

Electronic Theses and Dissertations, 2020-

2020

Improving Air Pollution Exposure Estimation Using Cell Phone Location Data and Low-cost Sensors

Xiaonan Yu
University of Central Florida

 Part of the [Environmental Engineering Commons](#)
Find similar works at: <https://stars.library.ucf.edu/etd2020>
University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2020- by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Yu, Xiaonan, "Improving Air Pollution Exposure Estimation Using Cell Phone Location Data and Low-cost Sensors" (2020). *Electronic Theses and Dissertations, 2020-*. 462.
<https://stars.library.ucf.edu/etd2020/462>

IMPROVING AIR POLLUTION EXPOSURE ESTIMATION USING CELL
PHONE LOCATION DATA AND LOW-COST SENSORS

by

XIAONAN YU
M.S. Fudan University, 2013

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Civil, Environmental, and Construction Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2020

Major Professor: Haofei Yu

©2020 Xiaonan Yu

ABSTRACT

Human exposure estimation to air pollution plays an important role in epidemiological studies which are designed to reveal correlations between human exposures to certain air pollutants and certain diseases, such as asthma, cardiovascular disease and reproductive diseases. Traditionally, when people's mobile data is hard to get, home location is used to estimate people's exposures assuming that people stay at home all the time. Whereas, people move and it is more accurate to estimate people's exposures including people's mobility. In our study, we showcased two methods to obtain people's mobile data: Google Maps location history (GMLH) data and Call Detailed Record (CDR) data. GMLH data was compared with Global Positioning System (GPS) data from four aspects: 1) spatial movement of the subject; 2) time the subject spent at different microenvironments; 3) time the subject spent on driving; 4) subject's time-weighted exposures to ambient particulate matter. The results showed that compared with GPS data, GMLH data capture well the subject's spatial mobility with resolution of 200m * 200m or larger and successfully captured the time the subject spent at different microenvironments and the time on driving. Also, with GMLH data we were able to accurately estimate the subject's time-weighted exposure to ambient PM pollution. CDR data was used to estimate subjects' mobile exposures for five chosen pollutants (CO, NO₂, SO₂, O₃, and PM_{2.5}). And the correlation between difference between static exposures and mobile exposures with mobility level is also investigated. My study revealed that there is no substantial difference between home based exposure (HBE) and CDR based exposure (CDRE) at population level. But at individual level, difference between HBE and CDRE increased with mobility increased.

It was also found that HBE would likely under-estimate exposure to traffic-related pollutants (CO, NO₂ and PM_{2.5}) during afternoon rush-hour, but over-estimate exposures to ozone during mid-afternoon. As smartphone and Google Maps application are used widely, these two methods have huge potential on obtaining people's mobility data. My study also tested the relative accuracy and reliability of two brand commercial sensors (PurpleAir and Dylos). Results showed that PurpleAir has good relative accuracy and reliability, while Dylos has moderate relative accuracy and reliability.

Firstly, I would like to thank my advisor, who helps me a lot on my academic career. I also want to thank my committee members. They spend their time on my dissertation and presentation and I cherish their feedback. Also, I want to thank my parents and my boyfriend, who give me huge support on my study career. My boyfriend's parents care about me, too. I also want to thank my therapists, Kim and Britta, who support me emotionally and spiritually. I have to say I love UCF. It is a big school with 50000 students and a lot of resources. I love gym, psychological center and health center, three places that I frequently went to. My school spares no effort to support their students.

ACKNOWLEDGMENTS

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

TABLE OF CONTENTS

LIST OF FIGURES	x
LIST OF TABLES	xiv
LIST OF ABBREVIATIONS	xvi
CHAPTER ONE: INTRODUCTION.....	1
CHAPTER TWO: LITERATURE REVIEW	6
2.1 Travel Survey and Diary Data.....	7
2.2 GPS Data.....	9
2.3 CDR Data.....	12
2.4 App Data	14
2.5 Census Data	16
2.6 Others.....	17
2.7. Discussion.....	19
2.8 Sensor Data	20
CHAPTER THREE: GOOGLE MAPS LOCATION DATA WORK.....	23
3.1 Introduction.....	23
3.2 Materials and Methods.....	25
3.3 Results.....	29
3.4 Discussion.....	33
CHAPTER FOUR: WORK ON USING CDR DATA FOR EXPOSURE ESTIMATION	

.....	38
4.1 Introduction.....	38
4.2 Materials and Methods.....	39
4.2.1 Data Description and Study Area.....	39
4.2.2 Exposure Estimation	41
4.3 Results.....	44
4.3.1 Concentration Fields	44
4.3.2 Overall Correlations Between HBE and CDRE	45
4.3.3 The Impact of Mobility on Exposure Estimates	46
4.3.4 The Impact of Mobility on Exposure Classifications and Effect Estimates	51
4.4 Discussion.....	54
4.4.1 The Impact of Method Choices on Exposure Estimation	54
4.4.2 The Impact of Mobility on Exposure Estimation	55
4.4.3 Limitations	56
4.5 Conclusion	58
CHAPTER FIVE: WORK ON LOW-COST SENSORS	60
5.1 Introduction.....	60
5.2 Material and Methods	61
5.3 Results and Discussion	63
5.3.1 PurpleAir Sensor.....	63

5.3.2 Dylos Sensor	66
5.4 Limitations and Future Work	70
CHAPTER SIX: SUMMARY AND CONCLUSION	71
APPENDIX A: ADDITIONAL CONTENTS FOR CHAPTER THREE.....	73
A.1 Microenvironment.....	74
A.2 Location Data.....	74
A.3 Evaluation of Google Maps Data.....	76
APPENDIX B: ADDITIONAL CONTENTS FOR CHAPTER FOUR.....	78
B.1 Correlations Between HBE and CDRE.....	79
B.2 The Impact of Mobility on Exposure Classifications.....	81
B.3 Temporal Variations of Differences Between HBE and CDRE	85
B.4 Potential Exposure Misclassifications When Mobility Were Neglected.....	88
LIST OF REFERENCE	91

LIST OF FIGURES

Figure 3-1. Comparison of the estimated daily total (a-d) and weekly total (e-h) time the subject spent in each 1 km (a,e), 500 m (b,f), 200 m (c,g) and 100 m (d,h) resolution grid cell based on GPS versus Google Maps location data.....	30
Figure 4-1. The study area of Shenzhen, China.....	40
Figure 4-2. Spatial fields of concentrations of the five chosen pollutants as simulated by the CMAQ (a-e) and IDW (f-j) methods.	44
Figure 4-3. 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.....	46
Figure 4-4. Distributions of differences in exposure estimates between HBE and CDRE for the five chosen pollutants for both CMAQ and IDW methods. Relative exposure differences were calculated as $(HBE-CDRE)/CDRE$	48
Figure 4-5. 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 $HBE-CDRE$	50
Figure 4-6. Temporal variations of p-values from the Wilcoxon rank sum tests performed for 9 mobility groups between HBE and CDRE when CMAQ and IDW concentration field were applied. Results for group 1 are not shown. Dotted black line is $p = 0.05$	51
Figure 4-7. 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.....	52
Figure 4-8. 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.....	53
Figure 4-9. The impact of mobility on bias factors when CMAQ and IDW concentration fields were applied.....	54
Figure 5-1. Pictures of sensor deployment (left: internal layout of the shelter; right: external appearance of the shelter).	61
Figure 5-2. Location of Sydney site which is marked as L057-3002 at upright corner [178].....	63
Figure 5-3. Correlations between two channels in each PurpleAir sensor (a: PurpleAir 1; b: PurpleAir 2; c: PurpleAir 3).	64
Figure 5-4. Correlations between each PurpleAir sensor.....	65
Figure 5-5. Correlations between each Dylos sensor at time span of 1 minute.....	66
Figure 5-6. Correlations between D1 and D2 by aggregating time in different time spans (a:5minute; b:10minute; c:15minute; d:30minute; e:60minute; f:6hour; g:12hour; h:24hour).	67
Figure 5-7. Correlations between D2 and D3 by aggregating time in different time span.	68
Figure A-1. Satellite images of three microenvironments and corresponding rectangles (a is a post office, b and c are grocery stores).....	74
Figure A-2. The collected location data from Google Maps and the GPS logger near a	

roadway intersection inside the study domain.75

Figure A-3. Comparison of the estimated daily total (a-d) and weekly total (e-h) time the subject spent in each grid cell based on GPS versus Google Maps location data, for grid cells in which the subject spent 10 minutes or less during the total time represented (as determined by GPS logger data). Resolutions represented are 1 km (a,e), 500 m (b,f), 200 m (c,g) and 100 m (d,h).....76

Figure B-1. The directions of potential CO exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used.82

Figure B-2. The directions of potential NO₂ exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used.82

Figure B-3. The directions of potential SO₂ exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used.83

Figure B-4. The directions of potential O₃ exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used.83

Figure B-5. The directions of potential CO exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used.84

Figure B-6. The directions of potential NO₂ exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used.84

Figure B-7. The directions of potential SO₂ exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used.85

Figure B-8. The directions of potential O₃ exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used.85

Figure B-9. Temporal variations of average relative differences between HBE and CDRE when CMAQ concentration field were used. The relative differences were estimated as $(HBE-CDRE)/CDRE$86

Figure B-10. Temporal variations of average relative differences between HBE and CDRE when IDW concentration field were used. The relative differences were estimated as $(HBE-CDRE)/CDRE$87

Figure B-11. Temporal variations of average absolute relative differences between HBE and CDRE when IDW concentration field were used. The relative differences were estimated as $(HBE-CDRE)/CDRE$87

Figure B-12. Temporal variations of average absolute relative differences between HBE and CDRE when IDW concentration field were used. The relative differences were estimated as $(HBE-CDRE)/CDRE$88

LIST OF TABLES

Table 3-1. The estimated time (in minutes) the subject spent at each of the 10 microenvironments, as estimated using GPS and Google Maps location data.....	31
Table 3-2. The estimated time (in minutes) the subject spent driving, and time-weighted daily exposure to ambient PM (normalized to subject’s home location), as estimated using GPS and GMLH data.	32
Table 4-1. Comparison between HBE and CDRE estimate of NO ₂ for all ten groups with different mobility.	47
Table 5-1. Statistical data of correlations between D1, D2 and D3 by aggregating time in different time spans.....	69
Table B-1. Comparison between HBE and CDRE estimates of CO for all ten groups with different mobility.	79
Table B-2. Comparison between HBE and CDRE estimates of SO ₂ for all ten groups with different mobility.	79
Table B-3. Comparison between HBE and CDRE estimates of O ₃ for all ten groups with different mobility.....	80
Table B-4. Comparison between HBE and CDRE estimates of PM _{2.5} for all ten groups with different mobility.	81
Table B-5. Percentage of sample populations in each quartile that were classified into different quartiles when subject mobility was neglected in exposure estimation. Results shown are for CO.	88
Table B-6. Percentage of sample populations in each quartile that were classified into	

different quartiles when subject mobility was neglected in exposure estimation. Results shown are for NO₂89

Table B-7. Percentage of sample populations in each quartile that were classified into different quartiles when subject mobility was neglected in exposure estimation. Results shown are for SO₂89

Table B-8. Percentage of sample populations in each quartile that were classified into different quartiles when subject mobility was neglected in exposure estimation. Results shown are for O₃89

LIST OF ABBREVIATIONS

AOD	Aerosol Optical Depth
BC	Black Carbon
EPA	Environmental Protection Agency
GPS	Global Positioning System
MPLH	Mobile Phone Locating History
NHAPS	National Human Activity Pattern Survey
NO ₂	Nitrogen Dioxide
NO _x	Nitrogen Oxides
PB-PAH	Particle-Bound Polycyclic Aromatic Hydrocarbon
PM	Particulate Matter
PM _{2.5}	Particulate Matter with a Diameter of 2.5 Micrometers or Less
PM ₁₀	Particulate Matter with a Diameter of 10 Micrometers or Less
UFP	Ultrafine Particle

CHAPTER ONE: INTRODUCTION

Ambient air pollution has adverse human health impacts and nowadays, it is getting considerable attention all over the world. Many physical diseases are proved to be associated with air pollution exposure, such as respiratory diseases, reducing physical functions and cardiovascular illness [1-7]. Pope III et al. reviewed studies focusing on the effects of particulate matter (PM) on human health between 1997 and 2006, and concluded that a $20\mu\text{g}/\text{m}^3$ increase in PM_{10} exposure is associated with an 0.4% to 1.4% increase in the risk of all-cause mortality [6]. Gehring et al. investigated the associations between exposures to NO_2 , NO_x and $\text{PM}_{2.5}$ and lung function, and concluded that exposure to these air pollutants may result in lung function reduction in schoolchildren [5]. Kurt et al. found that exposures to PM, ozone and nitrogen oxides may exacerbate asthma, and increase the risk of lung cancer and respiratory infections [7]. Cohen et al. found that exposure to $\text{PM}_{2.5}$ contributed to 4.2 million pre-mature deaths, while exposure to ozone caused an additional 254,000 deaths in 2015 globally [3].

Not only human's physical health is impacted by air pollution, but also human's mental health. For instance, according to Zhang et al., exposure to air pollution may hinder people's cognitive performance as they grow older [8]. Newbury et al. found that exposure to nitrogen dioxide (NO_2) and nitrogen oxides (NO_x) together account for 60% of the association between urban residency and adolescent psychotic experiences instead of family socioeconomic status or family psychiatric history [9].

Epidemiological studies in air pollution focus on investigating the associations between air pollutant exposures and adverse human health effects. Through these studies, we could uncover the linkages between many health endpoints and exposure to air pollutants. For

instance, increase PM and ozone were associated with the risk of increased mortality [10] [11]. Increased hospital admission due to cardiovascular and respiratory illnesses were linked with daily PM_{2.5} level variations [12]. NO₂ was also found to be associated with cardiovascular mortality, acute myocardial infarction, and hospital admission for chronic obstructive pulmonary disease [13]. Furthermore, 3.4% of cardiovascular and 2% of respiratory deaths were found to be attributable to SO₂ levels higher than 10 µg/m³ in a city in Iran [14].

In epidemiological studies, exposure estimation is a critical step. Subsequent statistical analysis, aiming at understanding the associations between human exposure and adverse health impacts, is based on an accurate exposure estimation. Whereas, imprecise estimations of exposure would introduce undesired biases or errors in epidemiological analysis [15-17]. Yu et al. found that misclassification errors are likely to be substantial when exposure is not estimated correctly (i.e. neglecting people's mobility) [15]. Chen et al. used mobile-phone locating-request (MPL) big data to estimate population exposures to PM_{2.5} and compared them with exposures calculated by census data. They concluded that dynamics of population is an important factor in the estimation of inhaled PM_{2.5} mass estimation [18]. Hodgson et al. compared pregnant mothers' exposures at delivery locations, exposures at conception locations and exposures at residences. They concluded that researchers should consider errors introduced by residential mobility if they want to explore the subtle associations between exposures and health outcomes [19].

Overall, accurate assessment of human exposure to air pollution is essential for air pollution health studies. Previously, many scholars adopted people' home addresses (or their home zip code) to estimate their exposures to ambient air pollutants [20] [11]. In this method, subjects were assumed stay at residence place for 24 hours, which is apparently unrealistic. Recently, researchers found that accounting for people's mobility information could lead to

significant differences in exposure estimation [17-19, 21-23]. In Dhondt et al.'s study, the difference of NO₂ exposures between including and not including mobility information could be up to 15% for people who live in rural zones [24]. For population exposure, Ma et al. pointed out that one of the administrative and commercial districts of Beijing, China has a high population ratio (day to night) of 1.35, while a residential district of Beijing has a lower ratio of 0.84 [25]. Such population fluctuation in commuter cities would influence population exposure estimation. Tang et al. found that exposures calculated when mobility and infiltration factors were taken into consideration are approximately 20% lower than ambient exposures estimated at residential addresses [22]. Therefore, including subject's mobility are expected to lead to more accurate estimation of exposures.

However, including mobility in exposure estimation requires subjects' spatiotemporal mobility data. In prospective studies, people's mobility data is relatively easier to obtain, and several methods have been employed in the past studies to capture individual human movement, such as detailed activity diaries [26], global positioning system (GPS) devices [27-29] or smartphone based location tracking applications [30]. However, activity diaries are laborious, and using dedicated tracking devices can be expensive for large sample populations. For retrospective studies, historical mobility data are needed, but not all the methods used in prospective studies can be used in retrospective studies. Therefore, in much past retrospective studies, researchers used subjects' residential addresses to estimate their exposures, which as discussed previously, is problematic since the assumption is that people stay at home for the entire study period. Several recent studies have attempted to address this issue by using some other methods, including aggregated travel survey data [22], accounting for multiple locations of subjects [31-33], or utilizing travel demand models [34-36]. However, these methods either lack details at individual level, or are only approximations of a subject's trajectory. Methods

that can retrospectively collect detailed and actual mobility data at individual level remain scarce.

Smartphones are becoming ubiquitous globally, and almost all smartphone has the capability to collect and archive users' location data. Such location data are collected by carriers in the form of call detail records [15], by the smartphone's triangulation location technology (cellular positioning), GPS, Wi-Fi, and other means [37]. These smartphone location data have already been used extensively in many areas such as criminal investigations, commercial advertisements and transportation planning [38-41]. In the field of air pollution, the potentials of such data are just started being recognized [18, 23, 42-45]. However, most of these smartphone location data are collected at irregular intervals, leading to spatiotemporal sparseness [46, 47]. Therefore, whether these data properly characterize an individual's spatiotemporal mobility, and how can they be used in epidemiological studies remain under-investigated.

Here, we demonstrated the potentials of two methods to account for human mobility into exposure estimation for retrospective studies. Both two methods used data originated from cell phones. In one approach, mobility data collected from carriers, the call detail record (CDR), was used. In the other approach, mobility data collected by the Google Maps application from a subject's smartphone were used.

For the CDR approach, we investigated how mobility impacted people's exposure to air pollution at both population level and individual level by using a public available CDR dataset. We also divided our subjects into 10 groups based on their mobility level to study how increased mobility level impact their exposure estimation to air pollutants. Further, we compared two different methods for developing concentration fields and investigated how the choice of concentration field impacted exposure estimates when mobility is considered.

In the second approach, we showed the potential of using Google Maps location history data to characterize an individual's exposure to air pollution. In this study, we compared one subject's Google Maps location history data with GPS logger data to evaluate the accuracy of Google Maps location data regarding: 1) spatial movement of the subject; 2) the time the subject spent at different microenvironments; 3) the time the subject spent driving during the one-week time period; 4) the subject's time-weighted exposures to PM (using satellite-derived aerosol optical depth (AOD) measurements data).

Finally, we also evaluated the performance of six low-cost air quality sensors manufacture by two brands, to investigate their accuracy and precision. Results from this evaluation contribute to better understanding the usefulness of low-cost sensors in improving exposure estimation.

CHAPTER TWO: LITERATURE REVIEW

Ambient air pollution is drawing substantial attention globally due to its adverse human health impacts. A large body of literature has suggested that air pollution exposure is associated with a wide range of adverse impact on human's physical health, such as reducing lung functions, causing respiratory cardiovascular diseases and increasing the risk of mortality [1-7]. In addition to the adverse impact on physical health, air pollution also negatively impacts human's mental health, such as hindering people's cognitive performance as they get older and causing psychiatric experiences [8, 9].

Epidemiological studies on air pollution focus on investigating the associations between air pollutants and adverse human health effects. [10-14]. In epidemiological studies, the term 'exposure' refers to human's contact with a certain air pollutant. In epidemiological investigations, exposure estimation is a critical step, and an accurate exposure estimation is crucial for subsequent statistical analysis aimed to understand the associations between human exposure and adverse health impacts. Conversely, inaccurate exposure estimation would introduce uncertainty in epidemiological analysis, leading to undesired bias or errors [15-17]. Therefore, how to accurately assess human's exposure is essential for pollution health studies. Previously, many researchers uses subjects' home addresses or zip code as subjects' only location information to estimate their exposures [20] [11]. This method assumes that people only stay at home, which is obviously unrealistic. Much recent researches also demonstrate that the inclusion of people's mobility could lead to significant differences in exposure estimation [15, 17-19, 21-23, 25].

Including mobility in exposure estimation requires subjects' spatiotemporal mobility data. How to obtain mobility data is a challenging issue in epidemiological studies. In this

chapter we discussed the six most common six methods and a few other less commonly used methods for obtaining mobility data. The strengths and weaknesses are also discussed.

The methods discussed here including travel survey (diary) data, global positioning system (GPS) data, call detail records (CDR) data, smartphone application data, census data and others. Some scholars also collect supplemental data to improve location estimates, such as GPS data with temperature [48] and GPS data with travel survey [49]. Each method has his own strengths and weaknesses in various aspects. Almost all methods can be used for prospective studies, but not all can be applied on retrospective studies. Table 2-1 lists studies reviewed for this chapter.

2.1 Travel Survey and Diary Data

Travel survey and diary data are similar methods, which involves asking subjects to reflect when and where they have been and their corresponding activities using a questionnaire or diary. Neither travel survey nor diary rely on positioning devices (such as GPS). Even though diary data provides more detailed information than survey data [50], they also share some similar characteristics. Electronic diaries were often used for subjects to report their times, locations, and activities [51].

Depends on the purpose of the survey, researchers may need to design their own survey questionnaire and conduct the survey by themselves. For instance, from 1992 to 1994 the U.S. Environmental Protection Agency (EPA) conducted a survey to collect exposure-related human activities in U.S. from 9386 people, which is called National Human Activity Pattern Survey (NHAPS). The survey was conducted through telephone interview and the respondents answered some personal and exposure-related questions and recalled their 24-h retrospective time-activities [26]. In some studies, researchers can choose to perform re-analysis of existing

dataset collected from an existing survey. For, instance, Tang et al. used data derived from a previous survey, the Travel Characterizes Survey (TCS) 2011 published by the HK Transport department. They screened 89385 subjects from 101385 original samples. The survey data includes subjects' age, sex, occupation, trip start location, trip end location, transport mode, number of trips made, duration of the trip on a weekday [22].

The diary method often requires subjects to write down their time-activity diary by themselves [51, 52], sometime with a time resolution of 15 minutes [51]. When studying children, parents can prepare the diary for their children. For instance, Elgethun et al. asked 31 children's mother to write down their children's time-location records. But after comparing the accuracy of diary method with GPS method they found 48% of the children's time-location information was misclassified by their parents using diary method [53].

Travel survey is often used for retrospective studies and people's mobility data can be extracted from existing surveys. Travel survey and diary data may contain personal information, such as age, sex, occupation et al., which can be utilized in further stratification analysis, such as exposure of people in different age. Survey and diary can cover various population subgroups, especially those who don't have access to cellphones, such as adults or children.

The weakness of travel survey and diary methods are well documented. 'Respondent error' is often the biggest concern [53]. The action 'recall' is suitable for short-term survey, and an individual may not accurately remember activities occurred long time ago. Meanwhile, the action of 'take down a diary' also is suitable for recording short-term behavior, since it's unlikely that an individual could keep a continuous recording of their time, location and activity for an extended time period. For travel survey, since subjects are required to recall their time, location and activity, recall bias could happen. For travel diary, participants may forget to record an entry in the diary or fill in an entry as a 'best-guess', which could also contribute to

bias [53, 54].

In addition, survey or diary method is often laborious and expensive. As a results, participants often need assistance to complete the questionnaire which add burdens to researchers. When subjects need assistance to complete the diary, bias could occur [52]. For practical reasons, the time subject spent in one place may be rounded to a larger number (i.e. 5minute activity can be rounded up to 15 minute) by participants [53], which are expected to introduce uncertainty to the study. It has been shown in some studies that the dairy method could overestimate or underestimate subjects' time spent in transit when compared with GPS data [48, 55]. The survey questionnaires are often created in one language and not all the subjects who answer the questionnaires are native speakers [50]. Therefore, bias may be introduced when a subject is an nan-native speakers [56]. Low literacy could also impede the completion of a diary [53]. Other issues with the survey or diary methods include incomplete or missing information, and privacy concerns.

If travel survey and dairy is the only method to obtain subject's mobility data, the bias discussed above could be unneglectable. Therefore, survey or dairy methods should be used with caution.

