SITE-SPECIFIC PM2.5 ESTIMATION AT THREE URBAN SCALES

by

Yogita Yashawant Karale

APPROVED BY SUPERVISORY COMMITTEE:

May Yuan, Chair

David Lary

Fang Qiu

Yongwan Chun

SITE-SPECIFIC PM2.5 ESTIMATION AT THREE URBAN SCALES

by

YOGITA YASHAWANT KARALE, B. TECH, M. TECH

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

GEOSPATIAL INFORMATION SCIENCES

THE UNIVERSITY OF TEXAS AT DALLAS

August 2021

ACKNOWLEDGMENTS

I thank my advisor, Dr. May Yuan, for her continuous support and guidance through each step of this journey. I must have tested her patience in the early stage of research while navigating a relatively new world of GIScience. Her questions challenged my thinking and helped me to look at the field of GIScience in a whole new way. I am incredibly grateful for her patience and prompt feedback which helped me steer through this dissertation. I am also grateful to Dr. David Lary, Dr. Fang Qiu, and Dr. Yongwan Chun for serving as my committee members and for their invaluable suggestions. Dr. Lary's advice helped me understand and process MODIS data better. Dr. Fang Qiu and Dr. Yongwan Chun provided continuous assistance throughout my time here at UTD. Discussions with them always helped me to understand the geospatial world better. I am also grateful to Dr. Anthony Cummings, Dr. Michael Tiefelsdorf, and Dr. Daniel Griffith for their mentoring and guidance through various courses.

I cannot thank enough UTD police for their assistance during my data collection exercise on campus. Many thanks to Lakitha Wijeratne for his help and time in calibrating the sensor. Special thanks to Brent whose collegial spirit and helpful nature brought all the GAIA lab members together and made the GAIA lab a lively place to research.

Finally, I would like to thank my family members for their encouragement and support. I am especially grateful to my husband without whose rock-solid support I could not have made it this far.

July 2021

SITE-SPECIFIC PM2.5 ESTIMATION AT THREE URBAN SCALES

Yogita Yashawant Karale, PhD
The University of Texas at Dallas, 2021

Supervising Professor: May Yuan

Fine particulate matter, also known as PM2.5, is one of the major risk factors to human health. Because of their small size, these particles travel deep within human lungs and pose a variety of health problems. A primary source of acquiring PM2.5 exposure is based on the nearest ground-level air quality monitoring station. However, these stations are often few and sparsely located due to their high costs for installation and maintenance. This study addresses three challenges related to PM2.5. First, the number of air-quality monitoring sites is insufficient to acquire the complex spatial variability of PM2.5. Therefore, in-situ ground observations fail to characterize PM2.5 distribution, and hence exposure, adequately. The shortfall calls for models capable of estimating PM2.5 at unmonitored locations. Satellite-based Aerosol Optical Depth (AOD) serves as a proxy to estimate PM2.5. Second, although satellite data can supplement PM2.5 estimates at unmonitored locations, the spatial resolutions of satellite-based estimates of PM2.5 are in the order of kilometers. These spatial grains are too coarse to capture PM2.5's spatial variation caused by contextual geographic factors such as buildings, and subsequently the estimates' applicabilities to support environmental exposome on health effects. Third, the current standards measure PM2.5 in terms of mass per volume, but findings from some recent studies suggest that

alternative measures of PM2.5 are also strongly associated with adverse health outcomes. However, observations in terms of these measures are not available.

The dissertation research aimed to address the three challenges in three studies. The first study evaluated the potential of the Convolutional Neural Network (CNN) approach to downscale PM2.5 using satellite-based AOD and meteorological data using Dallas-Fort Worth as a case study. The study developed a model capable of estimating PM2.5 corresponding to the hour of satellite overpass time and examined environmental predictors commonly available for all monitored or non-monitored locations. In particular, the study investigated the effect of the spatial extent to which predictors from the surrounding area influenced the PM2.5 estimates at a location. The results showed that the proposed CNN model effectively estimates PM2.5 concentration with correlation coefficient (R) of 0.87 and root mean squared error (RMSE) of 2.57 $\mu g/m^3$. Moreover, spatially lagged variables from a wider area around an estimation location improved the model performance. As most monitoring stations were in open areas, data from these stations could not be used to examine the effect of contextual factors, such as the building on PM2.5. The second study evaluated the effects of contextual geographic factors on PM2.5 in mass per volume (i.e., standard measures) in pedestrian-friendly areas on the University of Texas at Dallas campus. The study used a mobile sensor to collect spatial and temporal fine-resolution PM2.5 data on the campus. The study found very low spatial variation in the study area less than $1km^2$. Furthermore, weather-related variables played a dominant role in PM2.5 distribution as temporal variation over-powered spatial variation in PM2.5 data. The study employed a fixed effect model to assess the effect of time-invariant building morphological characteristics on PM2.5 and found that building's morphological characteristics explained

