2016

Personalized Air Quality Sensing: A Case Study Analysis in Singapore

Meredith McCormack-Mager
mmccorm2@wellesley.edu

WELLESLEY COLLEGE

Personalized Air Quality Sensing: A Case Study Analysis in Singapore

by

Meredith McCormack-Mager

Submitted in Partial Fulfillment
of the
Prerequisite for Honors
in the Department of Mathematics
Wellesley College

Supervisors: Dr. Marguerite Nyhan, Massachussetts Institute of Technology
Dr. Jonathan Tannenhauser, Wellesley College

May 2016

Department of Mathematics
Wellesley College

for Undergraduate Honors

Abstract

by Meredith McCormack-Mager

Singapore's current air quality sensing system tracks only background pollution levels using a handful of stationary sensors, missing localized air pollution information despite traffic emissions being the top cause of air pollution in the city. A novel approach to air pollution data collection using personal mobile sensors is analyzed, and is found to provide additional information about individual exposure to air pollutants. Using descriptive statistics and hypothesis testing, this thesis demonstrates that this personalized sensing technique detects higher air pollution levels and more variance in air quality in the Jurong East neighborhood of Singapore, offering a more specific picture of the air pollution experienced by citizens commuting within the area. Personalized sensing enables additional spatial and temporal analysis, and it is shown using spatial interpolation and data visualization that inclusion of these factors in air quality analysis gives a more specific picture of pollution levels in real time. Personalized sensing of air quality thus contributes highly relevant data about local air pollution levels and personal exposure, which have great potential to provide real-time, localized air quality predictions that the Singaporean government and citizens can use to improve their public and personal health.

Acknowledgements

I would like to begin by thanking my advisors, Jonathan Tannenhauser and Marguerite Nyhan, for their wisdom and encouragement throughout this process. Both deserve pages of praise for the time and dedication they have given to this thesis, and it is unfortunate that I must limit my gratitude to one paragraph. To Jonathan, thank you for your consistent faith in my abilities, for your continued excitement, positivity, and understanding through the twists and turns of this process, and for challenging me to ask and answer questions I would not otherwise have thought possible. To Maggie, thank you for taking a chance on a student who knew next to nothing about environmental science or statistics when she started this project and thank you for the engaging discussions, the extensive writing advice, and all of the tea.

I give thanks to Oscar Fernandez, Casey Pattanayak, Jonathan Tannenhauser, and Brian Tjaden for agreeing to read my entire thesis and serve as my thesis committee. I appreciate the time and thought you will spend reviewing this body of work and coming up with questions for my thesis defense.

I thank the many people who have contributed to this project in small and large ways. Thank you to Kevin Li for his kriging calculations and prioritization of this project despite his impressive courseload. Thanks are due to Casey Pattanayak for enabling me to discover a passion for statistics and for always making time to consult with me about this project. I thank Beth deSombre for helping me to develop my understanding of the underlying theory behind air pollution, Susan Monaghan for her insightful ideas about the impact of this project on income inequality, and Joel Gewirtz for assisting me in coming to a better understanding of haze in Singapore.

I am grateful to everyone who has supported me in this process, and to Wellesley College and Massachusetts Institute of Technology for making this research possible. Special thanks go to Cassidy Tanner for her meticulous editing, for her unwavering faith and encouragement, and for letting me eat all of her snacks. Many thanks go also to my parents, to Ethan Ackelsberg, to Megan Chen, and to Angellena Berberich-Eerebout for their moral support through the ups and downs of this undertaking. Thank you to the entire Wellesley Math Department, especially Alexander Diesl for coordinating the honors process, Melanie Chamberlain for her optimism, and my fellow honor theses students for their encouragement and advice. And thanks to all of my family, friends, and teammates who have kept me positive and excited about my thesis this whole year.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The current approach to air quality data collection in Singapore is limited in the scope of information it can provide to researchers and citizens. This thesis analyzes a new, personalized approach to collection of carbon monoxide and nitrogen dioxide pollution data and discusses the value of this method as an addition to the present approach used by the Singaporean government.

1.1 Effects of Air Pollution on Human Health

The World Health Organization considers air pollution to be the world's largest environmental health risk. Poor air quality contributes to higher prevalence of stroke, cancer, and ischaemic heart disease, and exacerbates the spread of communicable diseases. In some cases, it even leads to death (WHO, 2014).

In particular, carbon monoxide (CO) exposure can cause a range of health effects. Continuous low concentrations of carbon monoxide cause negative cardiovascular and neurobehavioral effects, while high, acute doses of CO lead to carbon monoxide poisoning. Carbon monoxide poisoning can result in myocardial impairment and pulmonary edema, among other disorders (Raub et al. 2000).

Nitrogen dioxide (NO_2), on the other hand, does not have any proven negative health effects in healthy adults, but has been found to weaken lung capacity in asthmatic children (Smith et al., 2000). Additionally, nitrogen dioxide is a chemical precursor to more harmful secondary pollutants including nitric acid, ozone, and particulate matter, which are well documented as having serious negative effects on human health and on the environment (WHO, 2003).

1.2 Dosimetry Prediction

Pollutants can have varied effects on the human body depending on their concentration and path of absorption into the bloodstream. The quantification of airborne pollutants entering and settling in the respiratory tract (from which they absorb into the bloodstream) is referred to as dosimetry, and can be predicted using ambient air quality information and knowledge of the path of a pollutant in the human airway.[1] This path is determined by multiple factors, which also contribute to whether the pollutant enters and deposits in the respiratory tract at all. These factors include size, concentration, and hygroscopicity of the pollutant, as well as the breathing rate of the individual intaking the pollutant. The pollutant is either exhaled or deposited into one or more of the extrathoracic, tracheobronchial, and alveolar regions of the lungs (McCreddin, 2014). The total amount of deposited pollutant is called the lung deposited dose, and is calculated using the standard International Commission on Radiological Protection model (ICRP, 1994).

Prediction of the lung deposited dose enables anticipation of negative health effects and their consequences. On a statistical level, the long term effects of continuous low-concentration air pollution in individuals can be anticipated from established correlations between pollutant concentration and heart rate variability, an indicator for cardiac stress (Nyhan et al 2013). For a government, dosimetry can be important in predicting the burden on the health care system from air pollutant related illnesses. This knowledge can then be used to streamline diagnosis and care, as well as to properly allocate resources towards air pollution abatement and treatment of pollution-associated illnesses.

1.3 Municipal Air Pollution

Although some air pollutants occur naturally, the vast majority of air pollution is a result of anthropogenic activity (Kampa and Castanas 2008). Nitrogen dioxide (NO_2) results from NO_x gases released in the combustion process of motor vehicles. In fact, nitrogen dioxide is regarded as an indicator of traffic pollution (WHO, 2003). Carbon monoxide is similarly produced by vehicle emissions, and outdoor concentrations of CO are highest in congested traffic, industrial areas, and parking garages and tunnels with poor ventilation (Raub et al. 2000). In megacities, where human concentration is particularly high, traffic congestion is pervasive, and residential communities sometimes overlap with industrial areas, there is especial danger of increased air pollution and the resulting increase in health and economic burdens for the city's many inhabitants.

[1] On occasion, dosimetry is used to quantify the effects of an airborne pollutant on a community. However, in this thesis discussion of dosimetry will be restricted to individuals.

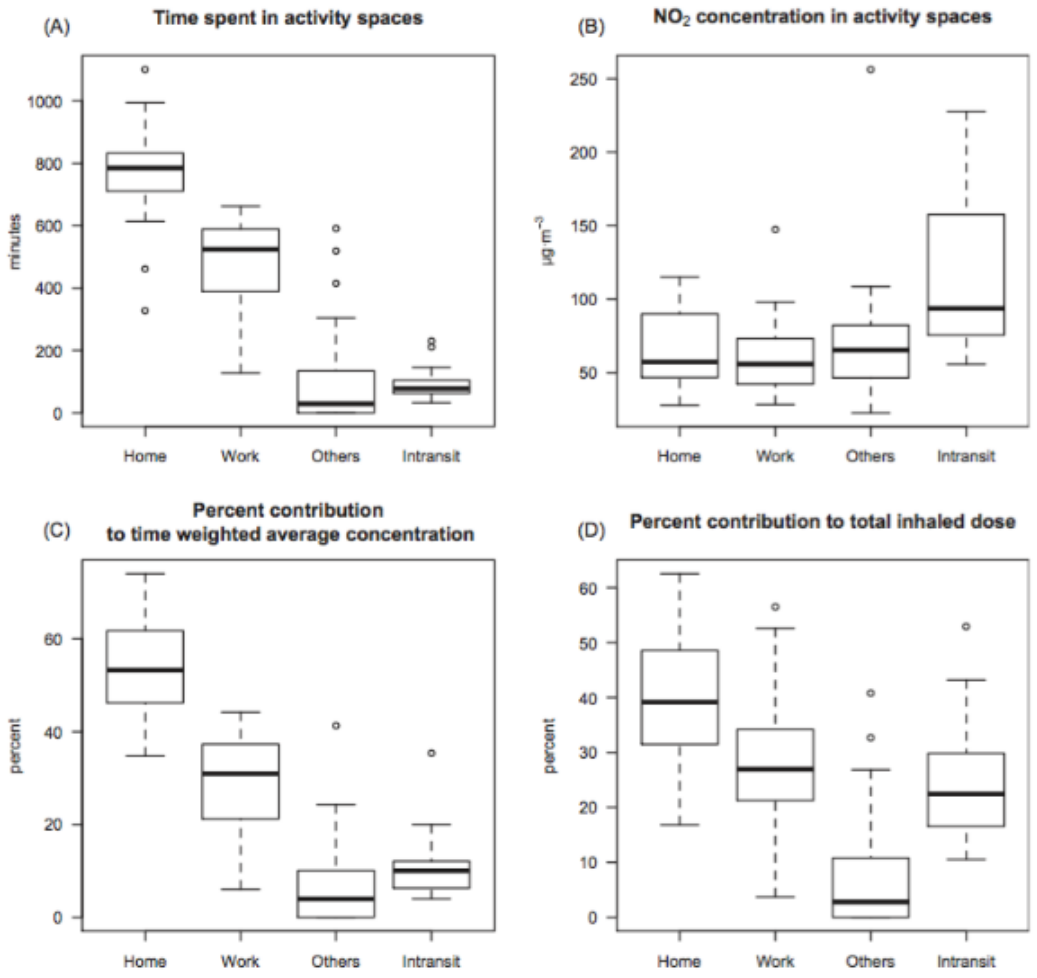


FIGURE 1.1: Boxplots emphasizing that the small amount of time individuals spend in transit each day contributes the most to their total air pollution exposure and thus total inhaled dose of pollutants.(Figure source: deNazelle et al., 2015)

Among individuals, transport activities contribute the most to variability in air pollution exposure between people from otherwise similar backgrounds (Dons, Evi, et al., 2011). A study by DeNazelle et al. in Barcelona, Spain found that commuting accounted for 6 percent of participants' time and 24 percent of their air pollution exposure (2013, Figure 1.1). This suggests that reevaluating the causes of poor air quality that affect people during the short period of their day in which they are commuting could have a huge impact on their overall pollutant exposure, and thus health.

1.4 Air Pollution in Singapore

Air quality in Singapore is regulated by the National Environmental Agency (NEA), which records data on carbon monoxide and nitrogen dioxide levels (among other pollutants) throughout the city-state. The NEA relies on fourteen stationary sensors distributed about the city to report air pollution data every hour on the hour. According to these sensors, Singapore has better air quality than its neighbors, and consistently meets the World Health Organization's international goals for air quality (NEA, 2016). However, stationary sensing lacks some of the power of personalized sensing in determining individual air pollution exposure, as discussed in the next section. Instead, stationary sensing is most useful for detecting background air pollution, namely the delocalized baseline pollution levels experienced by the whole city as a result of polluting factors such as haze.[2]

1.5 Personalized Approach to Air Pollution Monitoring

The rise of mobile phone usage enables real-time, localized spatio-temporal tracking on a large scale. Coupled with increased big data analysis capability, citizen crowdsourcing via passive mobile transmission has great potential to provide comprehensive environmental data, as proposed by Reis et al. (2015).

Background air pollution sensing establishes helpful context for the general pollution that a city experiences, but misses the local variance in pollution levels between neighborhoods or even between city blocks. A localized view of air quality includes both background and specific information about air pollution, and provides greater ammunition for the discovery of air pollution sources. If those sources can be reliably pinpointed to an intersection or business, then governments can more effectively construct and enforce air quality improvement strategies (Heimann et al., 2015).

For individuals, real time data on a local level enables citizens to make real time decisions about their travel patterns throughout their neighborhoods. This is especially important in commuting, where it was found using personalized sensing that activity patterns are significant determinants of personal exposure. The same study also found that exposure estimates using personalized sensing can be highly different from estimates of background air pollution and posited that measurements not accounting for mobility can have elevated error (DeNazelle et al., 2013).

[2]Singapore experiences a significant amount of haze every summer, when industrial forest fires in nearby Indonesia send polluting haze drifting over the city-state (Cochrane 2015).

1.6 Objectives

The overall aim of this study is to determine whether a personalized approach to data collection and analysis of air quality in Singapore contributes information and understanding not registered by the government's current air quality measurement methodology. Additional, specific objectives are as follows:

1. Use computational and data cleaning techniques to prepare government and personalized sensor data for accurate analysis and comparison (Chapter 3). Due to the differing natures of the two data collection methods and their recording techniques, it is critical that the data sets be transformed to be statistically comparable.

2. Using descriptive statistics and data visualization, summarize unique characteristics of personalized sensor data to comprehensively demonstrate the capabilities of the personalized air quality monitoring approach (Chapter 3). Investigate temporal and spatial characteristics of personalized sensor data that differentiate this method from the government's stationary sensing and discuss how these differences do or do not provide additional actionable information to researchers and citizens.

3. Calculate inhaled dose and lung-deposited dose from carbon monoxide and nitrogen dioxide exposure recorded by personalized sensors (Chapter 3). This information most directly contributes to predictions of human health effects, and is therefore useful for discussion of new public health analysis and policies resulting from localized personal sensor data.

4. Compare government-collected air quality data to personalized sensor-collected air quality data and determine whether these data sets tell the same story about air pollution in Singapore (Chapters 4, 5). Using relevant statistical tests and spatial interpolation techniques, visualize and analyze differences between the two data sets and discuss the implications of their differences and similarities.

This study finds that the personalized sensor data contribute beneficial additional information about air pollution levels in Singapore. The personalized sensor data are higher and more variant in concentration level than the data registered by the stationary government monitors. Additionally, the spatial tags and granular temporal information unique to the personalized sensor data enable a more comprehensive perspective of the pollutant data's behavior. Ultimately, this thesis concludes that personalized sensors are an effective tool for use by the Singaporean government and health care system to estimate localized pollution levels in real time and respond to specific air pollutant threats effectively.

Chapter 2

Background

This main aim of this chapter is to introduce the two methods of air pollution data collection under study. This will begin with a discussion of the personalized sensor collection method implemented by researchers from the MIT Airscapes Singapore project, followed by an overview of the air quality data collection method currently used by the Singaporean government. Additionally, this chapter reviews, by way of environmental science background, the process of inhaled air pollution dose calculations. The theory behind hypothesis testing, specifically permutation testing and Wilcoxon rank-sum testing, is included in Chapter 4 (see also Ramsey and Schafer, 2012), and explanation of the kriging method for spatial interpolation is provided in Chapter 5.

2.1 Smart Sensors for Environmental and Human Health Research

The Singapore Airscapes study was organized and conducted by environmental science researchers at MIT's *Sense*able City urban development lab (Nyhan et al., 2015). The Airscapes project sought to apply a distributed network of moving environmental sensors (Figure 2.1) in an environmental health research study, as proposed in previous literature (Reis et al., 2015). This methodology had been used in other urban environmental research with success (deNazelle et al., 2013). The main objective of the Singapore Airscapes project (of which this analysis is the final segment) is to determine whether comparing personalized air pollution exposure information to exposure as monitored municipally by the National Environment Agency would give meaningful data about the state of air pollution as experienced by the citizens of Singapore.

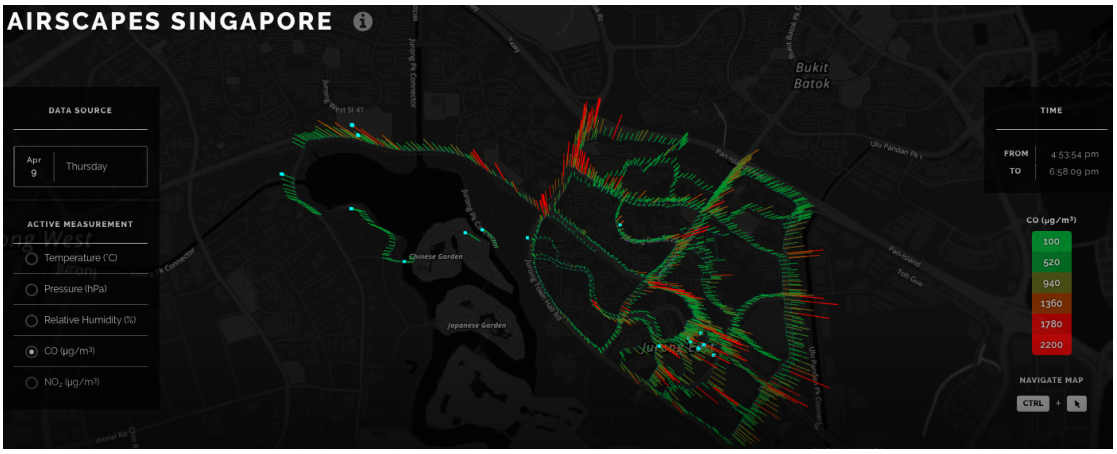


FIGURE 2.1: Still from Airscapes Singapore website depicting the tracing of carbon monoxide concentration paths on Jurong East. Taller, redder lines express high concentration, while shorter, greener lines correspond to low concentration. The square, blue dots represent the distributed network of sensors moving throughout Jurong East.