2.2 GPS Data

GPS refers to global positioning system. A GPS device records time, latitude and longitude simultaneously according to a pre-set time interval. GPS data is expected to have higher spatial and temporal resolution than survey and diary data which are mentioned above. It has been shown that survey and dairy method could overestimate or underestimate subjects' time spent in transit when compared with GPS data [48, 55]. With GPS data, the subjects' in-vehicle travel trips, home location and work location can be extracted. GPS device is worn by

subjects, such as in vests or in a backpack carried by subjects [53, 57]. However, this method requires subjects to be highly cooperative, willing to carry a GPS device with them all the time, and charge the battery regularly. For instance, the participants in Jun Wu's study were asked to carry a portable GPS device when they were awake for 1 week. Since the battery of the GPS device only lasts for 17 hours, participants were asked to turn on the device when they woke up and turn off the device and charge it at the end of their day [55]. GPS is a passive recording method and does not need participants to recall their locations, which eliminate the recall bias existed in the survey and diary method.

In certain locations, such as indoor or underground parking lots, the signal of GPS can be weak [54], leading to missing location data during the study period. Additional data cleaning procedures would be needed, such as replacing missing data or relocating unrealistic points [58]. Incorporating other factors, such as temperature, could help mitigating such issue to some extent. For instance, Nethery et al. applied temperature to adjust the subject's location classification. If GPS data indicates the subject's activity is located outside home, but the recorded temperature is shown 21°C, which is not expected outside, then the activity will be reclassified as indoor [48]. Carrying a GPS device constantly with subjects may also interfere with their behaviors and daily life. If participants are given GPS devices to collect their location data, the interference with participants' daily life should be considered, and should be reduced to an acceptable level [56].

Subjects' noncompliance would lead to incomplete GPS recordings, such as forget to shut down or recharge the device at night. In Jun Wu's study, GPS data is only valid for half of the expected days [55]. Though GPS data does not contain personal information, such as gender, age, economic status. Combining GPS with questionnaire, individual information can be obtained. In peri-urban South India, Sanchez et al. gave 47 participants (24 women, 23 men)

GPS devices and investigated the difference of mobility pattern between men and women. Different from women in the city, women in the peri-urban area spent 4 hours more staying at home than men, which supported the importance of stratified analysis in exposure estimations [59]. The GPS method can only be used in prospective studies where participants are given GPS devices before the data collection begins.

In order to improve the accuracy of spatiotemporal resolution and exposure estimation, the GPS method can be combined with daily activity diary method [49] [60]. This combinations allows better estimates of the time a subject spent in different microenvironments, especially in transportation [49]. Since diary requires relatively high cooperation of the participants, involving GPS method will also mitigate the error introduced by the participates [61]. For instance, Buonanno et al. investigated 24 couple exposures to UFP (ultra-fine particulate matter). Among them, all man had a full-time job while all woman stayed home. The subjects were given a GPS device and an UFP monitor, and they were asked to fill a travel diary. Combining all these methods, the authors found that the average exposure to UFP for women was higher than men during summer and winter. The activity that lead to the highest exposure for women was cooking, while for men it was transporting [49]. GPS can also be combines with other data such as acceleration. For instance, Dewulf et al. adopted accelerometer to obtain subjects' acceleration which can indicate subjects' transport mode (such as cycling and driving) and physical activity (such as light and heavy physical activity). With these data, subjects' ventilation rate can be assessed, and the inhaled dose of NO₂ can be estimated. The authors found that, if incorporating ventilation rate, the inhaled dose of NO₂ was 12% larger than dose that only including subjects' mobility [62].

2.3 CDR Data

CDR refers to 'Call Detail Record' which is recorded by carriers' towers when cellphones are connected to the tower, such as turning on, turning off, sending messages or making a call. In 2016, Dewulf et al. applied this method in air pollution exposure estimation [42], and it has gained much popularity since then. The sample size of existing studies that used CDR data ranges from thousand to as large as millions [42, 44]. Apart from using census data, the CDR method can also be used to perform population level analysis [63-66] while also accounting for individual mobility patterns, which attracted attention from many researchers. Since health policies mainly focus on the health improvements on population, instead at individual level, population level exposure estimations are of great interests [63]. In the study of Picornell et al., mobility data were extracted for the entire population of Madrid, Spain using CDR data, and population exposure to NO₂ was assessed [64]. Nyhan et al. used CDR data to calculate population-weighted exposures to PM_{2.5} in New York city while accounting for mobility [44]. In Roma, Italy, the variability of population exposures to NO₂ could reach to 50% at downtown region where population density exceeds 1000 people/km² during daytime. Such population variation cannot be identified when using census data [63].

Compared with exposures calculated using census data or home addresses, exposures estimated using CDR data is considered to be more reliable [15, 63, 67]. Nowadays, the vast majority of population owns mobile phones, even in rural areas [54], thus CDR data is expected to be widely available. Unlike the GPS method, this approach needs no additional devices. CDR data is also passively collected and doesn't intrude participants' personal life, so subjects' behaviors are not disturbed [42]. CDR data can be applied to long-term study extending for years. This approach can be used for both prospective and retrospective epidemiological studies.

Usually, demographic information (such as gender, age and education level) of the

mobile phone owner can't be obtained through CDR data, since those would be concealed by the provider [54]. In Claudio Gariazzo's study, demographic information are available, but they were stored in a separated database and cannot be linked to other database [63]. The two databases reduced the risk of leaking personal information.

However, there are a few weaknesses associated with using CDR data in exposure estimation. The spatiotemporal resolution of CDR data is relatively low compared with GPS data [42]. Location information as recorded in CDR are the locations of carrier's towers, but not the actual location of subscribers. In rural areas, the coverage of towers is less compared with urban region, thus leading to more error in location data. This method may also exclude some population subgroups who have no or limited access to mobile phones, such as children and certain elderly people. CDR data also contains private information, therefore public perception on the use of this approach needs to be carefully considered. Jones et al. studied public view on using CDR in health research. Without providing any background information, 62% of the participants (N=61) were willing to share their anonymous CDR data. After a workshop explaining measures for privacy protection, the proportion increased to 80%. The author concluded that people are generally willing to share their anonymized CDR data as long as they were well informed and safeguarded [68].

CDR data have also been used in research fields other than air quality. For instance, Thomas et al. used CDR data to assess per capita pharmaceutical use and illicit drugs use. They concluded that the results were only possible by using dynamic population (instead of static population) [65]. Furletti et al. were able to identify events within an urban area by examining the size of population at a particular time and location using CDR data. The identified events including religious event, concert performance or special holiday events, which provides useful information for urban managers and decision makers [66].

2.4 App Data

The widespread use of smartphone, smartphone application (herein referred to as App), and location-based services, has enabled a promising approach to collect mobility data for a large population at individual level [69]. We collectively name this type of data “App” data since all location data are collected by one or more smartphone applications. This method has been shown to improve exposure estimation considerably compared with the home address based approach [18, 69]. Smartphone Apps that have been used in exposure estimation studies include Tencent Apps, ExpoApp, CalFit and Google Maps. Among them ExpoApp and CalFit are, at least partially, designed for exposure estimation studies. Though not designed specifically for research purposes, Tencent Apps and Google Maps do collect location data from individuals and these Apps have been installed by a large population. Chen et al. used big data derived from Tencent big data platform in China (the data mainly come from two apps developed by Tencent, Wechat and QQ) to estimate population exposures to PM_{2.5} and compare them with exposures calculated using census data. And he concluded that if the dynamics of population movements were not considered, the bias of exposure assessment would reach over 100% across different temporal scales [18]. Gonzalez et al. developed an exposure assessment application called ExpoApp, which can estimate the time participants spent in various microenvironments, the degree of participants’ physical activity, and their exposures to pollutants and green space, by integrating various data sources including location, acceleration, monitoring and individual information [70]. They have also demonstrated the reliability of this approach [70, 71]. Nieuwenhuijsen et al. provided 54 school children a smartphone with CalFit software pre-installed to obtain their location information and physical activity level. In addition, a personal sampler is also provided to measure their exposures to black carbon (BC) [72]. Gonzalez et al. compared CalFit with GPS and found that CalFit worked better on

providing the information of which microenvironment the participant stayed [73].

The App method does not require additional devices, but only requires installing an app on the participant's smartphone. Once the app is developed, it can also be used by other researchers [70]. Like CDR method, the data collection process cause minimum interference to the subjects' daily life. The confidentiality of the data collected through App can and should be guaranteed, such as using asymmetric encryption [70].

However, the smartphone applications will increase the burden on the smartphone's battery. For ExpoApp and Calfit, the subjects need to install the App at the beginning of the studies. If they are also required to carry portable devices to measure the pollutant concentration in microenvironments, the study may become laborious and costly.

A very promising method among all the Apps mentioned previously is using the Google Maps app. CDR data records the locations of towers, and GPS data often has missing points under indoor circumstances. But because Google Maps data is a combination of GPS, Wi-Fi and cellular positioning [74], it can still obtain location data in indoor situations . However, despite its popularity, Google Maps is not installed on every smartphone, and among those who installed Google Maps, not all of them have enabled the 'Location History' feature.

Given its popularity, Google Maps is widely used, and thus large amount of location data are being collected from the users. In order to evaluate the accuracy of using Google Maps location data to characterize human mobility patterns, Ruktanonchai et al. compared Google Maps location history (GMLH) data with GPS tracker data and found that GMLH data is equivalent with GPS data within 100 meters [75]. Su et al. estimated one subject's exposure to NO_x by using GMLH data as well as indoor-outdoor air exchange ratio. With GMLH data and indoor/outdoor ratio considered, exposure estimates was up to 359% higher than the exposure estimated using home address alone [76].

GMLH data also has several other strengths. They may be available retrospectively for several months or even years, which can be difficult to achieve for other methods, such as survey (diary) method or GPS data. GMLH data can also capture more international location data during traveling [75]. GMLH data are passively collected and are easy to retrieve. Anyone who installed the Google Maps application on their smartphone, signed-in with their Google account, and enabled the Location History feature, will be able to easily download his/her own location data by using ‘Google takeout’ [74]. GMLH data can also be used in both prospective and retrospective studies. GMLH records location data at an average interval of one minute [75], therefore, temporal resolution is relatively high.

However, only users have access to their own GMLH data. It is expected to be difficult to use this method to evaluate mobility for a large population given the logistic issues, and potential sampling biases associated with participants recruitment [75]. In addition, people who don’t have Google Maps installed on their smartphone will not have GMLH data available. This subpopulation group will include population such as children, elderly, and people with low socioeconomic status.

2.5 Census Data

Census data has been widely used in past studies for the exposure estimation. Though this approach ignores individual mobility, it still has advantages when compared with other methods. First, though census data cannot be used to people’s mobility, it can be used to determine population changes over time. For instance, in Anna Rosofsky’s study, population in urban census blocks increased by 5.3%, while population in rural census blocks decreased by 4.1% in Massachusetts from 2000 to 2010 [77]. This phenomenon may indicate a general trend of population moving from rural to urban areas, and such information is useful for long-term

population exposure estimation. Another instance is that Aleksandropoulou V. et al. used census data to investigate the trends in population exposure to PM₁₀ and PM_{2.5} during 10 years in Greece [78].

Second, census data contains many detailed information, such as urban/rural groups, racial/ethnic groups, income status and education levels. With such stratified information, we can explore the potential disproportionate distribution of air pollution exposures among different groups. While this information can be collected from subjects in the methods mentioned above, when it comes to large population of people, census data method would be much more easier. For instance, with census data of Massachusetts state, Rosofsky et al. found that urban non-Hispanic black populations have the highest annual population weighted PM_{2.5} concentrations. And urban Hispanic populations have the highest annual population-weighted NO₂ concentrations [77].

Third, though census data do not contain detailed mobility data, population mobility can be approximated using census data. For instance, Nathan et al. adopted subjects' home address and work address from the Israeli Central Bureau of Statistics based on 2008 census. After simulating subjects' trajectories, the authors obtained the exposures accounting for people's mobility [23, 43]. Reis et al. used census to obtain workday population distribution in UK and calculated population exposures including both residential and workday location. Their results showed that exposure to NO₂ including work locations could lead to a 2% exposure increase when compared with exposure estimates not including work locations [79].

2.6 Others

In addition to above mentioned methods, there are also other methods available for estimating the impact of mobility on exposures. The associations between pregnant women's

exposures and infants' health is of interests for some researchers. Some researchers used mothers' home addresses to estimate exposure during fetal growth [80]. However, studies suggested that residential mobility should be considered if we want to explore the subtle associations between maternal exposures and birth outcomes [19]. However, in the study of Chen et al., the estimated exposures to ozone and PM₁₀ using birth certificates' address is found to be representative of exposures estimated using multiple maternal addresses (due to moving of mothers) [31]. Therefore, birth certificate data would still be valuable for mobile exposure estimation to some extent. When using birth certificate data, the 'mother' is the only subject available.

In a study focusing on the associations between air pollution exposure and psychotic outcomes, the participants' air pollution exposures were calculated by averaging the pollutant level of three places, home addresses and 2 commonly visited locations [9]. However, it is still unknown whether the average exposures of these three locations can approximate personal exposures.

Salmon et al. utilized wearable camera to estimate their subjects' exposures to PM_{2.5} [81]. The pictures taken by cameras were manually labeled by to indicate different microenvironments [82]. This method mainly focused on the effect of various activities on exposures, rather than the effect of mobility.

Dhondt et al. used an activity-based model to simulate 5 million people's mobility based on 8800 persons' diary survey information. They compared health impacts estimated using modeled exposures with health impacts estimated using home addresses and the results showed only modest difference [24].

Ma et al. used subway smart card data to study diurnal dynamic changes of population in the city of Beijing, China. They found that one administrative and commercial districts of

Beijing has a high population ratio (day to night) of 1.35, while a residential district of the same city has a low ratio of 0.84. This finding could change the population exposure estimation considerably [25]. However, this method may only be applicable to cities with existing subway infrastructures.

Escobar et al. used a novel open mobile mapping tool to estimate the mobility of offline people in rural Amazon and investigated the impact of human mobility on malaria dynamics, without the availability of internet and mobile phone signal [83].

Transaction records of credit cards also contain mobility information and individuals' detailed attributes, such as age or gender. Lenormand et al. investigated the mobility pattern of different group of people in Barcelona and Madrid using this method [84].

Finally, social media nowadays affecting people's daily life in the society. Social media check-in information can be used to retrieve people's mobility data. For instance, Wu et al. used 15 million social media check-in records during one year in Shanghai, China to model people's mobility pattern [85]. Longxu Yan et.al used social media check-in data (Weibo) to study the impact of air pollution on people's activities. They found potential correlations between air pollution and change in people's behavior. Leisure-related activities are also found to be more affected by air pollution than work-related activities [86].

2.7. Discussion

Many methods are available to capture the impact of human mobility on their exposure to environmental risk factors. Choosing the most suitable method needs careful consideration on the strengths and weakness of each method. The CDR-based method is promising, it can be applied for retrospective studies. However, CDR data contains the locations of cellphone towers, but not the subscribers. The GPS method collects the locations of subjects, but it cannot

be used for retrospective studies. GPS signal can be weak indoor or underground, leading to inaccurate data. The Google Maps method is available in retrospective studies, recording subject's real location and less weakness when indoor. Though GMLH data is not available for the entire population since the Google Maps App is not installed on all smartphones. Hence this method is difficult to be applied for population level studies.

In prospective studies, the subjects' future location data will be collected. For retrospective studies, the subjects' historical location data are desired, which are generally more difficult to collect. Among all the methods discussed above, some of them can be used for both prospective and retrospective studies, such as CDR, Google Maps, Census data, mother's birth certificate and social media data. Methods such as travel diary are suitable for retrospective studies, while some methods can only be used for prospective studies, such as travel diary, GPS, App and sensor. The best method to use depends on research design.

Indoor environment is another concern when estimating people's overall exposures to air pollutants. People spend most of their time indoor. Ignoring indoor factor would lead to misclassification errors. Ouidir et al. has showed that exposures estimated only including mobility correlated poorly with exposures estimated including both mobility and indoor factor (r ranges between 0.03-0.05) [58]. This is particularly an issue for participants who preferably stay indoor.

Though there are still much need for further research, the brief review presented in this paper provides insights for accounting mobility in air pollution exposure estimation, and contributes to further understanding on this subject.

2.8 Sensor Data

Sensor data is not one of the methods to obtain people's mobility, but it is worth

discussing when accounting for people's indoor and outdoor factors. Strictly speaking, a person's exposure to air pollution consists of exposure to both indoor and outdoor air pollution, and indoor pollutant concentrations can be significantly different from outdoor concentrations [61, 87]. It has been shown that ambient ozone concentration alone is not a favorable proxy for estimating personal exposures, other factors such as meteorological conditions and window opening status should be taken into account [88]. This finding can also be applied for other pollutants. Often indoor pollutant concentration alone is not a suitable exposure surrogate, though there are exceptions: pregnant women who spend 60% of their time indoor at home have personal exposures that correlate well with indoor pollution concentration levels [89].

Personal sampling method can be used to measure pollutant concentration at the place of activity [90], such as in microenvironment and in transit, which may not be accurately estimated by using ambient concentrations. In past studies, devices such as personal samplers have been carried by subjects to collect pollutant samples around breath zone, and later transferred for lab analysis [91-93]. The time difference between sample collection and lab analysis could affect concentration measurements [94]. The advancement of sensor technology has led to the commercialization of sensors that are able to provide near-real-time pollutant concentration measurement, which helps to overcome this issue. Location data collection may no-longer necessary in this case for estimating the subject's total exposure. In Daniela Dias's review, monitoring is used as the most reliable and accurate method to estimate exposures by measuring concentrations within subjects' breathing zone [90].

As discussed previously, results from past studies have shown that exposures as obtained by sensor data can appropriately account for pollutant concentrations in indoor environments, which can be considerably different from outdoor concentrations [87]. Using sensor data, one can account for pollutant infiltration, and indoor sources such as cooking,

which can greatly impact exposure estimation [90, 95]. When using sensor that is capable of providing data in high temporal resolution, short-term or peak exposures can be identified, such as cooking and commuting [95]. Such information is expected to be helpful for the studies of acute illnesses.

However, all sensors are required to be calibrated with reference instruments as drifts may occur in the process of using them. Often calibration under lab environment is not sufficient, and field calibration is necessary[95]. For instance, Juana Maria Delgado-Saborit compared BC and NO₂ sensors with reference methods under field conditions and estimated R² of 0.89 and 0.63, respectively [95]. And measurement made by sensors of different brands may not correlate well with each other.

There are also many logistic issues associated with using personal sensors. The process of participant recruitment may be difficult [87], which will increase study cost. Since subjects must carry sensors with them almost everywhere and all the time, considerable attrition loss is expected for long-term study. This method also can't be utilized for population level studies due to the cost and time. If subjects only carry sensors with them but don't carry any positioning device, subjects would be asked to keep time activity diary if spatiotemporal activity pattern is necessary for the study, which would add burden to researchers since sensor data needs to be correlated with diary data [95]. The battery life of sensor is another concern and frequent charging may be necessary, which would increase burden for subjects. One way to solve this issue is to add an additional battery to the original sensor, which however would increase the weight of the sensor [95]. Finally, this method can only be used for prospective studies.

CHAPTER THREE: GOOGLE MAPS LOCATION DATA WORK

3.1 Introduction

Exposure to ambient air pollution has been linked with numerous adverse health effects [6, 96-99]. An appropriate characterization of human exposure is critical for air pollution studies [100-103]. One major source of uncertainty in accurately estimating individual air pollution exposure is that the location of each individual changes constantly. Since concentrations of many pollutants are known to vary substantially spatiotemporally, neglecting subject mobility is expected to introduce a considerable amount of uncertainty in exposure estimation [31, 104], consequently leading to potential exposure misclassification errors [17, 105] that can lead to biases in subsequent statistical analyses [106, 107].

In prospective studies, detailed activity diaries [26], GPS devices [27-29] or smartphone-based location tracking applications [30] have been used to record individual human movement. Some of the methods can be intractably expensive for large sample populations, and not all methods are feasible for retrospective studies in which subject mobility data from the past are needed. In past retrospective studies, exposures have typically been estimated only at the residential addresses of the subjects. A few studies have attempted to address this issue by using a number of methods, including relying on aggregated travel survey data [22], accounting for multiple addresses of subjects [31-33], or employing computationally-intensive travel demand models to simulate explicit travel paths for subjects [34-36]. However, the inferred mobility data from these methods are either lacking details at the individual level, or are only approximations. Retrospective studies that utilize detailed and actual mobility data of individual subjects remain scarce.

Smartphones are now ubiquitous and location data are continuously being collected and archived from virtually all smartphones by network carriers [15], and by various smartphone applications (herein referred to as apps) [37] such as the popular Google Maps app (Alphabet Inc., Mountain View, CA). While these “historical” location data are being used extensively for criminal investigations, commercial purposes, and in academic research in fields such as transportation analysis and planning [38-41], the potential of smartphone location data is just starting to be explored in environmental research related to air pollution [18, 23, 42-45]. However, it is commonly recognized that most smartphone location data are spatiotemporally sparse since they are usually collected at irregular intervals [46, 47]. Whether these data appropriately capture individual mobility and the applicability of such data in air pollution health studies remains under-investigated.

As one of the most downloaded apps, and the most used navigation app on both iOS and Android platforms [108], Google Maps is regularly used by billions of smartphone users worldwide [109]. The app contains a feature called “Location History”, that when enabled, will continuously and passively collect location data from an individual’s smartphone using technologies including GPS, Wi-Fi and cellular positioning. In this manuscript, we demonstrate the potential of such Google Maps location history (GMLH) data in air pollution exposure estimation. We first compared GMLH data from a single smartphone with detailed location data recorded using a co-located reference GPS data logger to evaluate the accuracy of GMLH data in capturing the 1) spatial movement of the subject; 2) the time the subject spent at different microenvironments; and 3) the time the subject spent driving during the one-week time period. Using satellite-derived aerosol optical depth (AOD) measurements, we then compared the subject’s time-weighted exposures to ambient particulate matter as estimated using GMLH, the reference GPS data, and home address alone, to investigate the applicability of GMLH data in

exposure estimation for air pollution. Further, we conducted an online survey to assess the potential availability of GMLH data among smartphone users in the US.

3.2 Materials and Methods

As a proof-of-concept study, a single subject carried an Android smartphone with the Google Maps application installed, and a co-located reference GPS data logger for a one-week time period between April 26 and May 2, 2018. The GPS logger used here was the QStarz's BT-Q1000XT GPS Travel Recorder (Qstarz International Co., Ltd.), which has been successfully applied in several previous health studies to characterize individual mobility [29, 110-114]. The logger was pre-configured to record GPS coordinates every 10 seconds. The vibration sensor was enabled on the logger, which temporarily shuts down the device when no vibration is detected for more than 10 minutes and restarts the device when vibration is detected. The Android device used was the Google Nexus 6P (circa 2015) running on Android version 8.1.0 operating system. When the location history feature is enabled, the Google Maps application will continuously collect location data from the smartphone at varying frequencies. The collected data are stored in the "cloud" and are linked to the Google account that is synced with the Google Maps application. The archived location data can be viewed/edited in Google Maps Timeline and can be retrieved easily using Google Takeout (a tool that can be found on the website) with the user's Google account. Both Timeline and Takeout are services provided by Google (Alphabet Inc. Mountain View, CA).

During the week of data collection, Android location services, location history, Wi-Fi, and cellular data were enabled on the subject's Android smartphone. The subject did not open the Google Maps application. The purpose here was to capture a baseline scenario for data collection since more GMLH location data are expected to be collected when the Google Maps

application is being used. The subject conducted his routine travel activities during the week. To additionally test location data collection during non-routine travel, the subject travelled to a nearby tourist destination on day 3, and to a nearby regulatory air monitoring station on day 7 (45 and 80 miles from the subject's home address, respectively). At the end of the week, location data from the GPS data logger were retrieved using the QStarz proprietary software (QTravl, version 1.49). Data points identified as "drift" by the software, which are mostly repeated location records collected when the subject stays stationary for an extended time, were removed by the software prior to subsequent data analyses. Location data collected by the Google Maps application were retrieved using Google Takeout.

In order to compare the smartphone and GPS logger datasets, we linearly interpolated both GPS and Android location data to a uniform 10-second interval spanning the entire week. This approach assumes that the subject moves in a straight line with a constant speed between each pair of temporally consecutive location points. We do not expect the removed "drift" GPS data to impact the results of such interpolation since they are mostly repeated location records collected when the subject stays stationary.

