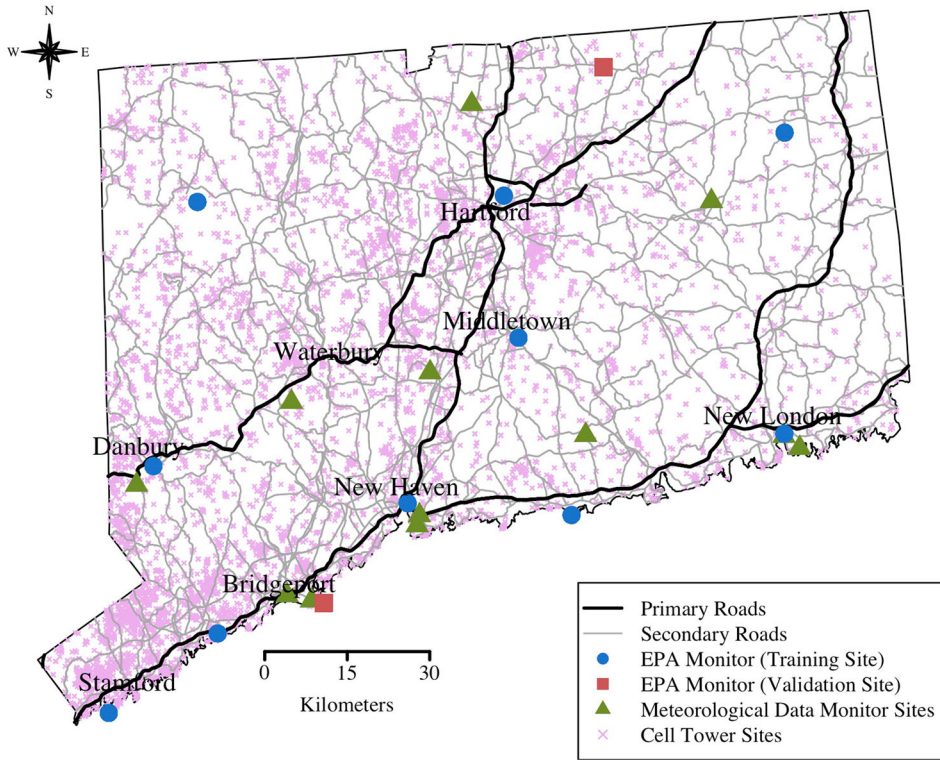
Figure 1.   Locations of primary and secondary roads; EPA and meteorological monitors; and cell-tower locations for the state of Connecticut (from OpenCelliD (https://opencellid.org)).
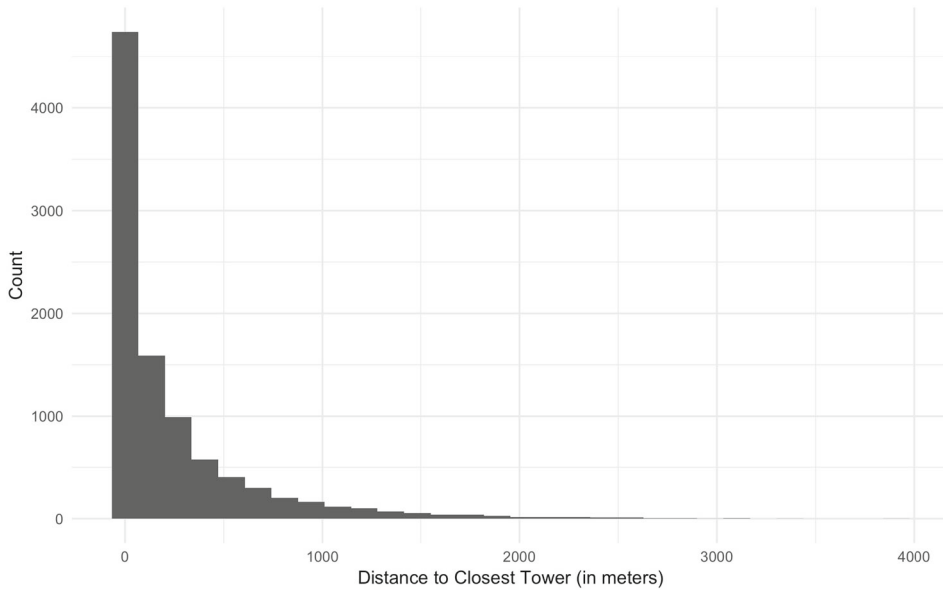

Figure 2.   Distribution of Haversine distance (in meters) to the nearest cell tower.
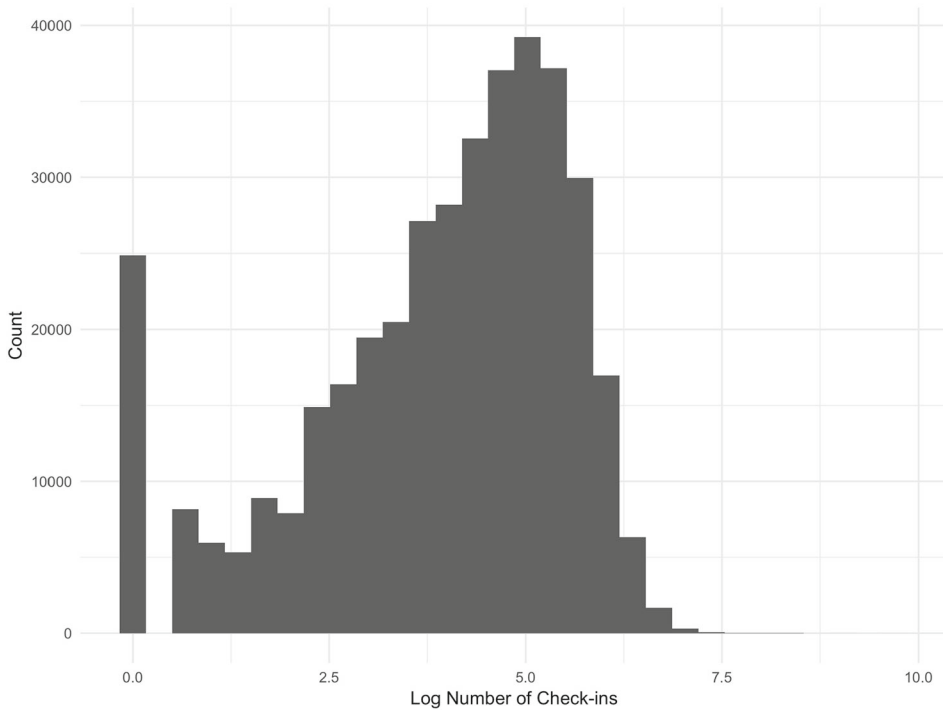
Figure 3.   A histogram of the log number of hand-offs.

