Biometrics 000, 1–22 DOI: 000
000 0000

Quantifying personal exposure to air pollution from smartphone-based location data

F. Finazzi

Department of Management, Information and Production Engineering, University of Bergamo,
Viale Marconi, 5 - 24044 Dalmine, Italy
email: francesco.finazzi@unibg.it

and

L. Paci

Department of Statistical Sciences, Università Cattolica del Sacro Cuore,
Largo Gemelli, 1 - 20123 Milan, Italy
email: lucia.paci@unicatt.it

Summary: Personal exposure assessment is a challenging task that requires both measurements of the state of the environment as well as individual's movements. In this paper, we show how location data collected by smartphone applications can be exploited to quantify the personal exposure of a large group of people to air pollution. A Bayesian approach that blends air quality monitoring data with individual location data is proposed to assess the individual exposure over time, under uncertainty on both the pollutant level and the individual location. A comparison with personal exposure obtained assuming fixed locations for the individuals is also provided. Location data collected by the Earthquake Network research project are employed to quantify the dynamic personal exposure to fine particulate matter of around 2500 people living in Santiago (Chile) over a 4-month period. For around 30% of individuals, the personal exposure based on people movements emerges significantly different over the static exposure. On the basis of this result and thanks to a simulation study, we claim that, even when the individual location is known with non-negligible error, it helps to better assess personal exposure to air pollution. The approach is flexible and can be adopted to quantify the personal exposure based on any location-aware smartphone application.

Key words: Dynamic models; Markov Chain Monte Carlo; Particulate matter; Space-time modeling.

1. Introduction

Several epidemiological studies have shown that air pollution represents a major global environmental risk to human health (Hoek et al., 2013; Di et al., 2017). The assessment of human exposure to air pollution usually relies on ambient exposure, namely the pollutant concentration to which people are exposed when outdoor. Ambient exposure is routinely evaluated through pollutant measurements collected by monitoring stations. Since stations are usually sparse across space, pollutant measurements are customary interpolated across space and over time (Lee and Shaddick, 2010; Berrocal et al., 2011; Paci et al., 2013).

Traditionally, population-wide exposure is derived by weighting the pollutant level with the population density (Finazzi et al., 2013; Fassò et al., 2016) or by matching pollutant concentrations from fixed monitoring stations with residence addresses of the population, coming from static census data (Cesaroni et al., 2013). However, according to the World Health Organization (WHO; WHO 2005), most people spend about 20% of their time away from their residence and approximately 4% in transit. Thus, accounting for people movements becomes crucial for accurate exposure assessment and for mitigating misleading results in environmental health research (Park and Kwan, 2017). This takes us to the issue of determining *personal* exposure, defined as the pollutant concentration corresponding to individual location at a given time. We refer the reader to Steinle et al. (2013) for a comprehensive review of personal exposure. In summary, assessing personal exposure requires to locate the individual across space and over time. Moreover, the temporal resolution at which the location is known should be high enough to describe movements which are relevant with respect to the pollutant spatio-temporal variability.

GPS technology may help to collect individual locations at high spatio-temporal resolution. However, people are not usually equipped with devices able to continuously collect and store

GPS-based locations. Since they are cost and time intensive, studies in which individuals are tracked using GPS are limited in the number of people monitored (Chaix et al., 2013).

A viable alternative is to exploit the ubiquity of smartphones to locate people. Telephone companies routinely collect and store this kind of information but, due to obvious privacy reasons, they are rarely keen to provide it. When available, it is often aggregated at area level (Liu et al., 2013; Nyhan et al., 2016; Gariazzo et al., 2016).

Nonetheless, smartphones have geolocation capabilities and by installing a suitable smartphone application (app hereafter) it would be possible to track the smartphone owner. Although this solution looks promising, it has serious drawbacks. First of all, apps cannot be forcibly installed on people smartphones and any study on personal exposure would require to convince a group of people to install the app. Secondly, and more importantly, monitoring the smartphone location at high temporal resolution using the GPS has an impact on the smartphone battery duration. Any power hungry app is likely to be removed by the smartphone owner within days (if not killed by the smartphone operating system).

The alternative strategy, adopted in this paper, is to make use of the location data collected by smartphone apps that are not intended for people tracking. Popular examples are social networks. These apps rarely enable the GPS and the user-app interaction is rarely constant in time. This suggests that the quality of the location data may be lower when expressed in terms of spatial accuracy and temporal frequency/regularity. Nonetheless, this solution allows a large number of people to be "monitored" over a possibly long period.

In this paper, we consider location data collected by the app of the Earthquake Network (EQN) research project (`www.earthquakenetwork.it`) that implements a worldwide earthquake early warning system based on networks of smartphones (Finazzi, 2016; Finazzi and Fassò, 2017). EQN requires to know the location of the smartphones with the app installed for detecting earthquakes in real time. During the last 6, around 5 million people have taken part to the project.

Specifically, the app sends a *heartbeat signal* to a central server every around 30 minutes, if the smartphone is active and Internet is available. Each smartphone has a unique ID which is used to identify the smartphone. The heartbeat signal sent by the smartphone to the server is described by ID, time-stamp, latitude, longitude and precision of the spatial location. Therefore, smartphone owners who keep the app installed are good candidates for personal exposure assessment.

When the aim is to derive a personal exposure, this kind of data set has the advantage of covering a relatively large number of people but it also has some challenges: i) smartphones are not guaranteed to be active 24/7. Therefore, the individual location may be unavailable for long periods of time; ii) since the GPS receiver is usually off, depending on the ability of the smartphone to geolocate itself, the location precision may be low; iii) there is no guarantee that the smartphone owner is located where the smartphone is. However, how many times your smartphone is more than 10 meters away from you?

The contribution of this work is twofold. First, we provide a statistical methodology for assessing personal exposure by combining air quality data with smartphone location data. A Bayesian approach is proposed to overcome challenges in smartphone data and exploit such information to evaluate personal exposure to air pollution. Key elements of the methodology are the statistical model for learning the individual location and the statistical model used to interpolate the pollutant surface across space and over time. In particular, the pollutant concentration is modeled via a space-time model while the individual location is modeled through a dynamic linear model able to handle daily patterns of people movements, location uncertainty and missing data with a limited computational burden. Under a Bayesian perspective, it turns out that the posterior predictive distribution of the pollutant concentration at individual location over time yields to our desired personal exposure, that accounts for uncertainty on both pollution level and individual location. Posterior inference is based on Markov Chain Monte Carlo (MCMC) methods. To our knowledge, this is the first work that provides fully uncertainty quantification of personal exposure to air pollution by blending air quality data and smartphone location data.