2.2 Air Pollutant Exposure Monitoring

This section introduces the distributed network of air quality sensors used by the Airscapes Singapore research team to collect the personalized data (2.2.1) and the municipal air quality monitoring strategy used by the Singaporean government to collect the stationary, background data (2.2.2), and discusses preprocessing techniques.

Personalized and government air pollution data were collected concurrently in the Jurong East neighborhood of Singapore during April 2015. The two methods of data collection differed in frequency of sampling, content sampled, and data preprocessing procedure, though both collections included carbon monoxide and nitrogen dioxide information for every sampling time point. Key points about the data collection (including special attention to the differences between the two methods) are highlighted in this section.

2.2.1 Personalized Sensor Air Quality Monitoring

Eleven portable air quality sensors were used to collect the personalized data set in the Jurong East neighborhood of Singapore (Figure 2.2) during the period of April 1st, 2015 to April 10th, 2015.

The study used FER Air Quality Sensors (Oletic and Bilas, 2015) to collect the personalized readings. The sensors are approximately 1.5in x 3in x 1in and easily clip on to belts or pockets. They communicated via bluetooth with participants' cell phones (Figure 2.3), which pinged the sensors every 20 seconds to request air quality and weather observations. These observations included the humidity (%), temperature (^{o}C), pressure

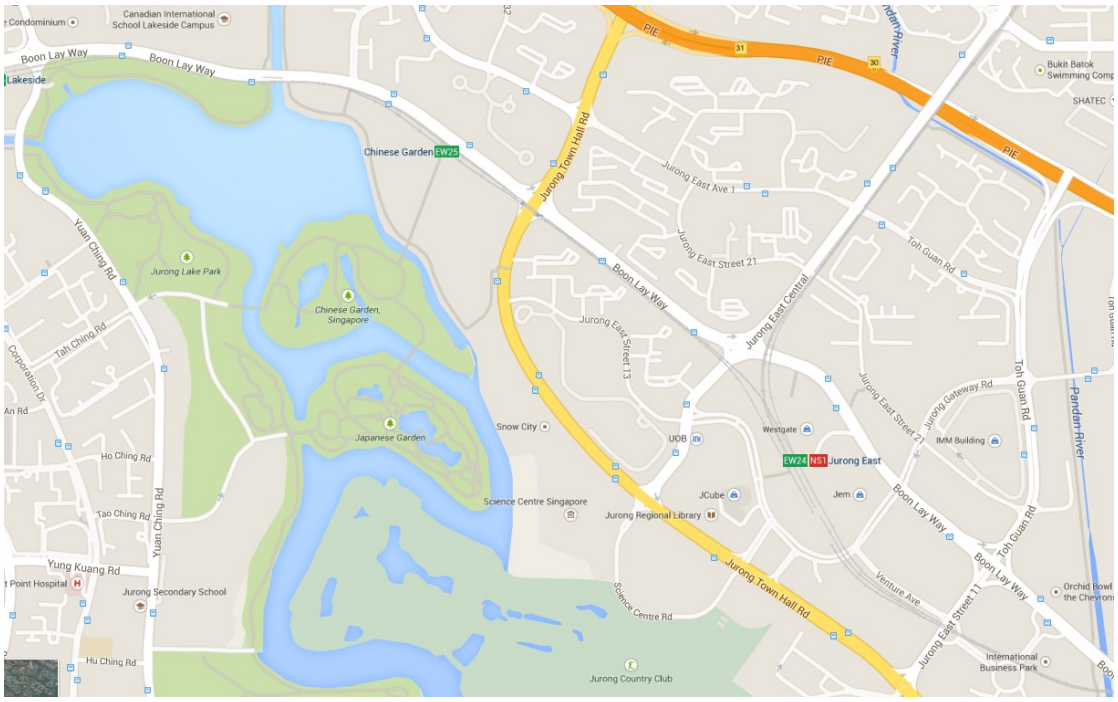


FIGURE 2.2: This map of Jurong East was given to participants in the study, who were then directed as to which route they should walk during their data collection period.

(hPa), CO level ($\mu g/m^3$), and NO_2 level ($\mu g/m^3$) that the sensor was experiencing at that moment. This data was then uploaded in real time to the research database.

University students from Singapore were recruited to carry the FER air quality sensors around Jurong East for 2-hour intervals in April 2015. The intervals were split into morning and afternoon sessions, and since multiple sensors were used, some intervals overlapped temporally. The participants were given maps of Jurong East (Figure 2.2) and directed where to walk. Each participant was given a different route in order to cover as much of the neighborhood as possible in the week-long study.

As a result of this multi-sensor testing, the personalized data was not distributed regularly over time and space. Additionally, the personalized sensors were deployed in an area with frequently changing conditions (e.g. traffic changes, appearance and disappearance of smokers).

2.2.2 Government Air Quality Monitoring

The government of Singapore collects data using fourteen static sensors positioned around the city. The city-state is divided into five monitoring regions: North, South, East, West, and Central. Jurong East, the neighborhood of Singapore on which the study focuses, is located in between the West, South, and Central monitoring regions

FIGURE 2.3: Participants were given portable sensors (shown right) that transmitted pollutant concentrations and descriptive air statistics via bluetooth to their cellphones every 20 seconds.

(Figure 2.4).[1] On the Singapore National Environmental Association (NEA) website, Jurong East is listed as being in the region covered by the West sensor (Figure 2.5). However, Jurong East is on the edge of the West Region, so for the purposes of this study, the neighborhood will be considered to be equally distant from the West, South, and Central sensors. Unlike the personalized data, government monitors are not mobile, and so experience a more consistent environment.

Each government monitor collects readings of the levels of carbon monoxide, nitrogen dioxide, sulfur dioxide, and multiple types of particulate pollutants. The carbon monoxide (CO) levels were recorded in mg/m^3 and the nitrogen dioxide (NO_2) levels were recorded in $\mu g/m^3$. The NO_2 levels were recorded precisely as collected every hour on the hour, whereas the CO levels were recorded as moving 8-hour averages of CO level readings and were posted every hour on the hour. This is because the international WHO standards for recommended maximum carbon monoxide exposure are based on 8-hour averages. For nitrogen dioxide, the standard is based upon hourly readings. In particular, these targets are a 10 mg/m^3 8-hour average (or 30 mg/m^3 1-hour average) for CO and a 200 $\mu g/m^3$ 1-hour average for NO_2. The data was collected via screenshots

[1] PSI is a Singapore-specific metric derived by the NEA that rates a linear combination of pollutant levels (carbon monoxide, nitrogen dioxide, particulate matter, and others) to determine overall air quality.

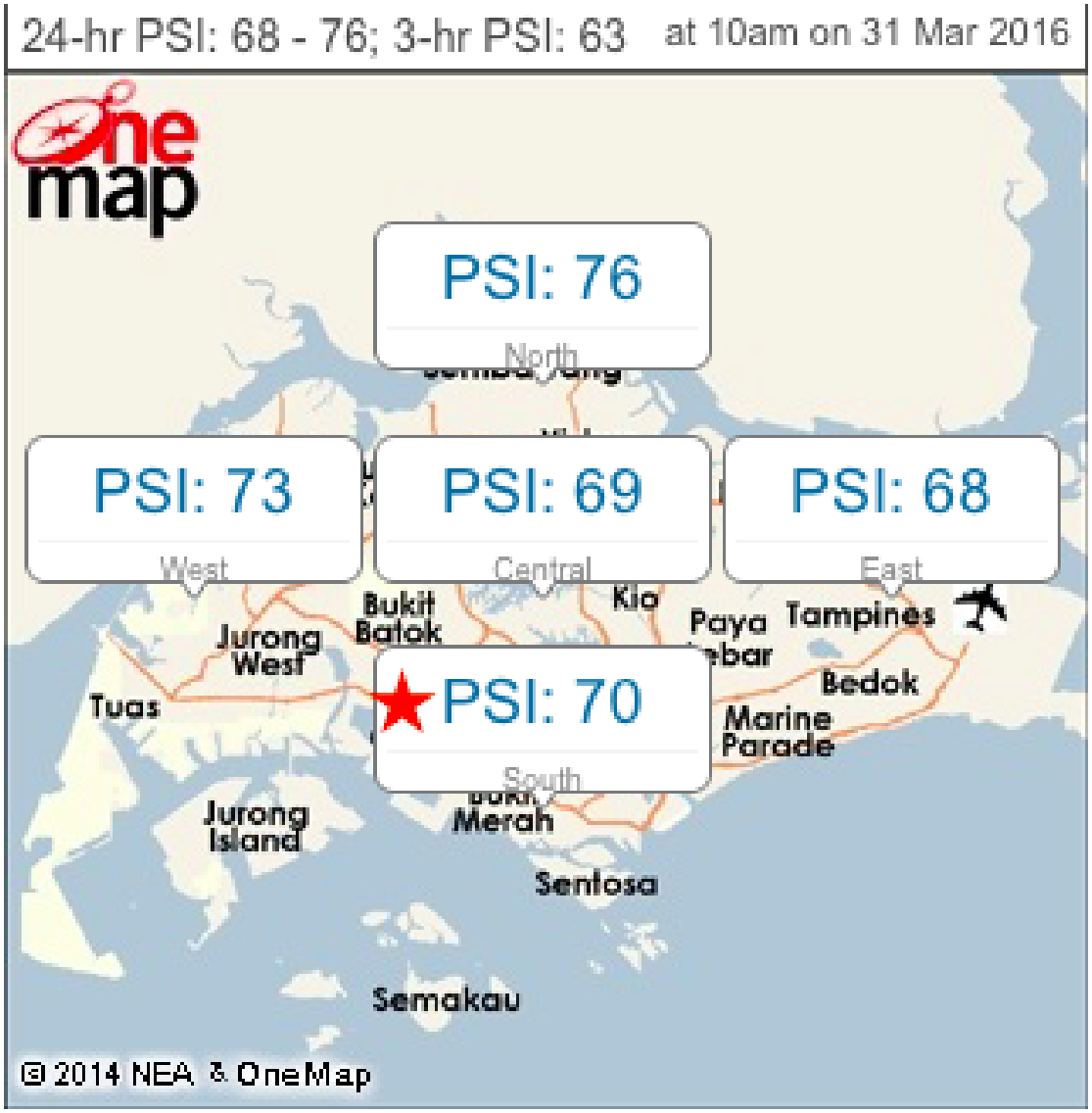


FIGURE 2.4: This map (from the Singapore National Environmental Agency website) indicates the location of the five government monitoring regions. The red star marks the locations of Jurong East. The PSI readings rate the haze on the day the photo was taken (31 March 2016).

of the NEA website taken every evening during the time period of April 4-10, 2015. Figure 2.6 shows one such screenshot from April 4th that includes the CO data from the first half of that day.

2.3 Inhaled Doses of Air Pollution

Inhaled air pollutant dose analysis provides a more accurate assessment of personal exposure than simply examining exposure concentrations in the micro-vicinity of the subjects studied (Nyhan et al, 2014). Using the personalized sensor data, a specific

Regions for Air Quality Reporting

Region	Town Centres /Areas
North	Admirality, Kranji, Woodlands, Sembawang, Yishun, Yio Chu Kang, Seletar, Sengkang
South	Holland, Queenstown, Bukit Merah, Telok Blangah, Pasir Panjang, Sentosa, Bukit Timah, Newton, Orchard, City, Marina South
East	Serangoon, Punggol, Hougang, Tampines, Pasir Ris, Loyang, Simei, Kallang, Katong, East Coast, Macpherson, Bedok, Pulau Ubin, Pulau Tekong
West	Lim Chu Kang, Choa Chu Kang, Bukit Panjang, Tuas, Jurong East, Jurong West, Jurong Industrial Estate, Bukit Batok, Hillview, West Coast, Clementi
Central	Thomson, Marymount, Sin Ming, Ang Mo Kio, Bishan, Serangoon Gardens, MacRitchie, Toa Payoh

FIGURE 2.5: This table is provided on the NEA website to guide users towards the sensor that is most appropriate to check for their neighborhood. It lists Jurong East as being best represented by the West sensor.

8-hr Carbon monoxide (mg/m^3) Readings on 04 Apr 2015

View reading for: 8-hr Carbon Monoxide ▾

Time	1am	2am	3am	4am	5am	6am	7am	8am	9am	10am	11am	12pm
North	0.5(5)	0.5(5)	0.5(5)	0.6(6)	0.6(6)	0.6(6)	0.7(7)	0.7(7)	0.8(8)	0.8(8)	0.8(8)	0.7(7)
South	0.5(5)	0.5(5)	0.5(5)	0.5(5)	0.5(5)	0.6(6)	0.6(6)	0.6(6)	0.6(6)	0.6(6)	0.5(5)	0.5(5)
East	0.6(6)	0.7(7)	0.7(7)	0.8(8)	0.8(8)	0.9(9)	0.9(9)	0.9(9)	0.9(9)	0.8(8)	0.7(7)	0.7(7)
West	0.4(4)	0.5(5)	0.6(6)	0.7(7)	0.8(8)	0.9(9)	0.9(9)	1(10)	1(10)	1(10)	1(10)	0.9(9)
Central	0.3(3)	0.4(4)	0.4(4)	0.5(5)	0.6(6)	0.7(7)	0.7(7)	0.8(8)	0.8(8)	0.7(7)	0.7(7)	0.6(6)
Time	1pm	2pm	3pm	4pm	5pm	6pm	7pm	8pm	9pm	10pm	11pm	12am

FIGURE 2.6: This screenshot includes the government-collected CO level readings (in mg/m^3) from 1am to 12pm on April 4th, 2015. The displayed values are the 8-hour averages finishing at the time-point listed at the top of each column and collected at the location designated by the row. The pertinent values are the ones not in parentheses.

analysis can be conducted of precisely how much pollution is inhaled by a commuting individual based on their activity, gender, and localized air pollution reading.

To measure minute ventilation (breathing rate in L/min), this thesis refers to the following empirical model developed by Zuurbier, et al. (2011) and later adopted in Nyhan, et al. (2014):

$$b = \exp(c + m * H)$$

where b is minute ventilation (L/min), H is subject heart rate, c is the equation intercept (1.03 for males, 0.57 for females), and m is the slope of the equation (0.021 for males, 0.023 for females). Assuming a fixed heart rate, cumulative minute ventilation is based solely on time exposed. Analysis of a subject's velocity allows for an assumed heart rate, which can then be used to calculate new cumulative minute ventilation based on time and type of activity (either using the above formula or previous data collected about respiratory rate (McCreddin, 2014), shown in Figure 2.7).

		Resting		**Sitting awake**		**Light exercise**		**Heavy exercise**	
Maximal workload (%):		8		12		32		64	
Gender (Male/Female):		**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
Breathing Parameters:	V_T (L)	0.625	0.444	0.75	0.464	1.25	0.992	1.923	1.364
	B (m^3h^{-1})	0.45	0.32	0.54	0.39	1.5	1.25	3.0	2.7
	f_R (min^{-1})	12	12	12	14	20	21	26	33

FIGURE 2.7: This table (from McCreddin, 2014) describes the expected breathing rate B in m^3/h (middle row) of individuals engaged in various activities. Note that in this case, breathing rate is expressed in units of m^3/h.

Using the information depicted in Figure 2.7, one can further calculate inhaled dose (the amount of pollutant inhaled) as follows:

$$\text{Inhaled Dose} = C * B * t$$

where C is pollutant concentration level ($\mu g/m^3$), B is breathing rate (L/min), and t is exposure duration (min).

Lung-deposited dose (the amount of inhaled dose that settles in the lungs) can also be calculated using the minute ventilate data and the International Commission on Radiological Protection (ICRP) dosimetry model for inhaled pollutants (ICRP, 1994). This thesis does not pursue lung-deposited dose analysis, since it does not contribute to the main comparison analysis, but suggests this analysis for future studies as being useful for predicting more specifically the health repercussions indicated by personalized data sensing.

Chapter 3

Descriptive Statistics

While powerful hypothesis testing tools will be used in Chapter 4 to rigorously compare the personalized and government data, much can be learned from basic descriptive statistics and visualization of both data sets. This chapter introduces data from the personalized (3.1) and government (3.5) sensors, and highlights their similarities and differences (3.6). The role of data visualization in statistical analysis is further explored and the additional challenges of introducing space and time considerations in air quality research are discussed (3.2). Additionally, the chapter delves more deeply into the personalized data set to draw conclusions about factors that correlate with air pollution levels (3.3) and to examine inhaled dose results from the personalized data (3.4). The overarching aim of Chapter 3 is to demonstrate the potential of personalized sensing techniques in providing comprehensive information about air quality in Jurong East and to begin to characterize how the results of these techniques differ from the government monitor data.