33.22% variation in the fixed effects in the model. Furthermore, openness in the direction of wind elevated the PM2.5 concentration. The third study investigated the potential of AOD to downscale Particle Number (PN) concentration, an alternative measure of PM2.5, and the effect of building morphology on PN concentration using PN measurements collected across the streets of San Francisco by the Google streetcar. The study showed that AOD remained useful to estimate street-level PN concentration across five different particle sizes. The subsequent analysis of variable importance revealed that AOD and AOD-related variables were more important than building morphology but less important than meteorological variables in the estimation of PN concentration.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iv

ABSTRACT v

LIST OF FIGURES x

LIST OF TABLES xii

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 MODELING THE PM2.5 IN AN URBAN ENVIRONMENT 7

2.1 Introduction 7

2.2 Materials and methods 14

2.3 Results and Discussion 27

2.4 Conclusion 30

CHAPTER 3 MICROSCALE DYNAMICS OF PM2.5: A CASE STUDY AT THE UNIVERSITY OF TEXAS AT DALLAS 33

3.1 Introduction 33

3.2 Data and Methods 36

3.3 Results 54

3.4 Discussion 63

3.5 Conclusion 65

CHAPTER 4 STREET-LEVEL QUANTIFICATION OF ALTERNATIVE PM2.5 MEASURE 67

4.1 Introduction 67

4.2 Data and methods 70

4.3 Results and Discussion 80

4.4 Conclusion 86

CHAPTER 5 CONCLUSION 88

APPENDIX SUPPLEMENTARY TABLE 93

REFERENCES 94

BIOGRAPHICAL SKETCH 106

CURRICULUM VITAE 107

LIST OF FIGURES

Figure 2.1: Distribution of PM2.5 stations in the study area 15
Figure 2.2: Flowchart of data and methodology 20
Figure 2.3: Different sized input data grids centered on PM2.5 station 22
Figure 2.4: The study's CNN architecture 23
Figure 2.5: Sample image 25
Figure 2.6: Augmented images 26
Figure 2.7: Correlation coefficient and RMSE for CNN with varying grid size 28
Figure 2.8: Correlation coefficient and RMSE for CNN with varying grid size without data 29
Figure 2.9: Percent change in R and RMSE in models without data augmentation 30
Figure 3.1: UTD Campus with Data Collection Paths A, B, C, D and E 37
Figure 3.2: Flowchart of data preparation 39
Figure 3.3: Buffer and buildings used in building morphology calculation 42
Figure 3.4: Mean building area 45
Figure 3.5: Number of buildings per 1000 m^2 46
Figure 3.6: Mean building height 47
Figure 3.7: Average distance between nearest neighbhor buildings 48
Figure 3.8: Building coverage ratio 49
Figure 3.9: Distribution of PM2.5 (a) raw PM2.5 observations (b) Transformed PM2.5 data. 55
Figure 3.10: Distribution of PM2.5 across all paths during each data collection run 57
Figure 3.11: PM2.5 at each path during data-collection runs 57
Figure 3.12: Correlation between fixed effects and building morphological variables 62

Figure 4.1: street segments and PN assignments 72
Figure 4.2: Data collection frequency across street segments in San Francisco 74
Figure 4.3: Street segments with overlapping data acquisition time within ± 30 minutes of MODIS overpass time....75
Figure 4.4: Neural network architecture....78
Figure 4.5: Correlation coefficient for different sized PNs across different models 82
Figure 4.6: RMSE for different sized PNs across different models....83
Figure 4.7: SHAP summary plot for PN1....84
Figure 4.8: SHAP summary plot for PN2....84
Figure 4.9: SHAP summary plot for PN3....85
Figure 4.10: SHAP summary plot for PN4....85
Figure 4.11: SHAP summary plot for PN5....86

