Quantifying the Impact of Daily Mobility on Errors in Air Pollution Exposure Estimation Using Mobile Phone Location Data

Xiaonan Yu[1], Cesunica Ivey[2], Zhijiong Huang[3], Sashikanth Gurram[4], Vijayaraghavan Sivaraman[4], Huizhong Shen[5], Naveen Eluru[1], Samiul Hasan[1], Lucas Henneman[6], Guoliang Shi[7], Hongliang Zhang[8], Haofei Yu[1,*], Junyu Zheng[3]

[1]Department of Civil, Environmental, and Construction Engineering. University of Central Florida. Orlando, FL. USA
[2]Department of Chemical and Environmental Engineering, University of California Riverside. Riverside, CA. USA
[3]Inisitute for Environmental and Climate Research, Jinan University, Guangzhou, China.
[4]AirSage, Atlanta, GA. USA
[5]School of Civil and Environmental Engineering. Georgia Institute of Technology. Atlanta, GA. USA.
[6]T.H. Chan School of Public Health, Harvard University. Cambridge, MA. USA
[7]College of Environmental Science and Engineering, Nankai University. Tianjin, China.
[8]Department of Environmental Science and Engineering, Fudan University. Shanghai, China

*Corresponding authors
Haofei Yu, Ph.D.
Address: 12800 Pegasus Drive Suite 211. Orlando, FL 32816
E-mail: haofei.yu@ucf.edu
Phone: +1 (407) 823-1309
Fax: +1 (407) 823-3315

Abstract

One major source of uncertainty in accurately estimating human exposure to air pollution is that human subjects move spatiotemporally, and such mobility is usually not considered in exposure estimation. How such mobility impacts exposure estimates at the population and individual level, particularly for subjects with different levels of mobility, remains under-investigated. In addition, a wide range of methods have been used in the past to develop air pollutant concentration fields for related health studies. How the choices of methods impact results of exposure estimation, especially when detailed mobility information is considered, is still largely unknown. In this study, by using a publicly available large cell phone location dataset containing over 35 million location records collected from 310,989 subjects, we investigated the impact of individual subjects' mobility on their estimated exposures for five chosen ambient pollutants (CO, NO_2, SO_2, O_3 and $PM_{2.5}$). We also estimated exposures separately for 10 groups of subjects with different levels of mobility to explore how increased mobility impacted their exposure estimates. Further, we applied and compared two methods to develop concentration fields for exposure estimation, including one based on CMAQ model outputs, and the other based on the interpolated observed pollutant concentrations using the inverse distance weighting (IDW) method. Our results suggest that detailed mobility information does not have a significant influence on mean population exposure estimate in our sample population, although impacts can be substantial at the individual level. Additionally, exposure classification error due to the use of home-location data only increased for subjects that exhibited higher

levels of mobility. Omitting mobility could result in underestimation of exposures to traffic-related pollutants particularly during afternoon rush-hour, and overestimate exposures to ozone especially during mid-afternoon. Between CMAQ and IDW, we found that the IDW method generates smooth concentration fields that were not suitable for exposure estimation with detailed mobility data. Therefore, the method for developing air pollution concentration fields when detailed mobility data were to be applied should be chosen carefully. Our findings have important implications for future air pollution health studies.

Keywords

Air pollution exposure; exposure misclassification; human mobility; cell phone location data; call detail record

1. Introduction

Exposure to air pollution is the second leading cause of non-communicable disease worldwide [1]. It is also associated with more than 4 million premature deaths annually [2, 3] and numerous other negative health consequences [4-10]. An accurate estimation of human exposure to air pollution is critical for assessing the potential connections between air pollution exposure and certain health outcomes, and for quantifying the health impacts of air pollution [11-14]. In many prior air pollution health studies, human exposure to air pollution was estimated using concentration data collected or simulated at the location of subjects' home addresses [15, 16], or even at further aggregated zones such as census tract [17] or ZIP code level [18]. Detailed spatiotemporal movements of subjects, i.e. human mobility, were usually omitted due to lack of data. This home-based exposure (herein referred to as HBE), could introduce considerable amount of exposure classification errors [19-24], which could potentially bias subsequent statistical analyses [25, 26].

To address this issue, a variety of methods have been adopted, including utilizing travel surveys and diaries [19, 27], personal measurements [28, 29], accounting for multiple addresses (e.g., residential or work address) or full-day travel data [19, 24] during the temporal window of exposure [15, 25, 30, 31], tracking subjects using GPS-enabled surveys [22, 32], and employing a variety of modeling tools and techniques to account for mobility [21, 33]. Though prior results suggest exposure estimation errors due to the omission of mobility could differ among individuals with different mobility patterns [19, 24], the direction and magnitude of such errors remains under-investigated. Further, numerous methods have been used in the past to develop pollutant concentration fields for air pollution health studies, and the developed fields vary substantially spatially and temporally [34-36]. How the choices of method impact exposure estimates when human mobility is considered is still largely unknown.

In our exploratory study [23], we demonstrated the feasibility of using cell phone location dataset in air pollution exposure estimation using a relatively small sample population (n = 9,886). Here, build upon our previous work, we: 1) applied two methods to develop pollution concentration fields, and investigated the impact of different methods on exposure estimates when detailed mobility information were considered; 2) included a substantially larger sample population (n = 310,989), divided the entire population into 10 groups with varying mobility levels, and investigated how different mobility impact exposure estimates; 3) investigated the temporal variability of exposure estimates among groups with different mobility levels; 4) investigated how exposure classification errors change due to mobility; 5) quantified the impact of exposure classification errors on subsequent health effect estimations. Details on the methods used in this study are presented in the next section, followed by the results of the study and a discussion of the potential of the methods and data, as well as associated limitations.

2. Material and Methods

2.1. Data description and study area

The cell phone location data applied here are Call Detail Record (CDR) data collected by mobile network operators. CDR data are collected from cellphones when the phone communicates with a nearby cell towers, specifically, when a network subscriber's cell phone communicates with a nearby cell tower (such as phone call, text messaging, or mobile data request), a suite of information is generated and archived for billing purposes [37-39]. The archived information contains the identities of cell towers that handle the communication, and the tower locations are already known. CDR data contains tremendous amount of digital footprints for virtually all subscribers of the network, and it has been extensively used in criminal investigation [40, 41], the study of human mobility [39, 42, 43], and urban and transportation planning [44-46]. It's worth noting that location information contained in CDR data are not the locations of cellphone users, rather they are the locations of nearby cellphone tower that handled the user's wireless communication.

In this study, we obtained a publicly available CDR dataset for Shenzhen, China [38, 47]. Shenzhen is a major city located in the Guangdong Province (Figure 1). It has an area of 1,991 km^2 and over 12 million residents, making it one of the most populated cities worldwide. The original CDR dataset contains over 38 million location records collected from 414,271 anonymized Subscriber Identification Module (SIM) cards on one typical weekday in October 2013. We excluded SIM cards with no location data available at night (here defined as after 8 pm and before 7 am), which is required to infer potential home addresses. The filtered CDR dataset applied here has 35.6 million location records for 310,989 unique SIM cards (herein referred to as subjects), with an average of approximately 115 records per subject per day. All identifiers contained in the original CDR data were removed from this database, leaving only a randomized SIM card ID, a time stamp, and latitude and longitude. This information was used to construct daily mobility patterns for each subject.

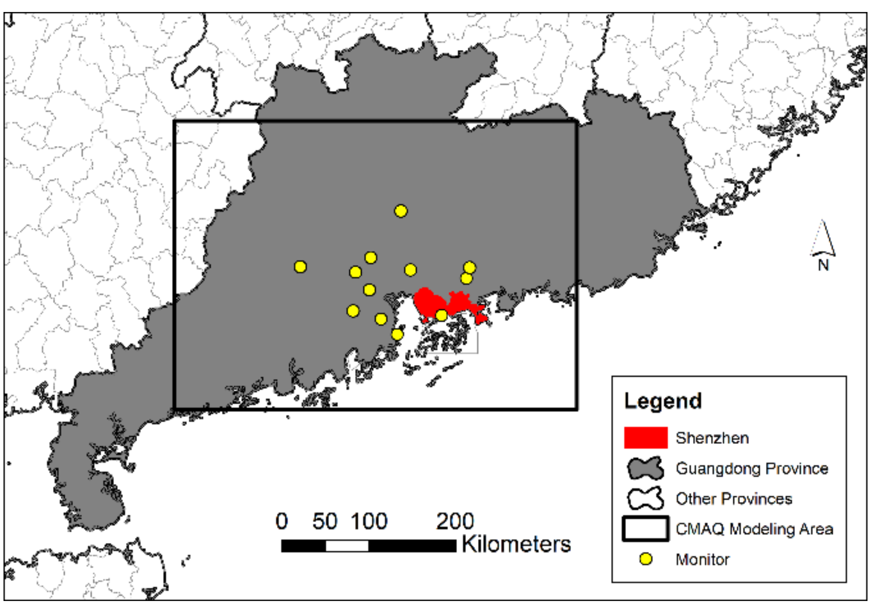

Figure 1. The study area of Shenzhen, China

2.2. Exposure estimation

Five pollutants were selected for this study, including carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ground-level ozone (O_3), and particulate matter with the aerodynamic diameter less than 2.5 µm ($PM_{2.5}$). All of these pollutants are important air pollutants regulated in both the United States (National Ambient Air Quality Standards) and China (GB3095-2012), and they are considered to pose harmful effects to human health and the environment, not only for the US and China, but also worldwide.