Our second contribution is to compare the resulting personal exposure with the traditional personal exposure obtained assuming that people are fixed in space, say at their "home" location. Hereafter, the former will be called *dynamic exposure* while the latter *static exposure*. We carried out an extensive simulation study (presented in the Supplementary materials) to show that the dynamic exposure offers a more accurate assessment of the personal exposure with respect to the static one.

In this work, we focus on fine particulate matter (PM_{10}), which is, quoting WHO, "a common proxy indicator for air pollution and it affects more people than any other pollutant". Indeed, both short and long term exposure to particulate matter are associated with increased human morbidity and mortality (Welty et al., 2009; Di et al., 2017). In particular, we study personal exposure to PM_{10} of 2502 smartphone users living in Santiago, Chile, where particulate pollution is a major public health issue (Valdés et al., 2012). Here, we show that differences between the static and the dynamic exposure do exist for about 30% of individuals employed in this study. Reasonably, higher dynamic exposure is highlighted for users with fixed location in low-polluted areas who routinely move to high-polluted areas.

The rest of the paper is organized as follows. Section 2 describes location data and air quality data employed in this work. Our methodology for assessing personal exposure is described in Section 3. First, a formal definition of personal exposure is introduced; then, stochastic models to analyze pollution data as well as location data are discussed. Computational details are presented in Section 4 while the comparison between the static and the dynamic personal exposure is given in Section 5. Section 6 illustrates our approach using the EQN data. Concluding remarks are provided in Section 7.

2. Data description

2.1 *Location data*

Smartphones with the EQN app installed are used in this study. As mentioned above, the smartphone app sends an heartbeat signal to the server containing information on ID, time-stamp, latitude, longitude and location precision. Due to changes in the way IDs were assigned to smartphones and due to the high variability in the number of users taking part in the project, only the period January 01, 2017 - April 30, 2017 is considered in this work. In fact, people tend to install the app soon after a major earthquake but they can also leave the project at any time. The above period is stable with respect to the number of people in the project, without peaks in the daily app installs or uninstalls.

We restrict the case study to the metropolitan area of Santiago, Chile. The study area covers roughly 40 × 55 kilometers. Due to its high seismicity, Chile has a high number of citizens taking part in the project and Santiago is the Chilean city with the most number of app users. Demographic summaries about Santiago population and app users living in Santiago are given in Web Appendix C. In particular, only users that have not spent more than 95% of time at their modal location are considered in this study. Indeed, we do not expect differences between static and dynamic personal exposure for users who rarely moved across the city over the period. As a result, we employ 2502 individuals for our study, which is a relatively large number when compared to other studies on personal exposure (Suarez et al., 2014). Note that the group of individuals is not claimed to be representative of the Santiago population and extending the results of this paper to the whole population of Santiago goes beyond the scope of our work. Relevant statistics on the data collected by smartphones are depicted in Web Figure 1.

Despite time-stamps are observed in continuous time, the period January 01, 2017 - April 30, 2017 is discretized with steps of 30 minutes starting from the midnight of January 1st, for a total of 5760 time steps. All the time-stamps are then related to the closest time step, with a maximum error of 15 minutes. Since the pollutant concentration is available on hourly scale, we argue that this approximation does not have a sensitive impact on the personal exposure assessment, while it greatly simplifies the modeling and expedites the computation.

2.2 *Air quality data*

Our focus is on Santiago area with pollutant concentrations derived from the city air quality monitoring network. According to Garcia-Chevesich et al. (2014), Santiago is "one of the cities with the most serious air pollution problems in the world" because of emissions from vehicles, manufacturing industries, fossil fuel and wood combustion, coupled with the fact that the city is surrounded by mountain ranges that prevent air drainage. Several epidemiological studies have provided evidence that air pollution exposure has serious health consequences for the population of Santiago (Franck et al., 2015; Prieto-Parra et al., 2017).

Chilean air quality is monitored by the National Air Quality Information System (SINCA). Within the geographic area covered by this study, SINCA currently runs 9 stations depicted in Figure 1. Here, air quality data consist in hourly PM_{10} concentrations measured at the 9 monitoring stations over the period January 01, 2017 - April 30, 2017, for a total of 2880 time steps; data are available at `https://sinca.mma.gob.cl`. Web Appendix A reports aggregated statistics for PM_{10} concentrations in $\mu g/m^3$ measured at each station as well as the relevant quantiles of the standard deviation of the PM_{10}; values suggest that the PM_{10} concentration is not constant across the area. This, in turn, means that the location of individuals matters in the assessment of their personal exposure.

3. Methodology

3.1 *Defining the personal exposure*

Let $\mathbf{s}_t = (s_{x,t}, s_{y,t})'$ denote a generic location over the geographic area $\mathcal{D}$ at time t, with $t = 1, \ldots, T$. The personal exposure of an individual known to be located at location $\mathbf{s}_t$ at time t, is defined as

$$E_t = y_t(\mathbf{s}_t), \tag{1}$$

where $y_t(\mathbf{s}_t) \geqslant 0$ is the pollutant concentration at the spatial location $\mathbf{s}_t$ and time t. In other words, the personal exposure corresponds to the pollutant concentration associated with individual location at time t. However, this definition is not operational since, in practice, both $\mathbf{s}_t$ and $y_t(\mathbf{s}_t)$ are unknown or known with error.

First, suppose that the individual location at time t is known, i.e., $\mathbf{s}_t$ is known. Let $\mathcal{Y}$ denote the set of pollutant measurements gathered at n monitoring stations over regularly time steps along a given period. Then, assessing the personal exposure translates in predicting the pollutant concentration at location $\mathbf{s}_t$, given the observed concentrations. In Bayesian words, we look for the posterior predictive distribution $p(y_t(\mathbf{s}_t) \mid \mathcal{Y})$ to evaluate the personal exposure for a given time; equivalently, we can write $p(E_t \mid \mathcal{Y})$. This predictive step is the Bayesian kriging operation; details on predictive sampling are deferred to Section 4.