As a phone moves through the network, it is "handed off" between towers. A hand-off is characterized by a unique, anonymous device identifier, a date and time, and the location of *the tower* to which the device was handed. Data on these hand-offs may be analyzed to evaluate traffic load, tower placement, and connection integrity. Sequences of hand-offs in time for a single device, which we will refer to as *trajectories*, can be used to locate the device, and the humans that carry them, up to the resolution of the order 10's to 1000's of meters, depend on the local cell-tower density.

This study considers devices from one major US carrier with at least one hand-off, where all hand-offs for the device during the period of July 18–24, 2016, were within the state of Connecticut. A total of approximately 50,000,000 hand-offs were recorded from approximately 400,000 unique devices. Given the sensitivity of these data, several steps were taken to ensure the privacy of all individuals. First, personally identifying characteristics are not included in these data. Data for the same device are linked using an anonymous unique identifier, rather than a telephone number, and the anonymization was performed by a party not involved in the data analysis. No demographic data are linked to any cell-phone user or used in this study. Second, all results are presented as aggregates. That is, no individual anonymous identifier was singled out for the study. By observing and reporting only on the aggregates, we protect the privacy of all individuals.

A histogram of the number of hand-offs per device is shown in Fig. 3. A total of approximately 25,000 devices had a single hand-off. The mode of the histogram where the log number of counts is greater than zero is centered around 5 and implies that most devices had approximately 150 hand-offs during the period considered.

3. DESCRIPTION OF POLLUTION MODEL

In order to determine hourly O_3 concentration exposure of users, we fitted a stationary Gaussian spatiotemporal model (Cressie and Wikle 2011) to observed hourly O_3 concentrations from July 6 to August 5, 2016, for the state of Connecticut, and used it to predict O_3 concentrations at approximately 10,000 cell-tower sites during the week of July 18–24, 2016.

3.1. DATA

We obtained data on observed hourly O_3 concentrations (in parts per billion–ppb) from the US Environmental Protection Agency (EPA) Air Quality System (AQS) Data Mart (US Environmental Protection Agency 2017) from July 6 to August 5, 2016, for the state of Connecticut. Figure 1 shows the location of the 12 O_3 monitoring sites in CT; we randomly selected 10 sites for model fitting and two sites for model validation (Stafford and Stratford). At each site, O_3 concentration data were available for 744 time points. Figure 4 shows the observed hourly O_3 concentrations at the 12 monitoring sites. The two validation sites are displayed in the bottom two panels. The mean hourly O_3 concentration over the 31 days across the 10 training sites was 35.9 ppb with a standard deviation of 19.4 ppb, while the mean hourly concentration at the two validation sites was 39.0 ppb with a standard deviation of 17.9 ppb.

In addition to data on the observed concentrations of O_3, we collected data on traffic and meteorological factors that could help explain the spatiotemporal variation observed in O_3 concentrations. Hourly temperature (in degrees Celsius) and wind speed (in meters per second) data were obtained from the National Oceanic and Atmospheric Administration (Quality Controlled Local Climatological Data) (US Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service, National Climatic Data Center 2017). Data were available from 12 weather stations in Connecticut (see Fig. 1). For each O_3 monitoring site, we assigned hourly temperature and wind speed based on the measurements recorded at the closest weather station (see supplementary material for discussion on this choice). For hours with missing data, the last available hour for which data were available was used. Ozone concentrations tend to be lower in urban areas as compared to their surrounding rural areas—a phenomenon known as ozone urban decrement (Munir et al. 2012). Using data on primary and secondary roads network in CT, available from the US Census Bureau (US Census Bureau 2017), we calculated the minimum distance to primary and secondary roads (in meters). These variables served as proxies for population density in a neighborhood of the O_3 monitoring sites to account for ozone urban decrement.

3.2. MODEL

As described by Cressie and Wikle (2011), a hierarchical spatiotemporal Gaussian process (GP) model can be specified in three stages: the observed data stage, the true underlying process stage, and the parameter stage. We can specify distributions for the data, process,
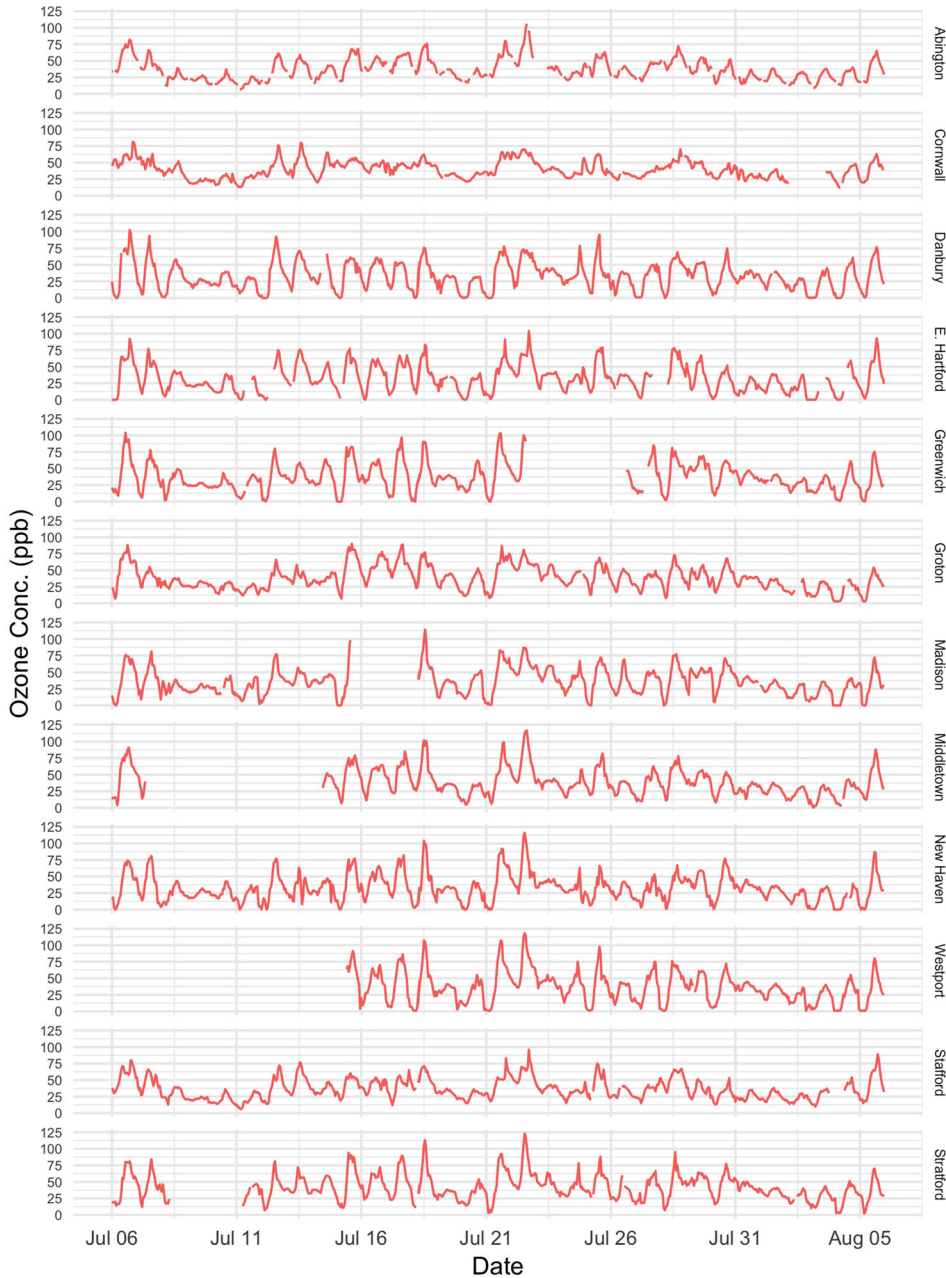
Figure 4. Time-series plot of the observed O_3 concentrations (ppb) at the 12 monitoring sites in CT from July 6 to August 5, 2016.