Similar to our previous study [23], we estimated all subjects' exposures to the five chosen pollutants using two methods: a static, home-based exposure (HBE) calculated by assuming all subjects stay at their corresponding home locations throughout the entire day; and a dynamic, CDR-based exposure (CDRE) calculated by matching detailed CDR location data with modeled pollutant concentrations at the corresponding locations. Specifically, HBE and CDRE are estimated as:

$$HBE = \frac{\sum_{h=1}^{n} C_{h,g}}{n}$$

$$CDRE = \frac{\sum_{h=1}^{n} \sum_{m=1}^{k} C_{h,m}}{n}$$

Where $C_{h,g}$ is pollutant concentration in hour h at the grid cell g where the corresponding subject' home is located; n is the total amount of hours in the study period (n = 24); $C_{h,m}$ is pollutant concentration in hour h at grid cell m where the subject is located within the corresponding hour. The subject may be located in k (k >= 1) grid cells in hour h. In the static method, each subject's home location was assumed to be their most frequent location at night (between 8 pm and 7 am), and we used modeled pollutant concentration data at their corresponding home location to estimate their exposures. In the dynamic method, the CDRE was estimated by arithmetically weighting concentrations at different locations where the subject visited based on the time (in hours) the subject spent at each location. If no location data was available for one specific hour, we assumed the subject stayed at the same location as in the previous hour. If location data was missing for the first hour (12 am – 1 am), the subject was assumed to be at their estimated home locations. For hours with multiple location records available, we used averaged concentration from all locations in the corresponding hour. We estimated HBE and CDRE for each subject separately.

Different from our previous study [23], we applied two approaches to develop spatiotemporal concentration fields of the five chosen pollutants: one based on outputs from the Community Multiscale Air Quality (CMAQ) model [48] for the corresponding day, and the other using the inverse distance weighting (IDW) method. Detailed information on CMAQ model configurations is available elsewhere [49]. To correct for potential model biases and errors, we fused hourly measurement data collected from 12 monitoring stations inside the CMAQ modeling domain (Figure 1) into CMAQ output by multiplying gridded hourly CMAQ fields with adjustment factors. The factors were calculated as the ratio between measured and modeled concentrations at the locations of each monitoring station, and then spatially interpolated to the center points of all CMAQ grid cells using kriging [34]. For the IDW method, we spatially interpolated hourly measurements from all monitoring stations inside the study area using inversed and squared distance as the weight. The spatial and temporal resolution of the concentration fields for both methods are 3 km and 1-hour, respectively. We acknowledge that an individual's exposure to air pollution occur at finer scales, we nonetheless still applied the aforementioned CMAQ and IDW fields mainly for two reasons: 1) Developing higher resolution pollution fields are not feasible in this study due to the limited availability of measurement data in the study area (Figure 1), and computational burden involved in running higher resolution CMAQ simulations; and 2) the location information in CDR contains the locations of cellphone towers close to the corresponding cellphone user. In addition, it's important to note that the aforementioned CMAQ and IDW methods are fundamentally different, and the results of exposure assessment are expected to be impacted substantially by the choice of methods.

To understand how different degrees of mobility impact exposure estimation, we further subdivided all subjects into 10 groups based on the number of unique CMAQ grid cells each individual subject visited during the day. The number of grid cells each subject visited in group 1 through 9 correspond to their respective group number, while all subjects that visited 10 or more unique grid cells were collectively assigned into group 10. Subjects in groups with larger group numbers are expected to have a high degree of mobility. We estimated HBE and CDRE separately for all 10 groups. While metrics, such as distance between home and work location [25], have been used in past studies. However, such information is not available in this study.

In epidemiological studies related to air pollution, subjects are frequently assigned to different groups based on their exposure levels (such as quartiles) [31, 50-53]. Statistical comparisons are then performed among these groups to investigate whether high exposure levels are associated with a higher incidence of certain health outcomes. The statistical analysis could be biased or confounded if subjects were misclassified into the wrong exposure group. To explore the impact of including detailed mobility data on exposure misclassification, we compared how subjects were assigned to four quartiles based on their CDRE and HBE. We define "misclassification" as the assignment of one subject, based on HBE, into a quartile that is different from CDRE-based quartile.

We performed the Wilcoxon rank sum test to examine whether the medians of CDRE and HBE exposure estimates are statistically different. We chose this test because the samples in this study are not normally distributed. Furthermore, we also calculated the expected bias factors to quantify potential biases in relative risk estimates when HBE was used [25, 54]. According to the classical error theory, exposure estimated using the home-based method may be expressed as:

$$Z = X + E \tag{1}$$

In equation 1, Z is exposure estimated using HBE; X is the true exposure value; and E is the error associated with the corresponding HBE. In this study, we use CDRE to represent X, and, based on our previous results, E is correlated with X *[23]*. Therefore, the following equation can be applied to calculate a bias factor [25, 54, 55]:

$$B = \frac{\sigma^2+\varphi}{\sigma^2+2\varphi+\omega^2} \tag{2}$$

In equation 2, B is the calculated bias factor; σ^2 is the variance of CDRE of all subjects; φ is the covariance between CDRE and errors in exposure estimation (calculated based on HBE-CDRE); and ω^2 is the variance of the errors in exposure estimation. The factor B represents the expected bias in relative risk estimates when the home-based method is applied. For example, a B factor of 0.75 suggests that applying the home-based method would lead to the relative risk being underestimated by 25%. It's also worth noting that the Wilcoxon rank sum test is a different statistical measure compared to the coefficient of determination

(R^2). The former intends to test equality, while the latter quantifies the proportion of variance contained in the dependent variable that can be predicted by the independent variable.

3. Results

3.1. Concentration fields

The spatial concentration fields of the five chosen pollutants simulated by the CMAQ and IDW methods differ considerably (Figure 2), especially for O_3, NO_2, and $PM_{2.5}$, where the latter two pollutants are known to have substantial primary contributions from transportation sectors. Due to the sparseness of monitor network, the IDW method generally results in smoother fields that lack spatial variabilities compared with the CMAQ method. The locations of monitoring stations can also be observed on the concentration fields as simulated by the IDW method (Figure S1).