In practice, also the individual location $\mathbf{s}_t$ is uncertain. However, we can make use of the spatial locations collected by the smartphone to learn the individual location along time. Let $\mathcal{S}$ be the set of smartphone spatial locations collected over the same period of $\mathcal{Y}$. Under the Bayesian framework, we use the posterior predictive density $p(\mathbf{s}_t \mid \mathcal{S})$ to infer the individual location at time t. Then, the personal exposure becomes a weighted average of the pollution level over the spatial domain, with weights given by the posterior predictive density of $\mathbf{s}_t$, that is

$$E_t^d = \int_{\mathcal{D}} y_t(\mathbf{s}_t) p(\mathbf{s}_t \mid \mathcal{S}) d\mathbf{s}_t. \tag{2}$$

We refer to definition in (2) to as the *dynamic* exposure. The integral in (2) cannot be evaluated explicitly but it is approximated numerically, as described in Section 4. Therefore, our goal becomes the posterior predictive distribution $p(E_t^d \mid \mathcal{Y}, \mathcal{S})$. In order to provide such posterior predictive distribution, a stochastic model for both pollutant measurements $\mathcal{Y}$ and smartphone locations $\mathcal{S}$ is needed. Here, $\mathcal{Y}$ and $\mathcal{S}$ are assumed to be independent. We discuss in Section 7 how to relax this assumption.

3.2 *Individual location model*

Learning individual locations at high temporal resolution is a challenging task because of the limits of GPS data discussed in the Introduction. An emerging approach used to predict individual's location relies on the reproducibility of human patterns, characterized by daily and weekly cyclical patterns (González et al., 2008; Song et al., 2010). For instance, Scellato et al. (2011) provided a spatio-temporal approach to predict arrival and residence times of users in their relevant places. Secchi et al. (2015) analyzed functional mobile data to identify sub-regions of the metropolitan area of Milan (Italy) sharing a similar pattern along time. However, such approaches are developed to understand people movement patterns rather than individual trajectories over space.

In principle, if we assume that the individual spends most of his/her time at few spatial locations (home, work, gym, etc.), then it is possible to group all the observed locations into a small number of clusters. For instance, Cho et al. (2011) employed a two-state mixture of Gaussian distributions centered at "home" and "work" locations to understand human motion from cell phone data and social networks. More generally, mixture models can be used to represent complex dynamic spatial distributions (Paci and Finazzi, 2018) and detect relevant places for individuals. However, model estimation is computational demanding and requires to estimate the number of clusters for each smartphone user, increasing the computational burden.

As an alternative, dynamic linear modeling has been successfully employed for animal tracking (Jonsen et al., 2005; Breed et al., 2012). Following this approach, we build a state space model that also incorporates cyclical patterns of people movements. We pursue model flexibility while saving computational feasibility.

The heartbeat signal sent by the smartphone at a given time t describes the smartphone location as the density of a Normal distribution, $N\left(\widetilde{\mathbf{s}}_t, \sigma_t^2 \mathbf{I}_2\right)$, where $\widetilde{\mathbf{s}}_t = (\widetilde{s}_{x,t}, \widetilde{s}_{y,t})'$ is the vector of latitude and longitude, $\mathbf{I}_2$ is the identity matrix and σ_t is the standard deviation estimated by the smartphone itself and assumed to be error free. Hence, we first assume a measurement error model such that the location provided by the smartphone is a noisy version of the "true" unknown location $\mathbf{s}_t$, that is:

$$\widetilde{\mathbf{s}}_t = \mathbf{s}_t + \boldsymbol{\varepsilon}_t, \tag{3}$$

where $\boldsymbol{\varepsilon}_t \sim N\left(\mathbf{0}, \sigma_t^2 \mathbf{I}_2\right)$ is the measurement error with known variance σ_t^2 (provided by the smartphone). Note that, given the time discretization, $\widetilde{\mathbf{s}}_t$ is missing if the smartphone has not sent the heartbeat signal in the interval $t \pm 15$ minutes. When this is the case, also σ_t^2 is missing since this information is included in the heartbeat signal along with $\widetilde{\mathbf{s}}_t$.

To model the latent location, we include fixed effects to describe daily and weekly routines of individuals. In fact, given a day of the week and a hour of the day, people tend to be at the same location and this information is used to reduce the predictive variance when σ_t^2 is high or the smartphone location is unavailable. In this work, distinction is made between working days and weekend days, as well as all the 24 hours in each day are supposed to have an effect on the individual location.

Hence, the latent individual location is modeled as

$$\begin{aligned} \mathbf{s}_t &= \sum_{h=1}^{48} \boldsymbol{\beta}_h \mathbb{1}\left(t \in h\right) + \mathbf{w}_t \\ \mathbf{w}_t &= \varphi\, \mathbf{w}_{t-1} + \boldsymbol{\eta}_t, \end{aligned} \tag{4}$$

where $\mathbb{1}$ is the indicator function and h (with abuse of notation) distinguishes the 24 hours in both working and weekend days such that the 2×1 vectors $\boldsymbol{\beta}_h$'s associated with the dummies capture the daily and weekly pattern; the hourly coefficients are constrained to sum to zero. Finally, φ is the autoregressive coefficient and $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\Lambda})$ is the innovation error. If available, any additional information useful to describe the individual location can be included. Model (3)-(4) is very flexible and feasible to be implemented for a large group of people.

Note that model (3)-(4) is specified independently for any smartphone user. In principle, this assumption may be relaxed for groups of people. For instance, the location of people living in the same house or working in the same place are related. However, learning the relationships among users is very complex since they likely vary within the day and over time. Moreover, the exploratory analysis displayed in Web Appendix A does not motivate the introduction of an interaction effect across users. Therefore, we rely here on the assumption that person/smartphone locations are independently observed across space and time.