Choosing the GPS data as the ground truth, we evaluated the capability of GMLH data in capturing: 1) the spatial mobility of the subject during the one-week time period; 2) the time the subject spent at different microenvironments; and 3) the time the subject spent driving. Here microenvironment is defined as a fixed-activity location that was not part of the subject's travel route. A total of 10 microenvironments were considered in this study, including the subject's home and work location, two grocery stores, three clinics, one postal office, one tourist destination, and an air quality monitoring station.

To evaluate the capture of spatial mobility, we divided the entire spatial extent of the subject's movement into grid cells of four different sizes: 1 km, 500 m, 200 m and 100 m. The

grid networks were overlaid with both sets of interpolated 10-second location data. If an interpolated data point was located within a grid cell, the subject was assumed to stay within the corresponding grid cell for the entire 10-second time period. The number of interpolated data points were counted for each grid cell, and the time (in minutes) the subject spent within each grid cell was calculated accordingly.

To evaluate the capture of time spent at each microenvironment, we first drew rectangles over the approximate perimeters of the corresponding microenvironment. We choose rectangular areas for this analysis because the perimeters of most microenvironment in this study are rectangular. If an interpolated data point was located within a rectangle, the subject was assumed to stay at the corresponding microenvironment for the entire 10-second time period. The time the subject spent at each microenvironment was then calculated accordingly. Further descriptions of the microenvironments are provided in supplemental materials.

To evaluate the capture of time spent driving, we first estimated the traveling speed of the subject for each 10-second interval by simply dividing the distance between two consecutive interpolated data points. The subject was assumed to be driving during the 10-second interval if the estimated speed was over 15 km/h. We did not rely on other contextual data such as roadway network here in consideration of the spatial accuracy of the location history data, and the fact that the subject also walked along roadways during the week. We recognize that this method will likely under-estimate actual driving time since the time the subject spent waiting at signalized intersections and during severe traffic congestions will not be counted toward the total driving time. We estimated the subject's time-weighted exposure to ambient particulate matter (PM) by combining gridded daily mobility data with weekly-averaged 1-km resolution aerosol optical depth (AOD) data for the corresponding week (re-

gridded to the same 1 km grid network). We used AOD data estimated by the Multiangle Implementation of Atmospheric Correction (MAIAC), an established satellite data product based on Moderate Resolution Imaging Spectroradiometer (MODIS) data. MAIAC is capable of extracting aerosol information over dark vegetated surfaces and bright desert areas. Further information on the operating principles of MAIAC can be found elsewhere [115-117]. For comparison to the home-based exposure estimate, AOD data were normalized by the retrievals from the grid cells where the subject's home is located. AOD data were missing for approximately 1.8% of the grid cells the subject visited during the week, and these missing data were estimated using natural neighborhood spatial interpolation. We applied weekly-averaged AOD data here due to the considerable amount of missing retrievals at the daily level. We did not develop actual PM concentration fields from the AOD data due to the short study duration (1-week) and the lack of available ground measurement data within the study domain.

Finally, we also conducted an online survey to assess the availability of GMLH data among smartphone users in the US. Survey participants were recruited using the Amazon Mechanical Turk platform. The survey consisted of a series of five questions and typically took less than 4 minutes to complete. Each participant received \$1 to compensate for their participation, and was asked to indicate their age, gender, and whether or not they own a smartphone and use the Google Maps app. Those who respond "Yes" to smartphone ownership and Google Maps usage were then asked to indicate whether or not they have GMLH data available. Graphical instructions were provided to help survey participants locate their GMLH data. This survey was approved by the Institutional Review Board at the University of Central Florida (IRB ID: STUDY00000281). Data from the survey were analyzed for demographic summary statistics and to determine the percentage of participants with available GMLH data.

3.3 Results

Over the course of 1 week, the subject's Google Maps application collected 2,224 location records. The sampling interval ranged from less than 1 second to 69 minutes, with an average interval of 4.5 minutes. The GPS data logger recorded 32,314 location records (with "drift" data removed). The GPS data logger sometimes recorded data in 1 second intervals despite the configuration of 10 s. For consistency, the 10 s interval data were still used for later analysis. In addition, due to an error with the built-in vibration sensor, the GPS logger failed to record location data for 1 hour 25 minutes during day 6, and for 7 hours 55 minutes during day 7. Data from both devices collected during these time periods were excluded from our comparison analyses. Further discussion of the location data collected are provided in supplemental materials.

The GMLH data captured well the spatial mobility of the subject within the one-week study period (Figure 3-1) when the interpolated data points were aggregated to 200 m or larger grid cells. For both 1 km and 500 m resolution, the coefficients of determination (R^2) between aggregated time spent (both daily and weekly totals) in each grid cell from the GPS and GMLH data points were all near perfect (rounded to 1.00), with the slope of linear regression of 0.99 and an intercept below 0.25. The fit decreased for 200 m resolution ($R^2 = 0.78$ and 0.90 for the daily and weekly total time, respectively), but performance was still good. However, when both GMLH and GPS data were aggregated to 100 m grid cells, the correlation largely disappeared ($R^2 = 0.085$ and 0.17), suggesting that the GMLH data cannot capture individual mobility at such a fine resolution. Additionally, the performance of GMLH data was less satisfactory, though still somewhat promising for 1 km resolution, for grid cells in which the subject spent less than 10 minutes based on GPS logger data (see supplemental materials for further discussion). These shorter residence grid cells mostly cover roadways that the subject drove

through. In this proof-of-concept study, we linearly interpolated the Google Maps location data, which contributed to the observed errors. Time estimates in these grid cells can be further improved in future studies by including travel route information that can be obtained using individual mobility modeling [118-123].

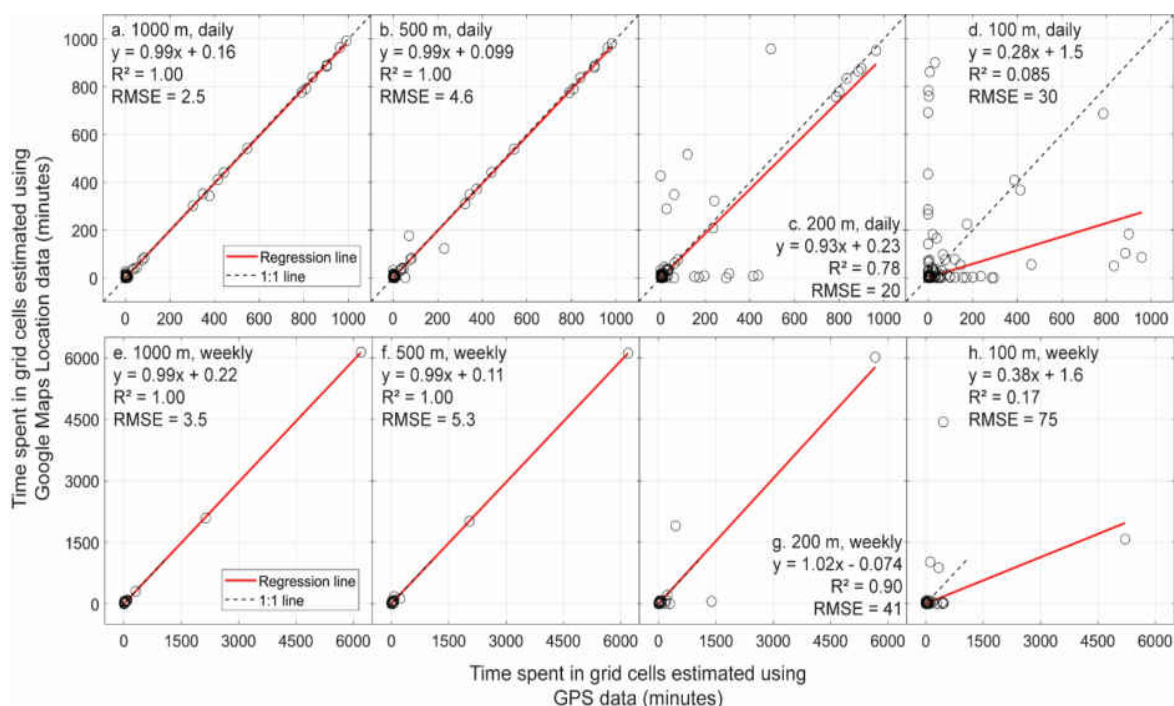


Figure 3-1. Comparison of the estimated daily total (a-d) and weekly total (e-h) time the subject spent in each 1 km (a,e), 500 m (b,f), 200 m (c,g) and 100 m (d,h) resolution grid cell based on GPS versus Google Maps location data.

Using the GPS data logger as the ground truth, we evaluated the time the subject spent in each of the 10 microenvironments as estimated using GMLH data (Table 1). During the one-week time period, Google Maps location data accurately captured the time the subject spent at all 10 microenvironments, with errors less than 3% for 8 of the 10 locations, approximately 5% for the location where the subject spent most of his time (#10), and only off by 2 minutes for microenvironment #4.

Table 3-1. The estimated time (in minutes) the subject spent at each of the 10 microenvironments, as estimated using GPS and Google Maps location data.

ID	Day 1		Day 2		Day 3		Day 4		Day 5		Day 6		Day 7		Total	
	GPS	GM	GPS	GM	GPS	GM	GPS	GM	GPS	GM	GPS	GM	GPS	GM	GPS	GM
1													13.2	13.0	13.2	13.0
2					300	300									300	300
3				0.17			4.33	3.83							4.33	4.00
4				0.17					29.2	29.0		1.50			29.2	30.7
5			4.50	4.83											4.50	4.83
6											34.0	32.5			34.0	32.5
7									82.0	80.3					82.0	80.3
8	76.8	76.3								0.17					76.8	76.5
9	419	421	549	550			446	446	381	382	14.7	6.50	399	400	2208	2205
10	897	893	836	834	983	977	647	960	902	898	804	803	788	788	5856	6153

GPS: Time estimated using GPS logger data; GM: Time estimated using Google Maps location history data. These times do not include the time the subject spent in travel. Additionally, note that 1 hour 25 minutes of data in day 6, and 7 hour 55 minute of data in day 7 were removed due to GPS logger malfunction.

Additionally, GMLH data reasonably captured the time the subject spent during driving (Table 2), with overall estimation errors of approximately 1.6%. These errors may be due in part to the relatively larger spatial inaccuracy of the collected Google Maps location data also documented in a recent study [124] (see supplemental materials for further discussion), which could impact the speed estimates, as well as the interpolation scheme. We expect the results would likely be further improved if individual mobility modeling [118-123] were to be performed, from which detailed travel route and driving time information can be obtained.

Table 3-2. The estimated time (in minutes) the subject spent driving, and time-weighted daily exposure to ambient PM (normalized to subject’s home location), as estimated using GPS and GMLH data.

Time spent driving (minutes)								
	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Total
GPS	49	50	114	28	47	91	137	516
Google Maps	39	41	133	29	39	77	152	508
Time-weighted exposure (normalized to subject’s home location)								
	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Average
Home address	1	1	1	1	1	1	1	1
GPS	1.09	1.11	1	1.08	1.09	1.06	1.13	1.08
Google Maps	1.1	1.11	1.01	1.09	1.1	1.08	1.14	1.09

The estimated time-weighted daily exposures to ambient PM using GMLH data were similar to the estimates using GPS data logger (Table 2). Based on the GPS data logger, the subject’s daily exposure varied between 1.00 (equal to the home-based exposure estimate) to 1.13 (13% higher than home-based exposure estimate). Exposure estimates using Google Maps location data closely approximate those estimated using GPS logger data, with a weekly average bias of only 0.053%.

We received 317 responses from the online survey to assess the availability of GMLH data among smartphone users in the US. A total of 33 responses were repeated responses recorded from 13 unique IP addresses. These responses were labeled as spam and were not included in data analysis. Among the remaining 284 valid responses, GMLH data were available for 61% (n = 174) of the sample, not available for 30% (n = 84), 5% (n = 14) were not sure whether or not they have GMLH data available, and 4% (n = 12) do not own a smartphone or do not use Google Maps on their smartphone. We note that 51% of the sample population were between 25-34 years old, which is a larger share for this age range than the overall population in the US.

3.4 Discussion

Our results suggest that GMLH data captured well the subject's spatial movement within the week of data collection when data were aggregated to 200m or larger sized grid cells. The GMLH data also captured the time the subject spent at different microenvironments, and the time the subject spent driving during the week. Utilizing GMLH data, we were also able to accurately capture the subject's time-weighted exposure to ambient PM pollution. Our results are consistent with Su et al. [125], who found that a single subject could be accurately identified in space-time more than 95% of the time, using GMLH data collected through the WiFi network (which is expected to collect less data than when GPS and cellular positioning are enabled). Furthermore, results from Ruktanonchai et al. [75], who focused primarily on trip mobility, show that GMLH data perform better than traditional travel diary data, particularly for capturing long-distance and international trips.

Additionally, evidence suggests that GMLH data may be available for a considerable portion of the population. Results from our online survey using the Amazon Mechanical Turk platform showed that 61% of the US survey sample (n = 284) had GMLH data available. This percentage is in the range of results obtained by Ruktanonchai et al. [75], who surveyed the availability of GMLH data among Android smartphone users in five countries (n = 250 per country for Japan, Mexico, UK, US and Brazil). They found that the availability of GMLH data ranged from 43% to 72%, only 5.6% – 17.5% of users knowingly disabled the service, and 24 – 51% of the survey sample were not sure whether they had GMLH data available. For a small group of individuals examined further (n = 25), 100% of those who were not sure (n = 7), were found to have GMLH data available. Therefore, we expect the actual availability of GMLH data to be even higher among Android smartphone users than that found here. Furthermore, according to market survey data from the Pew Research Center, approximately

77% of individuals in the US now own a smartphone. The percentage increases to 94% for people between 18-29 years old and 89% for those 30-49 years old [126]. As one of the most popular smartphone apps and the most-used navigation app worldwide [108, 109], Google Maps is regularly used by billions of people [109]. Therefore, we expect GMLH data to be available for a considerable portion of the general population. However, additional studies are needed to investigate the availability of GMLH data, particularly among different population subgroups.

In addition to being available for a large portion of the population, limited results suggest that GMLH data may be available for an extended time period. In this study, the subject's Google Maps location history dates back to May 2016 when the subject's smartphone was purchased and activated. In two and a half years, the subject's Google Maps application recorded approximately 235,000 location data points, with an average sampling interval of 5.5 minutes. Ruktanonchai et al [75] also investigated the temporal availability of GMLH data among a small group of individuals ($n = 25$), and found that GMLH data extended back, on average, for 556 days. Such an extended temporal data coverage further highlights the tremendous potential of GMLH data for improving air pollution exposure estimation in retrospective epidemiological studies.

There are also a few other potential advantages of using GMLH data for exposure estimation as compared with traditional approaches in collecting mobility data. First, GMLH data are actual mobility data collected from each individual. Traditional data collection methods for retrospective health studies (including aggregated travel surveys [22], accounting for residential mobility [31-33], or explicit travel simulation [34-36]) provide mobility data that either lack detail at the individual level, or are only approximations. Second, collecting GMLH data is comparatively easy and low-cost. GMLH data are automatically collected by

the Google Maps app when the Location History feature is enabled, which adds minimum burden on participants (and thus could potentially increase study participation and compliance). There is no need for external devices nor a dedicated app; the former can be costly, and the latter would require participants to install additional app(s) on their smartphone that could decrease battery life or smartphone performance. The GM app also requires no investigator resources for software purchase, usage, maintenance, or server hosting. Third, GMLH data are easy to retrieve, and users have full control over their own data. GMLH data can be retrieved using Google Takeout with a few mouse clicks and an account log in. Further, users can view or partially remove their location data in Google Maps Timeline using an intuitive graphical interface. Both Timeline and Takeout are services provided by Google that have the potential to increase participation of subjects who wish to share only a portion of their data. For other location data collection methods, such an accommodation would require the in-depth involvement of the investigators or would be costly to achieve. Fourth, GMLH data are passively collected and are not subject to biases often associated with widely-used location data collection approaches, such as the well-recognized recall bias in self-reported activity diaries [127]. Finally, GMLH data are collected from devices from which the user's Google Account was used to sign in. Even when the user switches to another device, GMLH data will continue to be collected and archived. Overall, this new approach to collect individual mobility information should allow us to overcome the current limitations associated with other data collection methods, thereby opening a new horizon for better investigation of the health impacts of air pollution.

Despite their promise for retrospective studies, there remain substantial limitations and concerns associated with using GMLH data for air pollution exposure and health research. Most importantly, location data collected from smartphones contains sensitive information on

individual users. Privacy concerns associated with smartphone location data have already attracted substantial attention from the media and the general public [128-131]. Therefore, it is clear that smartphone location data need to be handled carefully with privacy protection in mind. Studies on the public acceptability and legality of using this method are needed before launching a full-scale study, and this will aid in to establishing ethical research protocols for future research. Research using this data will require appropriate human subjects protections, including user consent before accessing or publishing the data. Second, the GMLH data presented here are collected from a single individual. However, there is little evidence to suggest that the technologies would not work with other test subjects, though the availability and accuracy of location data, and the amount and frequency of data collected by different smartphones, is likely to be different depending on user behavior, hardware, and other factors including the version of operating system, cellular coverage and signal strength, and Wi-Fi availability. Further research studies are also needed to determine whether the conclusions from this study can be generalized to a larger population. Third, our results show that the GMLH data in this study were not able to capture the subject's mobility at the highest resolution (100 m) investigated. However, the accuracy of GMLH data are expected to differ by the hardware of the smartphone. Here we used the Nexus 6P, which is an old version Android smartphone released in 2015. Further studies are needed to evaluate the spatial accuracy of GMLH data at finer resolution before they can be applied to characterize individual mobility at the fine scales that can be important for some exposures (e.g. near road exposures).

Finally, we acknowledge two important limitations to the findings of this study. First, due to lack of data, we applied MAIAC retrievals directly in exposure estimation. This is appropriate for relative comparisons of exposures, but because AOD is only a proxy of ambient PM concentrations, results should not be used as measures of actual exposures. Second, we did

not account for indoor/outdoor concentration differences at different microenvironments. Although this further limits the interpretation of our results as actual exposures, our purpose here was not to obtain accurate exposure estimates for the subject, but rather to demonstrate the applicability of GMLH data in exposure estimation. The subject lives in a suburban region where ambient concentrations of PM are relatively low, and works in a more urbanized region with higher concentration levels of PM. Therefore, the estimated time-weighted exposure is generally higher using GMLH and GPS data. Despite these limitations, results from this study highlight the tremendous potential of GMLH data to improve exposure estimation for retrospective epidemiological studies. Furthermore, understanding individual mobility is not only useful for epidemiological studies related to air pollution, but is also useful for many other academic research applications such as urban planning [132], transportation modeling [133], health intervention [110, 114], the spread of human diseases [134, 135] or computer viruses [136], and wireless network optimization [137].

CHAPTER FOUR: WORK ON USING CDR DATA FOR EXPOSURE ESTIMATION

4.1 Introduction

Exposure to air pollution is the second leading cause of non-communicable disease worldwide [138]. It is also associated with more than 4 million premature deaths annually [3, 139] and numerous other negative health consequences [1, 2, 6, 97-99, 140]. 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 [100-103]. 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 [79, 141], or even at further aggregated zones such as census tract [142] or ZIP code level [143]. 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 misclassification errors [15, 17, 43, 104, 144, 145], which could potentially bias subsequent statistical analyses [106, 107].

To address this issue, a variety of methods have been adopted, including utilizing travel surveys and diaries [26, 104], personal measurements [49, 60], accounting for multiple addresses (e.g., residential or work address) or full-day travel data [104, 145] during the temporal window of exposure [31, 79, 106, 146], tracking subjects using GPS-enabled surveys [72, 144], and employing a variety of modeling tools and techniques to account for mobility [17, 22]. Though prior results suggest exposure estimation errors due to the omission of mobility could differ among individuals with different mobility patterns [104, 145], 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 [147-149]. How the choices of method impact exposure estimates when human mobility is considered is still largely unknown.

In this study, by using a publicly available and an anonymized large cell phone location dataset [15], we investigated the influence of different levels of human mobility on exposure estimates to air pollution. We applied two different methods to investigate how the choices of method for developing pollution fields impact exposure estimates when mobility was considered. 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.

4.2 Materials and Methods

4.2.1 Data Description and Study Area

The cell phone location data applied here are Call Detail Record (CDR) data collected by mobile network operators. 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 [150-152]. 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 [153, 154], the study of human mobility [152, 155, 156], and urban and transportation planning [132, 157, 158].

In this study, we obtained a publicly available CDR dataset for Shenzhen, China [151]. Shenzhen is a major city located in the Guangdong Province (Figure 1). It has an area of 1,991 km² 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.

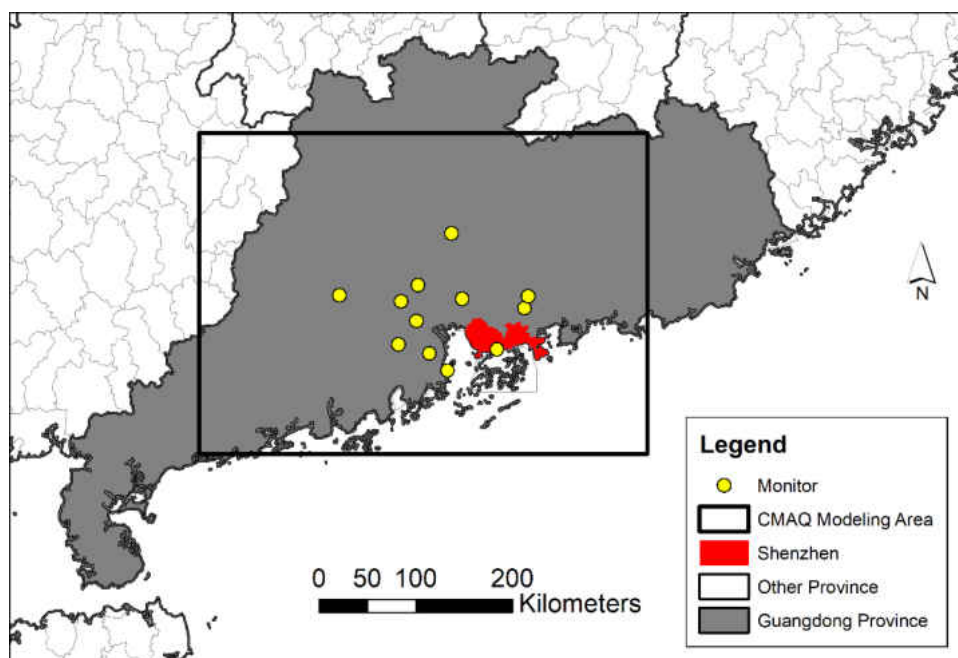


Figure 4-1. The study area of Shenzhen, China.

4.2.2 Exposure Estimation

Five pollutants were selected for this study, including carbon monoxide (CO), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), ground-level ozone (O₃), and particulate matter with the aerodynamic diameter less than 2.5 μm (PM_{2.5}). All 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.

Similar to our previous study [15], 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. 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.

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 [159] for the corresponding day, and the other using the inverse distance weighting (IDW) method. Detailed information on CMAQ model configurations is available elsewhere [160]. 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 [147]. 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.

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 [106], 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, 161-164]. 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 [106, 165]. According to the classical error theory, exposure estimated using the home-based method may be expressed as:

$$Z = X + E \quad (1)$$

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 [15]. Therefore, the following equation can be applied to calculate a bias factor [166]:

$$B = \frac{\sigma^2 + \varphi}{\sigma^2 + 2\varphi + \omega^2} \quad (2)$$

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%.

4.3 Results

4.3.1 Concentration Fields

The spatial concentration fields of the five chosen pollutants simulated by the CMAQ and IDW methods differ considerably (Figure 4-2), especially for O₃, NO₂, and PM_{2.5}, where the latter two pollutants are known to have substantial primary contributions from transportation sectors. The IDW method generally results in smoother fields that lack spatial variabilities compared with the CMAQ method.