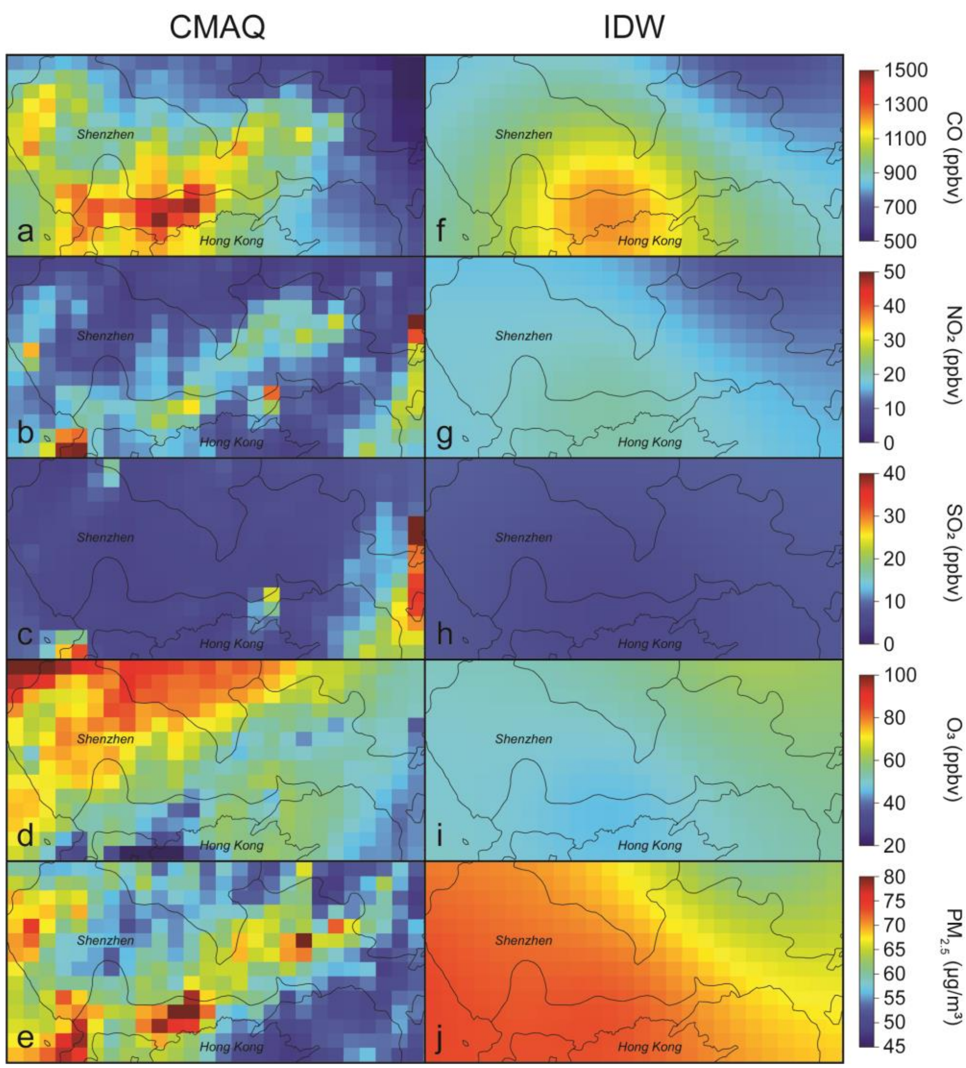

Figure 2. Spatial fields of concentrations of the five chosen pollutants as simulated by the CMAQ (a-e) and IDW (f-j) methods

3.2. Overall correlations between HBE and CDRE

Mean CMAQ-based HBE and CDRE estimates for all subjects were highly correlated with each other (Figure 3). The coefficient of determination (R^2) ranged from 0.95 (NO_2) to 0.98 (SO_2), with the slopes of linear regression close to 1, and intercepts were close to 0 for all pollutants. The estimated regression parameters are considerably different comparing with our previous study [23] (e.g: R^2 ranged between 0.65 to 0.76 in the previous study). We also observed many vertically aligned data points, suggesting many subjects had identical HBE but their CDRE was considerably different when individual mobility was considered. Additionally, a large number of data points were clustered near the 1:1 line, suggesting that a substantial portion of the subjects had similar HBE and CDRE.

Similar findings were also observed for IDW-based exposures (Figure 3), including the clustered data points along the 1:1 line, the high overall correlations between HBE and CDRE, and the varying CDRE estimates for many subjects with identical HBE estimates. However, the range of estimates for both HBE and CDRE were much smaller for the IDW exposures, particularly for NO_2, O_3 and $PM_{2.5}$, where the vast majority of data points were clustered within small concentration ranges. It's also worth noting that results of Wilcoxon rank sum tests show HBE and CDRE are overall statistically different for all pollutants.

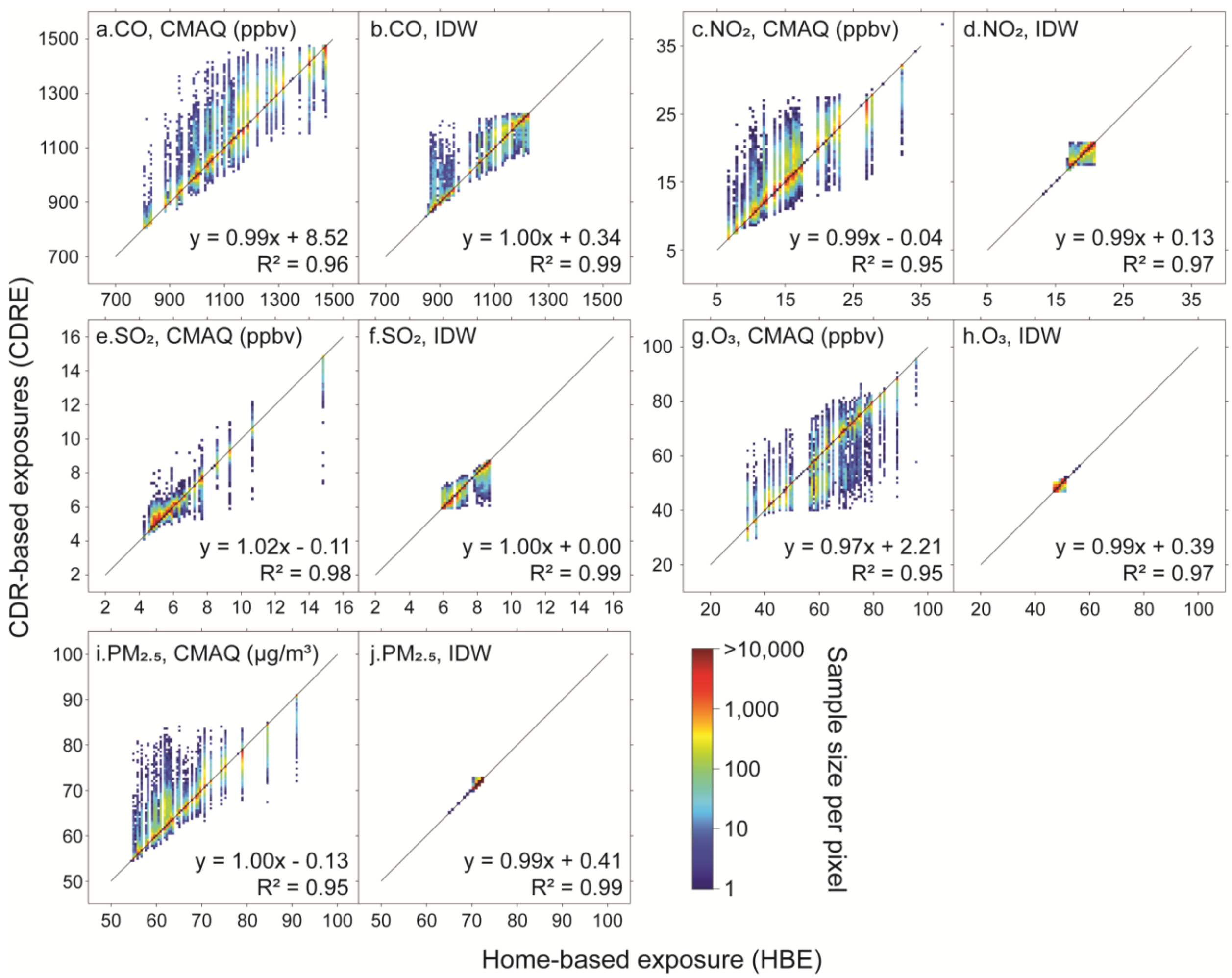

Figure 3. Linear correlations between HBE and CDRE estimates of the five chosen pollutants for all subjects based on CMAQ (a,c,e,g,i) and IDW (b,d,f,h,j) concentration fields. Pixels are color coded by sample size. The solid black line shown is the 1:1 line.

3.3. The impact of mobility on exposure estimates

We found that the correlations between HBE and CDRE estimates shrink with an increased degree of mobility (NO_2 presented in Table 2, other pollutants in Tables S2 through S5). Compared with CMAQ, the decreasing correlations between CDRE and HBE were smaller when IDW fields were used, with considerably smaller RMSE, MNB and MNE. For $PM_{2.5}$, as shown by the numbers presented in Table S5, the RMSE, MNB and MNE for the group with the highest degree of mobility (group 10) was only 5.4%, 6.7%, and 4.6%, respectively, of those when CMAQ fields were used. For example, the MNE for group 10 is 3.23% when CMAQ fields were used, but only 0.15% when IDW fields were used. The only exception is SO_2 (Table S3), for which the RMSE and MNE changed similarly between the CMAQ and IDW methods, though MNB is only 0.9% when the IDW method was applied.