and parameters for each stage. The R package spTimer by Bakar et al. (2015) can be used to efficiently fit GP models. Following the formulation of Bakar et al. (2015), let $Z(\mathbf{s}_i, t)$ denote the observed O_3 concentration at location $\mathbf{s}_i$, $i = 1, \ldots, 10$, and time point t, $t = 1, \ldots, 744$, and $Y(\mathbf{s}_i, t)$ denote the true underlying O_3 concentration at location

$\mathbf{s}_i$ at time t. Let $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_{10}, t))'$ and $\mathbf{Y}_t = (Y(\mathbf{s}_1, t), \ldots, Y(\mathbf{s}_{10}, t))'$ be the vectors of observed and true underlying O_3 concentrations, respectively, at time t. We define the nugget effect, or the pure error term, as $\boldsymbol{\epsilon}_t = (\epsilon(\mathbf{s}_1, t), \ldots, \epsilon(\mathbf{s}_{10}, t))'$ to be independent (across space and time) and normally distributed $\mathcal{N}\left(\mathbf{0}, \ \sigma_\epsilon^2 \mathbf{I}_{10}\right)$, where σ_ϵ^2 is the unknown pure error variance, and $\mathbf{I}_{10}$ is the 10×10 identity matrix. Similarly, we denote the spatial random effects with independent replicates in time as $\boldsymbol{\eta}_t = (\eta(\mathbf{s}_1, t), \ldots, \eta(\mathbf{s}_{10}, t))'$, assumed to be normally distributed $\mathcal{N}\left(\mathbf{0}, \ \sigma_\eta^2 S_\eta\right)$ and independent of $\boldsymbol{\epsilon}_t$, where σ_η^2 is the site-invariant spatial variance (also called the sill) and S_η is the spatial correlation matrix. We assume that the spatial correlation can be modeled by the exponential function, so that the covariance between two locations $\mathbf{s}_i$ and $\mathbf{s}_j$ is a function of the Euclidean distance d_{ij} between the sites, i.e., $Cov\left(\eta(\mathbf{s}_i, t), \eta(\mathbf{s}_j, t)\right) = \sigma_\eta^2 \cdot \exp^{-\phi d_{ij}}$. Further, let $\mathbf{X}_t$ be a 10×5 matrix of covariates (including a column of $1s$ for the intercept) and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_4)$ denote the 5×1 vector of unknown regression coefficients.

We can then specify the hierarchical GP model by:

$$\mathbf{Z}_t = \mathbf{Y}_t + \boldsymbol{\epsilon}_t \tag{1}$$

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\eta}_t \tag{2}$$

Details on fitting the model and obtaining parameter estimates using the package `spTimer`, including the full conditional distributions of the parameters, are given in Bakar et al. (2015). We used the default recommendations from the package `spTimer` for the initial values and the values of the hyper-parameters for the prior distributions. Specifically, we assigned flat normal priors centered at 0 with large variances (10^{10}) for the regression coefficients and flat inverse-gamma priors for the variance components σ_ϵ^2 and σ_η^2. The value for the spatial decay parameter (ϕ) in the spatial covariance matrix was fixed at $3/d_{\max}$, where $d_{\max}$ is the maximum distance between the ozone monitor sites. Various alternative fixed values were tested, as well as estimating ϕ using a uniform prior distribution. However, the predictive performance was best for the model that used the default fixed value of ϕ (0.0186). While the model is specified in the higher language R, the package `spTimer` performs calculations in the lower level language C for much faster computation. As per default, the Markov chain Monte Carlo was run for 4000 further iterations after discarding the first 1000 as burn-in. MCMC diagnostics were performed using package `coda` (Plummer et al. 2006); all MCMC chains had converged during the burn-in, and auto-correlation plots displayed independence between iterations. The residual plot also did not show any departures from normality.

Estimated parameters from the model are given in Table 1. As expected, higher temperature is associated with increased O_3 concentrations, while increased wind speed is associated with lower O_3 concentration. Minimum distance to primary and to secondary roads both have positive slopes, indicating the expected ozone urban decrement. All coefficients are statistically significant since none of the 95% credible intervals contain 0.

3.3. PREDICTION

Once the GP model has been fitted and posterior distributions for the unknown model parameters have been obtained, spatial prediction at a new location $\mathbf{s}_0$ (and temporal pre-

Table 1. Posterior mean, median, standard deviation, and 95% credible interval for the regression and covariance parameters.