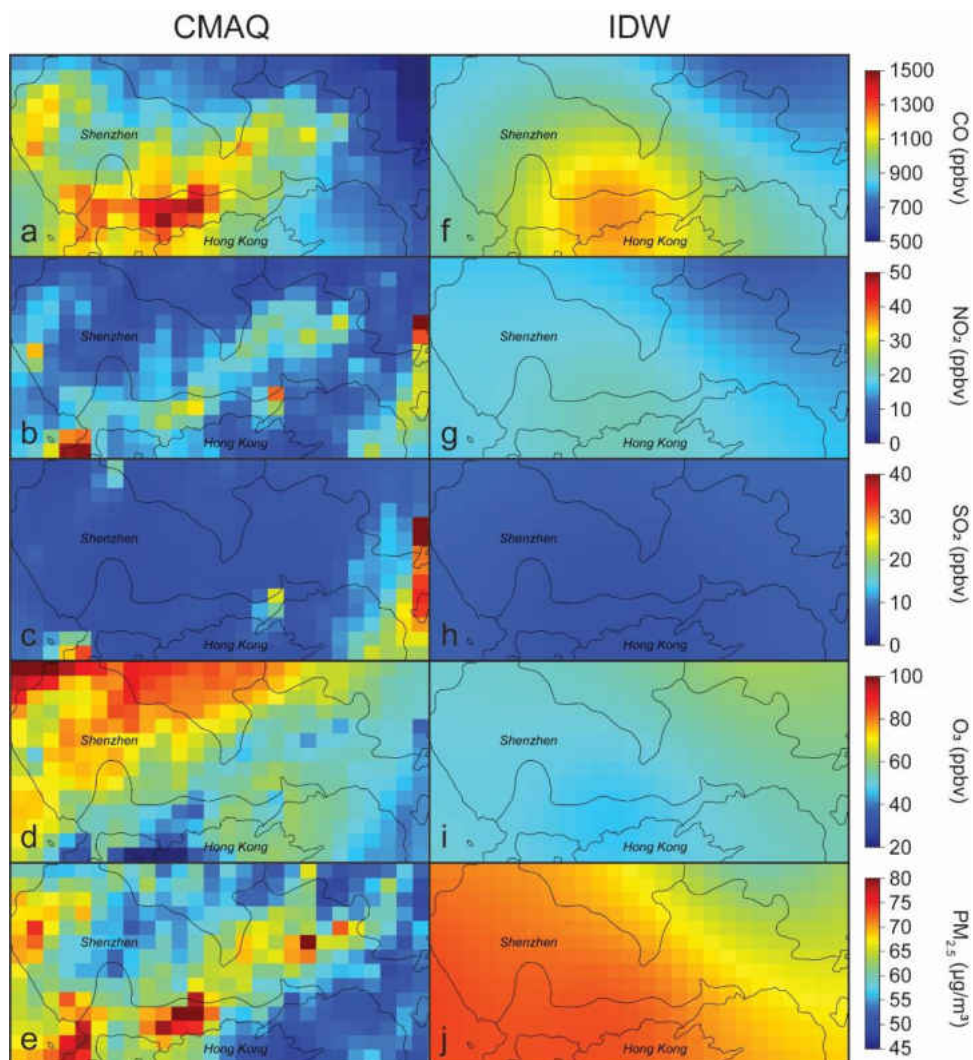


Figure 4-2. Spatial fields of concentrations of the five chosen pollutants as simulated by the CMAQ (a-e) and IDW (f-j) methods.

4.3.2 Overall Correlations Between HBE and CDRE

Mean CMAQ-based HBE and CDRE estimates for all subjects were highly correlated with each other (Figure 4-3). 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. Similar to our previous study [15], we 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 4-3), 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 $\text{PM}_{2.5}$, 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.

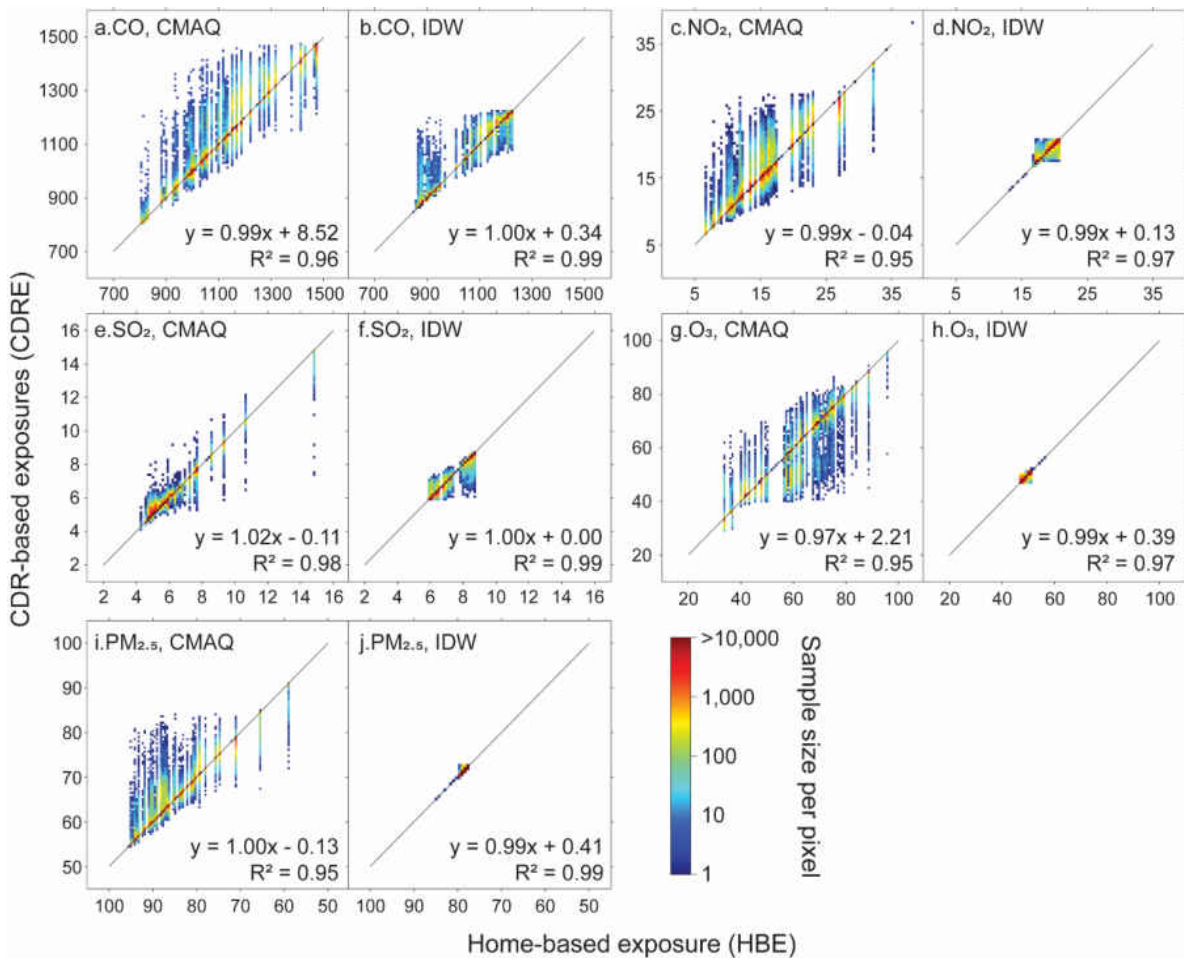


Figure 4-3. Linear correlations between HBE and CDRE estimates of the five chosen pollutants for all subjects based on CMAQ (a,c,e,g,i) and IDW (b,d,f,h,j) concentration fields. Pixels are color coded by sample size. The solid black line shown is the 1:1 line.

4.3.3 The Impact of Mobility on Exposure Estimates

We found that the correlations between HBE and CDRE estimates shrink with an increased degree of mobility (NO₂ presented in Table 4-1, other pollutants in Tables B-1 through B-4). With increased numbers of grid cells visited (representing greater mobility), correlations (R²) between HBE and CDRE showed a generally decreasing trend for all pollutants for both CMAQ and IDW fields, with generally increasing root-mean-squared-error (RMSE), mean normalized bias (MNB) and mean normalized error (MNE).

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} (Table B-4), 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. The only exception is SO₂ (Table B-2), 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 4-1. Comparison between HBE and CDRE estimate of NO₂ for all ten groups with different mobility.

		Group number									
		1	2	3	4	5	6	7	8	9	10
CMAQ	CDRE mean	16.1	16.6	16.7	16.8	16.7	16.3	15.9	15.9	15.6	15.6
	HBE mean	16.1	16.5	16.3	16.2	15.8	15.5	15.2	15.2	15.0	15.1
	^a RMSE	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	19.4	19.2	19.3	19.3	19.3	19.2	19.1	19.1	19.0	19.0
	HBE mean	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

^aRMSE: 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

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

^cMNE: mean normalized error. Calculated as $\frac{1}{N} \sum_{i=1}^N |\frac{HBE_i - CDRE_i}{CDRE_i}|$

^dR²: 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 4-1). 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).

With increased mobility, we generally observe more variability in the differences between HBE and CDRE estimates at the individual level (Figure 4-4), as indicated by the greater spread between the 25th and 75th percentile of exposure difference for groups with more grid cells visited. Across all pollutants and the two methods, the 50th percentile of exposure differences was consistently close to 0 for all mobility groups.

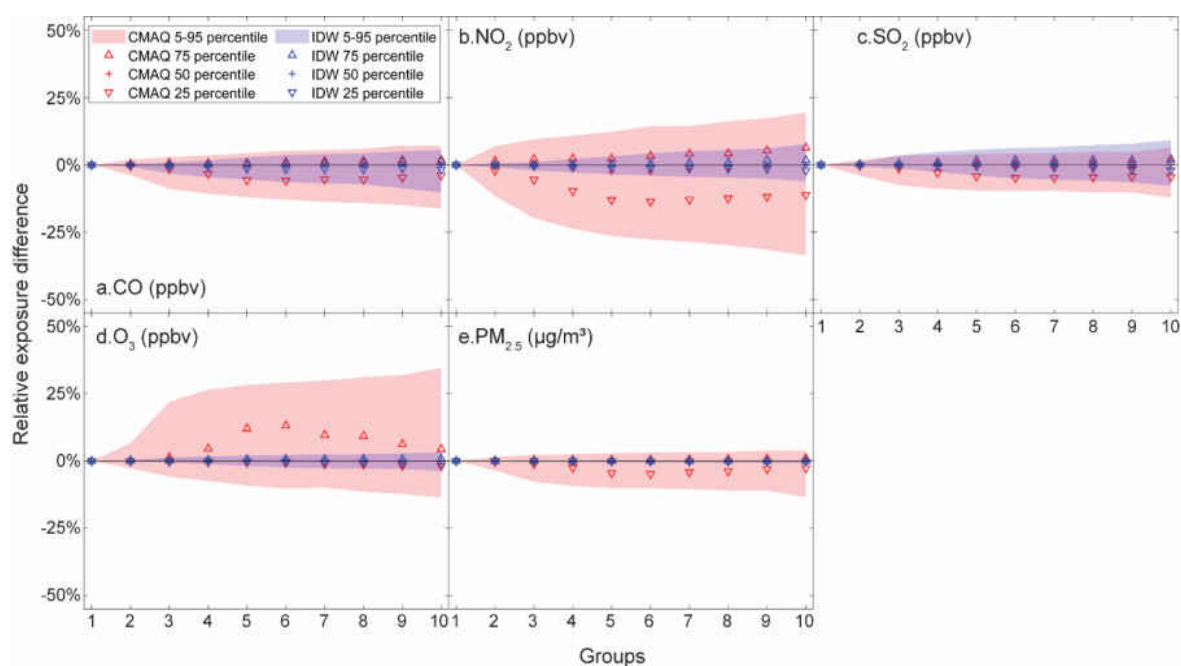


Figure 4-4. Distributions of differences in exposure estimates between HBE and CDRE for the five chosen pollutants for both CMAQ and IDW methods. Relative exposure differences were calculated as $(HBE-CDRE)/CDRE$.

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 NO₂, O₃ and PM_{2.5}. SO₂ again was the exception where exposure variability was similar between the two methods. Mobility had the greatest impact for NO₂ and O₃. When CMAQ concentration fields were applied, the observed differences were more negative (higher CDRE than HBE) for CO, NO₂ and PM_{2.5}, but were more positive (lower CDRE than HBE) for O₃. 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 4-5). The differences between CDRE and HBE were overall smaller toward mid-night and during early morning for all groups and were generally larger during daytime. The most significant differences between HBE and CDRE estimates occurred at different hours for different pollutants. When CMAQ concentration fields were applied, CO, NO₂ and PM_{2.5} exhibited the largest differences near the afternoon rush hour, though these differences dissipate quickly thereafter. For O₃, the largest differences occurred around mid-afternoon at 4 pm around when the highest ambient O₃ concentrations are expected. For SO₂, 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, NO₂ 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 O₃, 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 SO₂, and group 5 and 6 for other pollutants).

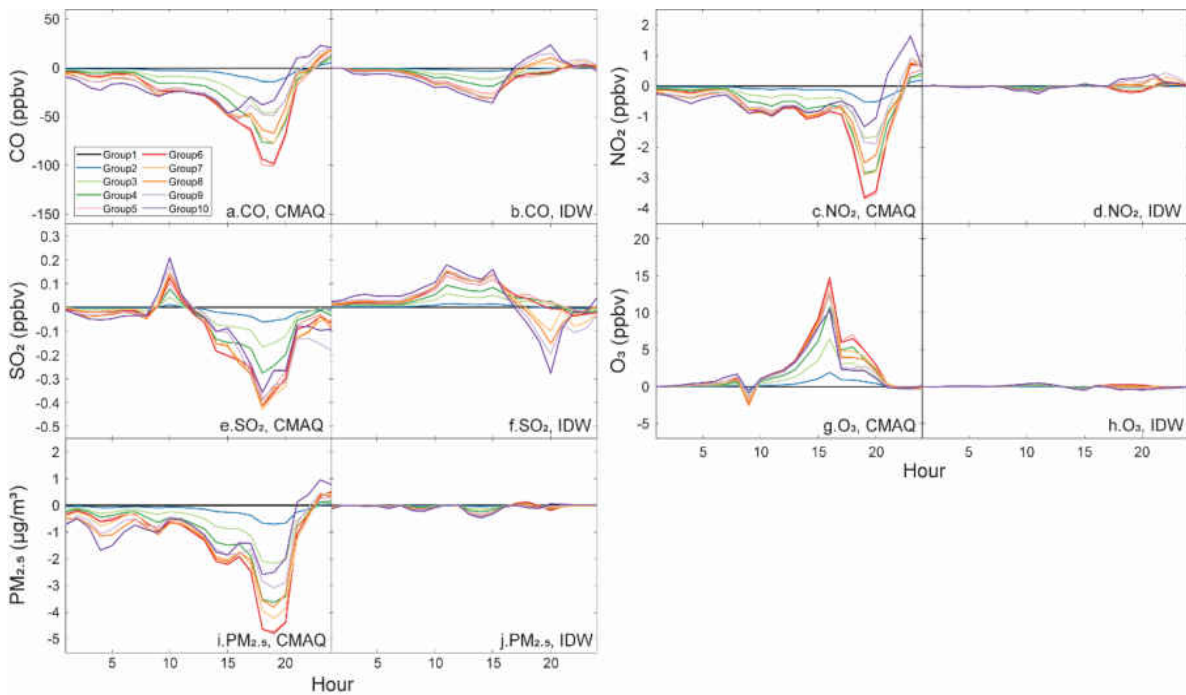


Figure 4-5. Temporal variations of exposure differences for all 10 mobility groups between HBE and CDRE when CMAQ and IDW concentration field were applied. Exposure differences were calculated as HBE-CDRE.

The temporal variations of exposure differences, however, were mostly not observed when IDW concentration fields were applied (Figure 4-5). We still observed generally larger differences during daytime (though smaller magnitude), but the consistent patterns of fluctuations as seen among CO, NO₂ and PM_{2.5} in Figure 4-5 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. The estimated p-value for each mobility group are presented in Figure 4-6. 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 SO₂, 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.

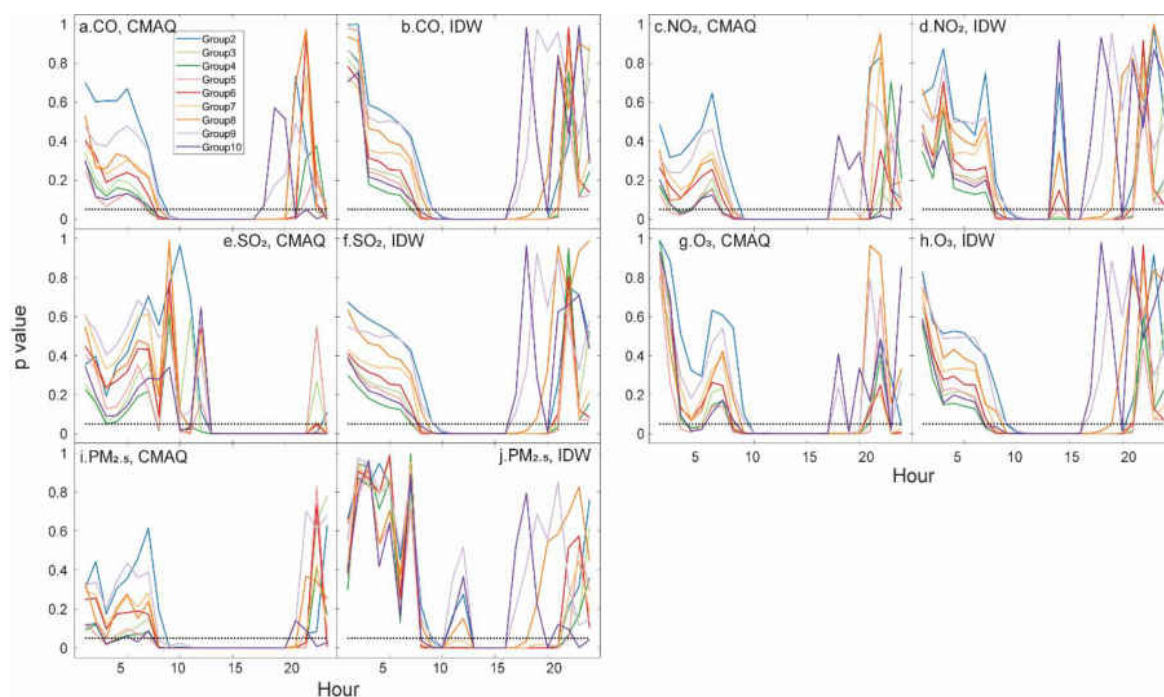


Figure 4-6. Temporal variations of p-values from the Wilcoxon rank sum tests performed for 9 mobility groups between HBE and CDRE when CMAQ and IDW concentration field were applied. Results for group 1 are not shown. Dotted black line is $p = 0.05$.

4.3.4 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 4-7 and 4-8, and results for other pollutants are presented in Figures B-1-B-8.

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 4-7), 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 4-7), 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 4-8). Similar findings can be observed for both CMAQ and IDW concentration fields for all other pollutants (Figures B-1-B-8). 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 B-1, B-5) and NO₂ (Figures B-2, B-6). This result was less consistent for SO₂ (Figures B-3, B-7) and somewhat reversed for O₃ (Figure B-4, B-8).

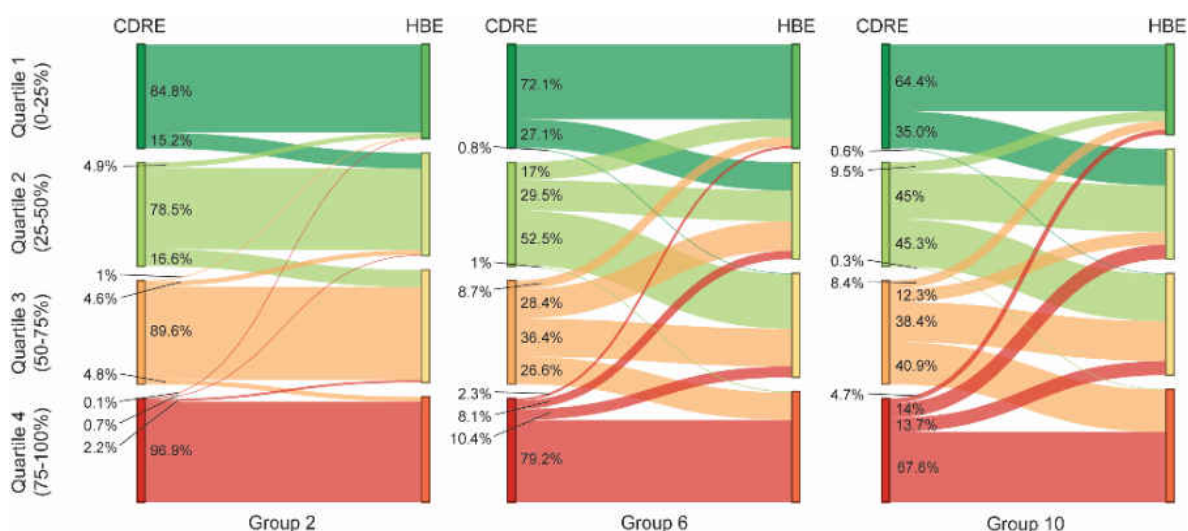


Figure 4-7. The directions of potential PM_{2.5} exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used. For simplification purposes only results for groups 2, 6 and 10 are presented.

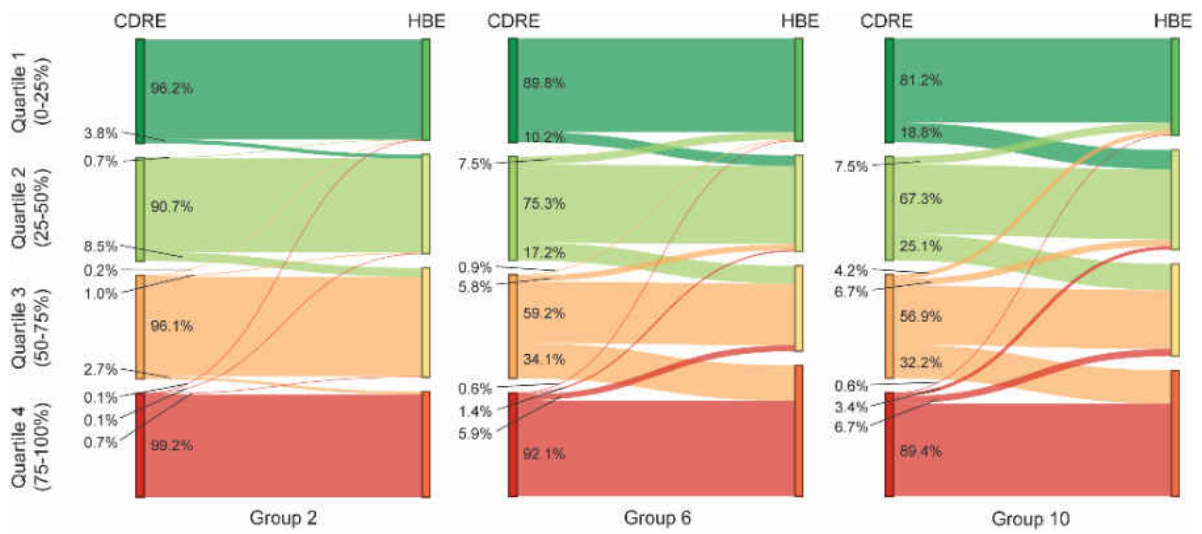


Figure 4-8. The directions of potential PM_{2.5} exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used. For simplification purposes only results for groups 2, 6 and 10 are presented.

The estimated bias factors for groups with different mobility levels are presented in Figure 9. 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 NO₂ and for group 10. This value suggests that the estimated relative risk for NO₂ will be underestimated by 33% when mobility was ignored during exposure estimation. Between CMAQ and IDW, the estimated bias factors are considerably different, especially for PM_{2.5}. For group 10, the bias is 0.70 when CMAQ fields are used, and 0.94 when IDW fields are used.