3.1 Introduction to Personalized Sensor Data

Personalized sensor data was collected over the period of April 1st, 2015 through April 10th, 2015. In total, 52 outings were recorded by 11 sensors. Temperature data was collected in degrees Celsius (^{o}C), pressure in hectopascals (hPa), relative humidity in percentages (%), NO_2 concentration in $\mu g/m^3$, and CO concentration in $\mu g/m^3$. Readings of each type were collected every 20 seconds during the outings. Most outings overlapped with others in time, and some included breaks in the data due either to data transmission failure or breaks taken by the participants (detailed discussion of the collection process can be found in Section 2.2.1). Parsing and cleaning of the data involved

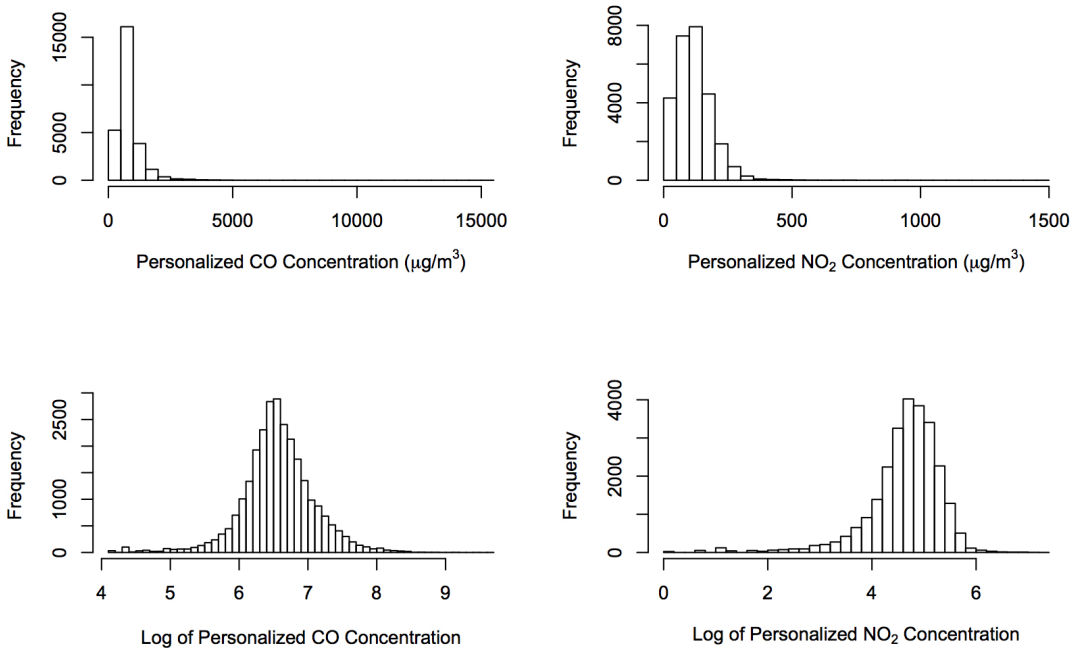


FIGURE 3.1: Histograms displaying the distributions of the personalized sensor data. The top panels demonstrate the distributions of the raw CO (left) and NO_2 (right) data, both of which are right-skewed distributions. The bottom panels indicate that in fact the CO data follows a log normal distribution and the NO_2 data is close to following a log normal distribution.

some discarding of unusable data (explained in 3.2.1). The final number of readings included in the analysis was 27,083 for CO and 27,067 for NO_2.

Both the NO_2 and CO readings collected by the personalized sensors follow right-skewed distributions (Figure 3.1). The CO data follows a log normal distribution with mean 6.56 and standard deviation 0.53 and the NO_2 data closely approximates a log normal distribution with mean 4.63 and standard deviation 0.72 (note that the NO_2 data more closely follows a square root normal distribution except for a dramatic spike on the left tail of the distribution, which is why a log normal distribution is nevertheless preferred). The CO concentration data had mean 813.76 $\mu g/m^3$ and median 700 $\mu g/m^3$, with minimum value 61 $\mu g/m^3$ and maximum value 15255 $\mu g/m^3$. The NO_2 concentration data had mean 117.93 $\mu g/m^3$ and median 110 $\mu g/m^3$, with minimum value 0 $\mu g/m^3$ and maximum value 1466 $\mu g/m^3$. These values are all consistent with the assertion that these data follow right-skewed distributions.

3.2 Spatio-Temporal Visualization and Analysis

Unlike the government monitor data, the personalized data provide localized details about the time, latitude, and longitude for which they were collected. These details are helpful for verifying the data, for visualizing and better understanding the breadth of the data, and for estimating values for nearby locations. The first two points (verification and visualization) will be addressed in this section, and the third (estimation) is covered in Chapter 5.

3.2.1 Time and Location Inconsistencies

In the data cleaning process, inconsistencies arose in both the personalized sensor data and government monitor data. These issues, and how they were addressed, are discussed in this section.

Using the spatial and temporal tags associated with each data point, it was determined that some personalized sensor data collected in the study had not been collected in Jurong East and should be discarded. Recall that the personalized sensor data was collected by participants roaming the streets of Jurong East with portable, bluetooth-enabled air pollution sensors engineered in Zagreb, Croatia (Oletic and Bilas, 2015). In the initial mapping of the data, it was discovered that some observations in the data set were geotagged in Croatia rather than Singapore. Checking the timestamps of these data confirmed that the sensors had not been cleared from the internal sensor logs ahead of the Airscapes Singapore study, and some of leftover Croatian data were mistakenly included in the sensor logs. These data were subsequently removed from the data set. Some data were geotagged as having been collected in Singapore, but outside of Jurong East. These data were found to have been accidentally collected in transport to and from the principal researcher's home and were also discarded. Other data were found to be geotagged in the South China Sea near Singapore. It is unlikely that any of the participants went swimming this far off the coast of Singapore during their participation in the study, so the geotagged locations were mostly likely in error. Since other data were found to have been collected outside of Jurong East, it could not be guaranteed that these data had been collected in Jurong East as opposed to Croatia or another area of Singapore, hence these data were also discarded. Finally, some sensor readings were geotagged at (0, 0) latitude and longitude, and were categorized as missing location data (although one cannot omit the possibility that some participants spent their week on a boat in the Atlantic Ocean). Most missing location data cases occurred at the beginning of a sensor's first outing in Singapore, and were likely a result of initial sensor calibration. Three other cases of missing location data not meeting this criterion were found, in all of

which the location data was missing for less than 2 minutes between two time intervals of data for which location information was present. Location interpolation was nontrivial due to the irregular walking patterns taken by the participants in those particular outings and the readings with missing values comprised only a very small portion of the data set, so this data was discarded. The discarding of data was implemented by imposing time and location boundaries on which data were allowed to be included in the analysis. The time bounds were April 1st through April 10th, 2015, the latitude bounds were 1.3225 N to 1.3455 N, and the longitude bounds were 103.72 E to 103.75 E.

The government data was guaranteed to have been collected in Singapore, due to the stationary nature of the sensors, but there did exist a notable error in the time stamps. The Singaporean National Environmental Agency (NEA) lists the air pollution averages for the previous hour every hour at a few minutes past the hour. The pollution averages for that hour remain up for the remainder of the day, and all readings from a previous day are replaced when the first hour of data is posted for the present day at 1am. Each night from April 1st to April 5th of 2015, the air pollution levels for the whole day were screenshotted, usually around 9pm. On the night of April 10th, 2015, however, the data was screenshotted at 12:25am of April 11th. A bug discovered in the NEA website caused the data to be listed as the data corresponding to April 11th and initially resulted in the mislabelling of the April 10th data. Hence, the data initially labelled April 11th were corrected to be labelled April 10th.

3.2.2 Visualization of Personalized Sensor Data

First, the data are examined with respect to space. Figure 3.2 demonstrates how air pollution levels within Jurong East depend on location, as some areas of the map are perennially higher in pollutant levels than others. The map also shows how air quality levels can vary greatly even in small geographic areas, such as at a particular traffic intersection.

Second, the data are examined with respect to time. Figure 3.3 displays the air pollution levels of every sensor outing during the period from 5pm to 7:15pm on April 2nd, 2015, and demonstrates great variation in the data over time for both CO and NO_2. However, Figure 3.3 also indicates a drawback to visualization of personalized sensing, which is that observation of too many data stories at once can overwhelm a viewer and lead to greater confusion. Instead, it is often easier to look at smaller portions of anecdotal evidence or use computational data analysis techniques (Chapters 4, 5) to derive meaning from the data. Figure 3.4 takes an anecdotal approach to statistics, and demonstrates that even among outings taken at the same time, there can be great variation in the

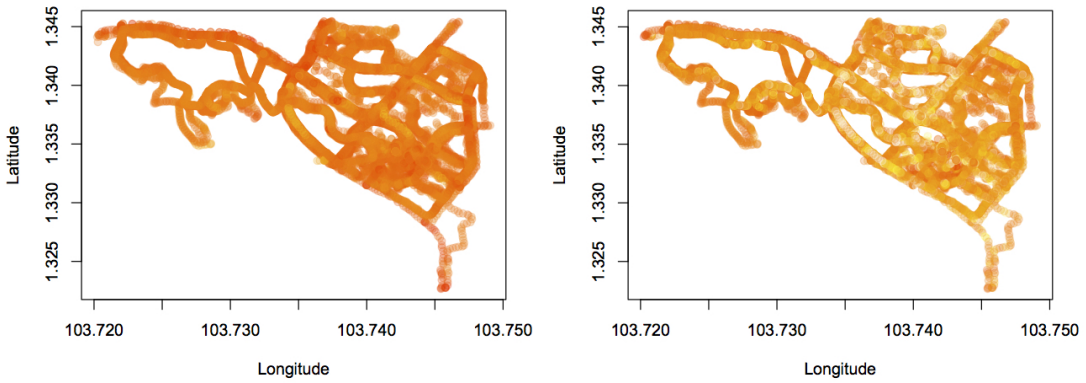


FIGURE 3.2: Maps of the CO (left) and NO_2 (right) levels detected by the personalized sensors during the study. Darker reds correspond to higher pollution levels, while lighter yellows correspond to lower pollution levels. Some areas of each map (e.g. the busy intersection in the top middle of the CO panel) demonstrate higher air pollution levels in general, while other areas (e.g. the twisting garden paths in the top left of the CO panel) experience lower pollutant levels in general). The color scales are different for each panel, and should not be compared.

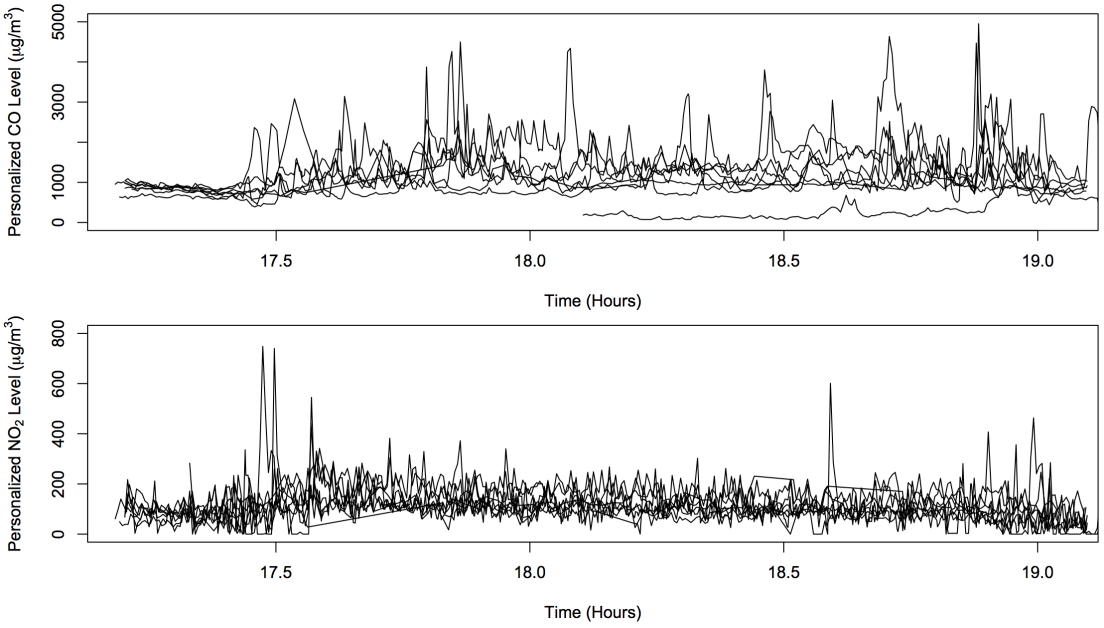


FIGURE 3.3: Time series representations of the personalized sensor data collected on April 2nd, 2015 between 5pm and 7:15pm. The x-axis denotes 24-hour time of day in hours and the y-axis represents pollutant concentration in $\mu g/m^3$.

pollution levels experienced by the different participants. In this case, the red sensor experienced much higher and more variable pollution levels for both CO and NO_2 than were experienced by the yellow sensor. This suggests that some other factor (perhaps geographical dependence) is at play in determining the air pollution exposure experienced by citizens of Jurong East.

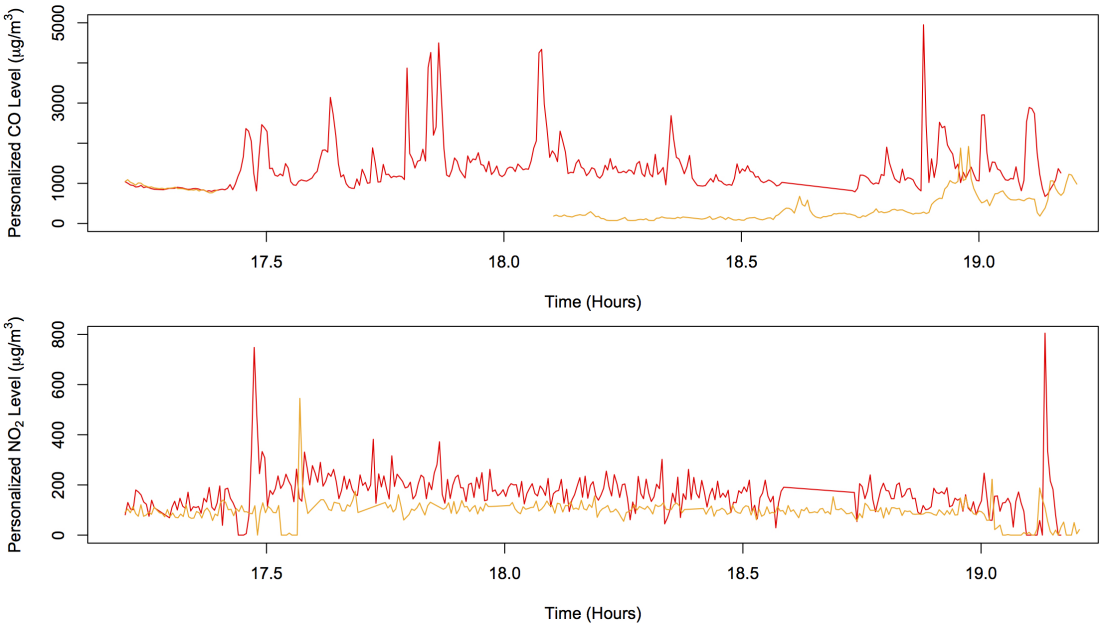


FIGURE 3.4: Time series of two personalized sensors' readings for the time between 5pm and 7:15pm on April 2nd, 2015. The orange sensor had missing CO data for the first hour of the sensor outing. Note the higher values and higher variance of the red sensor compared to the orange sensor for both the CO and NO_2 time series.

3.3 Correlated Factors

This study finds no correlation between air pollution and other air attributes such as air pressure, relative humidity, and temperature (Figures 3.5, 3.6). In fact, linear correlation coefficients for the relationships between CO or NO_2 and temperature, relative humidity, or pressure are all less than 0.060. Also, no correlation was found between CO and NO_2 (correlation coefficient: 0.028), suggesting that CO and NO_2 pollution stem from different sources. Concurrently extreme concentrations of both pollutants are almost never found (Figure 3.7), suggesting further that pollution levels are a result of individual episodes of high pollution of one particular gas (causes of which could be events such as smoking a cigarette or crossing a busy intersection).

3.4 Pollutant Exposure and Inhaled Doses

Air pollution experienced by individuals can be quantified in a number of ways. This section focuses in particular on dosimetry results relating to air pollution exposure and inhaled doses of air pollutants.

Investigation of cumulative exposure over time reaffirms the point made in Section 3.2.2 that participants experienced highly variable pollution levels during their outings. Additionally, it demonstrates that some participants were exposed to higher total amounts

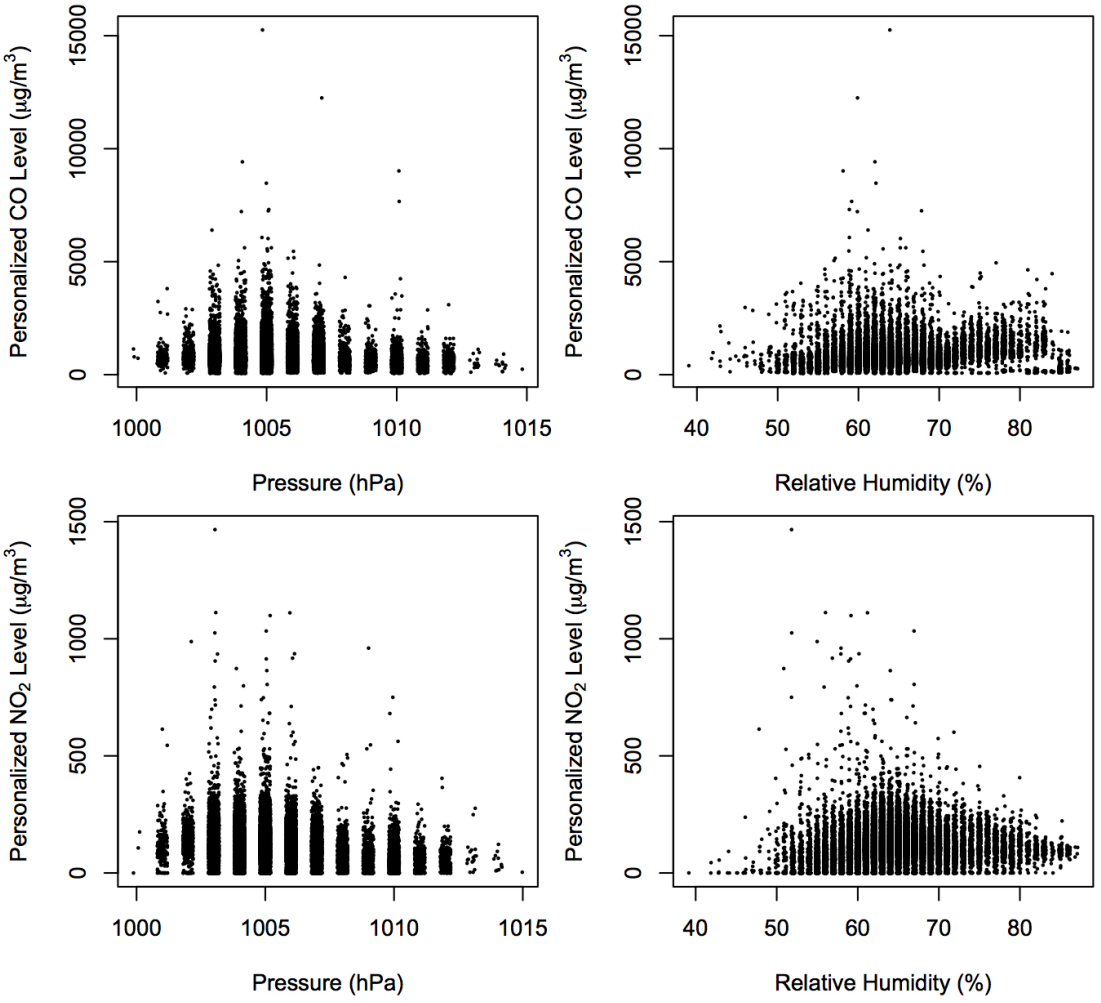


FIGURE 3.5: Scatterplots demonstrating a lack of correlation between gaseous air pollutants (CO and NO_2) and other air characteristics (air pressure and relative humidity).