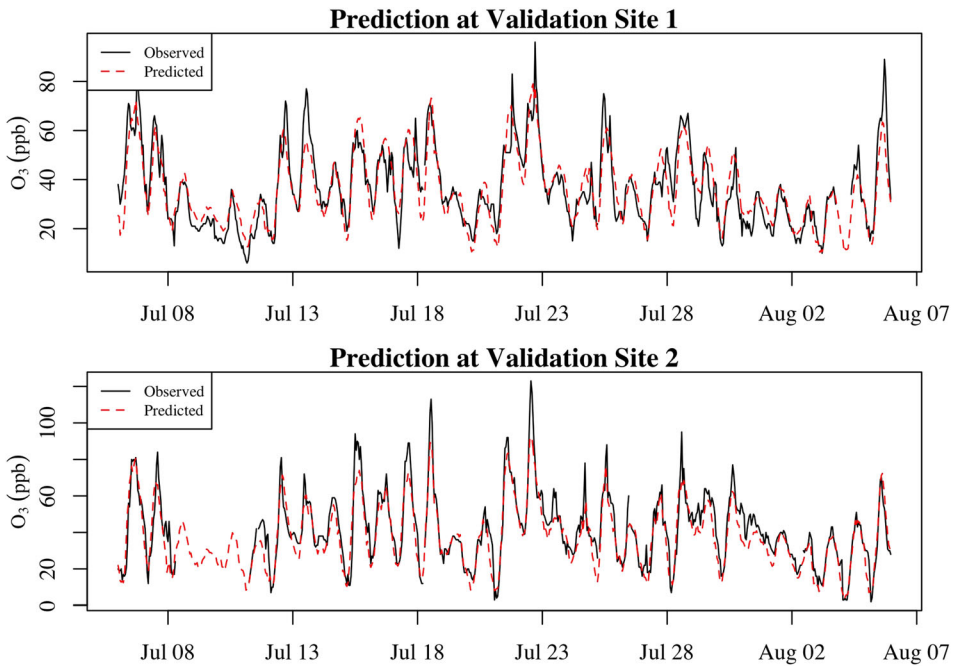| Parameter	Mean	Median	SD	95% CI
β_0 (Intercept)	10.68	10.66	1.27	(8.27, 13.2)
β_1 (Temperature)	1.022	1.023	0.05	(0.92, 1.13)
β_2 (Wind speed)	-0.411	-0.410	0.10	$(-0.60, -0.23)$
β_3 (Dist. to Prim. Rd)	0.0002	0.0002	0.0000	(0.0001, 0.0002)
β_4 (Dist. to Sec. Rd)	0.0005	0.0005	0.0001	(0.0003, 0.0007)
σ_ϵ^2 (Nugget)	13.23	13.21	0.79	(11.8, 14.8)
σ_η^2 (Sill)	174.88	174.92	2.87	(169.3, 180.4)

Figure 5. Comparison of the observed (black solid) and predicted (red dashed) O_3 concentrations (ppb) at the two validation sites (Color figure online).

diction at a future time point t') can be obtained using the posterior predictive distribution for $Z(\mathbf{s}_0, t')$. The function `predict` in the package `spTimer` provides spatiotemporal predictions (see Bakar et al. 2015 for technical details). We predicted hourly O_3 concentrations for the 31-day period from July 6 to August 5, 2016, at the two validation EPA O_3 monitoring sites, as well as at the approximately 10,000 cell-tower sites.

Figure 5 shows a time-series comparison of the observed and predicted hourly O_3 concentrations at the two validation sites. At both sites, the predicted concentrations are very similar to the observed concentrations. The root-mean-square error at the validation sites is 7.68, while the relative bias is -0.0461 and the relative mean separation is 0.1811. The small values for the relative bias and relative mean separation suggest a fairly good model fit. Figure 6 shows a prediction map for hourly O_3 concentrations across CT at midnight, 6 am,
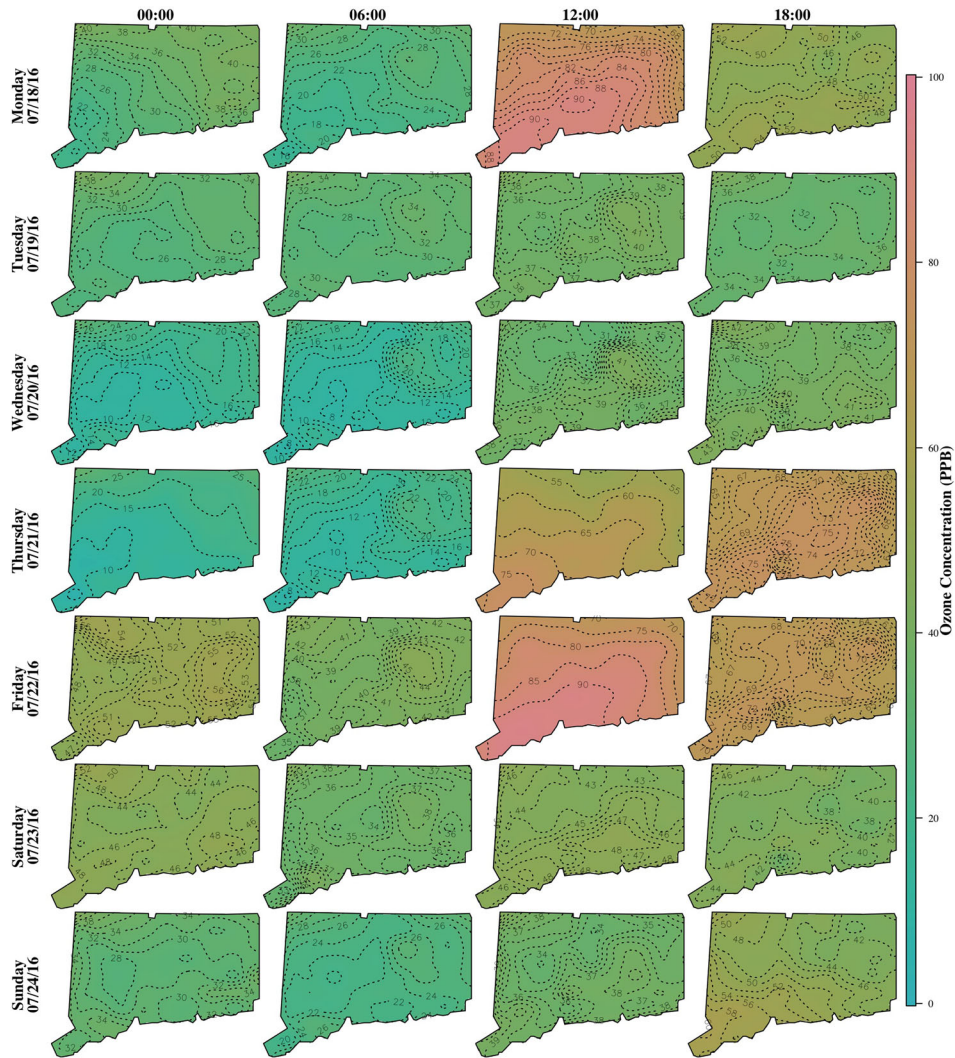
Figure 6.   Prediction of ozone concentrations at midnight, 6 AM, noon, and 6 PM for each of the 7 days in the study (Color figure online).

noon, and 6 pm on seven consecutive days from July 18–24, 2016. It shows that the distribution of O_3 concentration is somewhat homogeneous across space, but changes considerably throughout the day. There also appears to be a concentration gradient in the southwest to northeast direction, particularly during the daytime hours, with higher concentrations in the southwestern part of the state.