3.3 *Pollutant concentration model*

In this work, hourly PM_{10} concentrations measured by the monitoring network are modeled through a hierarchical space-time model, that represents the gold standard when analyzing air quality data (see e.g., Sahu et al. 2006; Cocchi et al. 2007; Finazzi et al. 2013). Let $\widetilde{y}_{t'}(\mathbf{s})$ be the PM_{10} concentration collected at location $\mathbf{s}$ and generic time t' ($t' = 1, \ldots, T'$). Then, the model is given by:

$$\begin{aligned} \widetilde{y}_{t'}(\mathbf{s}) &= \gamma_{t'} + \upsilon_{t'}(\mathbf{s}) + \epsilon_{t'}(\mathbf{s}) \\ \upsilon_{t'}(\mathbf{s}) &= \phi\, \upsilon_{t'-1}(\mathbf{s}) + \zeta_{t'}(\mathbf{s}), \end{aligned} \tag{5}$$

where $\epsilon_{t'}(\mathbf{s})$ are white noise errors, normally distributed with zero mean and variance τ^2, and $\zeta_{t'}(\mathbf{s})$ are independent-in-time spatial random effects coming from a zero mean Gaussian process with isotropic spatial covariance function of form $C(\mathbf{s}, \mathbf{s} + h; \theta) = \delta^2 \rho(h; \theta)$. Here, $\rho(h; \theta)$ is the exponential correlation function depending on decay parameter θ.

Apart from the time-varying constant $\gamma_{t'}$, model (5) does not include additional fixed effects since no covariates are available here. This means that the spatio-temporal variability of the pollutant across space and over time is only induced by what observed at the monitoring stations. However, when high-resolution covariates are available for downscaling the pollutant, they can be easily integrated in (5).

4. Computational details

The Bayesian hierarchy of models (3)-(4) and (5) is completed by specifying the prior distribution for all model parameters, as detailed in Web Appendix D. The resulting posterior distributions are also reported in Web Appendix D. A Gibbs sampling scheme is adopted to approximate the joint posterior distribution under location model (3)-(4) as well as under pollutant model (5). The full conditional distributions of the model parameters are deferred to Web Appendix D.

Using MCMC methods, posterior sampling of the exposure at any time t is provided by composition as follows. According to Sections 3.1 and 3.3, we denote $y_t(\mathbf{s}) = \gamma_t + v_t(\mathbf{s})$ the underlying pollutant level. Then, a Monte Carlo integration is used to approximate the integral in (2) by discretizing the domain $\mathcal{D}$ into a regular grid of 0.015 degree, that is

$$E_t^d \approx \sum_{l=1}^{L} y_t(A_l) p(\mathbf{s}_t \in A_l \mid \mathcal{S}), \tag{6}$$

where A_l denotes a grid cell, with $l = 1, \dots, L$. This will require samples from the predictive distribution of the pollutant along with samples from the predictive distribution of the individual location. In particular, predictions of the pollutant over the grid are provided by the Bayesian kriging operation that requires sampling from the conditional distribution $p(v_t(A_l) \mid \mathbf{V})$. To clarify, we assume that the pollutant level is constant between t' and $t' + 1$ as well as within each grid cell A_l, so that the pollutant predictions are obtained at the centroid of each grid cell for each time t, with $t = 1, \dots, T$.

With regard to the predictive location, a posterior sample from $p(\mathbf{s}_t \mid \mathcal{S})$ is obtained, again, by composition. Then, the posterior probability of the individual to be located within grid cell A_l at time t is approximated as

$$p(\mathbf{s}_t \in A_l \mid \mathcal{S}) \approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left(\mathbf{s}_t^{(b)} \in A_l\right), \tag{7}$$

where $b = 1, \ldots, B$ is the MCMC sample and $\mathbb{1}$ is the indicator function. In summary, we implement the following algorithm to provide a posterior sample from $p(E_t^d \mid \mathcal{Y}, \mathcal{S})$:

(1) Draw a MCMC sample $\left\{\boldsymbol{\beta}^{(b)}, \mathbf{W}^{(b)}, \mathbf{G}^{(b)}, \boldsymbol{\Lambda}^{(b)}\right\}$ from the joint posterior distribution under model (4), with $b = 1, \ldots, B$ and compute $\mathbf{s}_t^{(b)} = \sum_{h=1}^{48} \boldsymbol{\beta}_h^{(b)} \mathbb{1}\,(t \in h) + \mathbf{w}_t^{(b)}$.

(2) Approximate $p(\mathbf{s}_t \in A_l \mid \mathcal{S})$ using equation (7).

(3) Draw a MCMC sample $\left\{\boldsymbol{\gamma}^{(k)}, \mathbf{V}^{(k)}, \phi^{(k)}, \lambda^{2(k)}, \tau^{2(k)}\right\}$ from the joint posterior distribution under model (5), with $k = 1, \ldots, K$.

(4) Draw a sample $v_t(A_l)^{(k)}$ from $p(v_t(A_l) \mid \mathbf{V})$ and compute $y_t(A_l)^{(k)} = \gamma_t^{(k)} + v_t(A_l)^{(k)}$.

(5) Finally, get a sample $E_t^{(k)}$ using equation (6).

As a result, a sample from the posterior distribution of the dynamic exposure E_t^d is obtained for any time t, i.e., half an hour. A MATLAB code (available as Supporting information) implements the MCMC scheme described above.

5. Dynamic vs static exposure

So far, we discussed how to obtain a dynamic personal exposure by exploiting the fact that people move across the space. It is of interest to compare such exposure with the personal exposure obtained assuming that the person is fixed in space, miming customary epidemiological studies based on population census data. In particular, we assume that the modal location over the study period, say $\hat{\mathbf{s}}$, represents the "home" location of the user. Then, the static personal exposure at time t is defined as

$$E_t^s = y_t(\hat{\mathbf{s}}), \tag{8}$$

that is the pollutant level for time t at the user's home location. Note that, with static exposure, the individual location is assumed to be known and constant over time. The posterior distribution of E_t^s is provided by kriging the pollutant level at site $\hat{\mathbf{s}}$ for any time. In other words, the posterior distribution of the static exposure (8) is the posterior predictive distribution of the PM$_{10}$ at the modal location $\hat{\mathbf{s}}$ for each time t, i.e., $p\left(E_t^s \mid \mathcal{Y}\right) = p\left(y_t(\hat{\mathbf{s}}) \mid \mathcal{Y}\right)$.

The extensive simulation study presented in Web Appendix F shows that, when the pollutant concentration varies over the city and the user visits places far from home during the day, the dynamic exposure provides a better estimate of the true personal exposure with respect to the static exposure.

Let $\Delta_t = E_t^d - E_t^s$ be the difference between the dynamic and the static personal exposure at time t. Then, we focus on the average difference over time, say

$$\Delta = \frac{1}{T}\sum_{t=1}^{T} \Delta_t. \tag{9}$$

The posterior distribution of Δ in (9) is, again, sampled by composition.