Table 2. Comparison between HBE and CDRE estimate of NO_2 for all ten groups with different mobility

		Group number									
		1	2	3	4	5	6	7	8	9	10
CMAQ	CDRE mean (ppbv)	16.1	16.6	16.7	16.8	16.7	16.3	15.9	15.9	15.6	15.6
	HBE mean (ppbv)	16.1	16.5	16.3	16.2	15.8	15.5	15.2	15.2	15.0	15.1
	[a]RMSE (ppbv)	0.00	1.16	1.79	2.16	2.50	2.60	2.62	2.74	2.78	3.02
	[b]MNB (%)	0.0%	-0.8%	-2.3%	-3.8%	-5.0%	-4.9%	-4.3%	-4.1%	-3.5%	-2.8%
	[c]MNE (%)	0.0%	3.6%	6.2%	8.1%	9.8%	10.5%	10.6%	10.8%	11.2%	11.9%
	[d]R^2	1.00	0.95	0.88	0.83	0.76	0.72	0.70	0.67	0.66	0.64
IDW	CDRE mean (ppbv)	19.4	19.2	19.3	19.3	19.3	19.2	19.1	19.1	19.0	19.0
	HBE mean (ppbv)	19.4	19.2	19.3	19.3	19.3	19.2	19.1	19.1	19.0	19.0
	[a]RMSE	0.00	0.23	0.35	0.43	0.49	0.56	0.62	0.62	0.67	0.72
	[b]MNB (%)	0.0%	0.0%	-0.1%	-0.1%	-0.2%	-0.1%	0.0%	0.0%	0.2%	0.4%
	[c]MNE (%)	0.0%	0.4%	0.8%	1.1%	1.4%	1.7%	1.9%	2.0%	2.3%	2.4%
	[d]R^2	1.00	0.98	0.94	0.92	0.88	0.85	0.81	0.81	0.78	0.75
	Sample size	167570	75313	32177	16350	8354	4617	2700	1562	916	1430

[a]RMSE: root mean squared error. Calculated as $[\frac{1}{N}\sum_{i=1}^{N}(HBE_i - CDRE_i)^2]^{1/2}$, where CDRE and HBE is the estimated exposures based on CDR and home-based method for the *i*th subject

[b]MNB: mean normalized bias. Calculated as $\frac{1}{N}\sum_{i=1}^{N}(\frac{HBE_i-CDRE_i}{CDRE_i})$

[c]MNE: mean normalized error. Calculated as $\frac{1}{N}\sum_{i=1}^{N}\left|\frac{HBE_i-CDRE_i}{CDRE_i}\right|$

[d]R^2: coefficient of determination between HBE and CDRE estimates in the corresponding group.

In this dataset, over half (54%) of all subjects stayed in the same 3 km grid cell throughout the entire day, and the majority (94%) of all subjects visited 4 or fewer grid cells (Table 2). Although subjects that were highly mobile (especially those who visited 6 and more grid cells) accounted for a relatively small fraction of the entire population, the sample sizes of all groups were still considerable due to the large overall sample population (sample size = 916 for the smallest group, group 9).

The impacts of mobility on exposure estimates differ by pollutant and by concentration fields used. Between CMAQ and IDW methods, the range of variability was considerably smaller when the IDW method was applied, particularly for NO_2, O_3 and $PM_{2.5}$. SO_2 again was the exception where exposure variability was similar between the two methods. Mobility had the greatest impact for NO_2 and O_3. When CMAQ concentration fields were applied, the observed differences were more negative (higher CDRE than HBE) for CO, NO_2 and $PM_{2.5}$, but were more positive (lower CDRE than HBE) for O_3. Such observations are not clearly visible when the IDW concentration fields were applied.

The impacts of mobility on exposures also differed by time of the day (Figure 4), with larger differences found during daytime for all groups, though the biggest difference occurred at different hours for different pollutants. When CMAQ concentration fields were applied, CO, NO_2 and $PM_{2.5}$ exhibited the largest differences near the afternoon rush hour, though these differences dissipates quickly thereafter. For O_3, the largest differences occurred around mid-afternoon at 4 pm around when the highest ambient O_3 concentrations are expected. For SO_2, we observed a slight peak in differences between HBE and CDRE at around 10 am. Additionally, the observed differences were mostly negative during daytime for CO, NO_2 and $PM_{2.5}$, suggesting the home-based method resulted in lower exposure estimates, although the differences changed to slightly positive toward mid-night. However, the exposure differences are mostly positive for O_3, indicating higher exposure estimates when the home-based method is used. When CMAQ concentration fields were applied, the biggest exposure differences were not observed for the group with the highest mobility (group 10), rather it was observed for subjects with moderate to high degree of mobility (group 7 for SO_2, and group 5 and 6 for other pollutants).

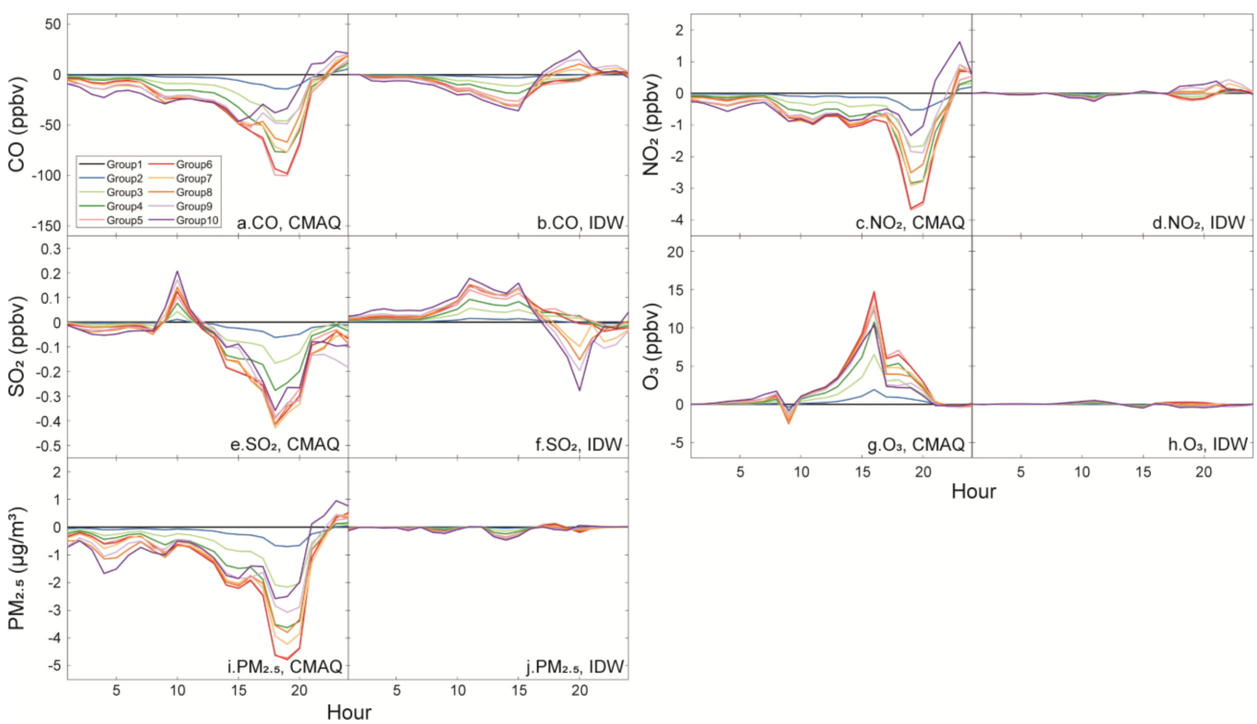

Figure 4. Temporal variations of exposure differences for all 10 mobility groups between HBE and CDRE when CMAQ and IDW concentration field were applied. Exposure differences were calculated as HBE-CDRE.

The temporal variations of exposure differences, however, were mostly not observed when IDW concentration fields were applied (Figure 4). We still observed generally larger differences during daytime (though smaller magnitude), but the consistent patterns of fluctuations as seen among CO, NO_2 and $PM_{2.5}$ in Figure 4 were not observed when IDW fields were applied. The biggest differences were observed at different hours for different pollutants and with no consistent directions. Exposure differences generally showed a consistent increasing trend with increased mobility.