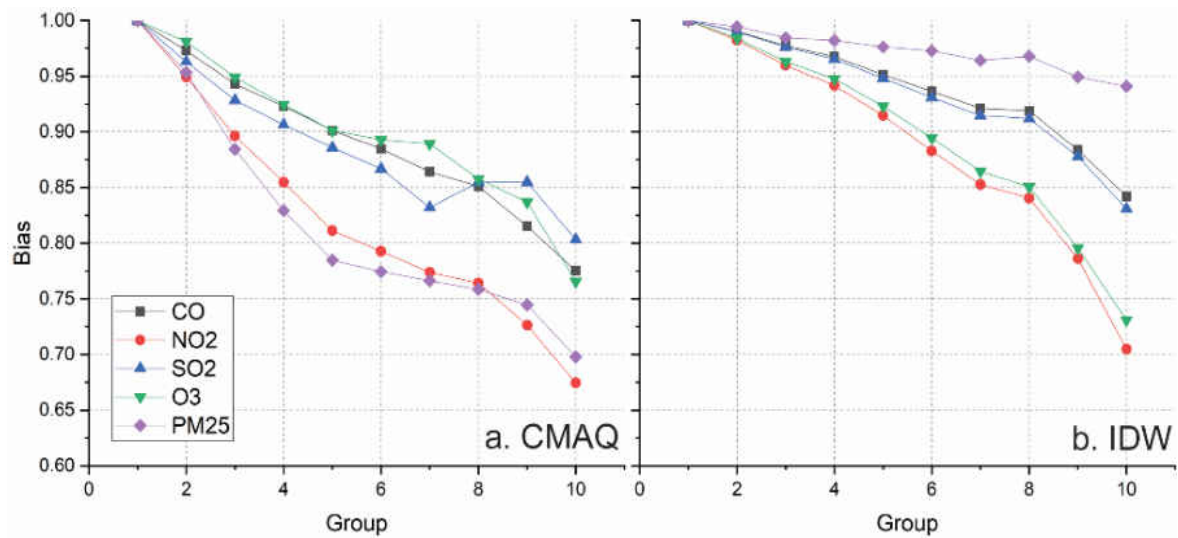


Figure 4-9. The impact of mobility on bias factors when CMAQ and IDW concentration fields were applied.

4.4 Discussion

4.4.1 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 4-2). Consequently, the estimated population average exposures (Table 4-1), the distributions of individual exposure estimates (Figures 4-3, 4-4), particularly among groups with different degrees of mobility (Figures 4-5, 4-6), and the impact of neglecting mobility on exposure estimates (Figures 4-7, 4-8), 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 [147]. 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. 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.

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

4.4.2 The Impact of Mobility on Exposure Estimation

At the population level, we did not find substantial differences between HBE and CDRE exposures, consistent with our previous study [15] and other studies [42, 44, 63, 165, 167]. The results suggested that the home-based method for exposure estimation is still informative when only average exposure estimates for a sufficiently large population are of interest [168].

We found that the impact of mobility on exposure estimates differed by time of day and by pollutants. 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, NO₂, 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 4-5). 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 4-5). The temporal differences in impacts of mobility on exposure have also been noted previously [167]. 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 4-2), suggesting that the home-based method captured less exposure variability among individuals with increased mobility [31]. Therefore, we expect larger exposure misclassification errors for subjects that are highly mobile, which is supported by our analysis on the potential exposure misclassifications based on HBE and CDRE (Figures 4-7, 4-8). It is also worth mentioning again that 71% of subjects (Figure 4-7) 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.

4.4.3 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 [150], 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 [169]), or other momentarily collected data such as Google Maps Location History data [170]. 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 workday within a typical week. Therefore, the spatiotemporal mobility patterns as represented in this CDR database represent the unique characteristics of the study region, such as the spatial distributions of population and land use types. However, the ambient concentrations of common air pollutants are expected to share similar spatial and temporal distribution patterns across the world for urban areas [171], thus we also do not expect our conclusions to change if the same study were to be performed in another city.

4.5 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, NO₂, SO₂, O₃, 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 considerable differences between population-averaged exposures as estimated with and without detailed mobility data. 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 misclassification 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 underestimate exposure to traffic-related pollutants (CO, NO₂ 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. Therefore, the IDW method is not suitable for air pollution exposure estimations when detailed mobility data are considered, if a dense measurement network is not available.

Our findings demonstrated the tremendous potentials of CDR data in air pollution exposure estimation for a large population. Despite the privacy concerns associated with using CDR data, our results have significant implications for future air pollution health studies in which subject mobility is important.

CHAPTER FIVE: WORK ON LOW-COST SENSORS

5.1 Introduction

Epidemiological studies in air pollution field reveal associations between adverse impacts of air pollutants on human health and human exposures to air pollutants [1-3]. Subject's exposures including mobility are calculated by combining two elements: location data of the places they visited and ambient concentrations distribution of air pollutants covering those places. Accurate ambient concentrations are essential for human exposure estimation. Traditionally, Federal Equivalent Method (FEM) and Federal Reference Method (FRM) are used to measure ambient concentrations [172], but they are huge, expensive and sparse [173]. In high-traffic areas, these kinds of monitors are even more rare. Also, ambient concentration distribution can vary considerably at a fine spatial level [172]. Therefore, these heavy instruments (FEM and FRM) are no longer able to meet our requirements. Nowadays, advanced technology is bringing low-cost sensors to researchers' attention. Low-cost sensors can be distributed on a large area and can highly improve the resolution level of ambient concentration. Low-cost sensors have huge potential in air pollution field and there are lots of studies using low-cost sensors to monitor air quality [174-176]. The low-cost and portable sensors can also be used to measure personal exposures to air pollutants by subjects carrying them for daily activities [57]. This approach can take both people's mobility and indoor/outdoor factor into consideration.

Obviously, sensors are not as accurate as FRMs and FEMs. Often, manufacturer's specifications of its low-cost sensors do not provide sufficient data for the desired utilization [174]. Also, it is difficult to combine outcomes from different studies or apply outcomes on another situation that differ from the original test [174]. Therefore, testing sensors for each trail becomes necessary.

Before deploying sensors or assigning sensors to subjects, we need to know the reliability and accuracy of the sensors that we choose. In this paper, we tested the reliability of two brands of commercially available sensors, one is ‘PurpleAir’, the other one is ‘Dylos’.

5.2 Material and Methods

We bought three PurpleAir (model: PurpleAir PA-II) sensors and three Dylos (model: Dylos DC 1700) sensors. PurpleAir uses two identical Plantower PMS5003 laser particle counter and adopts a complex algorithm to calculate mass concentration which is the final output. PurpleAir collects both concentration and particle counts and has two channels, while Dylos collects particle counts only (larger than $0.5\ \mu\text{m}$ (small bin) and larger than $2.5\ \mu\text{m}$ (large bin)) and has one channel. The number of $\text{PM}_{2.5}$ particles from Dylos were calculated by subtracting the number in small bin by the number in large bin. Figure 1. is the pictures of deployment of the sensors and the shelter.



Figure 5-1. Pictures of sensor deployment (left: internal layout of the shelter; right: external appearance of the shelter).

The South Coast Air Quality Management District (SCAQMD) publicized a sensor evaluation report of Dylos DC1700-PM. The report proved that Dylos DC1700-PM has high data recovery (~ 100%) and correlated moderately with two FEMs (beta attenuation monitor, GRIMM and BAM) on $PM_{2.5}$ in the field with $0.66 < R^2 < 0.68$ and $0.51 < R^2 < 0.55$, respectively [177]. The South Coast Air Quality Management District (SCAQMD) also publicized a sensor evaluation report of PurpleAir PA-II and proved that PurpleAir PA-II has a high data recovery (~ 95%). PurpleAir PA-II correlated moderately with two FEMs (beta attenuation monitor, GRIMM and BAM) on $PM_{2.5}$ in the field with $R^2 > 0.93$ and $R^2 > 0.86$, respectively [177]. Therefore, we didn't test the performance of the sensors with FRMs or FEMs. We only test the accuracy and reliability of the sensors.

We installed 3 PurpleAir sensors (P1, P2 and P3) and 3 Dylos sensors (D1, D2 and D3), at the same place: a monitoring site, Site 3002 (Latitude 27.965556, Longitude -82.230278), Sydney in Hillsborough, Florida from February 2018 to April 2018. The location is shown in figure.2. PurpleAir sensors were operated for 62 days and recorded one measurement per 20s (around 280,000 measurements in total). Dylos sensors were operated for 40 days and recorded one measurement per 60s (around 57,600 measurements in total).

For PurpleAir, we test the correlations between two channels in each sensor and correlations between three sensors. For Dylos, we test the correlations between three sensors and the correlations between three sensors by averaging measurements based on different aggregated time span, from 5 minutes to 24 hours.

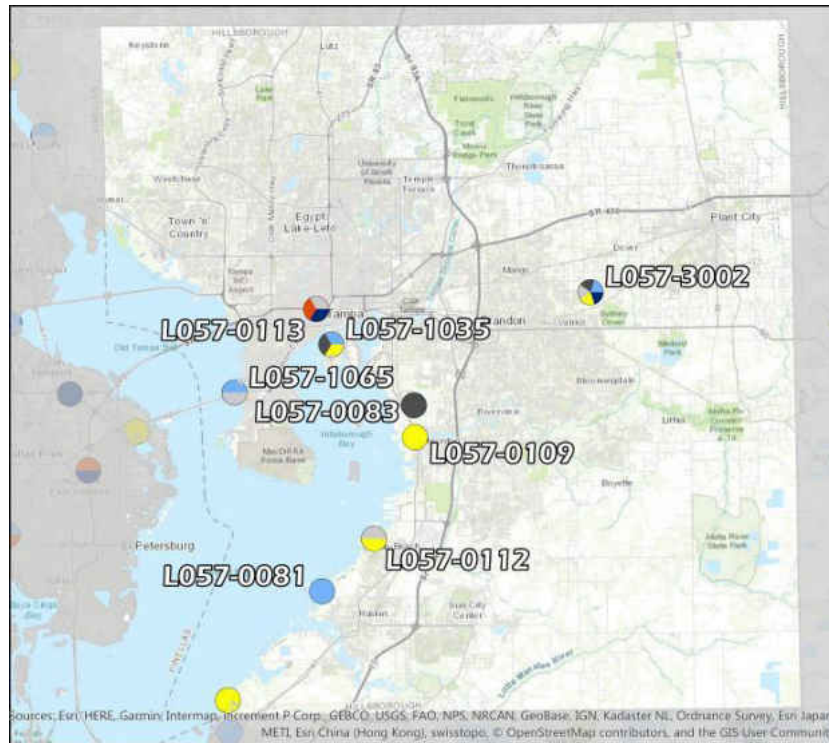


Figure 5-2. Location of Sydney site which is marked as L057-3002 at upright corner [178].

5.3 Results and Discussion

5.3.1 PurpleAir Sensor

Figure 1 represents the correlations between two channels in one PurpleAir sensor. The time span between two adjacent points is 20 seconds. Figure 1 shows that channel 1 and channel 2 correlate well with each other in every sensor (P1, P2 and P3). All R squares are above 0.985 and slopes are around 1 while intercepts are around 0. In these three sensors (P1, P2 and P3), each channel is proved to be reliable. The use of an average value of two channels would be more accurate than the use of only one value. Two channels also pose the advantage that if one channel is out of order, the other channel's values are still usable.

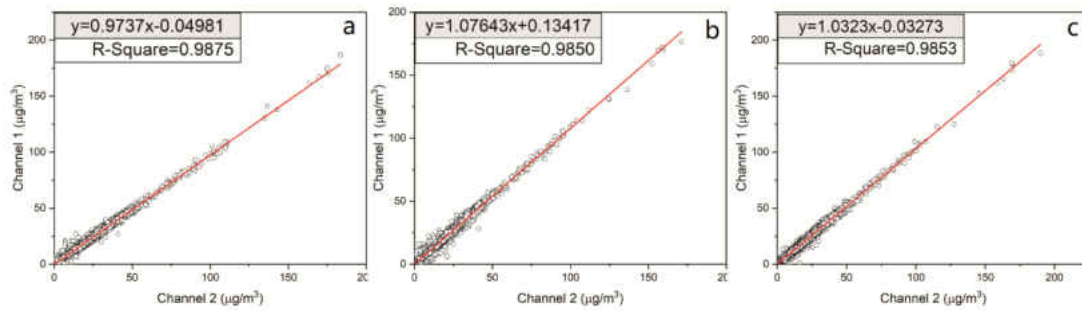


Figure 5-3. Correlations between two channels in each PurpleAir sensor (a: PurpleAir 1; b: PurpleAir 2; c: PurpleAir 3).

For PurpleAir, the average values of two channels of P1, P2 and P3 correlate well with each other by time span of 1 minute as shown in figure 2. All R squares are above 0.985 and slopes are around 1 while intercepts are around 0. Since we didn't calibrate them before use, the high R square values proved that P1, P2 and P3 are trustful. When a number of PurpleAir sensors will be deployed, it is possible that they don't need calibration before use, which highly reduce the workload of researchers. In the study of Wang et al., they found moderate correlation between individual sensors (Plantower PMS 7003) with $0.67 < R^2 < 0.92$ in outdoor situations [179]. In my study, all R squares are higher than theirs, above 0.98 in outdoor situations. The reason might be the difference of concentration values. In Kan Wang's study, average concentration is about $30 \mu\text{g}/\text{m}^3$, while in my study average concentration is about $10 \mu\text{g}/\text{m}^3$.

We installed all our sensors at the monitoring site of Sydney in Hillsborough, Florida in order to compare the measurements of our sensor with FEM instrument of Florida Department of Environmental Protection. However, due to some unknown reasons, the $\text{PM}_{2.5}$ values of FEM that we download from their website have very low correlations with either of our sensors. As our sensors are commercially trustful, it is hard for us to say whether the FEM

instrument output unreliable data or our sensors. Therefore, we didn't discuss the precision of our sensors compared to FEM instrument at the site.

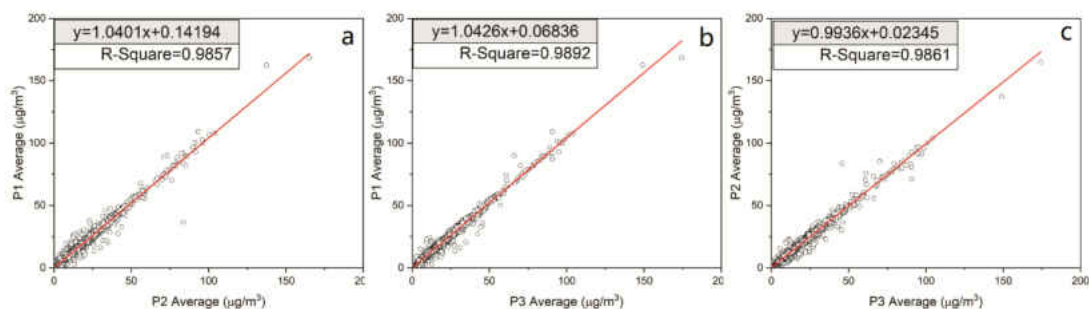


Figure 5-4. Correlations between each PurpleAir sensor.

In all, PurpleAir sensor is proved to be a reliable $PM_{2.5}$ detector. In the study of Gupta et al., PurpleAir sensors overestimate $PM_{2.5}$ concentrations consistently compared to FEM monitor in field situation. The relative difference is almost 0 when concentration is less than $10 \mu\text{g}/\text{m}^3$. The average relative bias is 35% when $PM_{2.5}$ values are between 15 and $50 \mu\text{g}/\text{m}^3$ [180]. In Kelly et al. study, Plantower sensors (PMS 1003/3003) also overestimate ambient concentrations and when ambient concentrations exceed $40 \mu\text{g}/\text{m}^3$, sensors' response become non-linear [173]. Kan Wang tested PMS 7003 and found that the sensor didn't need to be calibrated when relative humidity (RH) is less than 60% [179]. With all RH values less than 60% in our study, sensors may not need calibration. Even though without calibration the absolute accurate measurements cannot be obtained from PurpleAir sensor, the relative accurate measurements can be still useful, such as improving the satellite-derived $PM_{2.5}$ estimates [180], or inform the public [181]. PurpleAir sensor does not pose a screen on it, as shown in figure 1. The measured number can't be seen immediately, which limit its usage by communities or citizens who would like to use it for personal awareness of $PM_{2.5}$ levels.

5.3.2 Dylos Sensor

For Dylos, D1 doesn't correlate well with D2 and D3, but D2 and D3 correlate well as shown in figure 3. R square of D1 with D2 and D3 are 0.78 and 0.76, respectively. However, R square between D2 and D3 is 0.95. All the slopes are around 1. Intercept of D2 VS D3 is 5959.7, which is relatively small. Obviously, D1 is out of order, because it doesn't correlate well with neither D2 nor D3. But we can tell that D2 and D3 work well, because they correlate well with each other. D1 has many very small values, which leads D1 not correlate well with D2 and D3. But the reason keeps unknown. From our test, if some researchers want to use Dylos, it is highly possible that there would be some sensors that are not reliable. So that we do not recommend using Dylos sensors when attributing a number of sensors into field. Or we would highly recommend researchers to test every sensor before deploying them and eliminate the unreliable ones. For instance, in the study of Steinle et al., each volunteer carried a backpack which had Dylos DC1700 and a GPS device inside to measure personal exposures [57]. Since they used 17 Dylos sensors and each sensor was used independently, according to our research validation of each sensor is necessary.

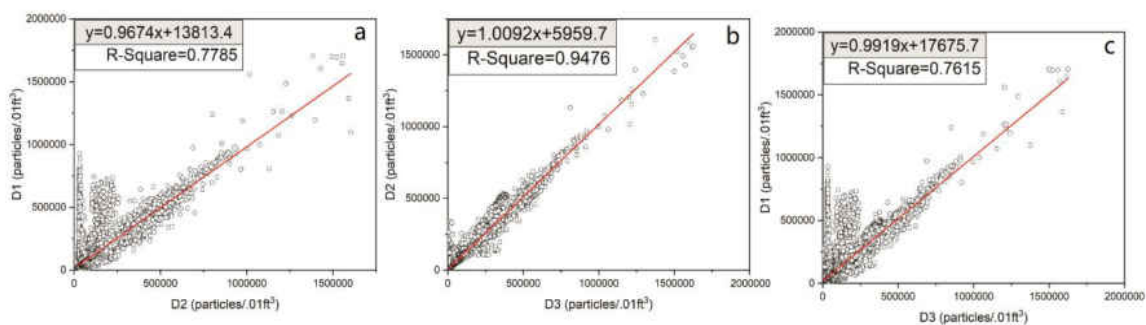


Figure 5-5. Correlations between each Dylos sensor at time span of 1 minute.

As the time span in Dylos is one minute and time setting of three different Dylos sensors may be not synchronous, time span of 1 minute might not be meaningful. Instead we aggregate time span into 5minute, 10minute, 15minute, 30minute, 60minute (1hour), 6hour, 12hour and 24hour (1day) and calculate the average value in each time span. And then we investigate the correlation between our three sensors. In figure 4, the results show that apart from figures aggregating by 24hour, others all have similar R square, ranging from 0.732 to 0.796. Whereas, R square of 24hour is 0.488. From 5min to 12h, ME and RMSE are decreasing (table 1). That means the bigger the time span is, the aggregated the data is. According to EPA, precision can be improved if data are averaged by aggregated time span [182]. But in our study, we didn't see this phenomenon.

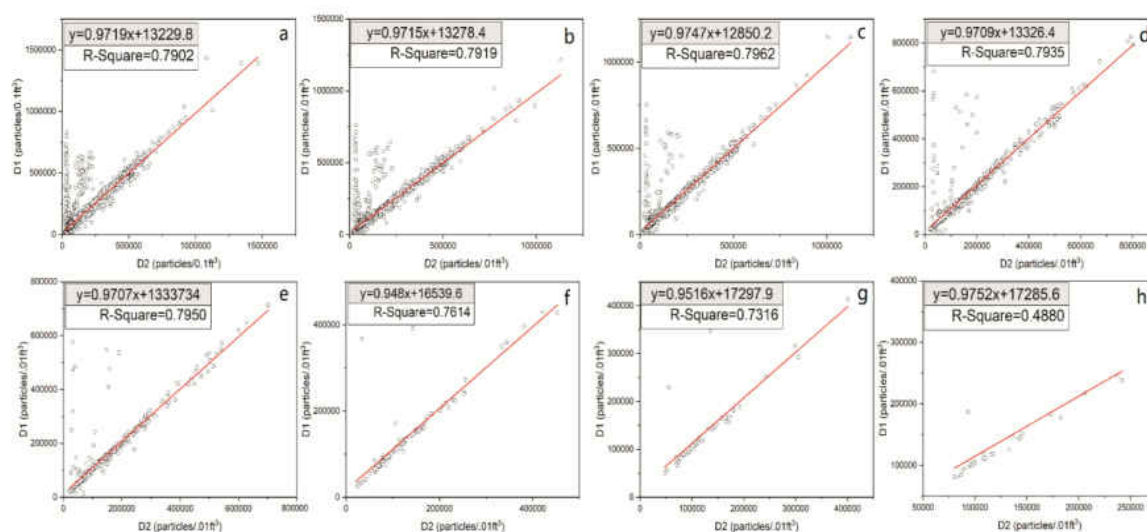


Figure 5-6. Correlations between D1 and D2 by aggregating time in different time spans (a:5minute; b:10minute; c:15minute; d:30minute; e:60minute; f:6hour; g:12hour; h:24hour).

In figure 5, we test the correlation between D2 and D3 in different aggregated time spans and the results show that apart from 24hour, all the R squares are similar, ranging from

0.929 to 0.952. Also from 5min to 12h, ME and RMSE are decreasing (see table 1). That means the bigger the time span is, the aggregated the data is. Also, for D2 and D2, who correlate well with each other, aggregated time span didn't increase their correlation level.

Higher resolution (such as 5 minute) shows similar R square with lower resolution (such as 12 hour), so we can use 5 min to aggregate time for Dylos. There are some occasions when high resolution is strongly needed for use. For instance, a sensor is installed near pollution source, such as the chimney of a power plant or beside traffics. Or if a sensor is banded on a car to test PM_{2.5} concentrations along the route, as the car moves very fast high temporal resolution is needed. Another application is to bond the sensor to an unmanned aerial vehicles (UAV) to detect pollutant concentrations [174]. Or if we want to calculate dynamic exposure of a certain group of human beings, such as taxi drivers or truck drivers [183], as they move quickly and cover a wide range of area, the high temporal resolution of concentration variation is called for.

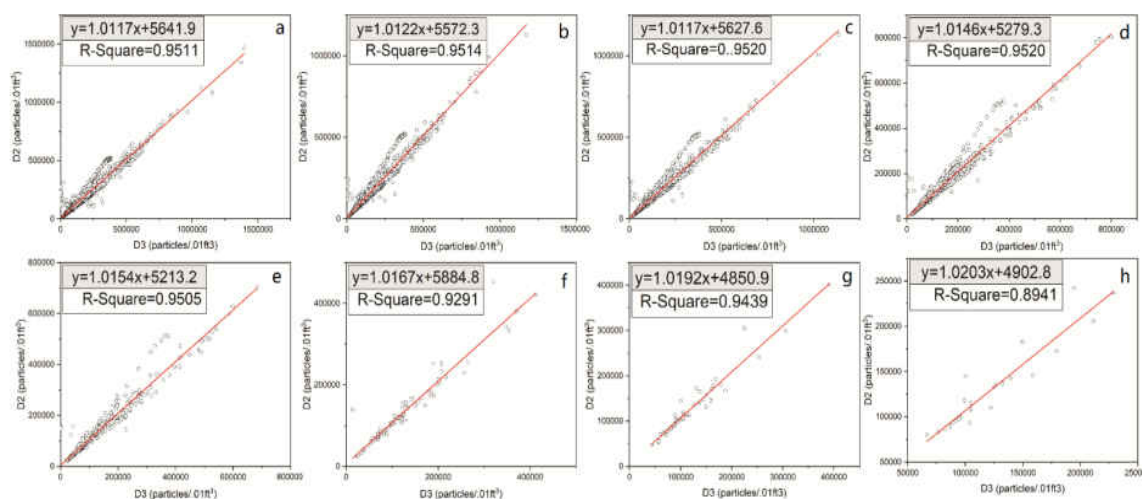


Figure 5-7. Correlations between D2 and D3 by aggregating time in different time span.

Table 5-1. Statistical data of correlations between D1, D2 and D3 by aggregating time in different time spans.