4. EXPOSURE DISTRIBUTIONS

As described in Sect. 2, a device hand-off is characterized by a unique, anonymous device identifier, a date and time, and the location of the tower to which the device was handed.

From sequences of hand-offs, sorted by date and time, the duration over which the device was checked into the tower was calculated. In addition, a new feature was calculated per device over the study duration—the tower that a device was checked into for the longest duration during the hours of 8:00 PM and 6:30 AM. We note that this new variable is correlated with the nighttime local area for working individuals not participating in shift work. However, they are distinct in that, due to de-identification we do not know the residence or occupational characteristics of the devices' owners.

For each tower location, the hourly O_3 concentration was estimated as described in Sect. 3. The pollution estimate can be merged with the mobility data using a join over the date, time, and cell-tower locations. The individual-level exposure estimate is then calculated as the amount of pollution at the tower location multiplied by the duration at that tower's catchment area.

As per the standards set by the EPA, we calculated the average 8-h maximum exposure for each individual for each of the 7 days by calculating the hourly average ozone concentration for each 8-h window during the day and selecting the maximum of these hourly averages for each day.

4.1. DISTRIBUTION OF THE DIFFERENCE IN THE AVERAGE 8-H MAX EXPOSURE

To assess the distribution of O_3 exposure assignment bias when mobility is not taken into account, we calculated the average 8-h max exposure for each device per day under two scenarios: (a) using their trajectories to determine the tower catchment area visited as they moved throughout the day (which we will refer to as the dynamic scenario) and (b) assuming that the individuals spent the entire day at their nighttime local area (which we will refer to as the static scenario), similar to what is typically done in epidemiologic studies (see Figure S.7 for a violin plot of these distributions). The exposure assignment bias for each device per day was then calculated as the difference between the dynamic and static average 8-h max exposure assignment for each day. The result is shown in Fig. 7. (A similar plot showing the distribution of average hourly difference between the dynamic and static scenarios based on a 24-h cumulative exposure—instead of the average 8-h max exposure—is given in Figure S.8, while Figure S.9 shows the distribution of hourly exposure assignment for the dynamic and static scenarios.) All of the estimated differences were within 80 ppb/h. As a reference, the US EPA ozone air quality standard is set at 70 ppb (averaged over an 8-h period); concentrations consistently exceeding this level are considered harmful to human health and welfare (US Environmental Protection Agency 2015). Since the mean and medians are close to zero, this result *validates* many of the current studies of ozone exposure for a large cross section of the population, i.e., the exposure assignment bias due to not taking mobility into account is not too large. However, we observe that the distribution of the differences has heavy tails and that the upper tails are longer than the lower tails. This suggests that the static scenario underestimates the true concentration more frequently than overestimating it. Additionally, it should be noted that differences are correlated for an individual device. That is, devices with large differences between the models at a given hour tend to have large differences at other hours. This implies that there are distinct subpopulations for which static scenarios are biased. This is further explored in subsequent sections.
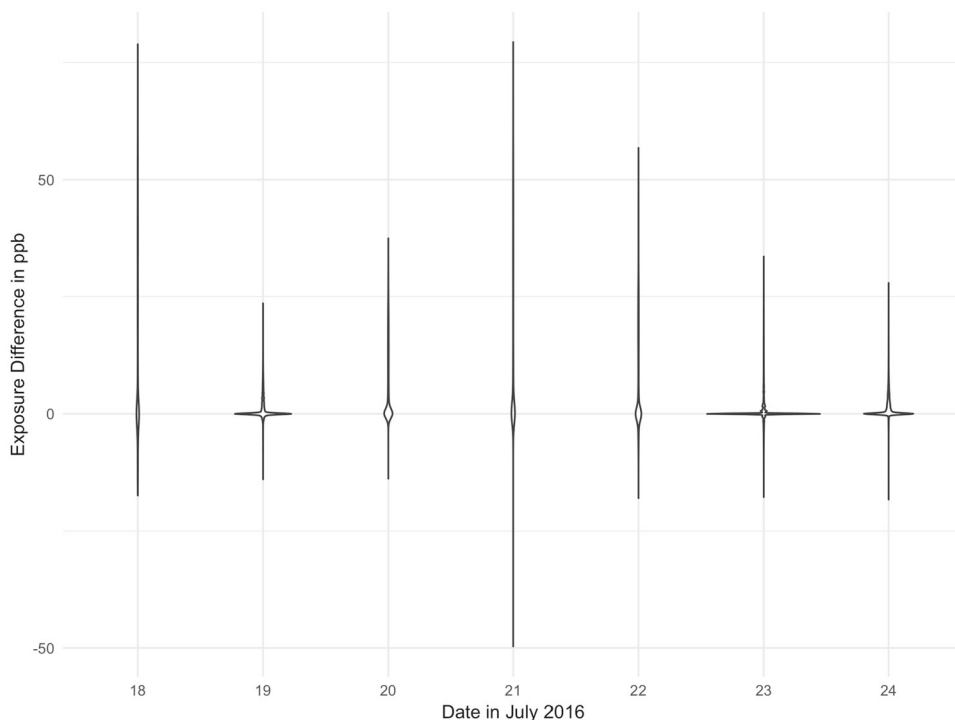
Figure 7.   Violin plot showing the distribution of average 8-h max ozone exposure difference (dynamic scenario exposure minus the static scenario exposure) for each day of the study window.