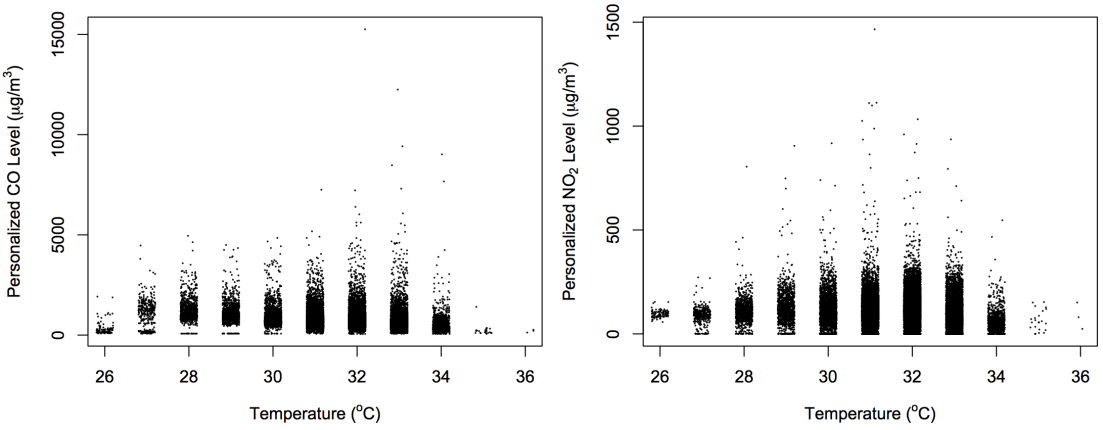


FIGURE 3.6: Scatterplots depicting a lack of correlation between gaseous air pollutants (CO and NO_2) and temperature.

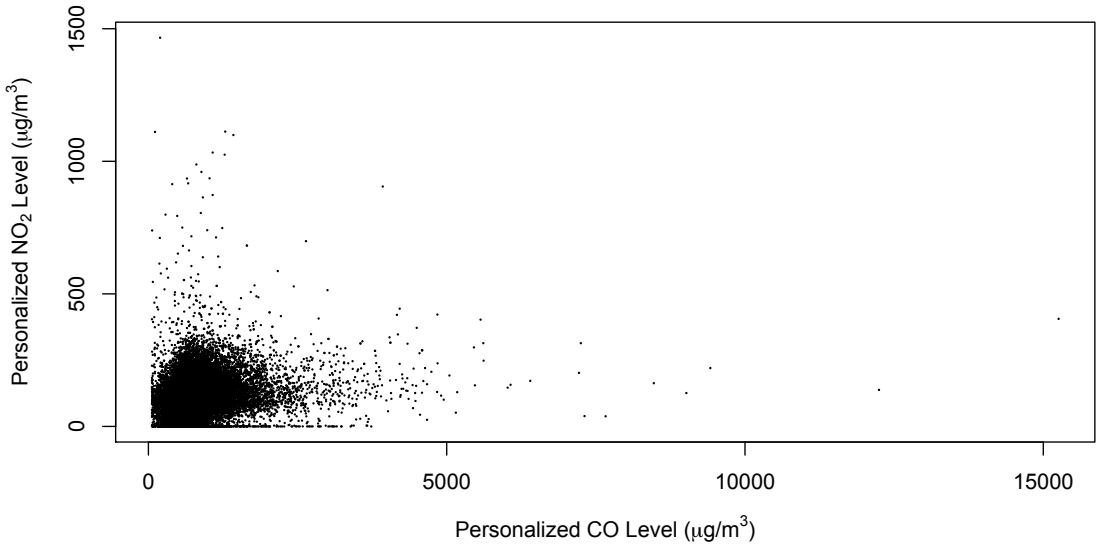


FIGURE 3.7: Scatterplot demonstrating a lack of correlation between CO level and NO_2 level. For typical readings, the data are distributed in a nonlinear clump and do not indicate any correlations. For especially high values of either pollutant, the other pollutant's concentration tends to remain in the typical range.

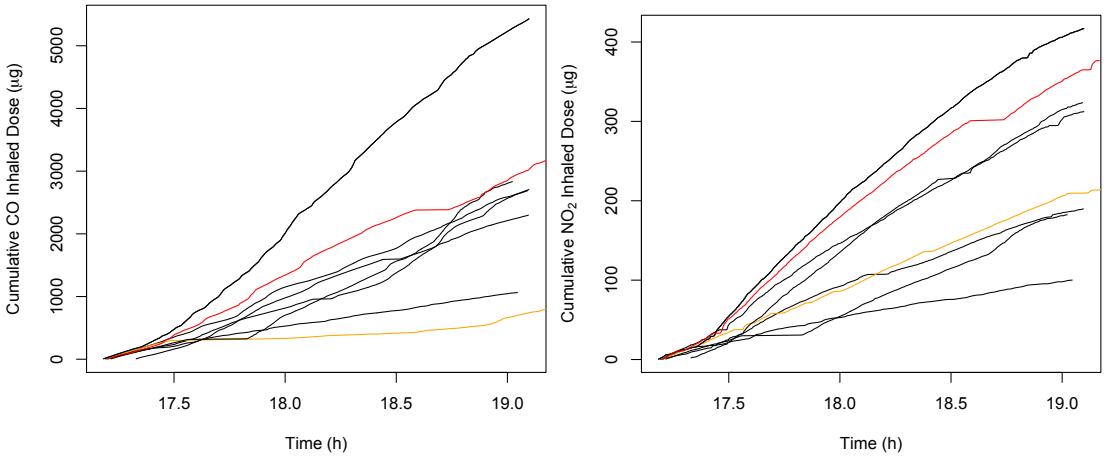


FIGURE 3.8: Graphs displaying the cumulative inhaled doses (in μg) of CO (left) and NO_2 (right) for a walking female during the first set of outings on April 2nd, 2015. Each line corresponds to a different outing taken in this period. The highlighted red and orange cumulative inhaled dose lines correspond to the orange and red exposure lines in Figure 3.4.

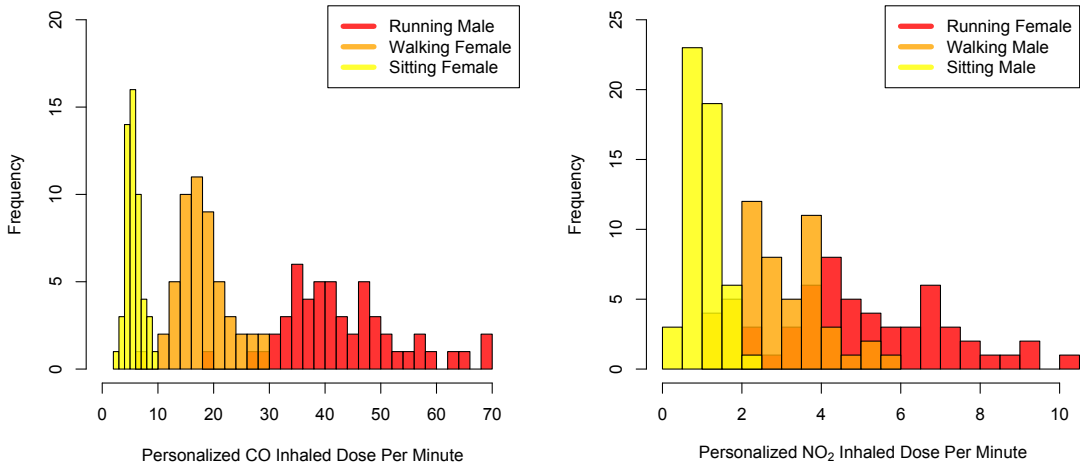


FIGURE 3.9: Histograms demonstrating the distribution of inhaled dose per minute averages for all outings in the Airscapes Singapore study. The collection of CO histograms (left) features the distribution of inhaled CO doses for a running male (red), walking female (orange), and sitting female (yellow), while the graph of NO_2 histograms (right) includes inhaled NO_2 doses for a running female (red), walking male (orange), and sitting male (yellow). The units for both sets of histograms are μg/min.

of air pollution than others (Figure 3.8). However, exposure information does not suffice to predict the total amount of air pollution inhaled by a person, which is a truer estimate of the severity of the health effects that follow air pollutant exposure. As discussed in Chapter 1, it is for this reason that inhaled dose calculations are of particular interest to environmental scientists and public health researchers. Inhaled dose calculations depend on respiratory intake rates, which differ statistically for males and females and for different activities. Males tend to have higher respiratory intake rates than females, and high intensity activities cause higher respiratory intake rates than low intensity ones (McCreddin, 2014). Table 3.1 displays the average inhaled dose rate (in μg/min) for each combination of air pollutant, sex, and activity (sitting, walking, or running). Figure 3.9 illustrates the distributions among all outings from the Airscapes Singapore study of average inhaled air pollution dose per minute. Note that higher intensity activities (whose average inhaled doses are consequently higher) similarly demonstrate more variance than those activities whose low intensity results in low inhaled doses. Mathematically, this is an effect of the inhaled dose model discussed in Chapter 2. With regards to public health, this presents a challenge for determining population inhaled doses without further information about population structure and habits.

TABLE 3.1: Inhaled Doses Per Minute (μg/min)

		Sitting	Walking	Running
CO	Female	5.29	16.95	36.62
	Male	7.32	20.34	40.69
NO_2	Female	0.77	2.46	5.31
	Male	1.06	2.95	5.90

3.5 Introduction to Stationary Government Monitor Data

Air quality data published online[1] by the Singaporean government were collected each day of April 4-10, 2015 by the Airscapes Singapore research team. These published data were the result of 1-hour and 8-hour pollution averages in each of the five regions of Singapore for NO_2 and CO respectively. A drawback of this method for sharing data with the public is that the data presented on the NEA website have low significant figures (with two for NO_2 and only one for CO), making any analyses less precise. In contrast, the personalized sensor data have up to five significant digits for every pollution reading (the only requirement on the personalized sensor pollution data is that the values all be integers). Over the time period of April 4-10, 2015, a total of 1530 readings were collected, 765 each corresponding to CO and NO_2.

Singapore runs fourteen municipal sensors, located in five different regions of the city. Information is not made public about the distribution of the sensors between the regions, and the published air pollution concentration levels are only provided for the regions rather than for each sensor.[2] The five regions appear to experience very different air pollution levels (especially with regards to NO_2 levels), demonstrating both strong heteroscedasticity and highly differing means (Figure 3.10).

To determine whether these differences are statistically significant, a Kruskal-Wallis test is conducted. The Kruskal-Wallis test is a non-parametric analog of the F-test that does not require distribution normality, a necessary requirement since the municipal monitor concentration level distributions are observed to be right-skewed (Figure 3.10). Instead, the Kruskal-Wallis test ranks the readings and uses the known rankings distribution to determine whether the distributions all originate from the same population. While the Kruskal-Wallis test does not require normality, it does require homoscedasticity (equal variances) and independence within and between distributions. Since the data are heteroscedastic in general, this test focuses only on the East, North, and South regions with respect to NO_2, since those distributions are observed to have the closest variances. Independence cannot be guaranteed given the possibility for serial or spatial

[1] http://www.nea.gov.sg

[2] It is not known how the pollution values are calculated within the regions, as this information is not provided on the NEA website.

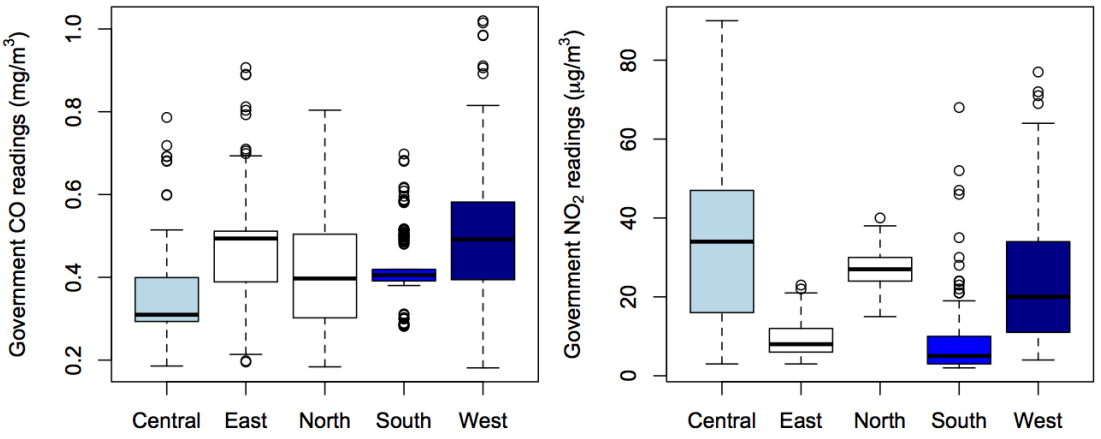


FIGURE 3.10: Boxplots displaying eight days of CO and NO_2 readings (3-10 April, 2015) from the five stationary government air pollution sensors in Singapore, each of which tells a very different story about air quality in the city state. The CO data in this figure are jittered to account for the discrete nature of the government collection, which reports only one significant figure for each CO reading (this jittering does, however, overestimate the variance of the readings) Jurong East is closest to the Central, South, and West sensors.

correlation, but is assumed for the purposes of this test (more discussion on this issue is pursued in Section 4.1). The null hypothesis for this test is that the distributions are all sampled from the same population and the alternative hypothesis is that they are not. The Kruskal-Wallis test returns a p-value of $5.2527*10^{-40}$, indicating that the null hypothesis should be rejected. This implies that the government data sets cannot all be considered representative of every location in Singapore.

In order to most closely approximate the true pollution levels of Jurong East, this analysis uses only the Central, West, and South sensors in its calculations, as those sensors are located relatively equidistantly from the Jurong East neighborhood. Using all sensors would not be appropriate due the observed differences in regional air pollution distribution.[3]

Once the East and North regions have been excluded, the government data consists of 918 readings, split evenly between CO and NO_2. The readings follow right-tailed right-skewed distributions (Figure 3.11). These distributions are numerically justified by noting that the CO mean of 0.42 mg/m^3 exceeds the median of 0.4 mg/m^3 and the NO_2 mean of 22.3 μg/m^3 is higher than the median of 15 μg/m^3. The CO data recorded by the government monitors near Jurong East range from 0.2 mg/m^3 (in the Central and West Regions) to 1.0 mg/m^3 (in the West Region), while the NO_2 data from the same

[3]These differences cannot be statistically verified using the typical method of an F-test or its non-parametric analogs such as Kruskal-Wallis due to the serial nature of the data. Air pollution levels are not randomly distributed in time, but rather depend on the air pollution levels present in previous samples. Hence, the requirement of within-sample independence necessary for most hypothesis tests is violated.

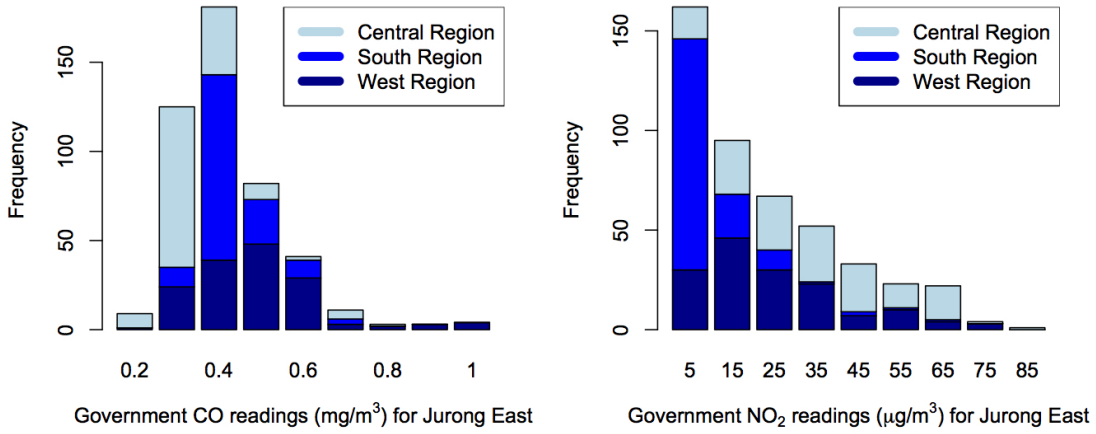


FIGURE 3.11: Histograms describing the government data from stationary sensors in the West, South, and Central Regions that were used to approximate the CO and NO_2 levels in Jurong East during the period of 3-10 April, 2015. The NO_2 data follow a right-skewed distribution, with the highest values coming from the Central Region and the lowest values coming mostly from the South Region. The CO data also follow an right-skewed distribution (though the distribution is much closer to a normal one), with the highest values coming from the West Region and the lowest from the Central Region.

area range from 2 μg/m^3 (in the South Region) to 90 μg/m^3 (in the Central Region). The spatial and temporal variance between the sensors is further discussed in Chapter 4, where it is noted that in addition to differences in the extreme values that have been remarked upon here, the pollutant concentration in certain regions is in fact always higher than in others.