We performed Wilcoxon rank sum tests to evaluate the differences between HBE and CDRE estimates for each mobility group. When CMAQ concentration fields were applied, most differences in HBE and CDRE estimates were statistically significant ($p < 0.05$) during normal business hours (9 am to 5 pm). The only exception is SO_2, for which HBE and CDRE estimates are statistically different between 1 pm to 10 pm. When IDW concentration fields were applied, HBE and CDRE estimates are still generally statistically different between 10 am to 5 pm, although with considerably greater variability.

3.4. The impact of mobility on exposure classifications and effect estimates

To investigate potential exposure misclassifications associated with omitting subject mobility, we investigated how subjects were assigned to different quartiles based on their HBE and CDRE estimates. Results for $PM_{2.5}$ are presented in Figures 5 and 6, and results for other pollutants are presented in Figures S2-S9.

We observed that a high percentage of the sample population was potentially misclassified into other quartiles, especially for groups with higher degrees of mobility. When CMAQ concentration fields were applied for $PM_{2.5}$ (Figure 5), more than half of the sample population in the middle quartiles (Q2 and Q3) were classified into different quartiles for groups 4 through 10 when individual mobility was omitted. The misclassification is especially prominent for the 2nd quartile of group 6 (Figure 5), for which 71% of subjects were misclassified into other quartiles when the home-based method was used. This finding was also observed when IDW fields were used, although the potential misclassifications were less severe, but still substantial (Figure 6). Similar findings can be observed for both CMAQ and IDW concentration fields for all other pollutants (Figures S2-S9). For subjects with moderate exposure levels (Q2 and Q3), generally more subjects were assigned to quartiles with higher exposures when the home-based method was used for CO (Figure S2, S6) and NO_2 (Figures S3, S7). This result was less consistent for SO_2 (Figures S4, S8) and somewhat reversed for O_3 (Figure S5, S9).

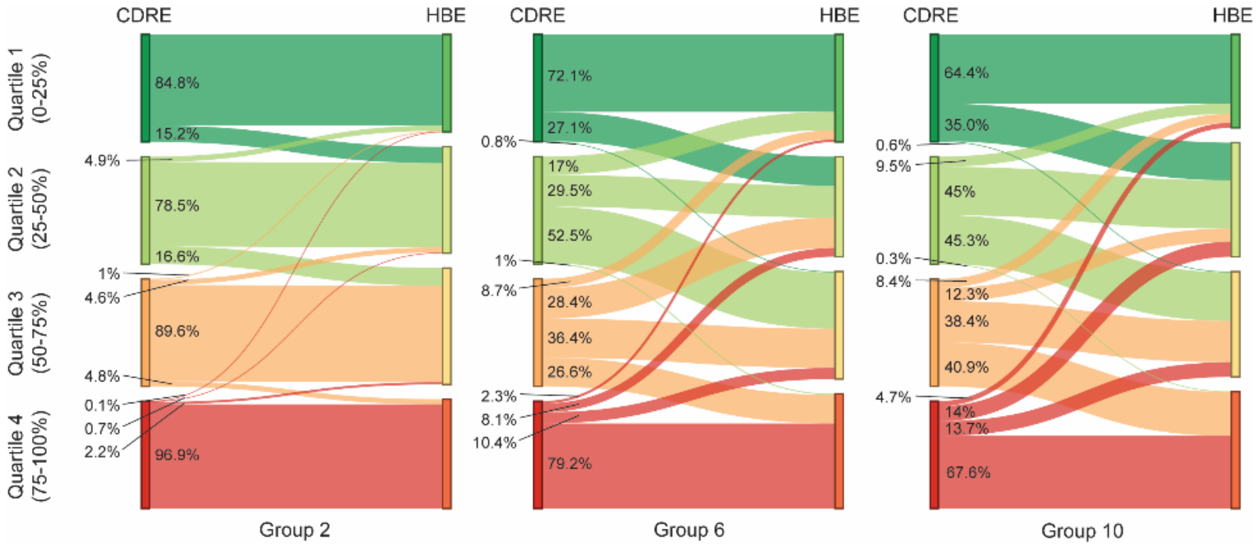

Figure 5. The directions of potential $PM_{2.5}$ exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used. For simplification purposes only results for groups 2, 6 and 10 are presented. Subjects in quartile 1 has the lowest exposures, and subjects in quartile 4 has the highest exposures.

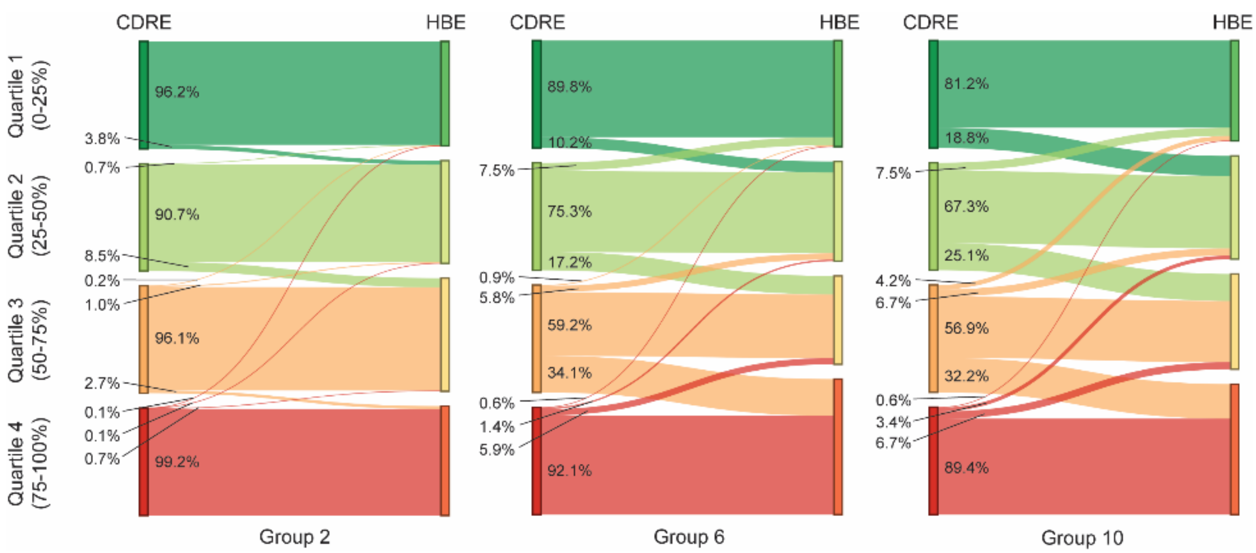

Figure 6. The directions of potential $PM_{2.5}$ exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used. For simplification purposes only results for groups 2, 6 and 10 are presented. Subjects in quartile 1 has the lowest exposures, and subjects in quartile 4 has the highest exposures.

The estimated bias factors for groups with different mobility levels are presented in Figure 7. With increased mobility, the estimated bias factors generally decrease regardless of concentration fields used. The smaller bias factor, a value of 0.67, is observed for NO_2 and for group 10. This value suggests that the estimated relative risk for NO_2 will be underestimated by 33% when mobility was ignored during exposure estimation. Between CMAQ and IDW, the estimated bias factors are relatively similar for NO_2, but are considerably different for other pollutants, especially for $PM_{2.5}$. For group 10, the bias for $PM_{2.5}$ is 0.70 when CMAQ fields are used, and 0.94 when IDW fields are used.