	D1 VS D2			D2 VS D3			D1 VS D3		
	R ²	ME ^a	RMSE ^b	R ²	ME ^a	RMSE ^b	R ²	ME ^a	RMSE ^b
5min	0.7902	14399	53388	0.9511	11653	24234	0.7962	21032	57473
10min	0.7919	13961	52632	0.9514	11525	23966	0.7700	20849	56819
15min	0.7962	13668	51930	0.9520	11437	23744	0.7737	20662	56236
30min	0.7935	13337	51164	0.9520	11297	23333	0.7707	20379	55456
1h	0.7950	13052	49466	0.9505	11261	23048	0.7718	20184	53806
6h	0.7614	12513	43818	0.9291	11302	22801	0.7286	19838	48921
12h	0.7316	13444	40148	0.9439	10664	17476	0.7203	19978	43687
24h	0.4880	15676	44796	0.8941	10815	15945	0.4971	22758	48048

^aME: mean error, calculated as $\frac{1}{N} \sum_{i=1}^N |x_i - y_i|$, where N is the number of one dataset (D1, D2 or D3), x_i is one value in one dataset (such as D1), y_i is the corresponding value in the other dataset (such as D2);

^bRMSE: Root Mean Square Error, calculated as $\sqrt{\frac{\sum_{i=1}^N (x_i - y_i)^2}{N}}$, where N is the number of one dataset (D1, D2 or D3), x_i is one value in one dataset (such as D1), y_i is the corresponding value in the other dataset (such as D2);

Even though we didn't validate Dylos sensors with FEM instruments, there are a number of studies proving that Dylos sensors correlate well with FEM instruments [184-186]. It is reasonable to employ our sensors without calibration. After all, in many research or project situations researchers are not able to calibrate their sensors. What's more, each Dylos DC1700 device weighs 1.2lb [187]. The light weight allows it to be carried easily. Such as in the study of Steinle et al., volunteers carried a backpack with Dylos sensor in daily activities [57]. Dylos DC1700 can also be used to quantify second-hand smoke (SHS) levels at home and Semple et al. proved this approach was valid [188]. Even though Dylos sensor has a screen, the data cannot be viewed in real time, which would not influence daily behaviors of subjects who are supposed to carry them.

5.4 Limitations and Future Work

One of the limitations of my study is lack of calibration with FEM, which is our original plan. But due to unknown reasons, the data derived from FEM are not valid. If we could compare our sensors' data with FEM's data, we could better understand our sensors' accuracy and could obtain an equation to calibrate sensor's measurements against FEM's readings. After calibration, sensors can be used in other places to measure PM_{2.5} concentrations in outdoor environments. Sensors could also be given to subjects in epidemiological studies to continuously measure personal PM_{2.5} concentrations and then subjects' exposure could be calculated easily. But subjects' microenvironments are often indoor environments. The accuracy of sensors would be different from outdoor environments [179]. It is better to find out the factor that made D1 not work, such as RH, because D1 is not supposed to perform poorly. After calibration, we could also fasten a sensor on vehicles to determine the air pollutant concentrations along the road.

CHAPTER SIX: SUMMARY AND CONCLUSION

My study showcased two approaches to obtain people's mobile data. One is using Google Maps application data, while the other one is using CDR data. My study also explored sensors' relative accuracy and reliability, which laid the foundation for the future usage of low-cost sensors for exposure estimation. In my first approach, I compared GMLH data with GPS data in four aspects: 1) spatial movement of the subject; 2) the time the subject spent at different microenvironments; 3) the time the subject spent driving during the one-week time period; 4) subject's time-weighted exposures to ambient particulate matter using AOD measurements. In my second approach, I used CDR data to investigate the impact of individual mobility on exposures for five chosen pollutants (CO, NO₂, SO₂, O₃, and PM_{2.5}). I divided our sample population into ten groups according to their degrees of mobility and compared exposures of each groups. I also compared two methods developing concentration fields for exposure estimation: CMAQ and IDW.

In my first approach, GMLH data was proved to capture well the subject's spatial mobility during the study period with resolution of 200m * 200m or larger. Compared with GPS logger data, GMLH data also successfully captured the time the subject spent at different microenvironments and the time the subject spent on driving. Also, with GMLH data we were able to accurately estimate the subject's time-weighted exposure to ambient PM pollution.

In my second approach, I found no considerable differences between exposures estimated with and without detailed mobility data at population level, which indicated the traditional home-based exposure estimation method is still meaningful when population level is considered. We also observed that at individual level difference between HBE and CDRE increased with mobility increased. It was also found that HBE would likely under-estimate

exposure to traffic-related pollutants (CO, NO₂ and PM_{2.5}) during afternoon rush-hour, but over-estimate exposures to ozone during mid-afternoon when ambient ozone concentrations were expected to be the highest. IDW method was found to fail to capture detailed concentration variations compared with CMAQ fields. Therefore, the IDW method is not suitable for air pollution exposure estimations when detailed mobility data are considered. My study demonstrated the tremendous potentials of CDR data in air pollution exposure estimation for a large population and significant implications for future air pollution health studies in which subject mobility is important.

In my sensor study, I found that two channels of PurpleAir correlated well with each other in each sensor. And the three PurpleAir correlate well with each other. Among three Dylos sensors, D1 is proved to be out of order, but D2 and D3 correlate well with each other. After comparison, 5

minute time span works for Dylos sensors. According to our results, we highly recommend test each Dylos sensor before use, while no need for PurpleAir sensors.

APPENDIX A: ADDITIONAL CONTENTS FOR CHAPTER THREE

A.1 Microenvironment

In this study, microenvironment is defined as a fixed activity location that is not part of a travel route where the subject visited during the week. A total of 10 microenvironments were considered in this study, including the subject's home and work location, two grocery stores, three clinics, one postal office, one tourist destination, and an air quality monitoring station.

Figure A-1 provides satellite images with their rectangles for three microenvironments. The areas of the 10 rectangles ranged from 0.01 km² (a grocery store) to 3.15 km² (the subject's work location).

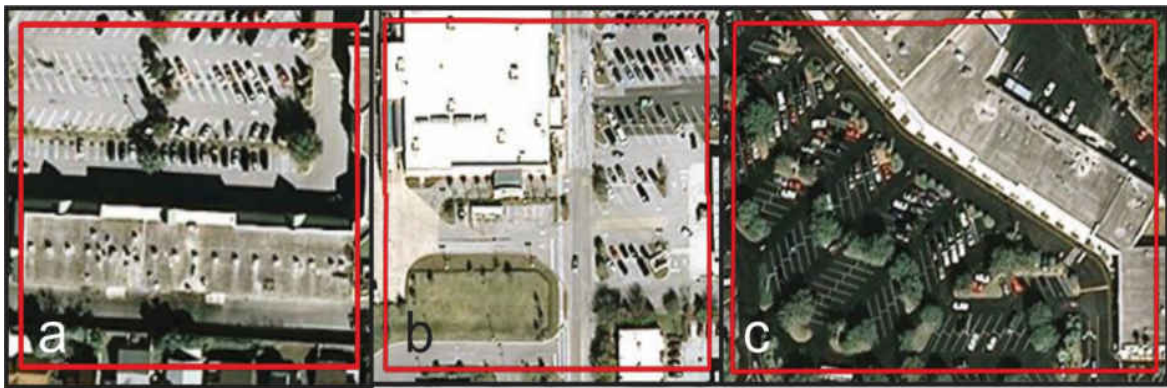


Figure A-1. Satellite images of three microenvironments and corresponding rectangles (a is a post office, b and c are grocery stores).

A.2 Location Data

Over the course of 1 week, the subject's Google Maps application collected 2,224 location records. The sampling interval ranged from less than 1 second to 69 minutes, with an average interval of 4.5 minutes. The GPS logger recorded 32,314 location records (with "drift" data removed). The collected location data from Google Maps and the GPS logger near a roadway intersection inside the study domain are presented in Figure A-2.

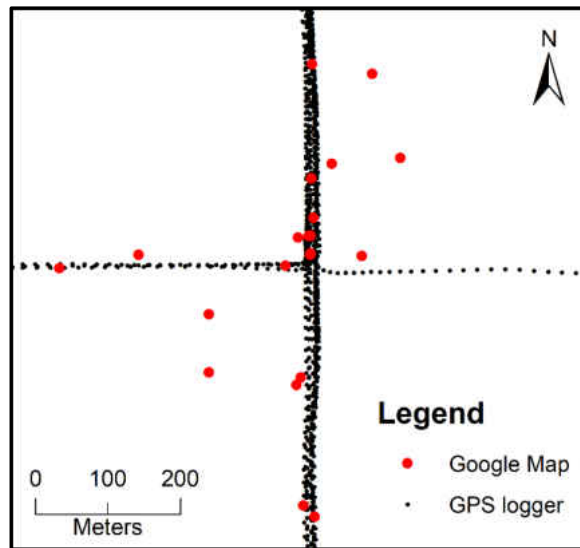


Figure A-2. The collected location data from Google Maps and the GPS logger near a roadway intersection inside the study domain.

We visually and qualitatively examined the spatial accuracy of both Google Maps and GPS logger data by overlaying the location data with reference roadway network data published by the Census Bureau. The GPS logger data were found to perform well in outdoor environments. The recorded outdoor location coordinates were accurate and had spatial errors generally less than 10 meters [189]. In indoor environments, the recorder did not perform as well due to poor GPS signal reception, and consequently the recorded location coordinates were less accurate, with spatial errors generally less than 100 meters. Since the recorder was configured to record every 10 seconds, the recorded location data contained many details. The spatial accuracy of Android location data, on the other hand, was less accurate, with spatial errors frequently exceeding 100 meters [190, 191]. Two data points were found to have spatial errors of over 2 km and 3 km, respectively, and both of the data points were collected while the subject was driving on a highway.

A.3 Evaluation of Google Maps Data

Using location data collected from the GPS logger as ground truth, we found that the Google Maps location data captured well the spatial mobility of the subject within the one-week study period, particularly in locations where the subject spent the vast majority of his time (the 10 microenvironments). However, Google Maps location data did not perform as well in grid cells in which the subject spent 10 minutes or less (as determined by GPS logger data), as indicated by the deteriorated R^2 values (Figure A-3). The R^2 value between the time the subject spent at different grid cells every day as estimated using GMLH and GPS data (Figure A-3 a-d) decreases from near perfect (as shown in the main manuscript) to 0.27 (1 km) and 0.16 (500 m) for grid cells where the subject spend 10 minutes or less, with minimum correlation between GMLH and GPS data at 200 m and 100 m resolution. These results suggest that GMLH data are less useful for characterizing individual mobility at fine scales for locations the subject only spent small amount of time.

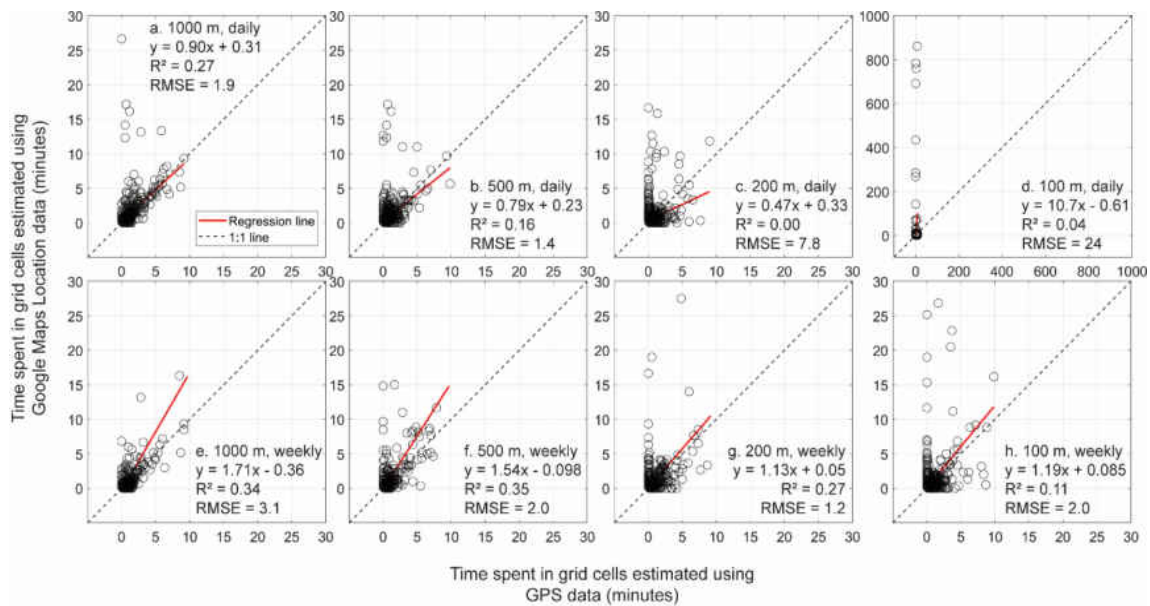


Figure A-3. Comparison of the estimated daily total (a-d) and weekly total (e-h) time the subject spent in each grid cell based on GPS versus Google Maps location data, for grid cells

in which the subject spent 10 minutes or less during the total time represented (as determined by GPS logger data). Resolutions represented are 1 km (a,e), 500 m (b,f), 200 m (c,g) and 100 m (d,h).

APPENDIX B: ADDITIONAL CONTENTS FOR CHAPTER FOUR

B.1 Correlations Between HBE and CDRE

Tables B-1 through B-4 provide correlations between HBE and CDRE estimates for CO, SO₂, O₃ and PM_{2.5}, exposure as estimated for 10 groups of subjects with increased degrees of mobility, when both CMAQ and IDW concentration fields were applied. Generally, the R² between HBE and CDRE showed a near monotonic decreasing trend for all pollutants for both CMAQ and IDW fields, with generally increasing mean absolute differences and standard deviations.

Table B-1. Comparison between HBE and CDRE estimates of CO for all ten groups with different mobility.

		Group number									
		1	2	3	4	5	6	7	8	9	10
CMAQ	CDRE mean	1149	1160	1165	1171	1167	1155	1141	1140	1132	1131
	HBE mean	1149	1156	1153	1150	1139	1127	1117	1115	1111	1112
	^a RMSE	0.0	29.8	49.9	61.4	72.5	75.9	76.6	81.3	83.3	89.6
	^b MNB	0.0%	-0.3%	-1.0%	-1.7%	-2.3%	-2.3%	-2.0%	-2.0%	-1.7%	-1.6%
	^c MNE	0.0%	1.1%	2.2%	3.1%	3.9%	4.2%	4.2%	4.5%	4.5%	4.8%
	^d R ²	1.00	0.97	0.92	0.88	0.82	0.80	0.78	0.76	0.74	0.73
IDW	CDRE mean	1079	1063	1067	1076	1076	1069	1061	1058	1051	1048
	HBE mean	1079	1062	1064	1070	1068	1060	1054	1051	1046	1045
	^a RMSE	0.0	13.7	23.6	28.9	33.9	37.6	39.8	39.6	43.6	46.1
	^b MNB	0.0%	-0.1%	-0.3%	-0.5%	-0.7%	-0.8%	-0.7%	-0.7%	-0.6%	-0.4%
	^c MNE	0.0%	0.4%	0.8%	1.2%	1.7%	1.9%	2.1%	2.2%	2.4%	2.5%
	^d R ²	1.00	0.99	0.97	0.96	0.94	0.92	0.91	0.91	0.89	0.88

^aRMSE: 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;

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

^cMNE: mean normalized error. Calculated as $\frac{1}{N} \sum_{i=1}^N |\frac{HBE_i - CDRE_i}{CDRE_i}|$

^dR²: coefficient of determination between HBE and CDRE estimates in the corresponding group.

Table B-2. Comparison between HBE and CDRE estimates of SO₂ for all ten groups with different mobility.

		Group number									
		1	2	3	4	5	6	7	8	9	10
CMAQ	CDRE mean	5.47	5.64	5.67	5.64	5.65	5.71	5.67	5.74	5.76	5.82
	HBE mean	5.47	5.63	5.63	5.57	5.56	5.61	5.58	5.64	5.66	5.73
	^a RMSE	0.00	0.15	0.22	0.25	0.28	0.31	0.34	0.33	0.33	0.37
	^b MNB	0.00	-	-	-	-	-	-	-	-	-
	%		0.29%	0.77%	1.33%	1.83%	1.97%	1.93%	1.98%	1.83%	1.93%
	^c MNE	0.00	1.06%	1.98%	2.65%	3.23%	3.54%	3.63%	3.69%	3.66%	4.10%
	%										
^d R ²	1.00	0.98	0.97	0.96	0.95	0.95	0.93	0.94	0.94	0.94	
IDW	CDRE mean	6.96	7.10	7.07	7.01	7.01	7.07	7.13	7.15	7.21	7.23
	HBE mean	6.96	7.10	7.09	7.04	7.05	7.11	7.16	7.18	7.23	7.23
	^a RMSE	0.00	0.10	0.17	0.20	0.24	0.27	0.29	0.28	0.32	0.34
	^b MNB	0.00	0.06%	0.27%	0.44%	0.61%	0.61%	0.43%	0.41%	0.20%	-
	%										0.02%
	^c MNE	0.00	0.47%	1.02%	1.48%	1.97%	2.28%	2.52%	2.63%	2.92%	3.10%
	%										
^d R ²	1.00	0.99	0.98	0.96	0.95	0.94	0.92	0.93	0.91	0.90	

^aRMSE: 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

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

^cMNE: mean normalized error. Calculated as $\frac{1}{N} \sum_{i=1}^N |\frac{HBE_i - CDRE_i}{CDRE_i}|$

^dR²: coefficient of determination between HBE and CDRE estimates in the corresponding group.

Table B-3. Comparison between HBE and CDRE estimates of O₃ for all ten groups with different mobility.

		Group number									
		1	2	3	4	5	6	7	8	9	10
CMAQ	CDRE mean	59.8	59.6	58.9	58.2	58.3	59.1	60.2	60.2	60.9	61.0
	HBE mean	59.8	59.9	60.0	60.0	60.8	61.7	62.5	62.3	62.7	62.6
	^a RMSE	0.00	2.51	4.40	5.60	6.61	6.93	6.80	7.19	7.14	7.67
	^b MNB	0.00%	0.64%	2.20%	3.78%	5.18%	5.31%	4.65%	4.55%	3.93%	3.50%
	^c MNE	0.00%	1.71%	4.15%	6.29%	8.16%	8.62%	8.15%	8.63%	8.30%	8.45%
	^d R ²	1.00	0.96	0.89	0.83	0.76	0.73	0.72	0.70	0.69	0.66
IDW	CDRE mean	48.3	48.5	48.5	48.4	48.4	48.5	48.7	48.7	48.8	48.9
	HBE mean	48.3	48.5	48.5	48.4	48.5	48.6	48.7	48.7	48.8	48.8
	^a RMSE	0.00	0.27	0.44	0.54	0.63	0.72	0.78	0.80	0.85	0.92
	^b MNB	0.00%	0.01%	0.07%	0.12%	0.17%	0.14%	0.06%	0.04%	-0.06%	-0.17%
	^c MNE	0.00%	0.19%	0.40%	0.58%	0.76%	0.90%	1.00%	1.06%	1.17%	1.23%
	^d R ²	1.00	0.99	0.98	0.96	0.95	0.94	0.92	0.93	0.91	0.90

^d R ²	1.00	0.98	0.95	0.92	0.88	0.85	0.81	0.81	0.77	0.75
-----------------------------	------	------	------	------	------	------	------	------	------	------

^aRMSE: 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

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

^cMNE: mean normalized error. Calculated as $\frac{1}{N}\sum_{i=1}^N|\frac{HBE_i - CDRE_i}{CDRE_i}|$

^dR²: coefficient of determination between HBE and CDRE estimates in the corresponding group.

Table B-4. Comparison between HBE and CDRE estimates of PM_{2.5} for all ten groups with different mobility.

		Group number									
		1	2	3	4	5	6	7	8	9	10
CMAQ	CDRE mean	64.9	66.0	66.2	66.2	66.0	65.6	65.1	65.3	65.1	65.3
	HBE mean	64.9	65.9	65.6	65.2	64.7	64.3	63.9	64.0	64.1	64.2
	^a RMSE	0.00	1.44	2.35	2.89	3.29	3.35	3.36	3.59	3.53	3.80
	^b MNB	0.00	-	-	-	-	-	-	-	-	-
		%	0.28%	0.90%	1.52%	2.03%	2.06%	1.92%	1.94%	1.62%	1.54%
	^c MNE	0.00	0.89%	1.79%	2.49%	3.08%	3.18%	3.10%	3.20%	3.08%	3.23%
	%										
	^d R ²	1.00	0.96	0.89	0.83	0.78	0.75	0.73	0.69	0.70	0.68
IDW	CDRE mean	72.1	72.1	72.1	72.2	72.2	72.1	72.1	72.1	72.1	72.1
	HBE mean	72.1	72.1	72.1	72.1	72.1	72.0	72.0	72.0	72.0	72.0
	^a RMSE	0.00	0.06	0.11	0.13	0.16	0.18	0.20	0.20	0.22	0.21
	^b MNB	0.00	-	-	-	-	-	-	-	-	-
		%	0.01%	0.03%	0.06%	0.08%	0.09%	0.10%	0.10%	0.11%	0.10%
	^c MNE	0.00	0.03%	0.06%	0.08%	0.11%	0.13%	0.14%	0.15%	0.15%	0.15%
	%										
	^d R ²	1.00	0.99	0.97	0.96	0.95	0.93	0.91	0.92	0.90	0.92

^aRMSE: 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

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

^cMNE: mean normalized error. Calculated as $\frac{1}{N}\sum_{i=1}^N|\frac{HBE_i - CDRE_i}{CDRE_i}|$

^dR²: coefficient of determination between HBE and CDRE estimates in the corresponding group.

B.2 The Impact of Mobility on Exposure Classifications

Figures B-1 through B-8 provide differences in exposure classifications (as quartiles) when subject mobility was omitted in the exposure estimation (i.e., when the home-based

method was used) for the 5 chosen pollutants and when both CMAQ and IDW fields were used.

For simplification purposes only results for groups 2, 6 and 10 are shown. Here quartile 1

includes the lowest exposures, and quartile 4 includes the highest exposures.



Figure B-1. The directions of potential CO exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used.

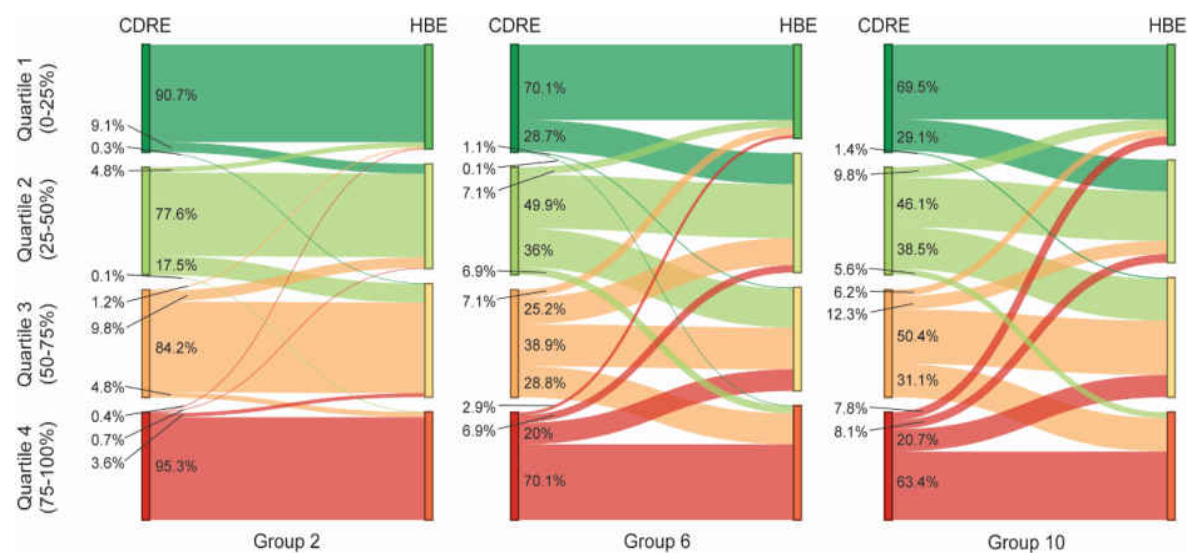


Figure B-2. The directions of potential NO₂ exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used.



Figure B-3. The directions of potential SO₂ exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used.