3.6 Summary

Although collected concurrently in the same city, the personalized sensor data and government monitor data differ greatly in their content (Sections 3.1, 3.5). The personalized sensor data set contains more finely tuned spatial and temporal information, and was collected by a mobile, distributed network of sensors as opposed to the stationary municipal monitor. Significantly more personalized sensor data were produced due to the smaller time steps of the personalized sensing technique. Averaging of the municipal monitors also contributed to a smaller government monitor data set (since multiple values were reduced to one average for each published air pollution concentration), and additionally is likely to have reduced the variance of the government data. Substantive data cleaning was required of both data sets, and aberrant data were corrected or discarded (Subsection 3.2.1).

Air pollution concentration levels from both data sets were found to follow right-tailed right-skewed distributions for CO and NO_2. In visualizing the personalized sensor data spatially and temporally, it was discovered that different routes and different areas of Jurong East experienced starkly different air pollution concentrations. Mapping the personalized sensor data enabled location of air-pollution hotspots (3.2.2). For the government monitor data, spatial examination revealed that not all regions of Singapore reflect the same air pollution concentration distribution, so the government data considered in this thesis's analysis were reduced to the subset of regions closest to Jurong East, namely the West, Central, and South regions (3.5).

No correlations were discovered between the air pollutants (CO and NO_2) and weather parameters (temperature, pressure, relative humidity), nor between CO and NO_2 (3.3). Cumulative exposure to and inhaled doses of air pollution were calculated based on the personalized sensor data, and confirm the conclusion that the high variance of pollution concentrations depends on the route taken through Jurong East (3.4).

Descriptive statistics, which have been the focus of this chapter, are often overlooked in favor of hypothesis testing (Chapter 4), but there can exist great benefit to proper treatment of descriptive analysis, as evidenced in this chapter. Visualization, in particular, is helpful in expressing the spatial and temporal content of a data set without having to make any of the assumptions required of hypothesis testing. However, to rigorously address the questions of comparison outlined in the study objectives (Section 1.6), hypothesis testing (Chapter 4) and predictive methods (Chapter 5) are necessary.

Chapter 4

Comparison of Personalized Sensing and Municipal Monitoring Techniques Using Hypothesis Testing

This chapter employs non-parametric hypothesis testing to compare the air pollution data sets collected by the personalized sensors and municipal monitors. Visual inspection indicates that the personalized air pollution data are higher and more variant than the government air pollution data (Figures 4.1, 4.2), motivating statistical tests. Two null hypotheses are proposed for these tests: that the government sampled data and the personalized sensor data were drawn from areas with the same air pollution levels and that the data (in both sets) are independent. The alternative hypothesis is that either the personalized dataset represents higher pollution levels than the government dataset or the independence assumption is violated.

4.1 Parametric vs. Non-Parametric Testing

In total, there were 27 hours in which data were collected from both government and personalized sensors. The personalized data did not cover the entirety of these hours, and there were overlapping personalized data collections during some time periods. The data collection varied spatially and temporally, and in both pollution measurement methods samples were collected in an ordered series.

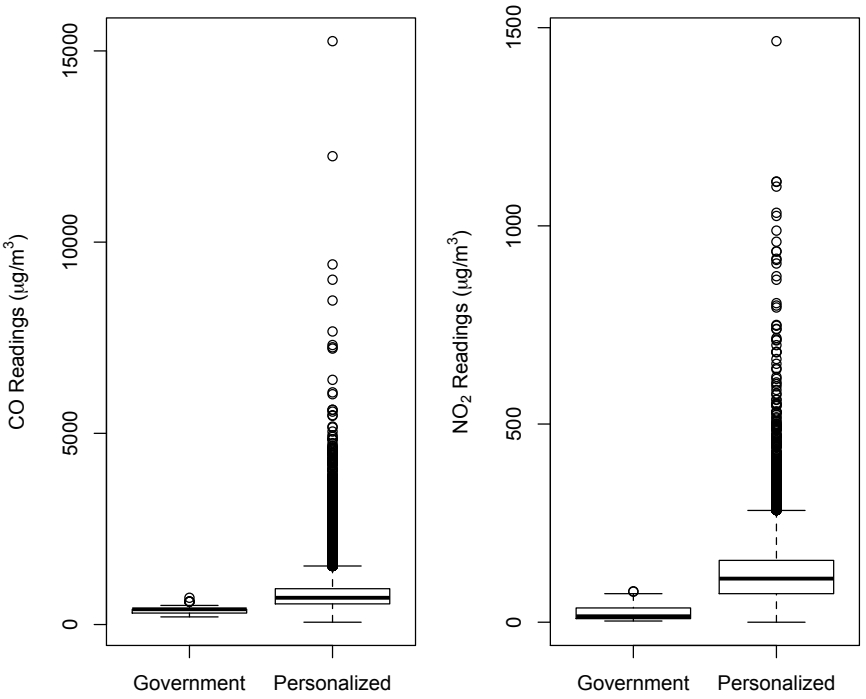


FIGURE 4.1: Boxplots of all CO and NO_2 pollution data. The personalized sensors recorded higher air pollutant concentrations than the government data, both in absolute values and on average, and detected significantly more variance in concentration levels.

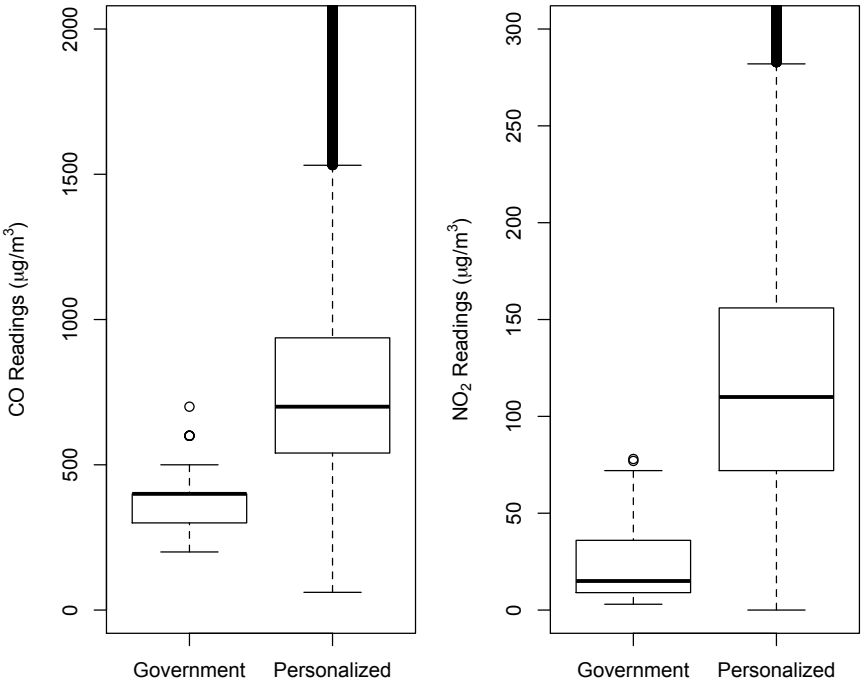


FIGURE 4.2: Zoomed in view of the boxplots of all CO and NO_2 pollution data (Figure 4.1). This closer view of the air pollution data highlights the substantial differences between the air pollution concentrations detected by the government monitors as opposed to the personalized sensors.

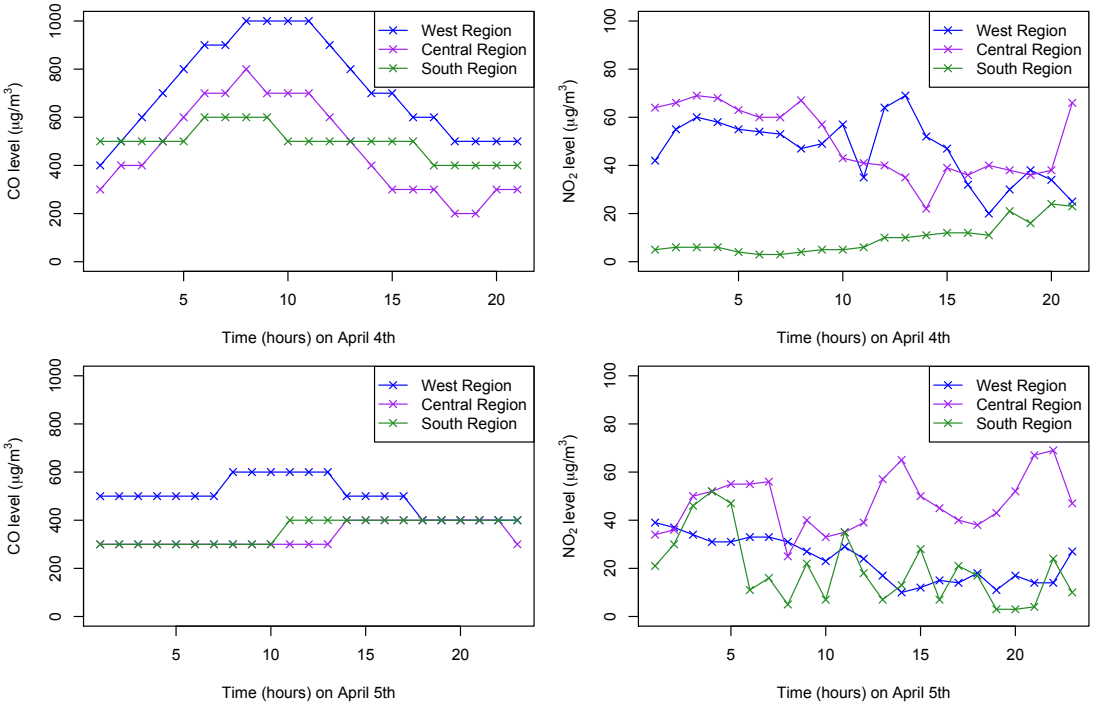


FIGURE 4.3: Plots of CO and NO_2 pollution time series from data collected by West, Central, and South government monitors on April 4th and 5th, 2015. NO_2 readings from the past hour were averaged every hour and CO readings from the previous eight hours were averaged every hour.

This sampling methodology runs the risk of creating serially correlated data sets. Municipally monitored air pollution data trends vary greatly by day (Figure 4.3), and there is no clear indication of consistently serially correlated data. Yet, a within-distribution independence assumption cannot be made without deeper investigation of air pollution dispersion rates. The personalized sensor data raises similar concerns about serial correlation and additional concerns regarding spatial correlation. Due to the changing environment caused by the participants' movements and the observation in Chapter 3 that air pollution levels can change quickly as one moves about a relatively small area, these data are possibly independent. Further investigation of pollution dispersion in environments like Singapore's is needed before making claims regarding this data's independence.[1] Hence, the independence assumptions necessary for hypothesis testing are not met and must be included in the null hypotheses of the tests.

Parametric tests are a popular choice for statistical analyses due to their computational efficiency and ability to make accurate predictions when their underlying assumptions are closely met. These assumptions generally include distribution normality, which is not met for either the personalized sensor or municipal monitor data. Hence, instead of using parametric tests to answer the question proposed in this chapter, two non-parametric

[1]Note that in future studies, a time series analysis of the data could be conducted to address this question.

hypothesis testing methods that do not rely on this assumption will be implemented, namely permutation testing and Wilcoxon rank-sum testing.

4.2 Permutation Test

The first test used is a permutation test using the difference in means of the sample distributions as the test statistic. Since there are too many data points to run a permutation test efficiently (permutation tests run in asymptotic factorial time on the number of observations), a subset is sampled from the set of possible permutations of the set of partitions of the data into group A (simulated government data) and group B (simulated personalized data). Only partitions that maintain the sizes of the original groups are considered in the analysis.

For example, suppose there are two government data points, a_1 and a_2, and three personalized data points, b_1, b_2, and b_3. A possible permutation might be the following:

Group A: a_1, b_3
Group B: a_2, b_1, b_2

In the CO permutation test, the number of personalized data points used was 27,083 and the number of government data points used was 54 (i.e. Group A had size 54 and Group B had size 27,083). In the NO_2 permutation test, there were 27,067 personalized data points used and 54 government data points used.

For each sample permutation, the test statistic is the mean of the group B values subtracted by the mean of the group A values. These statistics are used to construct a baseline distribution. The p-value is then computed as the proportion of simulated statistics in the baseline distribution that are more extreme than the true value. A left-sided p-value is used to reflect the one-sided alternative hypothesis that the mean of the personalized data is greater than the mean of the government data.

The p-value for a 100,000-partition permutation test on the CO data is 0. The p-value for a 100,000-partition permutation test on the NO_2 data is 0. The p-values for a full permutation test are below the sensitivity of this sampled permutation test, even with 100,000 simulated permutations, resulting in the zero-valued simulated p-values (Figure 4.4). These p-values indicate that the null hypotheses should be rejected in favor of the alternative hypothesis that the personalized data distribution has a higher mean than the government data distribution or that the independence assumption does not hold.

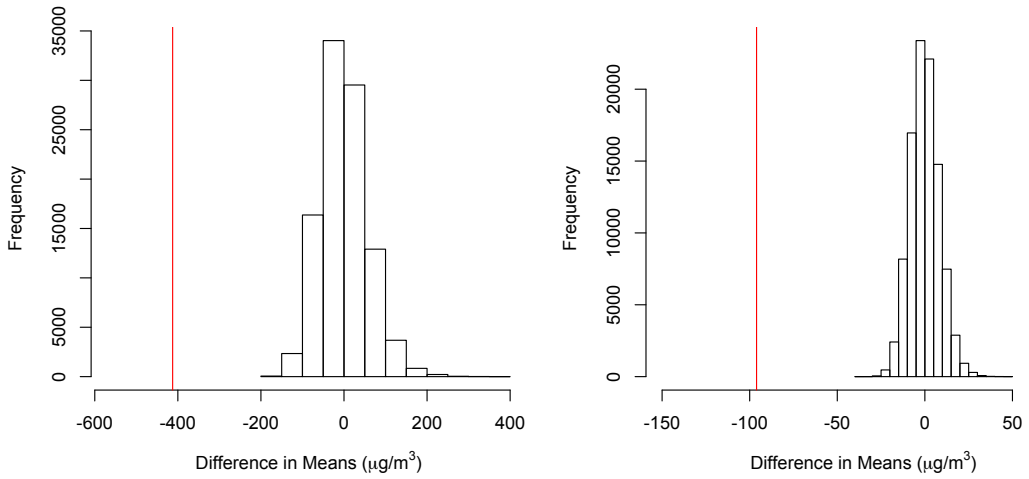


FIGURE 4.4: Histogram distributions of the difference in means in the CO (left) and NO_2 (right) permutation tests. The red lines indicate the real difference in means between the government and personalized data.

4.3 Wilcoxon Rank-Sum Test

Since the personalized sensor pollution readings follow right-skewed distributions, for each pollutant there exists a high number of extreme outliers (Figure 4.5).

Taking means of samples with outliers can misrepresent what a typical value looks like in the sample distribution. This problem is addressed translating the values into ranks and running Wilcoxon rank sum tests on the government and personalized pollution data for CO and NO_2. For the CO and NO_2 tests, the p-values are found to be $4.5305*10^{-33}$ and 0.08332 respectively.

The CO p-value again implies that the null hypotheses that the two data collection methods represent the same data should be rejected. However, the NO_2 p-value does not meet the standard 0.05 significance cutoff, so the test is inconclusive. Closer examination of the personalized sensor NO_2 data revealed that an irregular number (10 times the amount for any other value) of NO_2 concentration readings were zero (Figure 4.6), implying likely missing NO_2 concentration values in the personalized data set. When these irregular values were removed, the resulting p-value for the Wilcoxon rank sum test was $1.2655*10^{-44}$, supporting the conclusion from Section 4.2 that the null hypotheses should be rejected in favor of the alternative hypothesis.