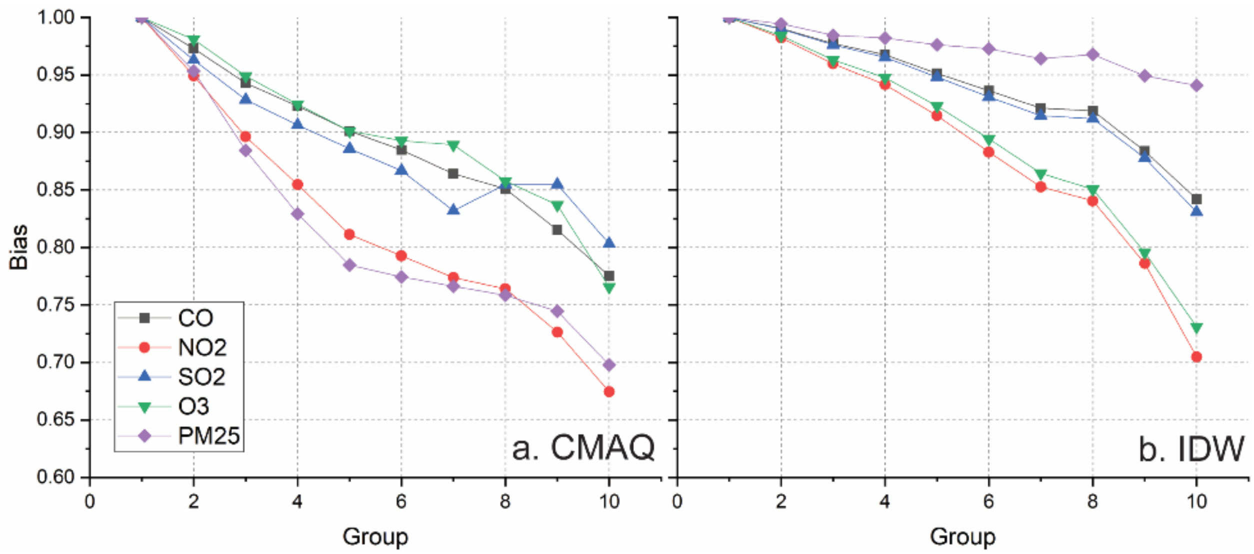

Figure 7. The impact of mobility on bias factors when CMAQ and IDW concentration fields were applied

4. Discussion

4.1. The impact of method choices on exposure estimation

An appropriate characterization of spatial concentration distributions of air pollutants is fundamental for air pollution exposure estimation. In this study, we applied two methods to develop air pollutant concentration fields: one based on outputs from the CMAQ model, and the other based on the IDW interpolation method. Spatial concentration fields developed using the two methods were considerably different from each other (Figure 2). This is expected because, as described previously, the two methods are fundamentally different, and both methods have their own strengths and weaknesses [34]. Consequently, the estimated population average exposures (Table 1), the distributions of individual exposure estimates (Figure 3), particularly among groups with different degrees of mobility (Figure 4), and the impact of neglecting mobility on exposure estimates (Figures 5-6), was different between the two methods. Such results were expected due to the different nature of the two methods. CMAQ is a mechanistic model that calculates ambient concentrations of air pollutants based on input emissions and meteorological data. IDW is an empirical spatial interpolation method that relies solely on available pollutant concentrations measured at discrete locations [34]. Pollution hotspots that are not captured by monitoring networks cannot be captured by the IDW method but may possibly be captured by the CMAQ model if appropriate emissions data are supplied. In this study, the monitoring network is sparse, and only 1 out of 12 monitor is located inside Shenzhen area (Figure 1). As a result, pollutant concentration fields developed using the IDW method were smooth and lacked the spatial concentration variabilities as observed in the CMAQ fields. Therefore, it's important to carefully select an appropriate method for developing pollutant concentration fields, particularly when the monitoring network is sparse.

When detailed mobility data were included, naturally, the appropriate characterization of spatial pollutant variability became even more important. In such applications, purely spatial interpolation methods, e.g., IDW, tessellation, or kriging, are also not ideal choices for developing pollutant concentration fields for study regions without an extensive monitoring network available [34]. These results highlighted the importance of choosing an appropriate method for developing pollutant concentration fields for exposure estimation purposes, particularly when detailed mobility data were included. Without an appropriate characterization of spatial pollutant concentration variations, exposure assessment may not significantly benefit from the inclusion of detailed mobility data at urban scale.

Subsequently, we will focus our discussion on results as obtained using the CMAQ concentration fields.

4.2. The impact of mobility on exposure estimation

In this study, the estimated regression parameters are considerably different from our previous study [23]. For example, the estimated R^2 ranged between 0.95 to 0.98 vs 0.65 to 0.76 in the previous study; and the slope ranged between 0.97 to 1.02 vs 0.60 to 0.72 previously. The seemingly contradictory findings can be explained by the difference in sample population. In our previous study, 9,886 subjects with the most amount of CDR data available were selected to explore the potential benefits of using CDR data in exposure estimation. The subjects were not randomly sampled, and with an average of approximately 463 records per subject per day (vs 115 records per subject per day in this study). The sample population in our previous study are relatively more mobile, and the subjects visited on average 2.3 grid cells over the study period (vs 1.9 grid cells in this study).

At the population level, we did not find substantial differences between HBE and CDRE exposures, consistent with our previous study [23] and other studies [54, 56-59]. The finding maybe partially explained by the fact that most subjects spent most of their time within the same grid, as indicated by the large number of data points clustered near the 1:1 line (Figure 3). Our results suggested that the home-based method for exposure estimation is still informative in the study region when only average exposure estimates for a sufficiently large population are of interest [60]. However, it's worth noting that several studies conducted in other cities [61, 62] have found that the population level exposure estimates are lower when individual mobility data were included in exposure estimation. The differences in findings may be partially due to the potentially different population mobility patterns among cities. Further studies are needed to investigate how our findings may vary among cities.

One of the main focus of this manuscript is on how different levels of mobility impact air pollution exposures. We found that the impact of mobility on exposure estimates differed by time of day and by pollutants (such analyses were not performed in our previous study [23]). Generally, the differences between HBE and CDRE were the smallest during early morning and midnight, a time when many subjects are expected to be at home. For traffic-related pollutants including CO, NO_2, and $PM_{2.5}$, we found that the home-based method likely underestimated subject exposures during daytime, especially near afternoon rush hour, when CMAQ concentration fields were used (Figure 4). Meanwhile, subject exposures to ozone may be over-estimated during daytime using HBE, with the highest error observed at around 4 pm, near the time when the highest ambient ozone concentrations are expected (Figure 4). The temporal differences in impacts of mobility on exposure have also been noted previously [57]. Interestingly, during peak hours, the most significant differences between HBE and CDRE were not observed for the group with the highest degree of mobility, rather the largest differences were observed on subjects with moderate to high degree of mobility (groups 5-7).

Our results showed that the impact of mobility on exposure could be substantial at the individual level, particularly for subjects that are highly mobile. Applying the home-based method yielded similar estimates for those who live close to where they travel throughout the day, although their actual exposure could be drastically different when individual mobility is considered. With an increased degree of mobility, we found that the correlations between HBE and CDRE decreased monotonically (Table 2), suggesting that the home-based method captured less exposure variability among individuals with increased mobility [31]. Therefore, we expect larger exposure classification errors for subjects that are highly mobile, which is supported by our analysis on the potential exposure misclassifications based on HBE and CDRE (Figures 5-6). It is also worth mentioning again that 71% of subjects (Figure 5) in the second quartile of group 6 were misclassified into different quartiles using HBE. These results suggest that the impact of traffic-related pollutants on human health may be larger than previously documented, and these findings may have significant implications for studies that rely on air pollution exposure estimation.

We found that ignoring mobility in exposure assessment could lead to up to 33% in underestimation of relative risk, though the magnitude of underestimation differs among pollutants (Figure 7). Between CMAQ and IDW, the results are also different, especially for $PM_{2.5}$, for which the largest estimated bias factor is only 0.94 when the IDW fields were applied (vs 0.70 for CMAQ field). These finding again demonstrated that the benefit of including detailed mobility data in exposure assessment may be reduced when the spatial variability of pollutant concentrations were not captured, and the method for developing pollution field need to be selected carefully when mobility data were to be included. The finding also have implications for future air pollution health studies.

4.3. Limitations

There are inherent limitations associated with this study. First, as with many CDR databases, the location data used in this study are not the exact location of the corresponding cell phone user, rather, they are the locations of the cell phone tower that handled the wireless communication, which are most likely the nearest tower to the cell phone user. However, we do not expect this limitation to substantially impact the findings for two reasons. 1) The study area is one of the most populated cities in the world with a well-known, densely distributed cell tower network. The CDR dataset contains over 1,000 locations of cell phone towers spread out across the study area. 2) We applied 3-km resolution concentration fields in exposure estimation. The retrieved concentration values are identical within one 3-km grid cell, and one cell phone user in Shenzhen is highly likely to have at least one cell tower within 3 km (see https://www.opencellid.org for more information on cell tower coverage in Shenzhen, China). Therefore, we do not expect the findings to change considerably even when the exact locations of all cell phone users are applied.

Second, CDR data comprise an “event-triggered” database. Location data are only collected when a cell phone communicates with nearby towers. Hence, CDR are temporally sparse in nature [37], and may not accurately capture the full spectrum of individual movements, especially for individuals who only use cell phones occasionally. Hence, exposures estimated using CDR may deviate from those estimated using a more complete location dataset such as those collected using dedicated applications (e.g. Dynamica [63]), or other momentarily collected data such as Google Maps Location History data [64]. However, in this study, our purpose is to compare the differences between exposure estimates with and without detailed mobility data applied. Given the large sample population in all 10 groups with different degrees of mobility, we do not expect the results to change even with an ideally complete mobility database.