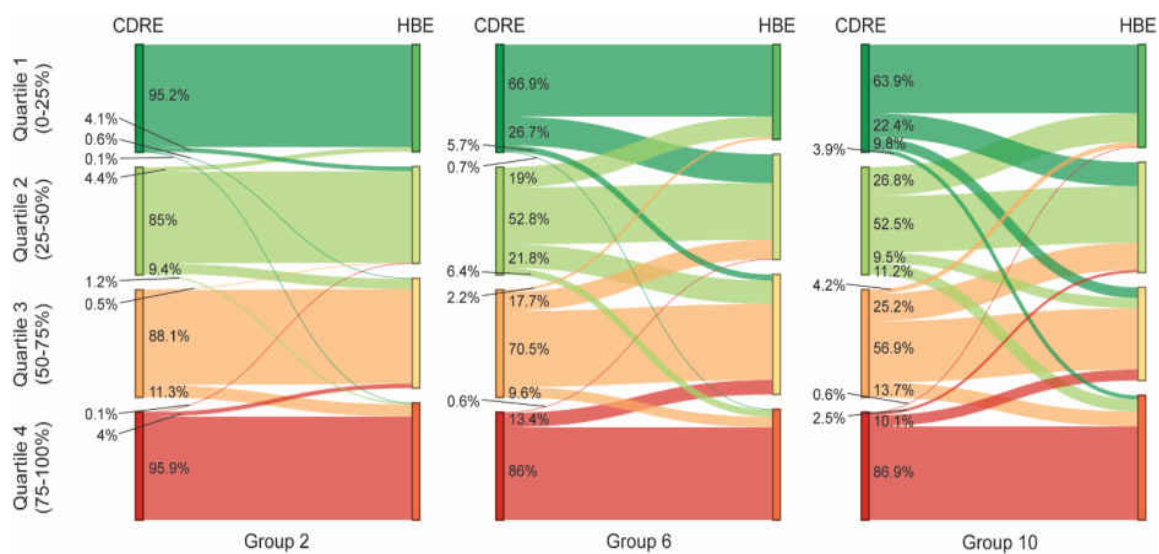


Figure B-4. The directions of potential O₃ exposure misclassifications when the home-based exposure estimation method was used and when CMAQ fields were used.



Figure B-5. The directions of potential CO exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used.

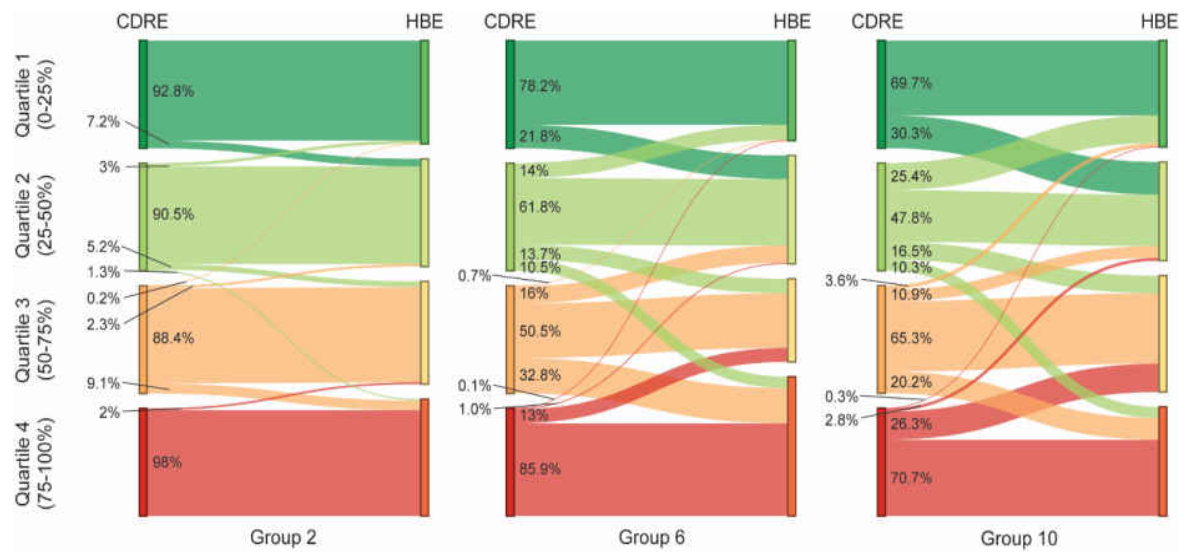


Figure B-6. The directions of potential NO₂ exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used.

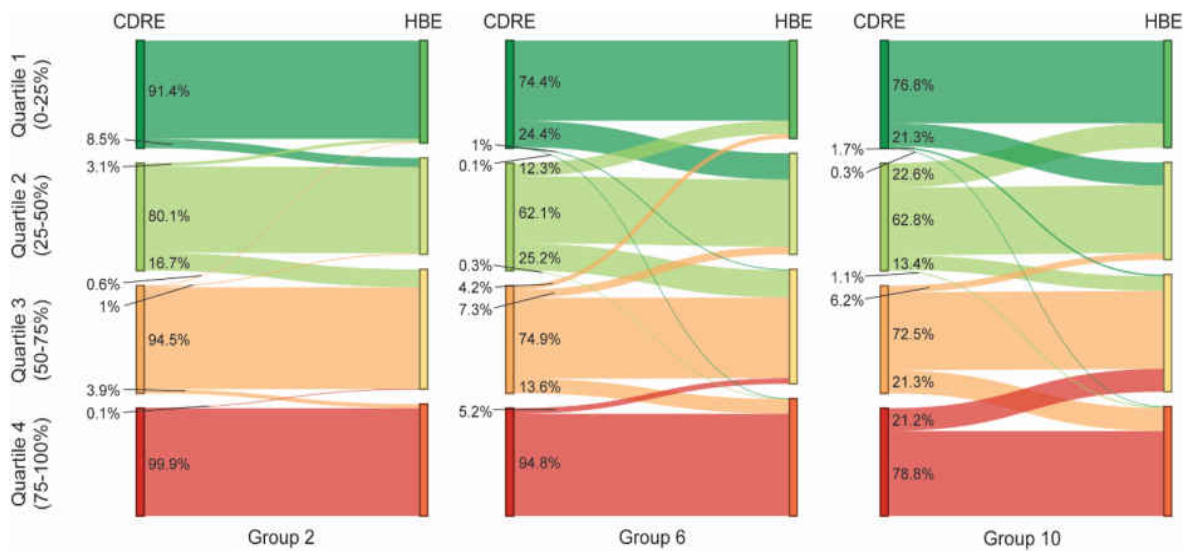


Figure B-7. The directions of potential SO₂ exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used.

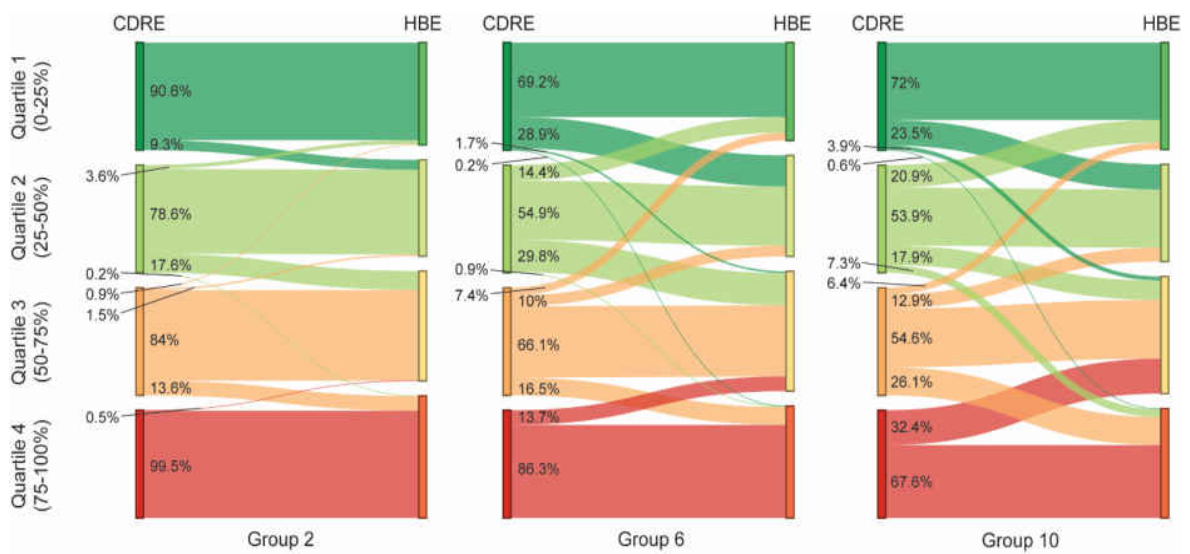


Figure B-8. The directions of potential O₃ exposure misclassifications when the home-based exposure estimation method was used and when IDW fields were used.

B.3 Temporal Variations of Differences Between HBE and CDRE

Figures B-9 through B-12 provides diurnal variations of the average relative differences and average absolute relative differences between HBE and CDRE for the 5 chosen pollutants when CMAQ and IDW fields were applied.

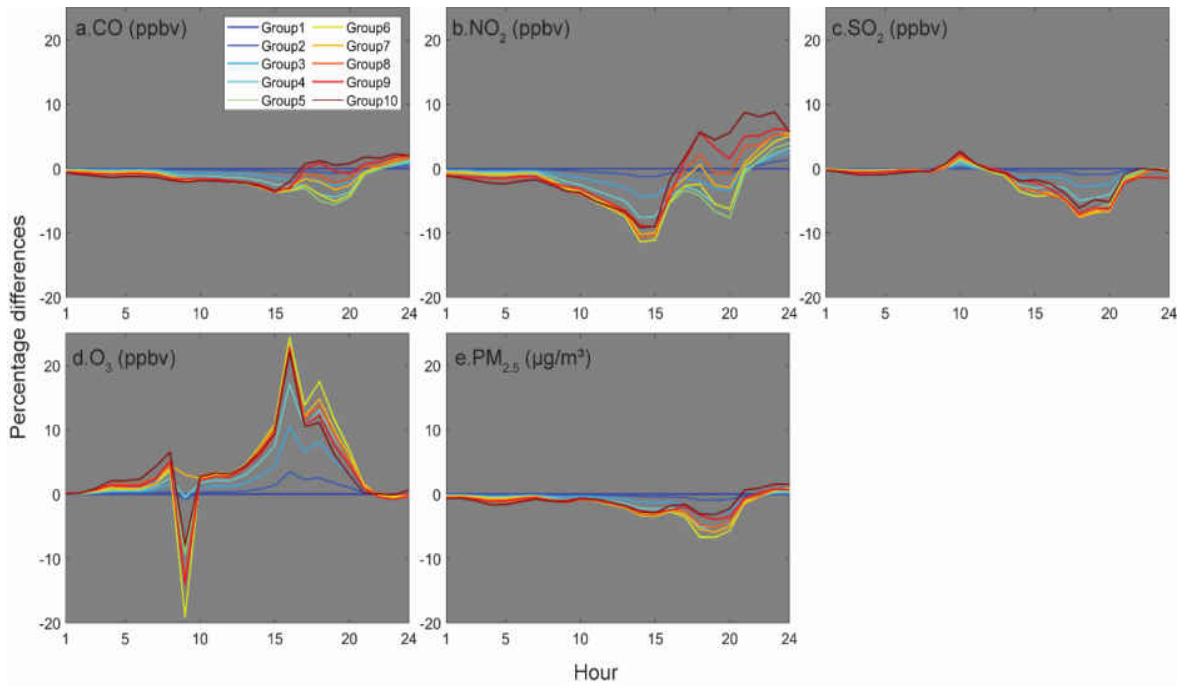


Figure B-9. Temporal variations of average relative differences between HBE and CDRE when CMAQ concentration field were used. The relative differences were estimated as $(HBE-CDRE)/CDRE$.

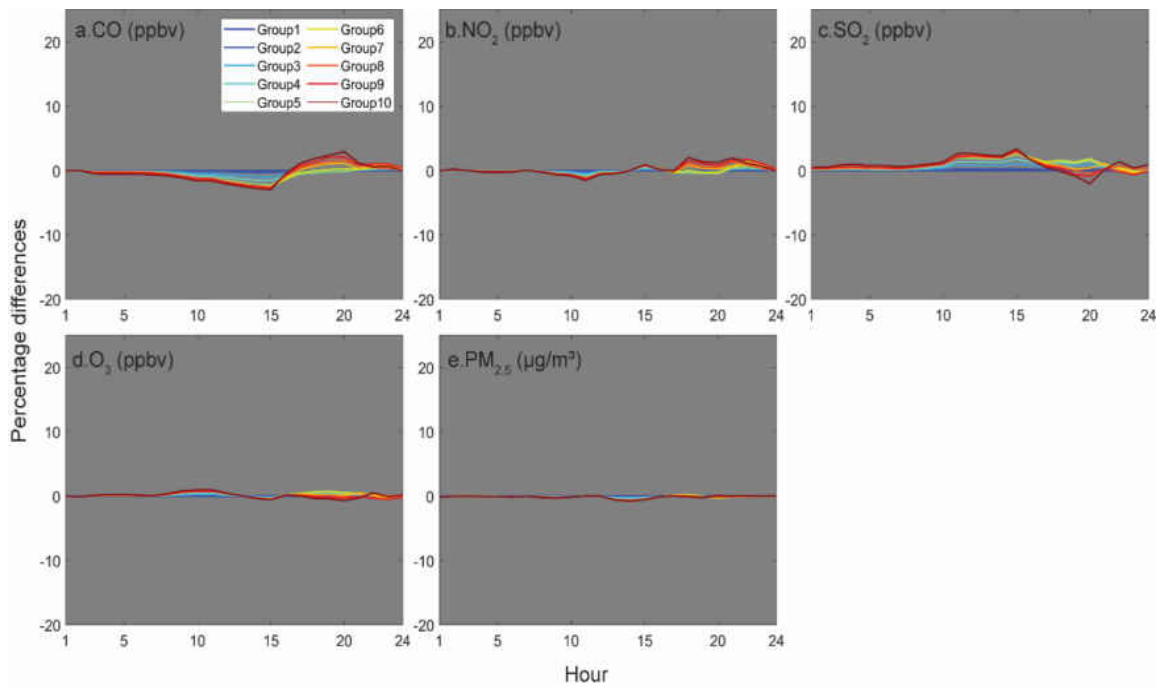


Figure B-10. Temporal variations of average relative differences between HBE and CDRE when IDW concentration field were used. The relative differences were estimated as $(\text{HBE}-\text{CDRE})/\text{CDRE}$.

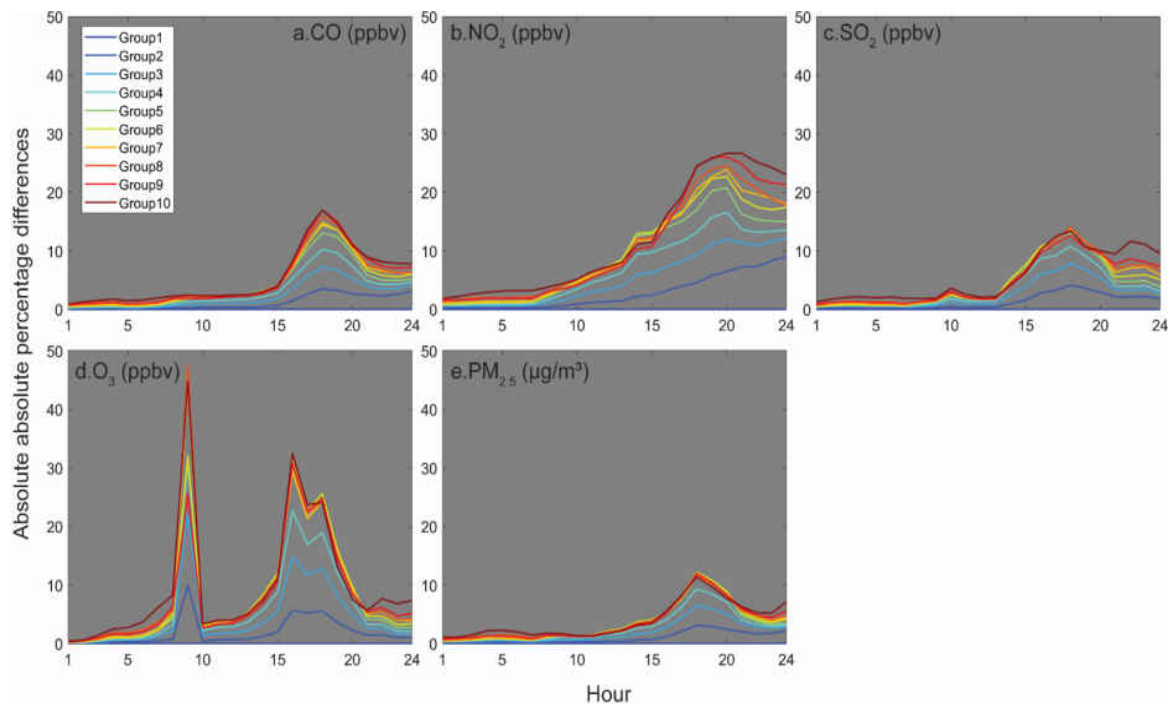


Figure B-11. Temporal variations of average absolute relative differences between HBE and CDRE when IDW concentration field were used. The relative differences were estimated as $(\text{HBE}-\text{CDRE})/\text{CDRE}$.

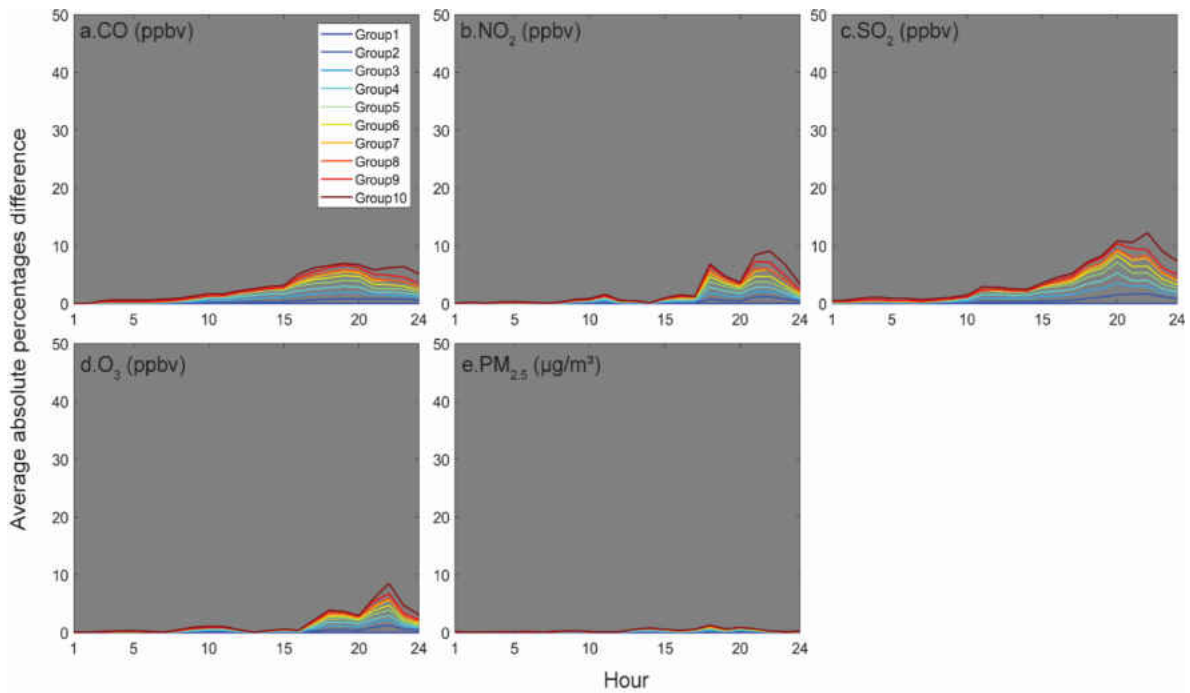


Figure B-12. Temporal variations of average absolute relative differences between HBE and CDRE when IDW concentration field were used. The relative differences were estimated as $(HBE-CDRE)/CDRE$.

B.4 Potential Exposure Misclassifications When Mobility Were Neglected

Tables B-5 through B-8 provides summed percentages of sample populations that were classified into different quartiles when their subject mobility were neglected, i.e. when the home-based exposure estimation method was used, for the 5 chosen pollutants and when both CMAQ and IDW fields were used.

Table B-5. Percentage of sample populations in each quartile that were classified into different quartiles when subject mobility was neglected in exposure estimation. Results shown are for CO.

	Grp 2	Grp 3	Grp 4	Grp 5	Grp 6	Grp 7	Grp 8	Grp 9	Grp 10
Q1 (0-25%)	7.0	11	14	13	15	17	27	24	26
Q2 (25-50%)	23	23	36	40	40	40	53	54	54
Q3 (50-75%)	7.9	21	32	61	57	47	54	50	56
CMAQ Q4 (75-100%)	4.4	10	13	10	15	23	26	28	30
IDW Q1 (0-25%)	5.7	12	8	15	14	13	15	22	32

Q2 (25-50%)	10	19	20	35	32	25	32	40	31
Q3 (50-75%)	11	15	24	20	46	40	37	34	32
Q4 (75-100%)	1.6	6.9	6.6	17	8.0	14	17	21	23

Table B-6. Percentage of sample populations in each quartile that were classified into different quartiles when subject mobility was neglected in exposure estimation. Results shown are for NO₂.

	Grp 2	Grp 3	Grp 4	Grp 5	Grp 6	Grp 7	Grp 8	Grp 9	Grp 10
Q1 (0-25%)	9.3	17	17	19	30	25	24	28	31
Q2 (25-50%)	22	37	46	58	50	54	53	52	54
Q3 (50-75%)	16	29	49	50	61	52	56	59	50
CMAQ Q4 (75-100%)	4.7	10	19	27	30	32	35	37	37
Q1 (0-25%)	7.2	7.7	15	20	22	22	20	31	30
Q2 (25-50%)	10	21	22	38	38	40	43	51	52
Q3 (50-75%)	12	13	26	23	50	46	42	35	35
IDW Q4 (75-100%)	2.0	7.5	9.4	21	14	20	25	27	29

Table B-7. Percentage of sample populations in each quartile that were classified into different quartiles when subject mobility was neglected in exposure estimation. Results shown are for SO₂.

	Grp 2	Grp 3	Grp 4	Grp 5	Grp 6	Grp 7	Grp 8	Grp 9	Grp 10
Q1 (0-25%)	18	23	40	41	40	41	38	31	34
Q2 (25-50%)	16	38	41	54	43	55	48	36	45
Q3 (50-75%)	45	46	42	40	46	40	26	30	28
CMAQ Q4 (75-100%)	2.9	5.3	6.3	7.8	5.3	6.8	10	7.9	8.1
Q1 (0-25%)	8.6	18	14	21	26	32	31	32	23
Q2 (25-50%)	20	19	32	25	38	28	32	28	37
Q3 (50-75%)	5.5	9.5	16	22	25	29	29	24	27
IDW Q4 (75-100%)	0.090	4.7	0.9	5.7	5.2	4.7	13	15	21

Table B-8. Percentage of sample populations in each quartile that were classified into different quartiles when subject mobility was neglected in exposure estimation. Results shown are for O₃.

	Grp 2	Grp 3	Grp 4	Grp 5	Grp 6	Grp 7	Grp 8	Grp 9	Grp 10
Q1 (0-25%)	4.8	14	18	25	33	38	36	31	36
Q2 (25-50%)	15	23	40	47	47	37	40	50	47
Q3 (50-75%)	12	18	26	40	30	34	44	39	43
CMAQ Q4 (75-100%)	4.1	6.3	8.3	9.2	14	13	14	17	13
Q1 (0-25%)	9.4	20	16	24	31	36	35	38	28
Q2 (25-50%)	21	22	34	30	45	35	36	37	46
IDW Q3 (50-75%)	16	14	23	33	34	39	45	43	45

Q4 (75-100%)		0.49	7.4	2.1	6.9	14	14	18	26	32
--------------	--	------	-----	-----	-----	----	----	----	----	----

LIST OF REFERENCE

1. de Zwart, F., et al., *Air pollution and performance-based physical functioning in Dutch older adults*. Environmental Health Perspectives (Online), 2018. **126**(1): p. 017009-1-017009-9.
2. 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.
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. Brugge, D., J.L. Durant, and C. Rioux, *Near-highway pollutants in motor vehicle exhaust: a review of epidemiologic evidence of cardiac and pulmonary health risks*. Environmental health, 2007. **6**(1): p. 23.
5. Gehring, U., et al., *Air pollution exposure and lung function in children: the ESCAPE project*. Environmental health perspectives, 2013. **121**(11-12): p. 1357-1364.
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. Kurt, O.K., J. Zhang, and K.E. Pinkerton, *Pulmonary health effects of air pollution*. Current opinion in pulmonary medicine, 2016. **22**(2): p. 138.
8. Zhang, X., X. Chen, and X. Zhang, *The impact of exposure to air pollution on cognitive performance*. Proceedings of the National Academy of Sciences, 2018. **115**(37): p. 9193-9197.