4.4 Dispersion Analysis

Sections 4.2 and 4.3 demonstrated that the government and personalized data describe two different populations with differing means, but their variances have not yet been

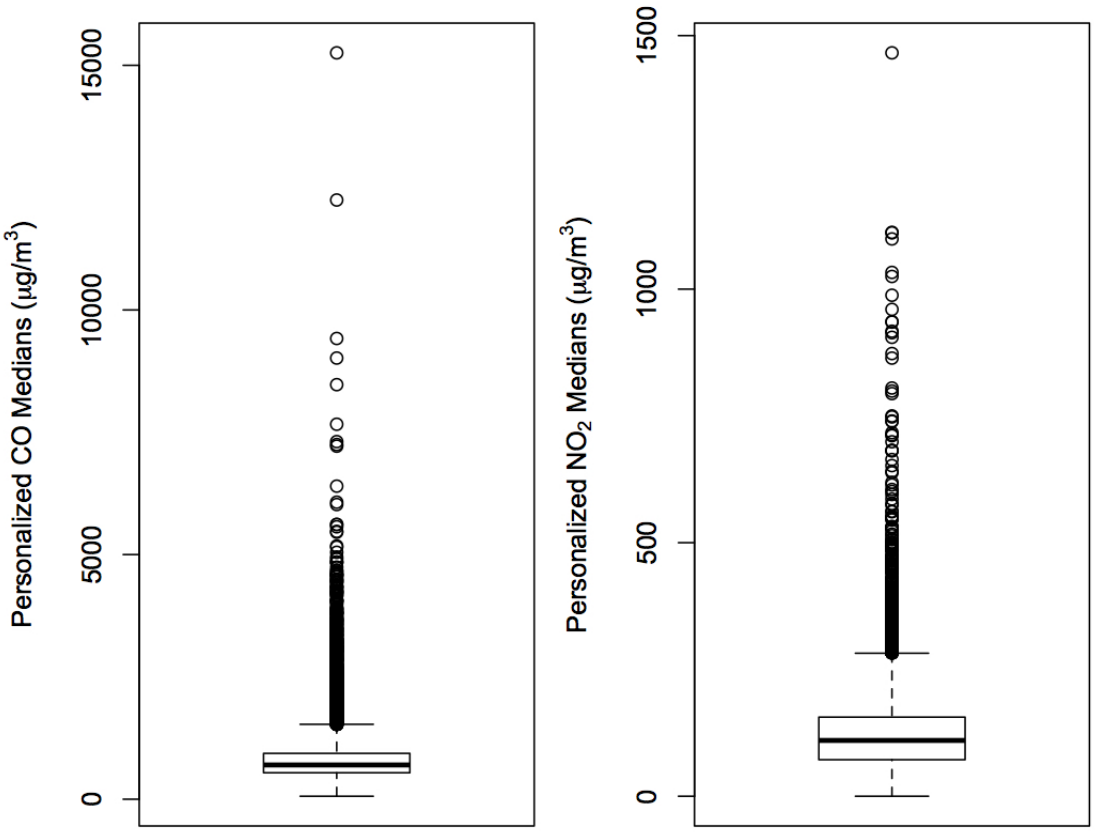


FIGURE 4.5: Both the CO and NO_2 personalized sensor readings follow right-skewed distributions, and contain a high number of outliers.

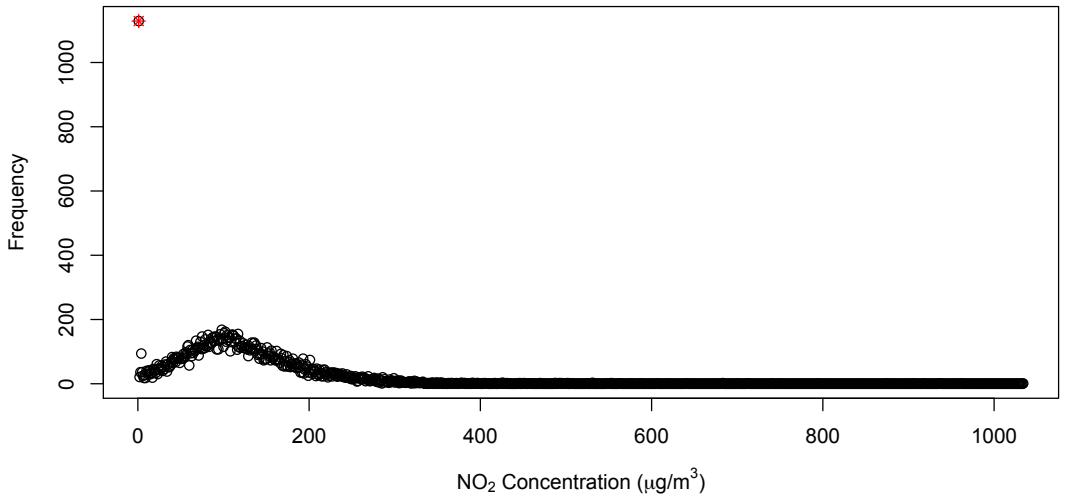


FIGURE 4.6: Graph depicting the frequency of each NO_2 concentration level observed by the personalized sensors for 4-10 April, 2015. The aberrant zero count is highlighted in red.

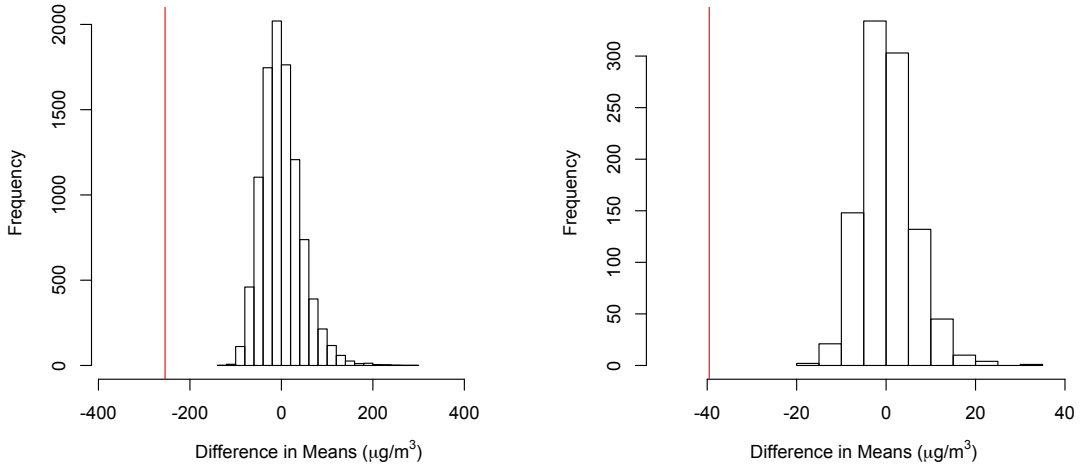


FIGURE 4.7: Histogram distributions of the difference in means of the absolute differences from the sample means in the CO data sets. The red line indicates the real difference in mean absolute differences for the government and personalized data.

discussed. Using the same permutation and rank sum tests, one may examine whether the variation in the populations follows the same trend. The null hypothesis in this case is that the variations in the two populations have the same distribution and that the variances are independent, and the alternative hypothesis is that the personalized data describes a population with a higher variance or that the variances are not independent.

The observations for the variance permutation test are the absolute differences between each data point and its population mean. The permutation test statistic is the mean of the simulated absolute differences from the personalized sensor data subtracted by the mean of the simulated absolute differences from the government monitor data. Again, a left-sided p-value is computed (to fit the alternative hypothesis), which turns out to be 0 for both the CO test and the NO_2 test (both 10,000 permutations). Hence, the null hypotheses are rejected in favor of the alternative hypotheses that the absolute differences from the means (and hence the variances) in the personalized data set are higher than the absolute differences in the government data (Figure 4.7) or the variances are not independent.

For the Wilcoxon rank sum test, absolute differences between the data points and their respective sample means are ranked and rank sum tests on these new data sets are conducted. The resulting left-sided p-values are $3.8751*10^{-30}$ for the CO test and $1.5002*10^{-23}$ for the NO_2 test. This is sufficient evidence in both cases to reject the null hypothesis that the populations have the same variance and that the variances are independent and instead embrace the alternative hypothesis.

4.5 Summary

This section proposed to answer the question of whether or not the personalized sensors and and municipal monitors had sampled from areas with the same air pollution distribution. Justification was given for use of non-parametric testing (namely permutation testing and Wilcoxon rank sum testing) to answer this question. Using these non-parametric tests, it was possible to determine that either the mean and variance of the personalized data were higher than the corresponding values for the government data or that the independence assumptions of the tests were violated by the data. The tests used in this section cannot fully separate the two branches of this conclusion, and further investigation into the spatial and temporal facets of these data (Chapter 5) is needed to solidly determine whether the personalized data are statistically higher and more variant than the government data.

Chapter 5

Spatial Analysis

In this chapter, random walks on a spatially interpolated field generated by personalized sensor data are used to generate new estimates of personal exposure. These values are then compared to estimates derived from the municipal monitor data and non-interpolated personalized sensor data. It is discovered in both comparisons that the interpolated personalized data yield higher concentration estimates.

The Singaporean government's stationary sensors do not have the capacity to describe local differences in air pollution within Jurong East. On the other hand, mobile sensing is uniquely powerful in that it intrinsically records spatial metadata using GPS, Wi-Fi, and cellular data transmission and includes this information with the air quality readings. The hypothesis testing conducted so far in this thesis (Chapter 4) has highlighted differences between the personalized sensor data and government monitor data, and demonstrated a need for examination of the spatial and temporal qualities of the air pollution data.

Spatial interpolation can be used to estimate pollutant concentrations across an area and over time using one or more precisely located data points. The personalized sensor observations yield highly localized readings of air quality, and are plentiful, making this data ideal for accurate spatial interpolation, since more data leads to more accurate estimates. This chapter compares these spatially interpolated estimates to the spatially-ignorant government monitor data and to the artificially spatially-ignorant personalized sensor data (i.e., the personalized sensor data as they were analyzed in Chapter 4 without consideration to within-neighborhood location).

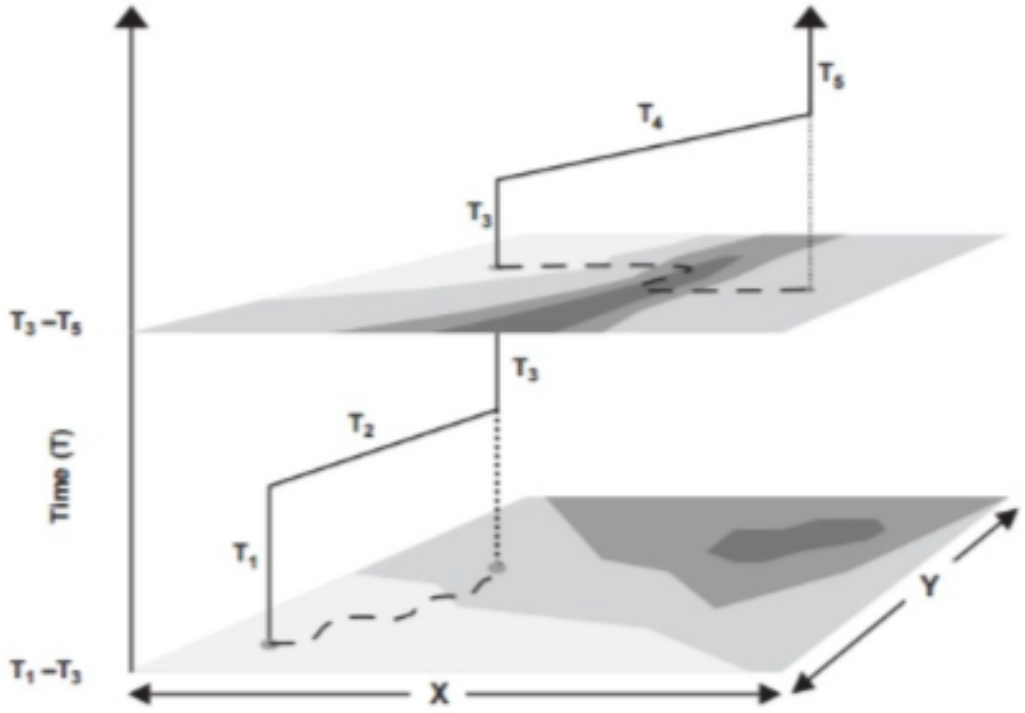


FIGURE 5.1: Example of a path through a spatially interpolated field with two location dimensions and one time dimension. In this path, a timestep T_n represents either a step parallel to the plane defined by the location axes or a step parallel to the time axis.(Nieuwenhuijsen, 2015)

5.1 Spatial and Temporal Interpolation to Produce Air Pollution Concentration Field

A subset of the personalized sensor data was used to construct an interpolated field on Jurong East.[1] This subset comprised the personalized sensor data collected on April 4th, 2015 from 16:54:21 to 19:10:41, and was chosen arbitrarily from the collection of subsets whose time intervals matched entirely with a corresponding timespan in the government data. This data was then interpolated on a $25 \times 25 \times 410$ grid across Jurong East and across the 2.25 hour timespan on April 4th using a Gaussian process called kriging. Spatially, each 25×25 slice of the grid represented a 2D footprint of the pollution levels over Jurong East. Temporally, the grid was divided into 410 timesteps, each 20 seconds apart.

[1] Only a subset of the data were used in this interpolation for efficiency reasons, as kriging (the process of interpolation described in this section) is computationally expensive.

5.1.1 Kriging

This subsection gives a brief overview of the process of ordinary kriging used to interpolate carbon monoxide values for the field described above using personalized sensor data. The kriging calculations were carried out by Kevin Li, an undergraduate student at MIT.

Kriging is a method of spatial interpolation that, given a set of known values with corresponding coordinates in a multidimensional space, estimates scalar values. In the case of this study, the known values are carbon monoxide concentration readings collected by a personalized sensor. Kriging assumes normality of the population from which the known values are drawn. While the CO values sampled by the personalized air pollution sensors on April 4th are log normal, this is largely due to a small collection of extreme values (Figure 5.2), and the distribution is in fact close to normal. Thus it is assumed for the purposes of spatial interpolation that the normality conditions for ordinary kriging are met.

Firstly, a regression function on pollutant concentration in the field is constructed from weighted averages of known pollutant concentration values in the time and space intervals that bound the interpolation field. In an ordinary kriging interpolation, as used in this study, the regression function is not initially known, but is required to be a constant function. The residuals for this regression are assumed to be normally distributed and are minimized by the regression process. Using the residuals from the known concentrations, a Gaussian process centered at 0 with variance and correlation matrix equal to those of the distribution of known residuals is constructed. It is from this distribution that interpolated values are assigned to the grid points at which concentration is not known (hence the initial requirement that the data are normally distributed). Finally, once the residuals for each point are known, the estimated CO concentration values are computed.

5.2 Random Walks Through CO Concentration Field

A random walk on an interpolated field over Jurong East produces another estimate of air pollution exposure, and iterations of these walks generate a distribution of estimates that one can use to predict the exposure a person would experience walking through the area.

As discussed in 5.1, the data were interpolated over a 25×25 grid for 410 distinct time steps. Though the times and grid point locations correspond to actual times and locations in Jurong East, a random walk can be constructed in abstract form in order to

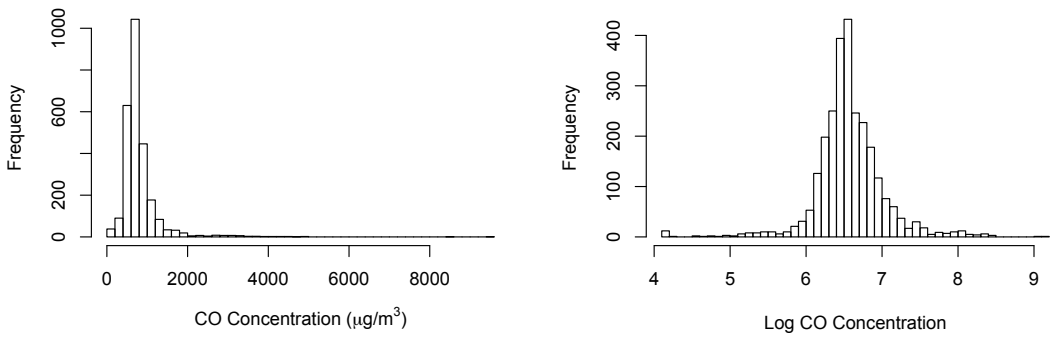


FIGURE 5.2: Histograms of CO distribution (left) and log CO distribution of April 4th personalized sensor data. The CO distribution is log normal based on the normal distribution of the log CO distribution, but is still relatively close to normal itself.

generalize this analysis to any time period and any grid on some rectangular neighborhood. It would be possible to construct a specific non-rectangular grid to reflect the walking paths in Jurong East, which would better estimate the exposure of a real person walking in the area. This type of non-rectangular grid would be useful for making predictions about the exposure a person would experience on their commute through Jurong East, but adds a great deal of complexity to the model. For the purposes of this general analysis that seeks to determine differences between background and local spatial sensing, such a specific grid is not necessary, and a rectangular grid is used for simplicity.

The rules of a random walk in this study are designed as follows:

1. The walker begins at a random point on the 25×25 grid.
2. The walker may not leave the 25×25 grid at any time.
3. At each timepoint, the walker may choose to remain on the same grid point or leave.
   (a) If the walker is not on an edge of the grid, they randomly and uniformly choose[2] two numbers $\Delta x, \Delta y \in \{-1, 0, 1\}$. A 1 corresponds to a step one grid point South and East respectively, a -1 corresponds to a step one grid point North or West respectively, and a 0 indicates that no step should be taken along that axis for the time point in question.
   (b) If the walker is on an edge of the grid, they are either in a corner or not in a corner. If they are not in a corner, then they follow the directions in

[2]It is possible to weight the steps such that the walker moves one way with more probability than others, but this adds complexity to the model unnecessary for the analysis at hand.