6. Analysis and results

We illustrate our approach by studying the personal exposure to PM$_{10}$ of 2502 users living in Santiago over the period January 01, 2017 - April, 30, 2017; see Section 2. Pollutant model (5) is fitted on PM$_{10}$ after a logarithmic transformation of the concentrations. Posterior summaries of model parameters are presented in Web Appendix E. Predictions are then back transformed to the original scale. Figure 1 displays the posterior mean surface of the PM$_{10}$ daily average (left panel) and the associated standard deviation map (right panel). Monitoring sites are superimposed. We note that average PM$_{10}$ is higher in the central area of the city while lower predictive variance is shown close to the stations, as expected.

[Figure 1 about here.]

Using data introduced in Section 2.1, we fit location model (3)-(4) for each smartphone user to provide its personal exposure, both static and dynamic. Web Appendix G offers the results of a validation study for the location model that shows the capability of the model to provide accurate location predictions.

As an illustration, Figure 2 displays the location tracking for a given smartphone user over a one-week period; credible intervals (CIs) are in shade. The plot makes evidence of the cyclical pattern of the user location over time. Also note that larger variability is associated with times at which the location is not provided by the smartphone. The top panel of Figure 3 shows the corresponding dynamic exposure of the same user over the same period; the uncertainty associated with the dynamic exposure results from blending pollutant (see Web Figure 6) and location uncertainties. For comparison, the bottom panel of Figure 3 displays the corresponding static exposure, i.e., assuming that the user was fixed at his/her modal location. Note that differences between the dynamic and static exposure are highlighted for some time steps as well as narrower CIs are associated with the dynamic exposure.

[Figure 2 about here.]

[Figure 3 about here.]

Figure 4 shows the posterior 95% CIs of average difference over time Δ for all users, ordered by posterior median. Roughly 30% of smartphone users exhibit a significant difference between dynamic and static exposure. In particular, 229 individuals show a significant negative difference up to 4.5 $\mu g/m^3$ every half an hour, on average, while 568 individuals are associated with a significant positive difference up to 6.0 $\mu g/m^3$ every half an hour, on average. This result is summarized in Figure 5 that illustrates all users at their modal-home location; main city roads are superimposed. Upward triangles on the left panel show users with a significantly higher dynamic exposure relative to the static one; downward triangles on the right panel show users with significantly lower dynamic exposure while dots on the mid panel show users with no significant difference between static and dynamic exposure. Reasonably, higher dynamic exposure is highlighted for users with modal location in low-polluted areas (see Figure 1) that routinely move to high-polluted areas along the observational time period. The opposite holds for people with modal location within the city center characterized by higher pollution level, on average. Supporting figures are presented in Web Appendix H.

[Figure 4 about here.]

[Figure 5 about here.]

7. Discussion

In this paper we proposed a methodology for quantifying personal exposure to air pollution by combining air quality data from monitoring networks and individual location data collected by smartphone apps. While air quality data are well studied in the literature, personal location data collected by smartphone apps are relatively recent and rarely available. Therefore, much effort has been dedicated to model the location data characterized by low sampling frequency, non-negligible error and cyclical patterns of individuals.

The methodology has been applied to a group of 2502 people in Santiago, taking part to EQN project. Location data collected by the smartphone app has been used to assess the dynamic exposure to PM_{10} over a 4-month period, uncertainty included. The dynamic exposure has been compared with the static exposure evaluated assuming that the individuals spend their entire time at a fixed location. We have showed that, for around 30% of the users, the smartphone-based exposure is significantly different from the static exposure. This can be helpful to tailor individual actions to mitigate the risk connected to air pollution.

We are aware that the modeling of both the pollutant level and the individual location can be improved to better explain their variability across space and over time. Nonetheless, our contribution has been to show how to blend such information sources to assess the personal exposure for a large group of persons.

As an extension, we can envision an interaction between people location and pollutant concentration. From one hand, we can account for the effect of air pollution on people movements in the location model. Indeed, some empirical studies have recently appeared in the literature with the aim of assessing whether people change their behavior in response to poor air quality (Welch et al., 2005; Borbet et al., 2018). Alerts reported by public/private agencies during severe pollution events may affect people behavior, since guidelines to reduce outdoor activities are often published during poor air quality episodes. For instance, sensitive persons (such as young children and/or older adults with underlying cardiac or pulmonary disease) may decide to avoid highly polluted areas. In this case, a further component can be introduced in model (4) to link individual movements to poor air quality episodes; e.g, a dummy variable that encodes whether the daily threshold has been exceeded, say the previous day. Moreover, people movements may depend on their perception about poor air quality (Semenza et al., 2008) rather than pollution measurements. Again, such information can be added to model (4) when available. On the other hand, the pollutant concentration may be driven by the presence of people in the area. Although it is not feasible to identify the effect of a single person to the air pollution, information about the dynamic spatial density of the population can be added to model (5), if available. Because of the set of smartphone users in our study is not representative of the population of the city, we only make inference at the individual level.

We acknowledge that personal exposure also arises from pollutant concentrations in indoor air; however, distinguishing between indoor and outdoor exposure is beyond the scope of this work. In fact, while smartphone location data may be used to understand when individuals are indoor or outdoor, indoor pollutant concentrations are hard to collect or assess.

In the future, miniaturized air quality sensors might be available, possibly on smartphones; however, several years are still needed for them to be spread enough for personal exposure assessment on large groups of people. Rather, location data are routinely collected by many smartphone apps similar to EQN app, for instance by social networks. In all these situations, personal exposure can be provided by employing the approach proposed in this work.

Acknowledgements

Funding for the authors was partially provided by PRIN2015 project 'Environmental processes and human activities: capturing their interactions via statistical methods', grant n. 20154X8K23. We thank the anonymous reviewers for their useful comments which have improved the paper.

References

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2011). Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics* **68,** 837–848.

Borbet, T. C., Gladson, L. A., and Cromar, K. R. (2018). Assessing air quality index awareness and use in Mexico City. *BMC Public Health* **18,** 538.

Breed, G. A., Costa, D. P., Jonsen, I. D., Robinson, P. W., and Mills-Flemming, J. (2012). State-space methods for more completely capturing behavioral dynamics from animal tracks. *Ecological Modelling* **235-236,** 49 – 58.