9. Newbury, J.B., et al., *Association of Air Pollution Exposure With Psychotic Experiences During Adolescence*. JAMA psychiatry, 2019.
10. Di, Q., et al., *Air pollution and mortality in the Medicare population*. New England Journal of Medicine, 2017. **376**(26): p. 2513-2522.
11. Di, Q., et al., *Association of short-term exposure to air pollution with mortality in older adults*. Jama, 2017. **318**(24): p. 2446-2456.
12. Dominici, F., et al., *Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases*. Jama, 2006. **295**(10): p. 1127-1134.
13. GOUDARZI, G.R., et al., *Estimation of health effects attributed to no2 exposure using airq model*. 2012.
14. Khaniabadi, Y.O., et al., *Human health risk assessment due to ambient PM10 and SO2 by an air quality modeling technique*. Process safety and environmental protection, 2017. **111**: p. 346-354.
15. Yu, H., et al., *Using cell phone location to assess misclassification errors in air pollution exposure estimation*. Environmental Pollution, 2018. **233**: p. 261-266.
16. Jerrett, M., et al., *Spatial analysis of air pollution and mortality in California*. American journal of respiratory and critical care medicine, 2013. **188**(5): p. 593-599.
17. 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.
18. Chen, B., et al., *Real-time estimation of population exposure to PM2.5 using mobile- and station-based big data*. International journal of environmental research and public health, 2018. **15**(4): p. 573.
19. Hodgson, S., et al., *Exposure misclassification due to residential mobility during*

- pregnancy*. International journal of hygiene and environmental health, 2015. **218**(4): p. 414-421.
20. Toledo-Corral, C., et al., *Effects of air pollution exposure on glucose metabolism in Los Angeles minority children*. Pediatric obesity, 2018. **13**(1): p. 54-62.
 21. Madsen, C., et al., *Ambient air pollution exposure, residential mobility and term birth weight in Oslo, Norway*. Environmental research, 2010. **110**(4): p. 363-371.
 22. 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.
 23. Shafran-Nathan, R., Yuval, and D.M. Broday, *Impacts of Personal Mobility and Diurnal Concentration Variability on Exposure Misclassification to Ambient Pollutants*. Environmental science & technology, 2018. **52**(6): p. 3520-3526.
 24. Dhondt, S., et al., *Health impact assessment of air pollution using a dynamic exposure profile: implications for exposure and health impact estimates*. Environmental impact assessment review, 2012. **36**: p. 42-51.
 25. Ma, Y., et al., *Modeling the hourly distribution of population at a high spatiotemporal resolution using subway smart card data: A case study in the central area of Beijing*. ISPRS International Journal of Geo-Information, 2017. **6**(5): p. 128.
 26. 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.
 27. Breen, M.S., et al., *GPS-based microenvironment tracker (MicroTrac) model to estimate time–location of individuals for air pollution exposure assessments: Model evaluation in central North Carolina*. Journal of Exposure Science and Environmental

- Epidemiology, 2014. **24**(4): p. 412.
28. Sloan, C.D., et al., *Elemental analysis of infant airborne particulate exposures*. Journal of Exposure Science and Environmental Epidemiology, 2017. **27**(5): p. 526.
 29. Good, N., et al., *The Fort Collins Commuter Study: Impact of route type and transport mode on personal exposure to multiple air pollutants*. Journal of exposure science and environmental epidemiology, 2016. **26**(4): p. 397.
 30. Glasgow, M.L., et al., *Using smartphones to collect time–activity data for long-term personal-level air pollution exposure assessment*. Journal of Exposure Science and Environmental Epidemiology, 2016. **26**(4): p. 356.
 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. Blaasaas, K.G., T. Tynes, and R.T. Lie, *Residence near power lines and the risk of birth defects*. Epidemiology, 2003. **14**(1): p. 95-98.
 33. Bell, M.L. and K. Belanger, *Review of research on residential mobility during pregnancy: consequences for assessment of prenatal environmental exposures*. Journal of Exposure Science and Environmental Epidemiology, 2012. **22**(5): p. 429.
 34. Gurram, S., *Understanding the Linkages between Urban Transportation Design and Population Exposure to Traffic-Related Air Pollution: Application of an Integrated Transportation and Air Pollution Modeling Framework to Tampa, FL*. 2017, University of South Florida.
 35. Shekarrizfard, M., et al., *Individual exposure to traffic related air pollution across land-use clusters*. Transportation Research Part D: Transport and Environment, 2016. **46**: p. 339-350.

36. Shekarrizfard, M., et al., *Regional assessment of exposure to traffic-related air pollution: Impacts of individual mobility and transit investment scenarios*. Sustainable Cities and Society, 2017. **29**: p. 68-76.
37. Zandbergen, P.A., *Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning*. Transactions in GIS, 2009. **13**(s1): p. 5-25.
38. Zhang, D., et al. *Dmodel: Online Taxicab Demand Model from Big Sensor Data in a Roving Sensor Network*. in *2014 IEEE International Congress on Big Data*. 2014.
39. Zhang, D., et al. *CallCab: A unified recommendation system for carpooling and regular taxicab services*. in *2013 IEEE International Conference on Big Data*. 2013.
40. 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: Seattle, Washington. p. 1-10.
41. Hasan, S., et al., *Spatiotemporal Patterns of Urban Human Mobility*. Journal of Statistical Physics, 2013. **151**(1): p. 304-318.
42. 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.
43. 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.
44. 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.
45. Marais, E.A. and C. Wiedinmyer. *Using Google Location History to track personal*

- exposure to air pollution.* in *AGU Fall Meeting Abstracts*. 2017.
46. Do, T.M.T., et al., *A probabilistic kernel method for human mobility prediction with smartphones.* *Pervasive and Mobile Computing*, 2015. **20**: p. 13-28.
 47. Jagadeesh, G.R. and T. Srikanthan. *Probabilistic map matching of sparse and noisy smartphone location data.* in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*. 2015. IEEE.
 48. Nethery, E., et al., *Using Global Positioning Systems (GPS) and temperature data to generate time-activity classifications for estimating personal exposure in air monitoring studies: an automated method.* *Environmental Health*, 2014. **13**(1): p. 33.
 49. 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.
 50. Freeman, N.C. and S.S. de Tejada, *Methods for collecting time/activity pattern information related to exposure to combustion products.* *Chemosphere*, 2002. **49**(9): p. 979-992.
 51. Wu, C.-F., et al., *Exposure assessment and modeling of particulate matter for asthmatic children using personal nephelometers.* *Atmospheric Environment*, 2005. **39**(19): p. 3457-3469.
 52. FREEMAN, N.C., et al., *Responses to the Region 5 NHEXAS time/activity diary.* *Journal of Exposure Science and Environmental Epidemiology*, 1999. **9**(5): p. 414.
 53. Elgethun, K., et al., *Comparison of global positioning system (GPS) tracking and parent-report diaries to characterize children's time–location patterns.* *Journal of Exposure Science and Environmental Epidemiology*, 2007. **17**(2): p. 196.
 54. Watanabe, C., *Health Impact of Urban Physicochemical Environment Considering the*

- Mobility of the People*, in *Health in Ecological Perspectives in the Anthropocene*. 2019, Springer. p. 13-27.
55. Wu, J., et al., *Travel patterns during pregnancy: comparison between Global Positioning System (GPS) tracking and questionnaire data*. *Environmental Health*, 2013. **12**(1): p. 86.
56. Stanley, K., et al., *How many days are enough?: capturing routine human mobility*. *International Journal of Geographical Information Science*, 2018. **32**(7): p. 1485-1504.
57. Steinle, S., et al., *Personal exposure monitoring of PM_{2.5} in indoor and outdoor microenvironments*. *Science of the Total Environment*, 2015. **508**: p. 383-394.
58. Ouidir, M., et al., *Estimation of exposure to atmospheric pollutants during pregnancy integrating space–time activity and indoor air levels: Does it make a difference?* *Environment international*, 2015. **84**: p. 161-173.
59. Sanchez, M., et al., *Predictors of daily mobility of adults in peri-urban South India*. *International journal of environmental research and public health*, 2017. **14**(7): p. 783.
60. Dons, E., et al., *Impact of time–activity patterns on personal exposure to black carbon*. *Atmospheric Environment*, 2011. **45**(21): p. 3594-3602.
61. Steinle, S., S. Reis, and C.E. Sabel, *Quantifying human exposure to air pollution—Moving from static monitoring to spatio-temporally resolved personal exposure assessment*. *Science of the Total Environment*, 2013. **443**: p. 184-193.
62. Dewulf, B., et al., *Dynamic assessment of inhaled air pollution using GPS and accelerometer data*. *Journal of Transport & Health*, 2016. **3**(1): p. 114-123.
63. 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.

64. Picornell, M., et al., *Population dynamics based on mobile phone data to improve air pollution exposure assessments*. Journal of exposure science & environmental epidemiology, 2019. **29**(2): p. 278.
65. Thomas, K.V., et al., *Use of mobile device data to better estimate dynamic population size for wastewater-based epidemiology*. Environmental science & technology, 2017. **51**(19): p. 11363-11370.
66. Furletti, B., et al., *Discovering and understanding city events with big data: the case of rome*. Information, 2017. **8**(3): p. 74.
67. Nyhan, M., et al., *Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data*. Journal of Exposure Science and Environmental Epidemiology., 2018.
68. Jones, K.H., et al., *Public Views on Using Mobile Phone Call Detail Records in Health Research: Qualitative Study*. JMIR mHealth and uHealth, 2019. **7**(1): p. e11730.
69. De Nazelle, A., et al., *Improving estimates of air pollution exposure through ubiquitous sensing technologies*. Environmental Pollution, 2013. **176**: p. 92-99.
70. Donaire-Gonzalez, D., et al., *ExpoApp: An integrated system to assess multiple personal environmental exposures*. Environment international, 2019. **126**: p. 494-503.
71. Pañella, P., et al., *Ultrafine particles and black carbon personal exposures in asthmatic and non-asthmatic children at school age*. Indoor air, 2017. **27**(5): p. 891-899.
72. 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.
73. Donaire-Gonzalez, D., et al., *Benefits of mobile phone technology for personal environmental monitoring*. JMIR mHealth and uHealth, 2016. **4**(4): p. e126.

74. 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. **252**: p. 924-930.
75. Ruktanonchai, N.W., et al., *Using Google Location History data to quantify fine-scale human mobility*. International journal of health geographics, 2018. **17**(1): p. 28.
76. Su, J.G., et al., *Integrating smart-phone based momentary location tracking with fixed site air quality monitoring for personal exposure assessment*. Science of The Total Environment, 2015. **506**: p. 518-526.
77. Rosofsky, A., et al., *Temporal trends in air pollution exposure inequality in Massachusetts*. Environmental research, 2018. **161**: p. 76-86.
78. Aleksandropoulou, V. and M. Lazaridis, *Trends in population exposure to particulate matter in urban areas of Greece during the last decade*. Science of the Total Environment, 2017. **581**: p. 399-412.
79. 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.
80. Guxens, M., et al., *Air pollution exposure during fetal life, brain morphology, and cognitive function in school-age children*. Biological psychiatry, 2018. **84**(4): p. 295-303.
81. Milà, C., et al., *When, Where, and What? Characterizing Personal PM_{2.5} Exposure in Periurban India by Integrating GPS, Wearable Camera, and Ambient and Personal Monitoring Data*. Environmental science & technology, 2018. **52**(22): p. 13481-13490.
82. Salmon, M., et al., *Wearable camera-derived microenvironments in relation to personal exposure to PM_{2.5}*. Environment international, 2018. **117**: p. 300-307.
83. Carrasco-Escobar, G., et al., *Use of open mobile mapping tool to assess human mobility*

- traceability in rural offline populations with contrasting malaria dynamics*. PeerJ, 2019. 7: p. e6298.
84. Lenormand, M., et al., *Influence of sociodemographic characteristics on human mobility*. Scientific reports, 2015. 5: p. 10075.
85. Wu, L., et al., *Intra-urban human mobility and activity transition: Evidence from social media check-in data*. PloS one, 2014. 9(5): p. e97010.
86. Yan, L., et al., *Exploring the effect of air pollution on social activity in China using geotagged social media check-in data*. Cities, 2018.
87. Mazaheri, M., et al., *Investigations into factors affecting personal exposure to particles in urban microenvironments using low-cost sensors*. Environment international, 2018. 120: p. 496-504.
88. Niu, Y., et al., *Estimation of personal ozone exposure using ambient concentrations and influencing factors*. Environment international, 2018. 117: p. 237-242.
89. Schembari, A., et al., *Personal, indoor and outdoor air pollution levels among pregnant women*. Atmospheric environment, 2013. 64: p. 287-295.
90. Dias, D. and O. Tchepel, *Spatial and temporal dynamics in air pollution exposure assessment*. International journal of environmental research and public health, 2018. 15(3): p. 558.
91. Hagenbjörk-Gustafsson, A., et al., *Determinants of personal exposure to some carcinogenic substances and nitrogen dioxide among the general population in five Swedish cities*. Journal of Exposure Science and Environmental Epidemiology, 2014. 24(4): p. 437.
92. Yazar, M., T. Bellander, and A.-S. Merritt, *Personal exposure to carcinogenic and toxic air pollutants in Stockholm, Sweden: A comparison over time*. Atmospheric

- environment, 2011. **45**(17): p. 2999-3004.
93. Carvalho, M.A., et al., *Associations of maternal personal exposure to air pollution on fetal weight and fetoplacental Doppler: a prospective cohort study*. Reproductive Toxicology, 2016. **62**: p. 9-17.
 94. Hannam, K., et al., *A comparison of population air pollution exposure estimation techniques with personal exposure estimates in a pregnant cohort*. Environmental Science: Processes & Impacts, 2013. **15**(8): p. 1562-1572.
 95. Delgado-Saborit, J.M., *Use of real-time sensors to characterise human exposures to combustion related pollutants*. Journal of Environmental Monitoring, 2012. **14**(7): p. 1824-1837.
 96. GBD 2016 Risk Factors Collaborators, *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*. Lancet, 2017. **390**(10100): p. 1345-1422.
 97. Kampa, M. and E. Castanas, *Human health effects of air pollution*. Environmental pollution, 2008. **151**(2): p. 362-367.
 98. Bernstein, J.A., et al., *Health effects of air pollution*. Journal of allergy and clinical immunology, 2004. **114**(5): p. 1116-1123.
 99. Kim, J., *Ambient air pollution: health hazards to children*. Pediatrics, 2004. **114**(6): p. 1699-1707.
 100. 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.
 101. 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.
102. 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.
103. 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.
104. 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.
105. Shafran-Nathan, R., et al., *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.
106. 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.
107. 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.
108. Panko, R. *The Popularity of Google Maps: Trends in Navigation Apps in 2018.* 2018 [cited 2019 January 21]; Available from: <https://themanifest.com/app-development/popularity-google-maps-trends-navigation-apps-2018>.
109. Popper, B. *Google announces over 2 billion monthly active devices on Android.* 2018

[cited 2019 January 21]; Available from:

<https://www.theverge.com/2017/5/17/15654454/android-reaches-2-billion-monthly-active-users>.

110. Gell, N.M., et al., *Efficacy, feasibility, and acceptability of a novel technology-based intervention to support physical activity in cancer survivors*. *Supportive Care in Cancer*, 2017. **25**(4): p. 1291-1300.
111. Schipperijn, J., et al., *Dynamic accuracy of GPS receivers for use in health research: a novel method to assess GPS accuracy in real-world settings*. *Frontiers in public health*, 2014. **2**: p. 21.
112. Kim, D., C.S. Smith, and E. Connors, *Travel Behavior of Blind Individuals before and after Receiving Orientation and Mobility Training*. 2016, Western Michigan University. Transportation Research Center for Livable Communities.
113. Vanwolleghem, G., et al., *Children's GPS-determined versus self-reported transport in leisure time and associations with parental perceptions of the neighborhood environment*. *International journal of health geographics*, 2016. **15**(1): p. 16.
114. Duncan, D.T., et al., *Feasibility and acceptability of global positioning system (GPS) methods to study the spatial contexts of substance use and sexual risk behaviors among young men who have sex with men in New York City: A p18 cohort sub-study*. *PloS one*, 2016. **11**(2): p. e0147520.
115. Lyapustin, A., et al., *Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm*. *Journal of Geophysical Research: Atmospheres*, 2011. **116**(D3).
116. Lyapustin, A., et al., *Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables*. *Journal of Geophysical Research: Atmospheres*, 2011. **116**(D3).

117. Lyapustin, A.I., et al., *Multi-angle implementation of atmospheric correction for MODIS (MAIAC): 3. Atmospheric correction*. Remote Sensing of Environment, 2012. **127**: p. 385-393.
118. Zhan, X., et al., *Urban link travel time estimation using large-scale taxi data with partial information*. Transportation Research Part C: Emerging Technologies, 2013. **33**: p. 37-49.
119. Yin, M., et al., *A generative model of urban activities from cellular data*. IEEE Transactions on Intelligent Transportation Systems, 2018. **19**(6): p. 1682-1696.
120. Hasan, S. and S.V. Ukkusuri, *Reconstructing Activity Location Sequences From Incomplete Check-In Data: A Semi-Markov Continuous-Time Bayesian Network Model*. IEEE Transactions on Intelligent Transportation Systems, 2018. **19**(3): p. 687-698.
121. Jiang, S., J. Ferreira, and M.C. González, *Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore*. IEEE Transactions on Big Data, 2017. **3**(2): p. 208-219.
122. Liao, L., et al., *Learning and inferring transportation routines*. Artificial Intelligence, 2007. **171**(5-6): p. 311-331.
123. Jiang, S., et al., *The TimeGeo modeling framework for urban mobility without travel surveys*. Proceedings of the National Academy of Sciences, 2016. **113**(37): p. E5370-E5378.
124. Macarulla Rodriguez, A., et al., *Google timeline accuracy assessment and error prediction*. Forensic sciences research, 2018. **3**(3): p. 240-255.
125. Su, J.G., et al., *Integrating smart-phone based momentary location tracking with fixed site air quality monitoring for personal exposure assessment*. Science of The Total Environment, 2015. **506-507**: p. 518-526.

126. Center, P.R., *Mobile Fact Sheet*. 2018, Pew Research Center.
127. Prince, S.A., et al., *A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review*. *The international journal of behavioral nutrition and physical activity*, 2008. **5**: p. 56-56.
128. Arthur, C., *iPhone keeps record of everywhere you go*, in *The Guardian*. 2011.
129. Zahradnik, F., *How to Find Your Location History in Google Maps or iPhone*, in *Lifewire*. 2018.
130. Smith, C., *The amount of data Google tracks from your Android phone is staggering*, in *BGR*. 2018.
131. Lomas, N., *Android devices seen covertly sending location data to Google*, in *TechCrunch*. 2017.
132. 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.
133. Kitamura, R., et al., *Micro-simulation of daily activity-travel patterns for travel demand forecasting*. *Transportation*, 2000. **27**(1): p. 25-51.
134. Colizza, V., et al., *Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions*. *PLoS medicine*, 2007. **4**(1): p. e13.
135. Hufnagel, L., D. Brockmann, and T. Geisel, *Forecast and control of epidemics in a globalized world*. *Proceedings of the National Academy of Sciences*, 2004. **101**(42): p. 15124-15129.
136. Kleinberg, J., *Computing: The wireless epidemic*. *Nature*, 2007. **449**(7160): p. 287.
137. Chaintreau, A., et al., *Impact of human mobility on opportunistic forwarding algorithms*. *IEEE Transactions on Mobile Computing*, 2007. **6**(6): p. 606-620.
138. 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.
139. 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.
 140. Gakidou, E., et al., *Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2013; 2016: a systematic analysis for the Global Burden of Disease Study 2016*. The Lancet. **390**(10100): p. 1345-1422.
 141. Zhang, S., et al., *Long-term effects of air pollution on ankle-brachial index*. Environment international, 2018. **118**: p. 17-25.
 142. 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.
 143. 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.
 144. 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.
 145. 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.
 146. 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.

147. Yu, H., et al., *Cross-comparison and evaluation of air pollution field estimation methods*. Atmospheric Environment, 2018. **179**: p. 49-60.
148. 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.
149. Bates, J.T., et al., *Source impact modeling of spatiotemporal trends in PM_{2.5} oxidative potential across the eastern United States*. Atmospheric environment, 2018. **193**: p. 158-167.
150. 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.
151. 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.
152. 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.
153. McMillan, J.E.R., W.B. Glisson, and M. Bromby. *Investigating the increase in mobile phone evidence in criminal activities*. in *System sciences (hicc), 2013 46th hawaii international conference on*. 2013. IEEE.
154. 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.
155. Becker, R., et al., *Human mobility characterization from cellular network data*.

- Communications of the ACM, 2013. **56**(1): p. 74-82.
156. Gonzalez, M.C., C.A. Hidalgo, and A.-L. Barabasi, *Understanding individual human mobility patterns*. nature, 2008. **453**(7196): p. 779.
 157. 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.
 158. 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.
 159. 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.
 160. 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.
 161. 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.
 162. 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.
 163. 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.
 164. 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.

165. 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.
166. Wacholder, S., *When measurement errors correlate with truth: surprising effects of nondifferential misclassification*. Epidemiology (Cambridge, Mass.), 1995. **6**(2): p. 157-161.
167. 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.
168. 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.
169. Fan, Y., et al., *SmarTrAC: A Smartphone Solution for Context-Aware Travel and Activity Capturing*. 2015, University of Minnesota: Minneapolis, MN.
170. 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.
171. Manning, M.I., et al., *Diurnal Patterns in Global Fine Particulate Matter Concentration*. Environmental Science & Technology Letters, 2018. **5**(11): p. 687-691.
172. Jiao, W., et al., *Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States*. Atmospheric Measurement Techniques, 2016. **9**(11): p. 5281-5292.
173. Kelly, K., et al., *Ambient and laboratory evaluation of a low-cost particulate matter sensor*. Environmental pollution, 2017. **221**: p. 491-500.

174. Morawska, L., et al., *Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?* Environment international, 2018. **116**: p. 286-299.
175. Schneider, P., et al., *Mapping urban air quality in near real-time using observations from low-cost sensors and model information.* Environment international, 2017. **106**: p. 234-247.
176. Guanochanga, B., et al. *Towards a Real-Time Air Pollution Monitoring Systems Implemented using Wireless Sensor Networks: Preliminary Results.* in *2018 IEEE Colombian Conference on Communications and Computing (COLCOM)*. 2018. IEEE.
177. *Field Evaluation of Dylos DC1700-PM.* Available from: <http://www.aqmd.gov/docs/default-source/aq-spec/field-evaluations/dylos-dc1700-pm---field-evaluation.pdf?sfvrsn=12>.
178. *Hillsborough County Air Monitoring.*
179. Wang, K., et al., *Evaluating the feasibility of a personal particle exposure monitor in outdoor and indoor microenvironments in Shanghai, China.* International journal of environmental health research, 2019. **29**(2): p. 209-220.
180. Gupta, P., et al., *Impact of California Fires on Local and Regional Air Quality: The Role of a Low-Cost Sensor Network and Satellite Observations.* GeoHealth, 2018.
181. Bulot, F.M., et al., *Long-term field comparison of multiple low-cost particulate matter sensors in an outdoor urban environment.* Scientific reports, 2019. **9**(1): p. 7497.
182. *How to Use Air Sensors: Air Sensor Guidebook.*
183. Zhu, Y., et al., *Comparing gravimetric and real-time sampling of PM_{2.5} concentrations inside truck cabins.* Journal of occupational and environmental hygiene, 2011. **8**(11): p. 662-672.

184. Han, I., E. Symanski, and T.H. Stock, *Feasibility of using low-cost portable particle monitors for measurement of fine and coarse particulate matter in urban ambient air*. Journal of the Air & Waste Management Association, 2017. **67**(3): p. 330-340.
185. Holstius, D.M., et al., *Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California*. Atmospheric Measurement Techniques, 2014. **7**(4): p. 1121-1131.
186. Manikonda, A., et al., *Laboratory assessment of low-cost PM monitors*. Journal of Aerosol Science, 2016. **102**: p. 29-40.
187. Jovašević-Stojanović, M., et al., *On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter*. Environmental Pollution, 2015. **206**: p. 696-704.
188. Semple, S., et al., *Using a new, low-cost air quality sensor to quantify second-hand smoke (SHS) levels in homes*. Tobacco control, 2015. **24**(2): p. 153-158.
189. Wu, J., et al., *Performances of different global positioning system devices for time-location tracking in air pollution epidemiological studies*. Environmental health insights, 2010. **4**: p. EHI. S6246.
190. Zandbergen, P.A. and S.J. Barbeau, *Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones*. The Journal of Navigation, 2011. **64**(3): p. 381-399.
191. Zandbergen, P.A., *Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning*. Transactions in GIS, 2009. **13**: p. 5-25.