Third, despite the relatively large population (N = 310,989) and number of location records (35.6 million), the CDR data used here are a randomly sampled subset from all cell phone users within the entire city of Shenzhen for one typical work day within a typical week. Therefore, the spatiotemporal mobility patterns as represented in this CDR database represent the unique characteristics of the study region. We do expect the patterns of population mobility, the spatiotemporally variability of air pollution concentrations, pollutant emissions, and meteorology conditions will vary across different cities. Further studies are needed to better understand how the findings from this study may change in another city.

Fourth, as described previously, due to the nature of CDR data, the availability of observations, and resources constrains, we applied air pollution concentration fields with 3 km spatial resolution and 1 h temporal resolution for estimating pollution exposures. We recognize that such coarse resolution may introduce uncertainties into related analyses and may also partially impact the findings, such as the impact of mobility on population-level exposure estimates (Figure 3) [61, 62]. Here, we performed an additional analysis to explore the impact of grid resolution on the classification of mobility levels. We split all 3 km CMAQ grid cells into 1.5 km grid cells and counted the number of unique grid cells each subject visited (Table 3). With increased grid resolution, a considerably higher fraction of population were assigned to higher mobility groups, especially for groups with the highest mobility levels (Groups 6 through 10). Such result exemplifies the need for fine-scale modeling, and further studies are needed to investigate how grid resolution impacts the results of exposure estimation with detailed mobility data. In addition, both CDR data and pollution fields are expected to contain uncertainties. What dataset contain greater amount of uncertainty remain unclear. Further studies are also needed to determine the impact of uncertainties on exposure outcomes.

Table 3. Subject population in each mobility group at 3 km and 1.5 km grid resolutions

	3 km grids	1.5 km grids	Change (%)
Group 1	167570	132847	-20.7%
Group 2	75313	72821	-3.3%
Group 3	32177	39341	22.3%
Group 4	16350	22689	38.8%
Group 5	8354	13918	66.6%
Group 6	4617	8845	91.6%
Group 7	2700	5886	118%
Group 8	1562	4105	163%
Group 9	916	2755	201%
Group 10	1430	7782	444%

Finally, it’s also worth noting that, the exposure estimates presented in this study are calculated using ambient pollutant concentrations. A subject’s exposure to indoor pollution was not considered here. The required data for estimating indoor pollution exposure (e.g.: type of micro-environment, pollution infiltration to indoor) are not available. In addition, due to the nature of CDR data, it’s difficult to precisely determine the location of micro-environment for each subject. For example, if one subject’s CDR data is located in close proximity to a major roadway, the investigator may not be able to determine whether the subject is driving on the roadway, or walking along the roadway, or even sitting inside a building next to the roadway.

5. Conclusion

In this study, we applied a large cell phone location database consisting of over 35 million location records from 310,989 subjects to investigate the impact of individual mobility on estimated ambient exposures for five chosen pollutants (CO, NO_2, SO_2, O_3, and $PM_{2.5}$). We further divided our sample population into ten groups with different degrees of mobility and compared exposures estimates for each group. We also applied and compared two methods to develop concentration fields for exposure estimation, including one based on output from the CMAQ model that was fused with observational data, and the other based on the spatial interpolation of observations using the inverse distance weighting method.

We found no substantial differences between population-averaged exposures as estimated with and without detailed mobility data (e.g.: exposure estimates differ by up to 5.4% for NO_2, Table 2). Thus, the traditional home-based exposure estimation method is still informative when only averaged exposures on a large population are needed. We observed generally increased variabilities in exposure estimates at the individual level with increased mobility. Exposure classification errors are also likely to increase with higher degrees of mobility, and could be substantial for groups of individuals that are highly mobile. We also examined the temporal variability of the differences between exposures as estimated with and without mobility data. We found the home-based method will likely under-estimate exposure to traffic-related pollutants (CO, NO_2 and $PM_{2.5}$) during day-time particularly during afternoon rush-hour, but also will likely over-estimate exposures to ground level ozone during mid-afternoon near the time when ambient ozone concentrations are expected to be the highest. These results suggest that mobility could be important for air pollution health studies for which obtaining accurate exposure estimates at individual level are critical, such as case-control studies or studies with a small sample size.

We found that the concentration fields developed using the IDW method failed to capture pollution hotspot as can be seen from the CMAQ fields, due primarily to the sparse monitoring network, and consequently limited measurement data available in the study domain. Therefore, the IDW method may not suitable for air pollution exposure estimations when detailed mobility data are considered, if a dense measurement network is not available. When detailed mobility data were to be applied in exposure estimation, the method for developing air pollution concentration fields should be selected carefully.

We also acknowledge that the CDR data applied in this study represent the unique characteristics of the study region, and further studies are needed to investigate how our findings could change among regions with different spatiotemporal patterns of population and pollution concentrations. Despite the limitation, overall, our results have significant implications for future air pollution health studies in which subject mobility is important.

Acknowledgements

We would like to acknowledge Dr. Desheng Zhang (Rutgers University) for providing the CDR data. This research was partially funded by the University of Central Florida startup grant.

References

1. Neira, M., A. Prüss-Ustün, and P. Mudu, *Reduce air pollution to beat NCDs: from recognition to action.* The Lancet, 2018. **392**(10154): p. 1178-1179.
2. Burnett, R., et al., *Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter.* Proceedings of the National Academy of Sciences, 2018. **115**(38): p. 9592-9597.
3. Cohen, A.J., et al., *Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015.* The Lancet, 2017. **389**(10082): p. 1907-1918.
4. Gakidou, E., et al., *Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016.* The Lancet. **390**(10100): p. 1345-1422.
5. Kampa, M. and E. Castanas, *Human health effects of air pollution.* Environmental pollution, 2008. **151**(2): p. 362-367.
6. Pope III, C.A. and D.W. Dockery, *Health effects of fine particulate air pollution: lines that connect.* Journal of the air & waste management association, 2006. **56**(6): p. 709-742.
7. Bernstein, J.A., et al., *Health effects of air pollution.* Journal of allergy and clinical immunology, 2004. **114**(5): p. 1116-1123.
8. Kim, J., *Ambient air pollution: health hazards to children.* Pediatrics, 2004. **114**(6): p. 1699-1707.
9. de Zwart, F., et al., *Air pollution and performance-based physical functioning in Dutch older adults.* Environmental Health Perspectives (Online), 2018. **126**(1 %@ 1552-9924).
10. Münzel, T., et al., *Environmental stressors and cardio-metabolic disease: part I–epidemiologic evidence supporting a role for noise and air pollution and effects of mitigation strategies.* European heart journal, 2017. **38**(8): p. 550-556.
11. Zhang, Z., et al., *Long-Term Exposure to Fine Particulate Matter, Blood Pressure, and Incident Hypertension in Taiwanese Adults.* Environmental health perspectives, 2018. **126**(1): p. 017008-017008.
12. Fann, N., et al., *Estimated Changes in Life Expectancy and Adult Mortality Resulting from Declining PM 2.5 Exposures in the Contiguous United States: 1980–2010.* Environmental Health Perspectives, 2017. **97003**: p. 1.
13. Malley, C.S., et al., *Preterm birth associated with maternal fine particulate matter exposure: a global, regional and national assessment.* Environment international, 2017. **101**: p. 173-182.
14. Chen, R., et al., *Fine Particulate Air Pollution and the Expression of microRNAs and Circulating Cytokines Relevant to Inflammation, Coagulation, and Vasoconstriction.* Environmental health perspectives, 2018. **126**(1): p. 017007-017007.
15. Reis, S., et al., *The influence of residential and workday population mobility on exposure to air pollution in the UK.* Environment international, 2018. **121**: p. 803-813.
16. Zhang, S., et al., *Long-term effects of air pollution on ankle-brachial index.* Environment international, 2018. **118**: p. 17-25.
17. Gray, S.C., S.E. Edwards, and M.L. Miranda, *Race, socioeconomic status, and air pollution exposure in North Carolina.* Environmental research, 2013. **126**: p. 152-158.
18. Cao, J., et al., *Association between long-term exposure to outdoor air pollution and mortality in China: a cohort study.* Journal of hazardous materials, 2011. **186**(2-3): p. 1594-1600.