Cesaroni, G., Badaloni, C., Gariazzo, C., Stafoggia, M., R, S., Davoli, M., and Forastiere, F. (2013). Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. *Environmental Health Perspectives* **121,** 324–331.

Chaix, B., Méline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., and et al. (2013). GPS tracking in neighborhood and health studies: A step forward for environmental exposure assessment, a step backward for causal inference? *Health & Place* **21,** 46 – 51.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090, New York, NY, USA. ACM.

Cocchi, D., Greco, F., and Trivisano, C. (2007). Hierarchical space-time modelling of PM10 pollution. *Atmospheric Environment* **41,** 532 – 542.

Di, Q., Dai, L., Wang, Y., Zanobetti, A., Choirat, C., Schwartz, J. D., and Dominici, F. (2017). Association of short-term exposure to air pollution with mortality in older adults. *Journal of the American Medical Association* **318,** 2446–2456.

Fassò, A., Finazzi, F., and Ndongo, F. (2016). European population exposure to airborne pollutants based on a multivariate spatio-temporal model. *Journal of Agricultural, Biological, and Environmental Statistics* **21,**.

Finazzi, F. (2016). The Earthquake Network project: toward a crowdsourced smartphone-based earthquake early warning system. *Bulletin of the Seismological Society of America* **106,** 1088–1099.

Finazzi, F. and Fassò, A. (2017). A statistical approach to crowdsourced smartphone-based earthquake early warning systems. *Stochastic Environmental Reseaerch Risk Assessment* **31,** 1649–1658.

Finazzi, F., Scott, E. M., and Fassò, A. (2013). A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62,** 287–308.

Franck, U., Leitte, A. M., and Suppan, P. (2015). Multifactorial airborne exposures and respiratory hospital admissions: The example of Santiago de Chile. *Science of The Total Environment* **502,** 114 – 121.

Garcia-Chevesich, P. A., Alvarado, S., Neary, D. G., Valdes, R., Valdes, J., Aguirre, J. J., and et al. (2014). Respiratory disease and particulate air pollution in Santiago Chile: Contribution of erosion particles from fine sediments. *Environmental Pollution* **187,** 202 – 205.

Gariazzo, C., Pelliccioni, A., and Bolignano, A. (2016). A dynamic urban air pollution population exposure assessment study using model and population density data derived by mobile phone traffic. *Atmospheric Environment* **131,** 289 – 300.

González, M. C., Hidalgo, C. A., and Barabási, A. (2008). Understanding individual human mobility patterns. *Nature* **453,** 779–782.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., and et al. (2013). Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental Health* **12,** 43.

Jonsen, I. D., Flemming, J. M., and Myers, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology* **86,** 2874–2880.

Lee, D. and Shaddick, G. (2010). Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics* **66,** 1238–1246.

Liu, H.-Y., Skjetne, E., and Kobernus, M. (2013). Mobile phone tracking: in support of modelling traffic-related air pollution contribution to individual exposure and its implications for public health impact assessment. *Environmental Health* **12,** 93.

Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., and et al. (2016). Exposure track: the impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. *Environmental Science & Technology* **50,** 9671–9681.

Paci, L. and Finazzi, F. (2018). Dynamic model-based clustering for spatio-temporal data. *Statistics and Computing* **28,** 359–374.

Paci, L., Gelfand, A. E., and Holland, D. (2013). Spatio-temporal modeling for real-time ozone forecasting. *Spatial Statistics* **4,** 79–93.

Park, Y. M. and Kwan, M.-P. (2017). Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. *Health & Place* **43,** 85–94.

Prieto-Parra, L., Yohannessen, K., Brea, C., Vidal, D., Ubilla, C. A., and Ruiz-Rudolph, P. (2017). Air pollution, PM2.5 composition, source factors, and respiratory symptoms in asthmatic and nonasthmatic children in Santiago, Chile. *Environment International* **101,** 190–200.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics* **11,** 61–86.

Scellato, S., Musolesi, M., Mascolo, C., Latora, V., and Campbell, A. T. (2011). Nextplace: A spatio-temporal prediction framework for pervasive systems. In Lyons, K., Hightower, J., and Huang, E. M., editors, *Pervasive Computing: 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15, 2011. Proceedings*, pages 152–169. Springer, Berlin, Heidelberg.

Secchi, P., Vantini, S., and Vitelli, V. (2015). Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Statistical Methods & Applications* **24,** 279–300.

Semenza, J. C., Wilson, D. J., Parra, J., Bontempo, B. D., Hart, M., Sailor, D. J., and George, L. A. (2008). Public perception and behavior change in relationship to hot weather and air pollution. *Environmental Research* **107,** 401 – 411.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science* **327,** 1018–1021.

Steinle, S., Reis, S., and Sabel, C. E. (2013). Quantifying human exposure to air pollution: Moving from static monitoring to spatio-temporally resolved personal exposure assessment. *Science of The Total Environment* **443,** 184–193.

Suarez, L., Mesias, S., Iglesias, V., Silva, C., Caceres, D. D., and Ruiz-Rudolph, P. (2014). Personal exposure to particulate matter in commuters using different transport modes (bus, bicycle, car and subway) in an assigned route in downtown Santiago, Chile. *Environmental Science: Processes & Impacts* **16,** 1309–1317.

Valdés, A., Zanobetti, A., Halonen, J. I., Cifuentes, L., Morata, D., and Schwartz, J. (2012). Elemental concentrations of ambient particles and cause specific mortality in Santiago, Chile: a time series study. *Environmental Health* **11,** 82.

Welch, E., Gu, X., and Kramer, L. (2005). The effects of ozone action day public advisories on train ridership in Chicago. *Transportation Research, Part D: Transport and Environment* **10,** 445–458.

Welty, L. J., Peng, R. D., Zeger, S. L., and Dominici, F. (2009). Bayesian Distributed Lag Models: Estimating effects of particulate matter air pollution on daily mortality. *Biometrics* **65,** 282–291.

WHO (2005). Principles of characterizing and applying human exposure models. IPCS harmonization project document 3, World Health Organization, Geneva.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article. Supporting information contains Web Tables and Figures referenced in Sections 2.1, 2.2 and 6 as well as MCMC details and the results of the simulation study. We also provide the MATLAB code that implements the methods described in Section 3 as supporting information.