4.2. WEEKEND VERSUS WEEKDAY DAILY EXPOSURE, ACCOUNTING FOR MOBILITY

Figure 8 shows the distribution of average hourly ozone exposure for individuals for each hour of the day for weekdays and weekends, taking mobility into account. Due to the size of the data, the plots are based on a random sample of 10,000 devices. The plots show that the minimum O_3 exposure is higher on the weekend (for every hour) as compared to the weekday, while the maximum exposure is higher on the weekdays. They also show that the range of average hourly O_3 exposure is greater on weekdays as compared to weekends. It should be noted that these results are likely primarily driven by the variation in O_3 concentrations seen on the weekday when compared to the weekend (as seen in Fig. 6), as opposed to the device mobility. The plots also reveal a generally unimodal exposure distribution during the late night/early morning hours, but a bimodal exposure distribution during the afternoon and evening hours. This again probably reflects the smaller spatial variation in O_3 concentrations during the late night and early morning hours, and greater spatial variation during the afternoon and evening hours. (Figure S.10 shows the distribution of the cumulative exposure by hour for weekdays and weekends, while Figure S.11 shows the distribution of the cumulative exposure at the end of the day for weekdays and weekends.)
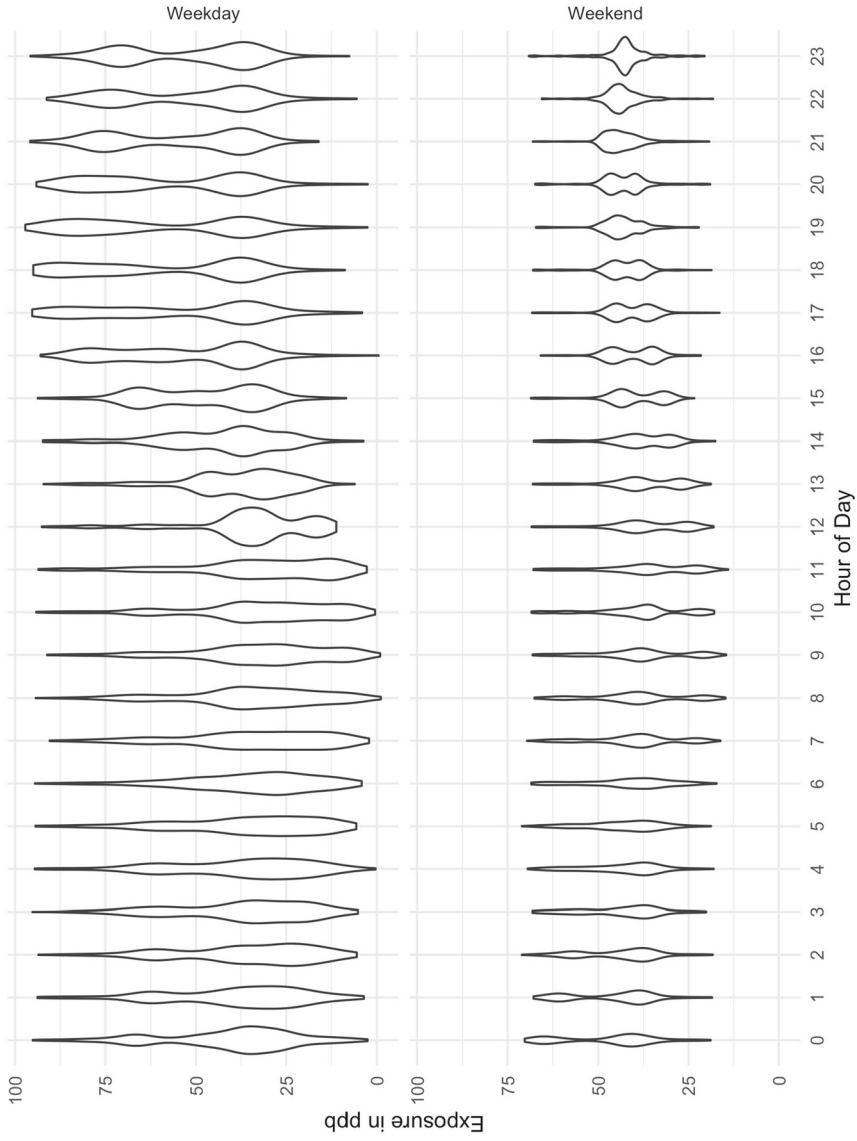
Figure 8.   Violin plots of the exposure distribution by hour of day for weekdays and weekends (using the dynamic scenario).

4.3. Areas of Poorly Predicted Exposure when Mobility is not Taken into Account

While the bias from not taking mobility into account is close to zero for most individuals, for some it is up to 80 ppb/h. While this bias may be negligible for a given day, cumulative exposure difference may result in very different health outcomes than what is predicted by models not taking mobility into account. In particular, individuals with an exposure assignment bias of around 80 ppb/h likely experience this assignment bias for 8–10 h per day on multiple days, resulting in vastly different cumulative exposure assignment biases.

To understand which class of individuals are under-served by existing models, device trajectories where the difference between exposure assessed using the dynamic versus static scenarios was in the highest and lowest 1%, corresponding to individuals with higher and lower exposure with respect to the static scenario, were extracted and a contour plot was created showing their nighttime local areas.
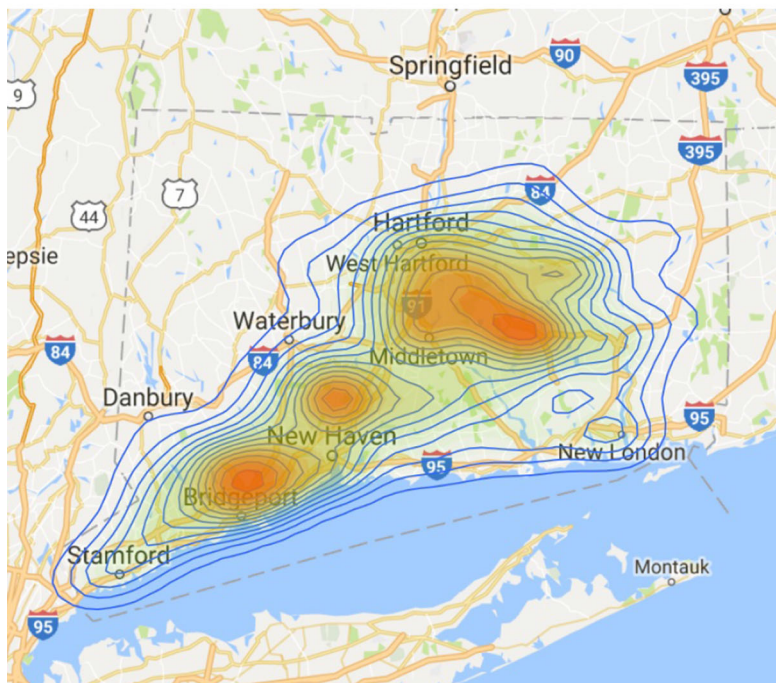
The nighttime local area of individuals receiving more exposure when mobility is taken into account is shown in Fig. 9a. The difference in exposure assessed based on the dynamic and static scenarios is generally due to individuals' mobility during daytime commuting hours. O_3 predictions at different times of the day given in Fig. 6 show that during the afternoon hours, O_3 concentrations are generally higher in the southwestern part of Connecticut (likely due to higher concentrations in New York City area) and decrease in a northeasterly direction. Comparing Fig. 9 to O_3 predictions shown in Fig. 6, we observe that many of the individuals identified as receiving a considerably larger exposure when taking mobility into account are likely suburban residents commuting into the associated urban areas (in a western or southern direction) where ozone concentrations are higher during the daytime hours. In particular, the cluster of individuals residing north and northeast of Hartford likely commute to Hartford during the day, while the cluster identified around Waterbury are likely residents that commute to Danbury during the day.