19. Gurram, S., A.L. Stuart, and A.R. Pinjari, *Impacts of travel activity and urbanicity on exposures to ambient oxides of nitrogen and on exposure disparities.* Air Quality, Atmosphere & Health, 2015. **8**(1): p. 97-114.
20. Shafran-Nathan, R., I. Levy, and D.M. Broday, *Exposure estimation errors to nitrogen oxides on a population scale due to daytime activity away from home.* Science of the Total Environment, 2017. **580**: p. 1401-1409.
21. Park, Y.M. and M.-P. Kwan, *Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored.* Health & place, 2017. **43**: p. 85-94.
22. Yoo, E., et al., *Geospatial estimation of individual exposure to air pollutants: Moving from static monitoring to activity-based dynamic exposure assessment.* Annals of the Association of American Geographers, 2015. **105**(5): p. 915-926.
23. Yu, H., et al., *Using cell phone location to assess misclassification errors in air pollution exposure estimation.* Environmental Pollution, 2018. **233**: p. 261-266.
24. Gurram, S., A.L. Stuart, and A.R. Pinjari, *Agent-based modeling to estimate exposures to urban air pollution from transportation: Exposure disparities and impacts of high-resolution data.* Computers, Environment and Urban Systems, 2019. **75**: p. 22-34.
25. Setton, E., et al., *The impact of daily mobility on exposure to traffic-related air pollution and health effect estimates.* Journal of Exposure Science and Environmental Epidemiology, 2011. **21**(1): p. 42.
26. Pennington, A.F., et al., *Measurement error in mobile source air pollution exposure estimates due to residential mobility during pregnancy.* Journal of Exposure Science and Environmental Epidemiology, 2017. **27**(5): p. 513.
27. Klepeis, N.E., et al., *The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants.* Journal of Exposure Science and Environmental Epidemiology, 2001. **11**(3): p. 231.
28. Dons, E., et al., *Impact of time–activity patterns on personal exposure to black carbon.* Atmospheric Environment, 2011. **45**(21): p. 3594-3602.
29. Buonanno, G., L. Stabile, and L. Morawska, *Personal exposure to ultrafine particles: the influence of time-activity patterns.* Science of the total environment, 2014. **468**: p. 903-907.
30. Bell, M.L., G. Banerjee, and G. Pereira, *Residential mobility of pregnant women and implications for assessment of spatially-varying environmental exposures.* Journal of exposure science & environmental epidemiology, 2018: p. 1.
31. Chen, L., et al., *Residential mobility during pregnancy and the potential for ambient air pollution exposure misclassification.* Environmental research, 2010. **110**(2): p. 162-168.
32. Nieuwenhuijsen, M.J., et al., *Variability in and agreement between modeled and personal continuously measured black carbon levels using novel smartphone and sensor technologies.* Environmental science & technology, 2015. **49**(5): p. 2977-2982.
33. Tang, R., et al., *Integrating travel behavior with land use regression to estimate dynamic air pollution exposure in Hong Kong.* Environment international, 2018. **113**: p. 100-108.
34. Yu, H., et al., *Cross-comparison and evaluation of air pollution field estimation methods.* Atmospheric Environment, 2018. **179**: p. 49-60.
35. Ivey, C.E., et al., *Development of PM_{2.5} source impact spatial fields using a hybrid source apportionment air quality model.* Geosci. Model Dev., 2015. **8**(7): p. 2153-2165.
36. Bates, J.T., et al., *Source impact modeling of spatiotemporal trends in PM2. 5 oxidative potential across the eastern United States.* Atmospheric environment, 2018. **193**: p. 158-167.
37. Zhao, Z., et al., *Understanding the bias of call detail records in human mobility research.* International Journal of Geographical Information Science, 2016. **30**(9): p. 1738-1762.

38. Zhang, D., et al. *coMobile: Real-time human mobility modeling at urban scale using multi-view learning*. in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2015. ACM.
39. Zhang, D., et al. *Exploring human mobility with multi-source data at extremely large metropolitan scales*. in *Proceedings of the 20th annual international conference on Mobile computing and networking*. 2014. ACM.
40. McMillan, J.E.R., W.B. Glisson, and M. Bromby. *Investigating the increase in mobile phone evidence in criminal activities*. in *System sciences (hicss), 2013 46th hawaii international conference on*. 2013. IEEE.
41. Kumar, M., M. Hanumanthappa, and T.S. Kumar. *Crime investigation and criminal network analysis using archive call detail records*. in *Advanced Computing (ICoAC), 2016 Eighth International Conference on*. 2017. IEEE.
42. Becker, R., et al., *Human mobility characterization from cellular network data.* Communications of the ACM, 2013. **56**(1): p. 74-82.
43. Gonzalez, M.C., C.A. Hidalgo, and A.-L. Barabasi, *Understanding individual human mobility patterns.* nature, 2008. **453**(7196): p. 779.
44. Becker, R.A., et al., *A tale of one city: Using cellular network data for urban planning.* IEEE Pervasive Computing, 2011. **10**(4): p. 18-26.
45. Wang, H., et al. *Transportation mode inference from anonymized and aggregated mobile phone call detail records*. in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. 2010. IEEE.
46. Iqbal, M.S., et al., *Development of origin–destination matrices using mobile phone call data.* Transportation Research Part C: Emerging Technologies, 2014. **40**: p. 63-74.
47. Zhang, D. *Desheng Zhang, Rutgers University*. 2020 [cited 2020 March 8]; Available from: https://www.cs.rutgers.edu/~dz220/data.html.
48. Byun, D. and K.L. Schere, *Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system.* Applied Mechanics Reviews, 2006. **59**(2): p. 51-77.
49. Che, W., et al., *Assessment of motor vehicle emission control policies using Model-3/CMAQ model for the Pearl River Delta region, China.* Atmospheric environment, 2011. **45**(9): p. 1740-1751.
50. Clark, N.A., et al., *Effect of early life exposure to air pollution on development of childhood asthma.* Environmental health perspectives, 2009. **118**(2): p. 284-290.
51. Dugandzic, R., et al., *The association between low level exposures to ambient air pollution and term low birth weight: a retrospective cohort study.* Environmental health, 2006. **5**(1): p. 3.
52. Mitchell, R. and F. Popham, *Effect of exposure to natural environment on health inequalities: an observational population study.* The Lancet, 2008. **372**(9650): p. 1655-1660.
53. Gauderman, W.J., et al., *Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study.* The Lancet, 2007. **369**(9561): p. 571-577.
54. Nyhan, M., et al., *Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data.* Journal of exposure science & environmental epidemiology, 2018.
55. Wacholder, S., *When measurement errors correlate with truth: surprising effects of nondifferential misclassification.* Epidemiology (Cambridge, Mass.), 1995. **6**(2): p. 157-161.
56. Dewulf, B., et al., *Dynamic assessment of exposure to air pollution using mobile phone data.* International journal of health geographics, 2016. **15**(1): p. 14.
57. Picornell, M., et al., *Population dynamics based on mobile phone data to improve air pollution exposure assessments.* Journal of exposure science & environmental epidemiology, 2018: p. 1.

58. Nyhan, M., et al., *"Exposure Track" · The Impact of Mobile-Device-Based Mobility Patterns on Quantifying Population Exposure to Air Pollution.* Environmental science & technology, 2016. **50**(17): p. 9671-9681.
59. Gariazzo, C., A. Pelliccioni, and A. Bolignano, *A dynamic urban air pollution population exposure assessment study using model and population density data derived by mobile phone traffic.* Atmospheric Environment, 2016. **131**: p. 289-300.
60. Nikkilä, A., et al., *Effects of incomplete residential histories on studies of environmental exposure with application to childhood leukaemia and background radiation.* Environmental Research, 2018. **166**: p. 466-472.
61. Singh, V., R.S. Sokhi, and J. Kukkonen, *An approach to predict population exposure to ambient air PM2.5 concentrations and its dependence on population activity for the megacity London.* Environmental Pollution, 2019: p. 113623.
62. Smith, J.D., et al., *London Hybrid Exposure Model: Improving Human Exposure Estimates to NO2 and PM2.5 in an Urban Setting.* Environmental Science & Technology, 2016. **50**(21): p. 11760-11768.
63. Fan, Y., et al., *SmarTrAC: A Smartphone Solution for Context-Aware Travel and Activity Capturing*. 2015, University of Minnesota: Minneapolis, MN.
64. Yu, X., et al., *On the accuracy and potential of Google Maps location history data to characterize individual mobility for air pollution health studies.* Environmental Pollution, 2019.