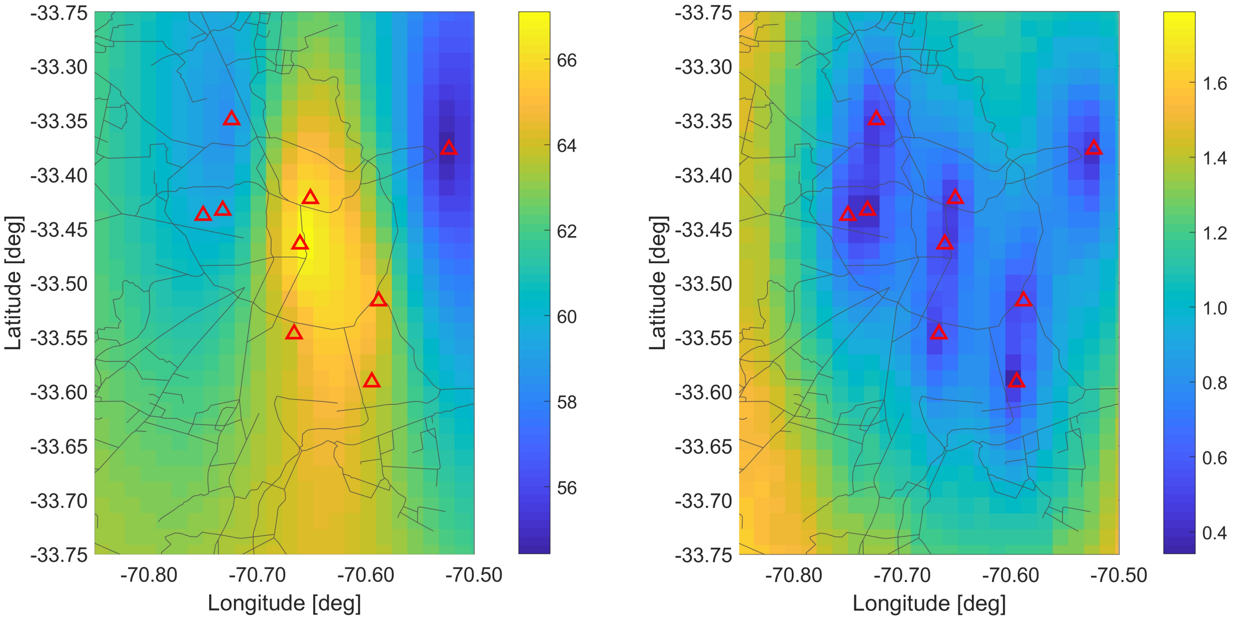

Figure 1. Posterior mean surface of the PM_{10} daily average in $\mu g/m^3$ (left panel) and standard deviation map in $\mu g/m^3$ (right panel). Triangles represent the monitoring stations. This figure appears in color in the electronic version of this article.

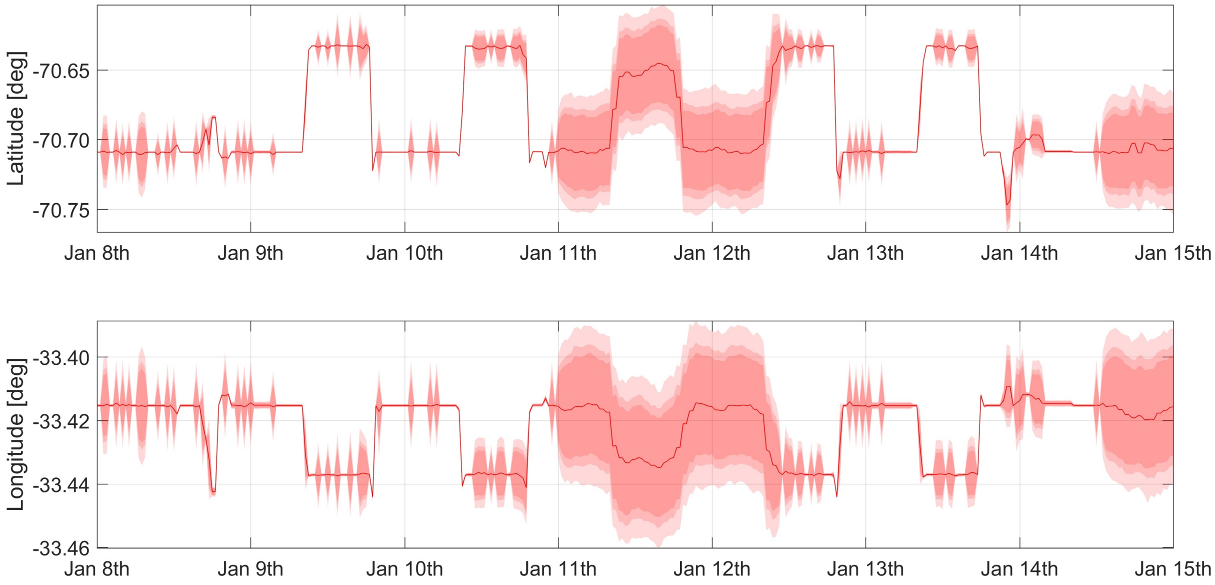

Figure 2. Location tracking for a given smartphone user over a one-week period; CIs are in shade. This figure appears in color in the electronic version of this article.