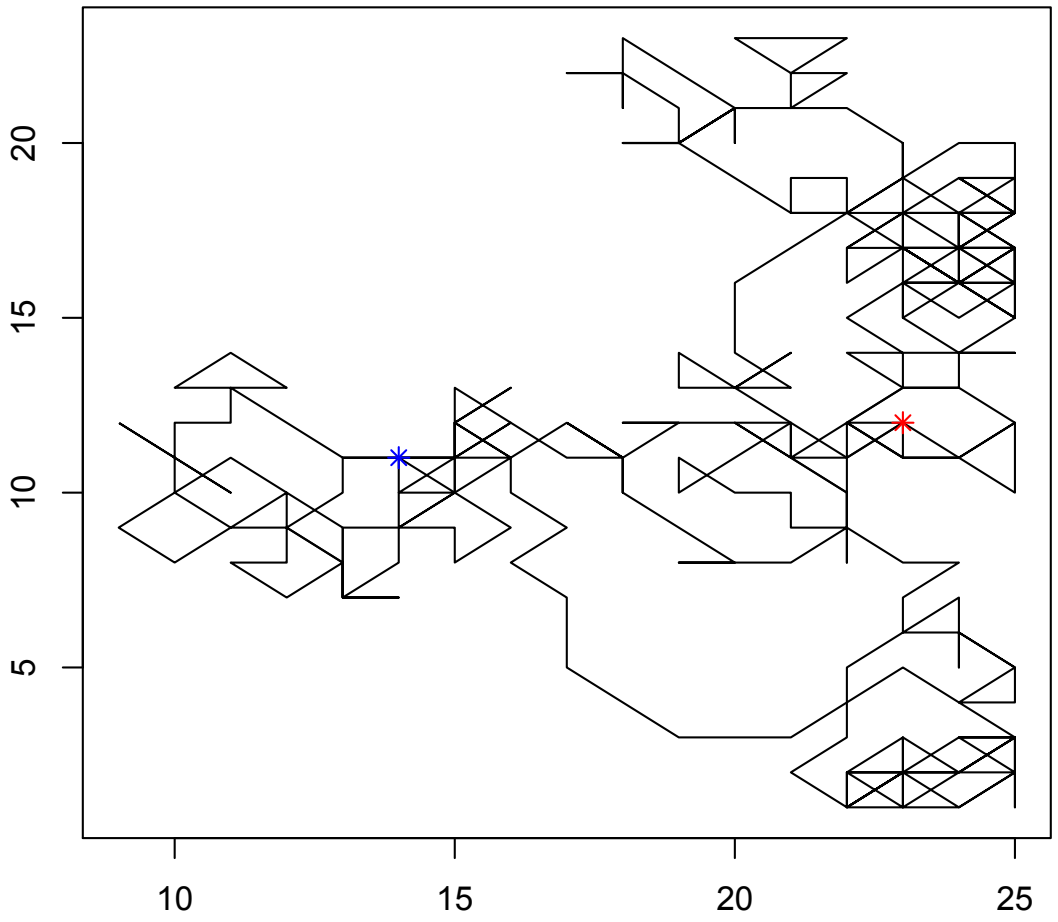


FIGURE 5.3: A possible random walk path on a 25 by 25 grid. The starting point is indicated in red and the ending point in blue.

step 3a for the axis (North-South or East-West) parallel to the edge. For any axis perpendicular to an edge, the walker randomly and uniformly chooses $q \in \{0, 1\}$. If $q = 0$, the walker does not move along this perpendicular axis. If $q = 1$, the walker moves one grid point away from the edge along the perpendicular axis.

The random walker keeps track of their random path through the abstract grid (Figure 5.3). Then, using these path coordinates and their corresponding time step values, a concentration vector containing the carbon monoxide levels that the walker would have experienced were the abstract random walk relocated to the time-dependent, spatially-interpolated air pollution field on Jurong East is calculated using the stored path coordinates and interpolated values. The predicted total exposure for the walk is then calculated as a Riemann sum with time steps 20 seconds apart, where each time step corresponds bijectively with a value in the concentration vector.

Example:
Suppose a 25-sided die is rolled twice and the results are 9 and 23 respectively. A walker is placed on a 25 by 25 grid at the gridpoint that is located 8 steps South and 22 steps East of the Northwest corner of the grid. The walker might take the first ten steps listed in Table 5.1.

Step	x-Coordinate	y-Coordinate	CO Level
1	9	23	869.00
2	8	24	895.25
3	7	24	923.80
4	8	25	939.17
5	9	25	951.29
6	10	25	953.38
7	10	24	950.67
8	10	25	947.70
9	10	25	947.09
10	10	24	945.08

TABLE 5.1: Example random walk coordinates and corresponding CO levels.

In the first step, the walker is placed at (9,23). In the second step, the walker moves diagonally Northeast by one step to (8, 24). In the third step, the walker moves North by one step to (7, 24), and so on. Note that in the ninth step, the walker does not move, but moves again in the tenth step. Consulting the data frame containing the interpolated personalized sensor data, CO levels are found for the first ten time and location points (Table 5.1).

Since each time step is 20 seconds apart, the total carbon monoxide exposure for the first ten steps is summed to be:

$$\begin{aligned}&(869.00 + 895.25 + 923.80 + 939.17 + 951.29 + 953.38\\ &\quad + 950.67 + 947.70 + 947.09 + 945.08)/3\\ &= 3107.477\mu\text{g*min/m}^3.\end{aligned}$$

5.3 Comparison of Random Walks and Spatially-Ignorant Estimates

This section investigates the effectiveness of spatial interpolation and random walks in providing new information about pollutant exposure compared to the government monitor data and to the personalized data as analyzed without consideration of its spatial components. To do so, a collection of 10000 random walks on the spatially interpolated field were run, yielding a right-skewed distribution of total CO exposures

(Figure 5.4). Next, the CO total exposure values for the spatially-ignorant personalized data and government data were calculated so that they could be compared to the data generated by the random walks.

In order to calculate the best spatially-ignorant personalized total exposure estimate to compare to the random walks, the non-interpolated personalized sensor data were narrowed to the 16:54:21 to 19:10:41 time period on April 4th, 2015. This data subset was then partitioned into 410 bins matching the 20-second time steps used in the spatial interpolation. These bins were averaged, and the CO concentration for the total time period was plotted as a step function with respect to time. This function was then integrated to produce the non-interpolated, personalized CO exposure estimate of 107949.1 μg*min/m^3.

The government monitor CO estimate for the April 4th time period is calculated similarly, though only four bins were used. Each bin corresponded to one of the hours from 4pm to 7pm inclusive on April 4th, 2015, and contained the pollution readings from the South, West, and Central government monitors. These bins were then averaged and the number of spatial interpolation time points per bin were counted (the first bin contained 17, the second and third bin contained 180 each, and the fourth bin contained 32). Using this information, the total CO exposure estimate was calculated to be: (17*466.6667 + 180*433.3333 + 180*366.6667 + 32*366.6667)/3 = 54555.56 μg*min/m^3.

Both the government monitor estimate and the spatially-ignorant personalized sensor estimate fall below all 10,000 estimates produced by the random walks on the spatially interpolated field (Figure 5.4). It was shown in Chapter 4 that the government monitor pollution readings tend to be much lower than the personalized sensor readings in general, so the government estimate being lower than an estimate derived from personalized sensor data is not surprising. However, it is quite interesting to note that the random walks estimates are unanimously higher than the the spatially-ignorant personalized sensor estimate. Hence, using random walks on spatially interpolated fields produced by the personalized sensor data is valuable in that it provides new context for CO distribution and citizen exposure in Singapore.

5.4 Summary

The personalized sensor data were used to interpolate a CO concentration field with respect to latitude, longitude, and time. A set of 100,000 random walks through this field was then generated, and, for each walk, the cumulative CO exposure of an individual tracing the walk through the concentration field was calculated. The distribution

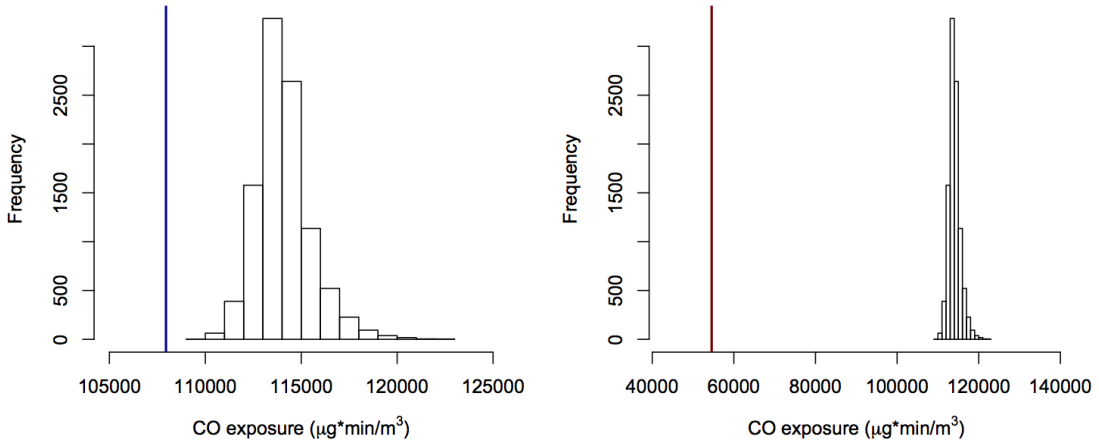


FIGURE 5.4: Histograms comparing the random walk data to the raw sensor data. The histograms pictured in both the left and right panels describe the distribution of the total carbon monoxide exposure calculated by taking random walks on the spatially interpolated field over the April 4th time period. The blue line indicates the carbon monoxide exposure calculated from the raw personalized sensor data without regards to position and the red line indicates the CO exposure calculated from the spatially-ignorant government data whose time steps were hours rather than 20 second chunks (as were the time steps for both kinds of personalized data).

of these random walk-generated cumulative exposure estimates was then compared to a cumulative exposure estimate for the same period derived from means of the government data, and to an analogous mean-derived estimate calculated from the personalized sensor data. In both cases, the estimates produced by the random walks through the interpolated CO field were uniformly higher than the mean-derived estimates.

Chapter 6

Discussion

This study investigated the extent to which personalized sensing of air pollution data collection provides pertinent information regarding individual exposure to air pollution in Singapore beyond the data from the presently-used stationary municipal monitors. Descriptive statistics, hypothesis testing, and data interpolation were used to quantitatively evaluate this question. Each of these analysis methods indicated that the personalized sensors provided new data about local air pollution concentration levels not observed by the government monitors.

Personalized sensor data were collected as part of the Airscapes Singapore study using a moving, distributed network of sensors, enabling many observations to be made concurrently and providing real-time, geotagged data to the Airscapes Singapore database. Municipal monitor data were extracted from the NEA website, which uploads retrospective pollution level averages every hour. These data sets then underwent data cleaning in which aberrant values were processed or eliminated. Each of the CO and NO_2 distributions from the government monitor and personalized sensor data were found to follow right-skewed log-normal distributions. Within each data set, air pollution concentration logs varied in magnitude depending on the locations where they were collected. Within the personalized sensor data, no correlations between air pollutants and temperature, relative humidity, and air pressure were found. The personalized sensor data were additionally used to calculate cumulative exposure to and inhaled doses of air pollution.

Comparison of personalized sensor data and municipal monitor data using non-parametric hypothesis tests indicated that the means and variances of the CO and NO_2 distributions of the personalized data were likely higher than those of the government data. A spatial and temporal interpolation of CO concentrations was constructed using a section of the personalized sensor data and random walks on this interpolated field were generated. These random walks were used to calculate cumulative exposure estimates that were

then compared to cumulative exposure estimates computed from the municipal monitor data and raw personalized sensor data without regard to spatial variability. The random walk exposure estimates were substantively higher than both exposure values that were calculated independently of locational consideration, confirming that the personalized sensor data were higher than the government data and demonstrating the importance of spatial and serial considerations in estimating air pollution exposure.

These findings demonstrate the potential of personalized sensor data collection and analysis methods in providing novel, detailed information about the state of air pollution on a localized level in Singapore. In particular, this analysis of the personalized sensor data gives a more specific view of how an individual commuter's experience of air pollution changes over time. This detailed analysis opens the door to an array of applications by researchers, citizens, and city officials to effect optimal health outcomes for the city's residents.

6.1 Applications

The personalized sensing approach is advantageous in that it collects air pollution data in high spatial and temporal resolution. This enables spatial interpolation and random walk analysis (Chapter 5) and provides helpful context when making the assumptions necessary for hypothesis testing (Chapter 4). Additionally, the unprocessed personalized sensor data (as opposed to the pre-averaged municipally monitored data) demonstrates more clearly the variance in pollution levels over small distances (Chapter 3). The results of cumulative exposure, inhaled doses (Chapter 3), and random walk analyses can be applied in reducing errors in air pollution related health impact predictions in individual commuters

Predictions about the health impacts of air pollution at a statistical and individual level can be used to forecast the health burden inflicted by air pollutants on Singapore and to optimize urban air pollution and transportation policies. Applications range from use in further research studies, to city-wide regulation ordinances, to neighborhood traffic reorganization. Detailed air pollution data is informative for researchers of pollution exposure science, environmental epidemiology, and urban development.

Real-time, spatially-specific information is of particular interest to smart-city research and the engineering of tools and services for city technologization. This research can be used to provide citizens of urban areas with constantly-updating information about evolving pollution hotspots. Armed with this knowledge, citizens would have the ability

to change their commuting behavior to minimize personal exposure to harmful air pollution and, subsequently, avoid the related negative health effects of such exposure. For example, these data were used to develop a commuter exposure assessment tool, where citizens of Jurong East, Singapore can observe the predicted cumulative exposure values for a set of predetermined paths through their neighborhood.[1]

Urban planning can be improved by the information provided by a distributed network of personalized sensors. Municipalities can use the real-time personalized concentration data to develop responsive strategies to reduce sources of air pollution. For instance, traffic emissions can be modified through use of adaptively-timed traffic lights and smart-city applications that reroute traffic in order to minimize instances of acute air pollution. Businesses or roads which are determined to be major sources of pollution can be moved away from areas that contain schools, are thickly settled, or are frequently used for major outdoor special events. A possible way to achieve this would be to implement congestion charging on high-polluting vehicles and businesses. Based on the data from the personalized sensors, low and ultra-low emissions zones could be identified, and charges, proportional to their pollutant emissions, would be levied on polluters within those zones. Alternatively, non-vehicular commuters could be moved away from sources of high air pollution by rerouting cycling lanes and pedestrian walkways to streets with lower air pollution. Using the personalized sensor air quality data, urban planners can optimally design and manage more sustainable cities, with a particular focus on improving public health.

Personalized sensor data can be used to improve the assessment of air pollution concentration levels on a city, neighborhood, or street level, which can inform regulation policies and municipal public health goals. Since pollution monitoring using personalized sensors focuses on highly localized areas, presently unknown sources of air pollution can be identified and targeted in pollution reduction efforts. Public health officials can better judge personal exposure using these sensors, and can publish quantified recommendations to clinics and hospitals about which air pollutants pose the biggest threat to the population so that these institutions are prepared to train their nurses accordingly. Additionally, medical professionals can use knowledge of air pollution hotspots derived from personalized sensor data to anticipate whether a patient may be suffering from pollution poisoning based on their areas of residence and work as well as their commuting strategy. Using these mechanisms, air pollution prevention and treatment of air pollution-related illnesses can be optimized using a personalized sensor data collection and associated spatial and personal exposure analyses.

[1] http://eoe.airscapes.io

6.2 Future Work

From a mathematical standpoint, several future directions are proposed. A time series regression analysis of the data would reduce the need to make assumptions about serial correlation in hypothesis testing. Similarly, kriging of log concentration would change the requirement for distribution normality to one of distribution log normality, enabling use of interpolation on more intervals in the data. If this sensing method were adopted by the Singaporean government, then an additional direction would be the development of a fast algorithm for real-time spatial interpolation of concentrations across Singapore, since ordinary kriging relies on static data and is computationally inefficient.

Future work from an urban planning perspective must include an analysis of how this data can be used to benefit underserved populations. Although there is great advantage to using personalized sensor data to determine which areas of Singapore are most prone to poor air quality, irresponsible dissemination of this data runs the risk of exacerbating cycles of poverty among low-income citizens. Were this information to be released without embarking upon concerted efforts to improve air quality in areas with high pollution, these neighborhoods would likely experience property devaluation. As highly-polluted areas became poorer, income inequality would increase and economic mobility of the low-income population in these neighborhoods would become more difficult. The consequences of this inequity would include continued greater pollution exposure, increased risk of illness and premature death, and an exacerbated struggle to find and retain employment for these already underserved populations. Environmental justice dictates that researchers, urban planners and city authorities involved in leading smart-city initiatives have a responsibility to prioritize pollution reduction in economically struggling areas in order to avoid such outcomes.

Chapter 7

Conclusions

The purpose of this study was to determine whether using a distributed network of personal sensors to characterize air quality in Singapore contributes information and understanding relevant to environmental health that supplements current governmental air quality measurement methods. To answer this question, air quality data collected by the stationary government monitors and by a distributed network of moving, personalized sensors were compared.

Computational and data cleaning techniques were successfully applied so that both data sets could be compared.

Descriptive statistics and data visualization techniques were used to illustrate how the unique temporal and spatial characteristics of personalized sensor data contributed novel, actionable information about the state of air pollution in Singapore, particularly at a localized level, that can be used to quantify and improve personal exposure to air pollution. Cumulative air pollution exposure and inhaled doses were calculated based on recordings from personalized sensors. This information provided context for a discussion of potential public health analyses and municipal policies to improve air quality using the localized personal sensor data. Government-collected air quality data were compared to personalized sensor-generated air quality data using relevant statistical tests, spatial interpolation techniques, and data visualization. The results of these comparisons were discussed, and it was determined that the air pollution concentration level distributions recorded by the two collection methods were fundamentally different.

Personalized sensor data were found to be highly effective in contributing additional information about air pollution levels in Singapore. It has been demonstrated that the personalized sensor data were higher and more variant in concentration level than the data registered by the stationary government monitors. Temporal and and spatial

analyses of the personalized sensor data were used to establish a more comprehensive picture of air quality that could not have been manifested by the stationary, hourly-updated government website. This thesis concludes that personalized sensors are an effective tool for use by Singaporean city officials to estimate localized pollution levels in real time and respond to specific air pollutant threats effectively and for use by individuals to make informed choices about their personal exposure to harmful, invisible pollutants. Development of megacities demands increased technologization of public health in order to a ensure future sustainability, and personalized pollution sensors are indispensable to that future.