The nighttime local area of individuals receiving less exposure when mobility is taken into account is shown in Fig. 9b. There are three distinct clusters: one is southeast of Hartford, one is north of New Haven, and one is north of Bridgeport. Many of these individuals are likely suburban residents commuting into the associated urban areas (in an eastern or northern direction) where ozone concentrations are lower during the daytime hours. In particular, the clusters southeast of Hartford and north of New Haven likely represent individuals that commute to Hartford and Middletown during the day, while the cluster north of Bridgeport perhaps represents individuals commuting northeast to New Haven.

These results suggest that the direction of mobility during daytime hours is an important factor determining the adequacy of the static scenario in accurately assessing individual exposure to O_3 concentration. Movement along the pollutant concentration gradient naturally results in a higher difference between actual exposure and the exposure modeled assuming static behavior.
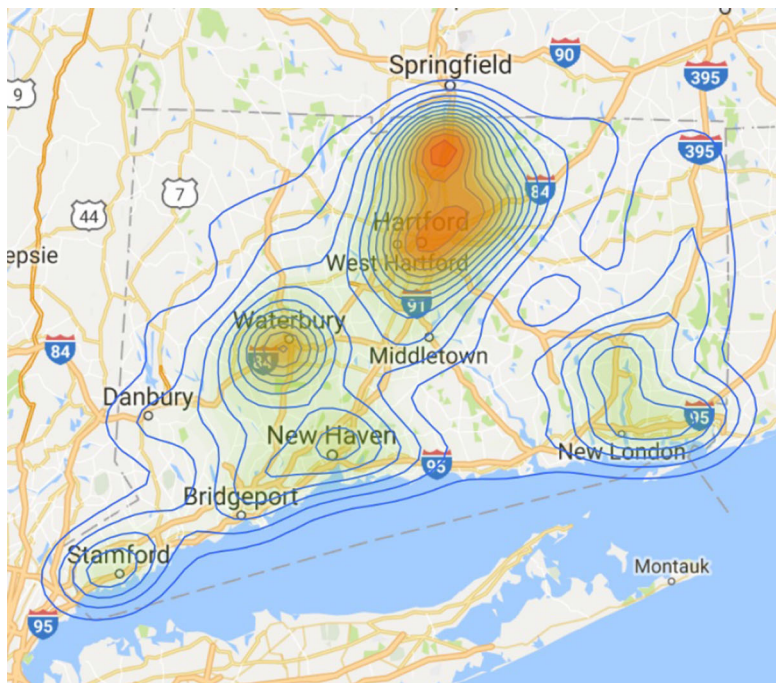
(a)

(b)

Figure 9.   Nighttime local areas of individuals with **a** higher exposure in the dynamic scenario compared to the static scenario (top 1%) and **b** lower exposure in the dynamic scenario compared to the static scenario (bottom 1%).

5. CONCLUSIONS

This paper integrates mobility data from cell-tower hand-offs with a pollution model to more realistically estimate ozone exposure during the period of July 18–24, 2016. These estimates were compared to those of a model where mobility data are not available, which is commonly seen in the literature. We show that the bias introduced by not taking mobility into account is minimal for the majority of individuals in the state of Connecticut, thereby validating many existing epidemiological studies examining the health impacts of exposure to O_3 on various outcomes. However, we also show that existing models do a poor job estimating exposure of individuals who routinely commute into and out of urban areas, particularly whose daytime movements follow the pollution concentration gradient, and, in these cases, mobility should be taken into account.

A key strength of our analysis lies in the use of mobility data on the individual level for a very large portion of the state's population, which allows us to more accurately capture real mobility patterns. However, there are a few limitations to this analysis as well.

The O_3 exposure model was developed using ten O_3 monitoring sites for the entire state of Connecticut. Given the spatial distribution of these sites, it is difficult to capture the small-scale spatial variability in O_3 concentrations. However, this may not be too big of a concern in this particular analysis since O_3 concentrations are known to be fairly homogeneous over short distances.

Data on mobility in this study were captured using hand-offs at cell towers and not using actual GPS coordinates. Therefore, the mobility trajectories provide an estimate of where cellular devices are at any given point and not their exact location. However, most cell towers are within 1000 m of each other, and therefore, cell device areas are generally within a 500-m buffer. This approach has the added advantage of protecting individuals' privacy in that their exact location is never known.

Due to practical reasons, we restricted our analysis exclusively to users who had at least one tower check-in during the 1-week window of analysis and stayed within the boundaries of CT during that week. This potentially has three issues related to generalizability of our results. First, users might not represent the general US population. We are not too concerned about this issue since given the competitive telecommunications market, there does not appear to be any strong evidence suggesting that this cellular service provider attracts a particular niche of customers. Additionally, an internal study found that their subscribers were well represented geographically. Second, individuals that spend the entire week within the state of CT do not include long-distance commuters. Long-distance commuters would be expected to have a much larger difference in exposure assignment between the dynamic and static scenarios, which would potentially lead to longer tails in the exposure difference distribution. Third, this study does not capture individuals who do not move within the 1-week period. However, this population is likely small and inhabiting an indoor space for the duration where they are not exposed to outdoor pollutants.

To our knowledge, this is the first study that compares exposure to a pollutant concentration based on the assumption of static behavior versus dynamic mobility for a significant portion (between 30 and 45%) of the domestic cell user population for an area while at the same time avoiding the bias associated with CDR records. While the results of this study

strengthen our confidence in the findings of epidemiologic studies looking at the adverse impact of O_3 concentration on various health outcomes, our analysis can be extended in two directions: looking at other states, or even the entire US, instead of just CT and over longer time duration, and analyzing other air pollutants, such as nitrogen oxides or small particulate matter, which are known to disperse quickly over short distances (Baldwin et al. 2015). For such pollutants, we expect there to be a significant difference in exposure assignment between the dynamic and static scenarios. However, modeling concentrations for pollutants that vary rapidly over short distances is challenging, as they require a rather dense network of pollutant monitors, which is generally not available. While data on human mobility are available to us in near real time, a bottleneck arises in accurate air pollutant modeling at fine spatial and temporal scales. Additionally, extending the analysis to other states or the entire US over longer time duration presents challenges of dealing with vast quantities of data for both the air pollution modeling part and human mobility. However, with rapid advancements in methods dealing with the analysis of Big Data, this may not be a big challenge in the future.

REFERENCES

Bakar, K. S., Sahu, S. K., et al. (2015). sptimer: Spatio-temporal Bayesian modelling using R. *Journal of Statistical Software*, 63(15):1–32.