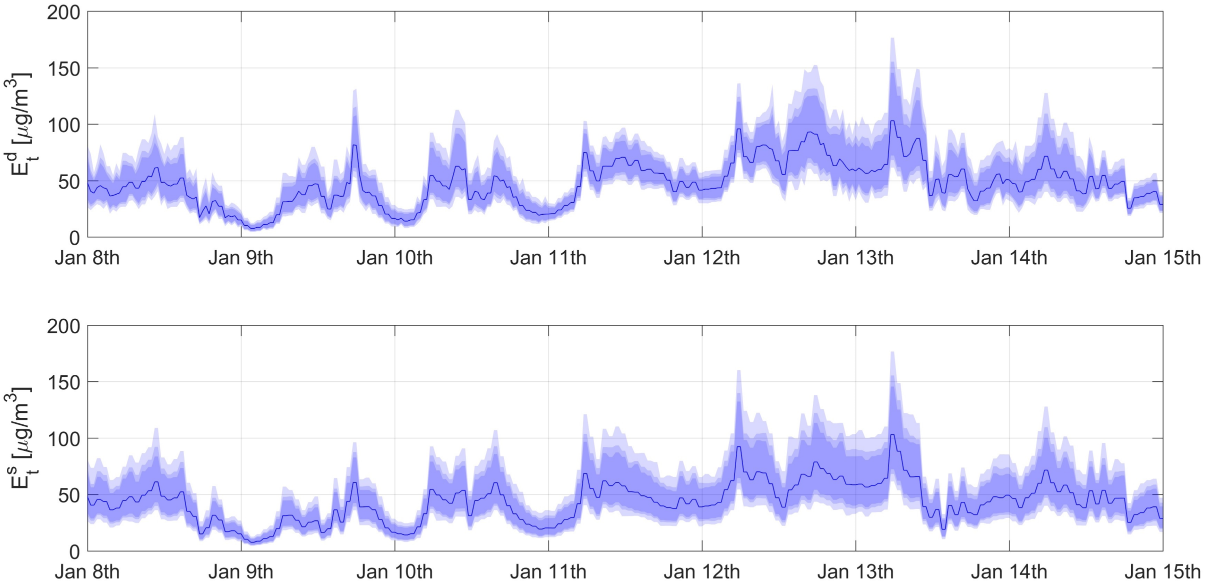

Figure 3. Posterior distributions of the dynamic (top panel) and static (bottom panel) exposure associated with user in Figure 2; CIs are in shade. This figure appears in color in the electronic version of this article.

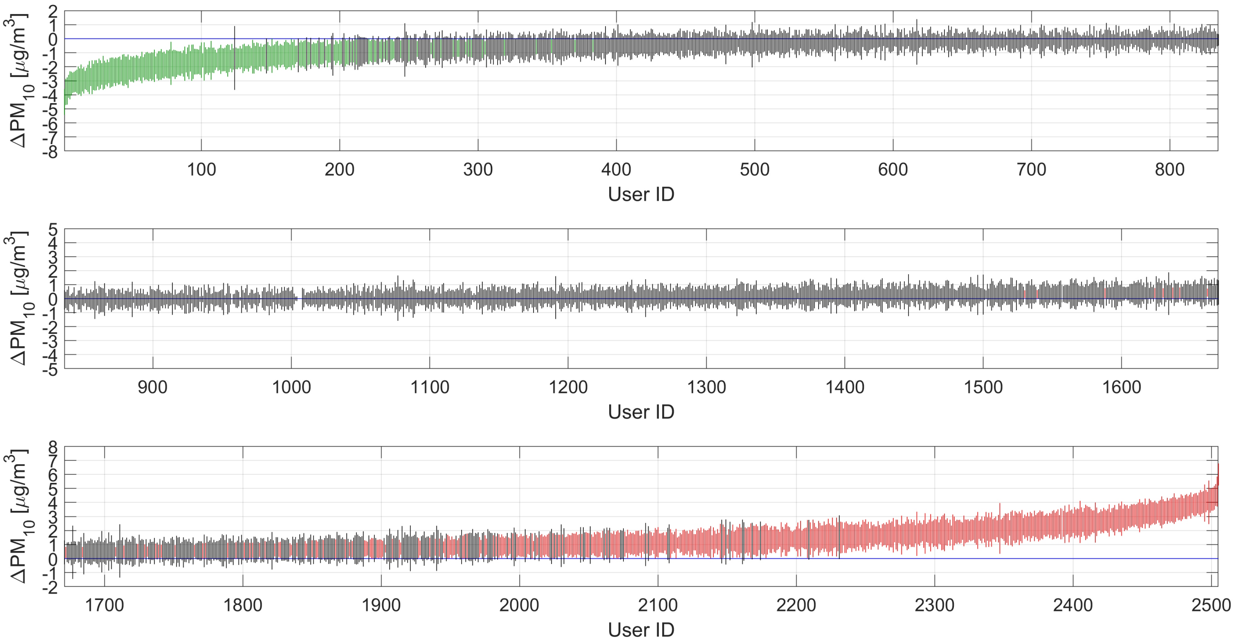

Figure 4. Posterior 95% CIs of Δ in (9) for all users ordered by posterior median. This figure appears in color in the electronic version of this article.

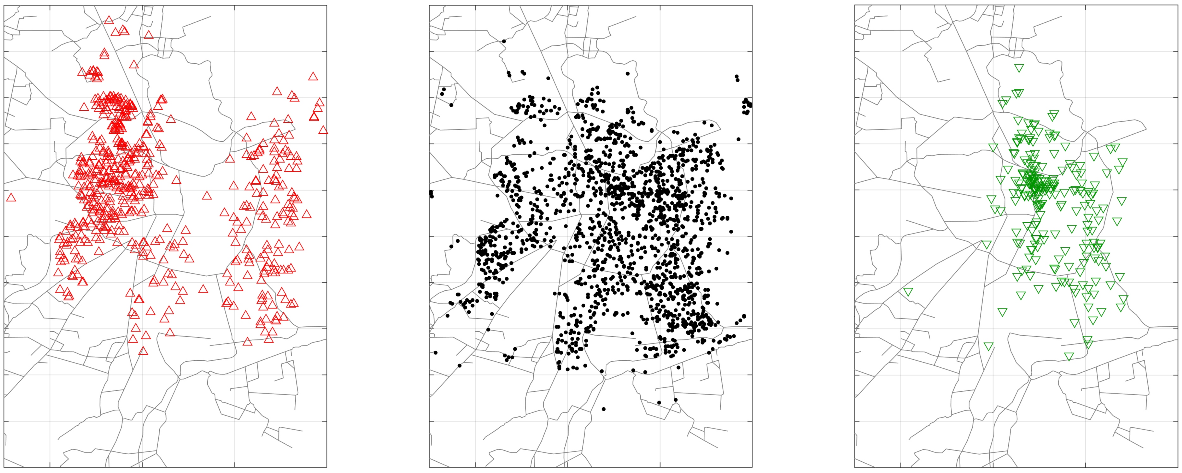

Figure 5. Modal locations of smartphone users: upward triangles on the left panel show users with a significantly higher dynamic exposure relative to the static one; downward triangles on the right panel show users with significantly lower dynamic exposure; dots on the mid panel show users with no significant difference between static and dynamic exposure. This figure appears in color in the electronic version of this article.