Appendix A

Air Quality Data

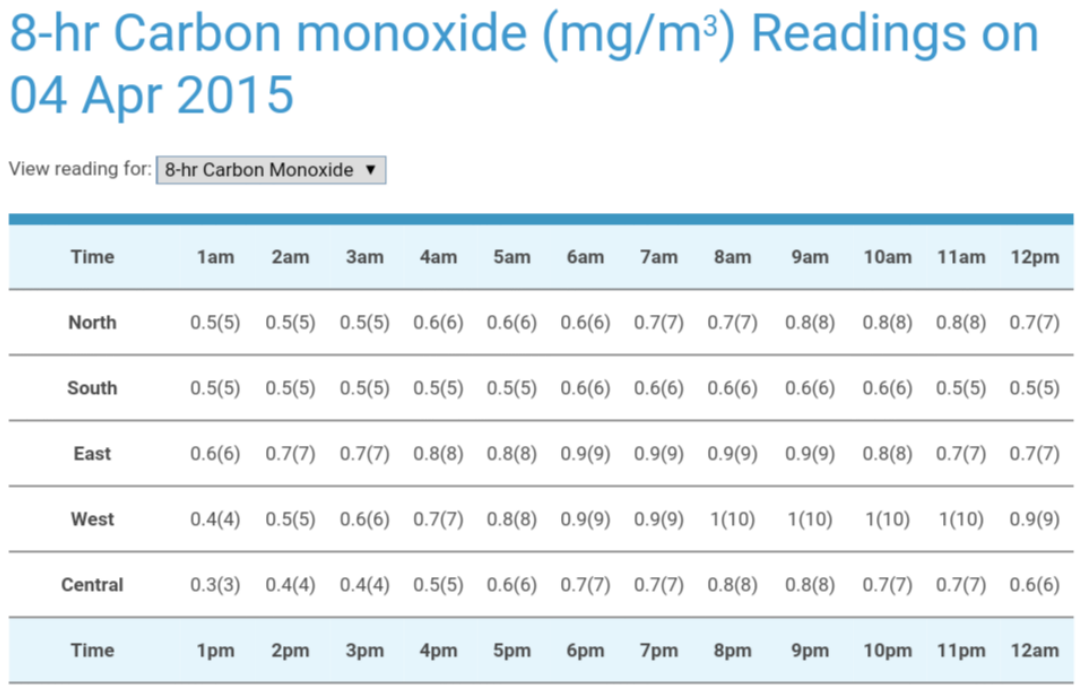
This appendix includes samples of the data used in this thesis. Some discussion of data parsing and cleaning is also included.

A.1 Personalized Sensor Data

The sensor data were transmitted via bluetooth to the participants' mobile phones, which then copied the air quality, location, and time data to files separated by day and sensor. In total, this amounted to 52 sensor log files, which were further split by reading type (CO, NO_2, pressure, temperature, humidity, time, latitude, longitude, and battery charge) and then concatenated into data frames by reading category. Included below is a sample of an original data file before it was split by reading type (Table A.1).

Table A.1: Personalized Sensor Data

date	time	type	reading	longitude	latitude
02.04.2015	17.20222	temperature	30	103.7427	1.333417
02.04.2015	17.20222	pressure	1003	103.7427	1.333417
02.04.2015	17.20222	batterys	90	103.7427	1.333417
02.04.2015	17.20250	co	855	103.7427	1.333417
02.04.2015	17.20278	humidity	61	103.7427	1.333417
02.04.2015	17.20306	no2	80	103.7427	1.333417
02.04.2015	17.20750	temperature	30	103.7427	1.333417
02.04.2015	17.20778	pressure	1006	103.7427	1.333417
02.04.2015	17.20778	batterys	89	103.7427	1.333417
02.04.2015	17.20806	co	876	103.7427	1.333417
02.04.2015	17.20833	humidity	61	103.7427	1.333417
02.04.2015	17.20861	no2	162	103.7427	1.333417



8-hr Carbon monoxide (mg/m³) Readings on 04 Apr 2015

View reading for: 8-hr Carbon Monoxide ▾

Time	1am	2am	3am	4am	5am	6am	7am	8am	9am	10am	11am	12pm
North	0.5(5)	0.5(5)	0.5(5)	0.6(6)	0.6(6)	0.6(6)	0.7(7)	0.7(7)	0.8(8)	0.8(8)	0.8(8)	0.7(7)
South	0.5(5)	0.5(5)	0.5(5)	0.5(5)	0.5(5)	0.6(6)	0.6(6)	0.6(6)	0.6(6)	0.6(6)	0.5(5)	0.5(5)
East	0.6(6)	0.7(7)	0.7(7)	0.8(8)	0.8(8)	0.9(9)	0.9(9)	0.9(9)	0.9(9)	0.8(8)	0.7(7)	0.7(7)
West	0.4(4)	0.5(5)	0.6(6)	0.7(7)	0.8(8)	0.9(9)	0.9(9)	1(10)	1(10)	1(10)	1(10)	0.9(9)
Central	0.3(3)	0.4(4)	0.4(4)	0.5(5)	0.6(6)	0.7(7)	0.7(7)	0.8(8)	0.8(8)	0.7(7)	0.7(7)	0.6(6)
Time	1pm	2pm	3pm	4pm	5pm	6pm	7pm	8pm	9pm	10pm	11pm	12am

FIGURE A.1: An example of a screenshot used to collect the government pollution data from the Singapore NEA website.

A.2 Government Sensor Data

The government stationary sensor data were collected using five sensors about which little information is known other than general region of placement (exact addresses of sensor locations are not public). These data were copied from the Singapore National Environmental Agency (NEA) website using mobile phone screenshots. Singapore's NEA posts the data hourly and they remain posted until 1am the following day. The data were then copied by hand into a spreadsheet, from which they were read into R and subsequently written out as a csv file. Included is one of the mobile phone screenshots that was initially used to collect the data (Figure A.1) and a sample of the data after they were read into csv format (Table A.2).

A.3 Spatially Interpolated Personalized Sensor Data

A portion of the personalized sensor data was spatially interpolated on a grid over Jurong East (Figure A.2), a sample of whose points are described in Table A.4. Interpolated carbon monoxide values were calculated for each point on the grid every 20 seconds from 16:54:21 to 19:10:41 on 4 April 2015. A sample of these values is included below (Table A.3).

Table A.2: Government Sensor Data

date	time	type	reading	location
4	1	NO2	33	North
4	1	NO2	5	South
4	1	NO2	20	East
4	1	NO2	42	West
4	1	NO2	64	Central
4	2	NO2	34	North
10	23	CO	0.3	Central
10	24	CO	0.3	North
10	24	CO	0.4	South
10	24	CO	0.5	East
10	24	CO	0.3	West
10	24	CO	0.3	Central

Table A.3: Spatially Interpolated (CO) Personalized Sensor Data

Timepoint\Gridpoint	X0	X0.1	X0.2	X0.3	X0.4	X0.5
1	869.00	869.00	869.00	869.00	869.00	869.00
2	895.25	895.25	895.25	895.25	895.25	895.25
3	923.80	923.80	923.80	923.80	923.80	923.80
4	939.17	939.17	939.17	939.17	939.17	939.17
5	951.29	951.29	951.29	951.29	951.29	951.29
6	953.38	953.38	953.38	953.38	953.38	953.38
7	950.67	950.67	950.67	950.67	950.67	950.67
8	947.70	947.70	947.70	947.70	947.70	947.70
9	947.09	947.09	947.09	947.09	947.09	947.09
10	945.08	945.08	945.08	945.08	945.08	945.08

Table A.4: Spatial Interpolation Gridpoints

Gridpoints	Longitude	Latitude
X0	1.332953	103.7269
X0.1	1.332953	103.7278
X0.2	1.332953	103.7287
X0.3	1.332953	103.7296
X0.4	1.332953	103.7305
X0.5	1.332953	103.7314

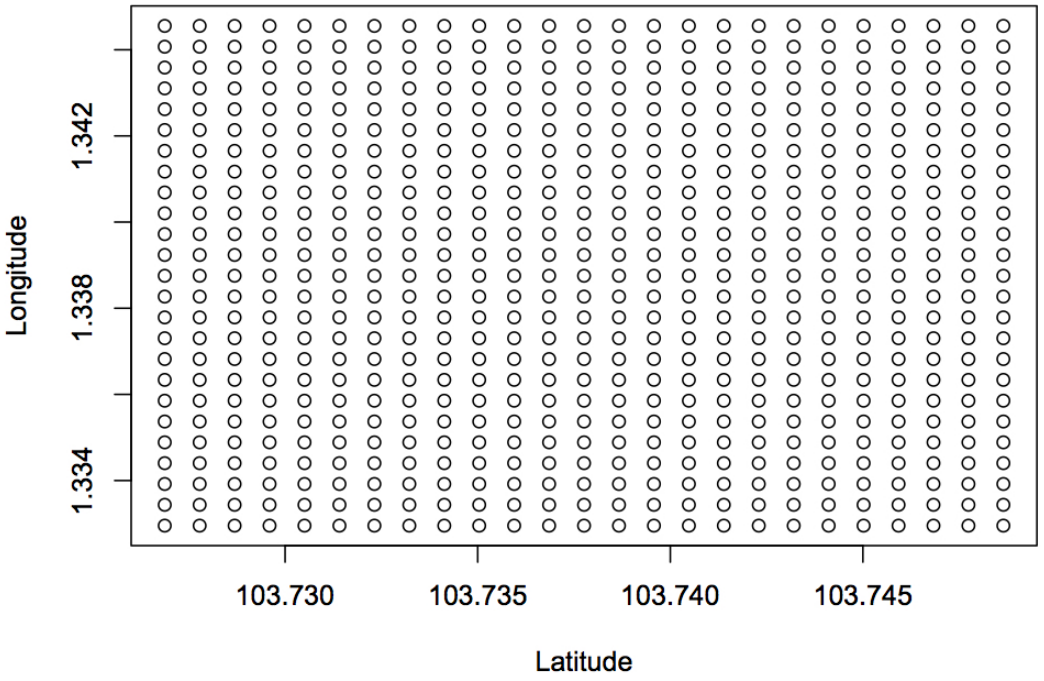


FIGURE A.2: This grid was overlaid onto Jurong East and used in the kriging spatial interpolation.

Appendix B

Algorithms Produced for Analyses

Samples of the original code written (in R) for this thesis are included.[1] In particular, the functions used to produce the hypothesis tests (Chapter 4) and random walks on the spatially interpolated field (Chapter 5) are provided below and described.

B.1 Permutation Testing

The function *perm.test2* takes in two datasets (*per* and *nea*), a number of permutations to make (*iter*), and bounds on the width of the histogram that is ultimately created (*xl*), and returns both a p-value for a permutation test and a visualization for that test. The vectors *per* and *nea* are concatenated to form the vector *data*, which is then permuted by resampling without replacement a dataset (s) of the same size as data. The first *length(nea)* values of s are categorized as simulated government data (whose mean is n) and the remaining *length(per)* values of s are categorized as simulated personalized data (whose mean is p). The test statistic for this test is computed to be n-p and is stored in *dist*. This process is repeated *iter* times. The true value of the personalized sensor mean subtracted from the government sensor mean is then computed and compared to the permutation distribution stored in *dist* to generate a p-value.

```
perm.test2 <- function(per, nea, iter, xl ){
dist <- rep(-3, iter)
data <- rep(-2, length(per) + length(nea))
```

[1] Only a sample of the code used in the analyses is provided here. Additional code is available upon request.

```
data[1:length(nea)] <- nea
len <- length(nea) + 1
data[len:length(data)] <- per
for (ii in 1:iter){
s <- sample(x = data, size = length(data), replace = FALSE)
n <- mean(s[1:length(nea)])
p <- mean(s[len:length(data)])
dist[ii] <- n-p
}
mp <- mean(per)
mn <- mean(nea)
real_value<- mn-mp
numless <- sum(dist <= real_value)

par(mar = c(4.2,4.5,1,1))
hist(dist, xlim = xl, xlab = expression("Difference in Means
("*mu*"g/"*m^3*")"), main = "")
abline(v = real_value, col = 'red')

return(numless/iter)
}
```

B.2 Wilcoxon Rank Sum Testing

For this analysis, the *wilcox.test* algorithm provided by R was used, with the alternative hypothesis set to require that the government sensor data be less than the personalized sensor data.

B.3 Random Walks on a Spatially Interpolated Field

The random walks each occur on a 25 by 25 grid over a span of 410 steps. The variables *ind1* and *ind2* keep track of the x and y coordinates respectively, and are initially set to random coordinate values in the 1 to 25 range. This starting point is considered the first step in the walk. Any adjacent grid point to the current point (including diagonals) is considered a valid next step. In the code, this is encapsulated in the portion of *random.walk* in which *delx* and *dely* are set. On a given axis (for x or y), the walker can move one step ahead (in the positive direction), one step behind (in the negative

direction), or no steps at all. To randomly determine the walker's movement, a number from -1 to 1 is generated, indicating positive, negative, or no movement. If the walker is on an edge of the grid (either at index 1 or 25), then their options are limited to movement one step away from the edge or no movement. In both the edge and non-edge cases, each movement choice is weighted equally. The walker is then moved accordingly at the end of the for loop, and their new coordinates are recorded in the arrays *xindex* and *yindex*. The function transforms these two arrays into the columns of a new data frame with dimensions 410 by 2 and returns the data frame.

```
random.walk <- function(){
 #create arrays in which to store indices
xindex = rep(-1, 410)
yindex = rep(-1, 410)

#generate random starting indices
ind1 = sample(25)[1]
ind2 = sample(25)[1]
xindex[1] = ind1
yindex[1] = ind2

#move about the adjacency matrix
for (t in 2:410){
delx = 0
if (ind1 == 1){
temp = sample(2)[1]
delx = temp - 1
} else if (ind1 == 25){
temp = sample(2)[1]
delx = temp - 2
} else {
temp = sample(3)[1]
delx = temp - 2
}
ind1 = ind1 + delx

dely = 0
if (ind2 == 1){
temp = sample(2)[1]
dely = temp - 1
```

```
} else if (ind2 == 25){
temp = sample(2)[1]
dely = temp - 2
} else {
temp = sample(3)[1]
dely = temp - 2
}
ind2 = ind2 + dely
xindex[t] = ind1
yindex[t] = ind2
}
indices = data.frame(xindex, yindex)
return(indices)
}
```

The function *get.rand.exposure* runs a random walk on a 25 by 25 grid by calling *random.walk*. Then, for each pair of coordinates returned by *random.walk* (stored as rows in a data frame), the function traces the given coordinates in matrix *coordMatrix*[2] to find the column indices of the gridpoint in the kriging output data frame concentrations (see Appendix A for a sample of this data frame). Once the column indices have been determined, the kriged values are accessed using the column indices and step numbers. These values are then summed and divided by three to reflect that exposure in this case is calculated by minute, while the steps are made in 20s jumps. The total exposure over the walk is returned.

```
get.rand.exposure <- function(){
#take random walk on indices
rw = random.walk()

#calculate exposure per 20s period
exp = rep(-1, 410)
for (t in 1:410){
index = coordMatrix[rw[t, 1], rw[t, 2]]
exp[t] = concentrations[t, index]
}

tot_exposure = sum(exp)*1/3 # sum to give total concentration/time
```

[2]The process for creating coordMatrix depends on the ordering of the grid by the kriging program.

```
return(tot_exposure)
}
```

The function *many.rand.exposure.walks* iterates the previous function *get.rand.exposure* *numIter* times and returns an array of exposure values from a set of random walks.

```
many.rand.exposure.walks <- function(numIter = 100){
result <- rep(-1, numIter)
for (ii in 1:numIter){
result[ii] = get_rand_exposure()
}
return(result)
}
```

Bibliography

NEA (National Environment Agency-Singapore). Air quality and targets, March 2016. URL http://www.nea.gov.sg/anti-pollution-radiation-protection/air-pollution-control.

Joe Cochrane. Southeast asia, choking on haze, struggles for a solution. *New York Times*, pages 4477–4479, 2015. URL http://www.nytimes.com/2015/10/09/world/asia/indonesia-forest-fires-haze-singapore-malaysia.html?emc=eta1&_r=0.

Audrey deNazelle et al. Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environmental Pollution*, 176:92–99, 2013.

B.J. Smith et al. Health effects of daily indoor nitrogen dioxide exposure in people with asthma. *European Respiratory Journal*, 16.5:879–885, 2000a.

Evi Dons et al. Impact of timeactivity patterns on personal exposure to black carbon. *Atmospheric Environment*, 45.21:3594–3602, 2011.

James A. Raub et al. Carbon monoxide poisoninga public health perspective. *Toxicology*, 145.1:1–14, 2000b.

Marguerite Nyhan et al. Comparison of particulate matter dose and acute heart rate variability response in cyclists, pedestrians, bus and train passengers. *Science of the Total Environment*, 468:821–831, 2014.

Stefan Reis et al. Integrating modelling and smart sensors for environmental and human health. *Environmental Modelling Software*, 74:238–246, 2015.

ICRP (International Commission for Radiological Protection). Annals of the icrp, human respiratory tract model for radiological protection, 24. *ICRP Publication*, 66:1–3, 1994.

Tom Dhaene Ivo Couckuyt and Piet Demeester. A matlab kriging toolbox: Getting started. *ooDACE toolbox*, 2013.

Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental pollution*, 151.2:362–367, 2008.

A. McCreddin. *Modelling personal exposure to particulate air pollution: An assessment of time-integrated activity modelling, Monte Carlo simulation artificial neural network approaches*. PhD thesis, Trinity College, University of Dublin, Ireland, 2014.

Mark J. Nieuwenhuijsen. *Exposure Assessment in Environmental Epidemiology*. Oxford University Press, USA, 2015.

Dinko Oletic and Vedran Bilas. Design of sensor node for air quality crowdsensing. *Sensors Applications Symposium (SAS), IEEE*, 2015.

WHO (World Health Organization). Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide: report on a who working group, bonn, germany 13-15 january 2003, 2003. URL http://www.who.int.

WHO (World Health Organization). 7 million premature deaths annually linked to air pollution, March 2014. URL http://www.who.int.

Fred Ramsey and Daniel Schafer. *The statistical sleuth: a course in methods of data analysis*. Cengage Learning, 2012.

Monika Martha Maria Zuurbier. *Commuters air pollution exposure and acute health effects*. Utrecht University, 2011.