Baldwin, N., Gilani, O., Raja, S., Batterman, S., Ganguly, R., Hopke, P., Berrocal, V., Robins, T., and Hoogterp, S. (2015). Factors affecting pollutant concentrations in the near-road environment. *Atmospheric Environment*, 115:223–235.

Bayir, M. A., Demirbas, M., and Eagle, N. (2009). Discovering spatiotemporal mobility profiles of cellphone users. In *World of Wireless, Mobile and Multimedia Networks & Workshops, 2009*, pages 1–9. IEEE.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82.

Benson, P. E. (1992). A review of the development and application of the CALINE3 and 4 models. *Atmospheric Environment. Part B. Urban Atmosphere*, 26(3):379–390.

Brauer, M., Hoek, G., Smit, H., De Jongste, J., Gerritsen, J., Postma, D. S., Kerkhof, M., and Brunekreef, B. (2007). Air pollution and development of asthma, allergy and infections in a birth cohort. *European Respiratory Journal*, 29(5):879–888.

Byun, D. and Schere, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (CMAQ) modeling system. *Applied Mechanics Reviews*, 59(2):51–77.

Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (2011). Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston metropolitan area. *IEEE Pervasive Computing*, 10(4):36–44.

Chen, C.-H., Xirasagar, S., and Lin, H.-C. (2006). Seasonality in adult asthma admissions, air pollutant levels, and climate: A population-based study. *Journal of Asthma*, 43(4):287–292.

Cressie, N. and Wikle, C. K. (2011). Statistics for spatio-temporal data. Hoboken.

Delamater, P. L., Finley, A. O., and Banerjee, S. (2012). An analysis of asthma hospitalizations, air pollution, and weather conditions in Los Angeles county, California. *Science of the Total Environment*, 425:110–118.

Gent, J. F., Triche, E. W., Holford, T. R., Belanger, K., Bracken, M. B., Beckett, W. S., and Leaderer, B. P. (2003). Association of low-level ozone and fine particles with respiratory symptoms in children with asthma. *Journal of the American Medical Association*, 290(14):1859–1867.

Jerrett, M., Burnett, R. T., Ma, R., Pope III, C. A., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. E., et al. (2005). Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*, pages 727–736.

Lu, S., Fang, Z., Zhang, X., Shaw, S.-L., Yin, L., Zhao, Z., and Yang, X. (2017). Understanding the representativeness of mobile phone location data in characterizing human mobility indicators. *ISPRS International Journal of Geo-Information*, 6(1):7.

Marques-Neto, H. T., Xavier, F. H., Xavier, W. Z., Malab, C. H. S., Ziviani, A., Silveira, L. M., and Almeida, J. M. (2018). Understanding human mobility and workload dynamics due to different large-scale events using mobile phone data. *Journal of Network and Systems Management*, pages 1–22.

Munir, S., Chen, H., and Ropkins, K. (2012). Modelling the impact of road traffic on ground level ozone concentration using a quantile regression approach. *Atmospheric environment*, 60:283–291.

Neidell, M. J. (2004). Air pollution, health, and socio-economic status: the effect of outdoor air quality on childhood asthma. *Journal of Health Economics*, 23(6):1209–1236.

Nyhan, M., Kloog, I., Britter, R., Ratti, C., and Koutrakis, P. (2018). Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data. *Journal of Exposure Science and Environmental Epidemiology Epidemiology*.

Özkaynak, H., Frey, H. C., Burke, J., and Pinder, R. W. (2009). Analysis of coupled model uncertainties in source-to-dose modeling of human exposures to ambient air pollution: A pm2. 5 case study. *Atmospheric environment*, 43(9):1641–1649.

Pedersen, M., Giorgis-Allemand, L., Bernard, C., Aguilera, I., Andersen, A.-M. N., Ballester, F., Beelen, R. M., Chatzi, L., Cirach, M., Danileviciute, A., et al. (2013). Ambient air pollution and low birthweight: a European cohort study (ESCAPE). *The Lancet Respiratory Medicine*, 1(9):695–704.

Peel, J. L., Tolbert, P. E., Klein, M., Metzger, K. B., Flanders, W. D., Todd, K., Mulholland, J. A., Ryan, P. B., and Frumkin, H. (2005). Ambient air pollution and respiratory emergency department visits. *Epidemiology*, 16(2):164–174.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Journal of the American Medical Association*, 287(9):1132–1141.

Pope III, C. A., Hansen, M. L., Long, R. W., Nielsen, K. R., Eatough, N. L., Wilson, W. E., and Eatough, D. J. (2004). Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. *Environmental Health Perspectives*, 112(3):339–345.

Rage, E., Siroux, V., Künzli, N., Pin, I., Kauffmann, F., et al. (2009). Air pollution and asthma severity in adults. *Occupational and Environmental Medicine*, 66(3):182–188.

Sacks, J. D., Rappold, A. G., Davis Jr, J. A., Richardson, D. B., Waller, A. E., and Luben, T. J. (2014). Influence of urbanicity and county characteristics on the association between ozone and asthma emergency department visits in North Carolina. *Environmental Health Perspectives*, 122(5):506–512.

Thuillier, E., Moalic, L., Lamrous, S., and Caminada, A. (2018). Clustering weekly patterns of human mobility through mobile phone data. *IEEE Transactions on Mobile Computing*, 17:817–830.

Turner, M. C., Jerrett, M., Pope III, C. A., Krewski, D., Gapstur, S. M., Diver, W. R., Beckerman, B. S., Marshall, J. D., Su, J., Crouse, D. L., et al. (2016). Long-term ozone exposure and mortality in a large prospective study. *American Journal of Respiratory and Critical Care Medicine*, 193(10):1134–1142.

US Census Bureau (2017). 2016 tiger/line shapefiles [internet database] available at. https://www.census.gov/geo/maps-data/data/tiger-line.html, Accessed September 20.

US Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service, National Climatic Data Center (2017). Quality controlled climatological database [internet database] available at. https://www.ncdc.noaa.gov/cdo-web/webservices/v2, Accessed September 20.

US Environmental Protection Agency (2015). National Ambient Air Quality Standards for Ozone (80 FR 65291). *Federal Register*, 80(206):65292–65468.

US Environmental Protection Agency (2017). Air quality system data mart [internet database] available at. http://www.epa.gov/ttn/airs/aqsdatamart, Accessed September 20.

Warren, J. L., Son, J.-Y., Pereira, G., Leaderer, B. P., and Bell, M. L. (2017). Investigating the impact of maternal residential mobility on identifying critical windows of susceptibility to ambient air pollution during pregnancy. *American journal of epidemiology*.

Zanobetti, A. and Schwartz, J. (2005). The effect of particulate air pollution on emergency admissions for myocardial infarction: a multicity case-crossover analysis. *Environmental Health Perspectives*, 113(8):978–982.