LIST OF TABLES

Table 2.1: List of predictors.....19

Table 2.2: AOD data availability.....24

Table 3.1: Summary of building morphological characteristics around data collection paths.....50

Table 3.2: Details of date and time of data collection rounds.....56

Table 3.3: Fixed effect model results with varying viewshed distance.....59

Table 3.4: Fixed effect model results.....59

Table 3.5: Spatial autocorrelation in errors of the fixed effect models.....61

Table 3.6: Results of regression between fixed effects and building morphological characteristics 63

Table 4.1: Particle size table.....71

Table 4.2: Summary statistics of building morphological characteristics.....77

Table 4.3: Range of particle number concentration across different particle sizes.....79

Table 4.4: PN estimation results for the base model and with the exclusion of AOD and building morphological characteristics.....81

Table A.1: Results across the neural networks with 8, 16 and 24 neurons for five particle bins..93

CHAPTER 1

INTRODUCTION

The dissertation study examines the distributions of PM2.5 at three urban scales: regional, local, and microenvironments. Scale is at the heart of many scientific inquiries. While some principles may be scaleless, most environmental, ecological, and social sciences are scale-dependent. Scale may apply to the processes of interest or the observations available for analysis. A mismatched scale of observations to the scale of processes will lead to biased or invalid answers to our inquiries. Theories are only applicable to problems within the expected spatiotemporal scales to assure their internal validity. Consequently, the spatiotemporal scales of a theory determine the kind of observations that should be used to guide hypothesis formulation and testing (Tate and Atkinson, 2001). Statistical techniques and sampling design need to have data at the proper scales that capture spatiotemporal variations manifested by the processes of interest (McGill, 2010).

In Geographic Information Science, the concept of scale has two significant meanings: the resolution of data and the extent of a study area (Goodchild, 2011). A smaller study area will need higher resolution data to capture the spatiotemporal variation generated by local processes. In practice, we often have data resolution too coarse to elicit the local variation and need to develop downscaling strategies to infer high-resolution information from low-resolution variables. Traditionally, downscaling methods are theoretically based on process dynamics or statistically based on variable relationships or stochastic assumptions (von Storch and Zorita,

2019). This dissertation research proposes the use of machine learning for downscaling to account for non-linear relationships among variables in space and time as neither theoretical nor statistical approach can adequately capture environmental complexity arising from intricate interactions among multiple variables. To date, environmental modeling is common at a global or regional scale, constrained mainly by the scale of observations. The United Nations estimated 58% of the world population lived in urban areas in 2018 and projected an increase to 68% by 2050 (United Nations, 2018). New spatiotemporal approaches to downscale data allow environmental modeling at the intra-urban scale where the environment interacts with daily human activities. Specifically, this dissertation research investigates a common, yet critical issue: air pollution and explores downscaling in two strategies: (1) downscaling with satellite and in-situ ground observations, (2) downscaling based on considerations of site characteristics and using measurements from mobile sampling.

Particulate matter (PM) is one of the major contributors to mortality due to air pollution. Several studies found a connection between PM measurement in mass per volume and short- and long-term health effects. In air quality standards, particulate matter is categorized into two size groups: known as PM2.5 and PM10. PM2.5 includes particles with a size less than or equal to 2.5 micrometers. Due to their smaller sizes, PM2.5 particles, when inhaled, can go deeper into the lungs, and pose various health problems related to breathing and lung functioning. In the long run, an increase of 10 $\mu g/m^3$ in PM2.5 raises the risk of mortality due to lung cancer and cardio-pulmonary diseases by 8% and 6%, respectively (Pope III *et al.*, 2002). Short-term exposure to PM2.5 increases the chances of hospitalization for cardiovascular and respiratory diseases in a

population aged above 65 (Dominici *et al.*, 2006) in the United States. Not only does PM2.5 cause several health problems, but it also worsens health conditions for people who are already suffering from respiratory health conditions like asthma.

In epidemiology, accurate PM2.5 estimation across space and time is important to assess exposure and establish connections between PM2.5 and various health problems. A primary source of acquiring PM2.5 exposure is based on the nearest air quality monitoring site. However, ground-level monitoring stations are often few and sparsely located due to their high costs for installation and maintenance. There are three major challenges to estimate PM2.5 distributions accurately. First, the number of air-quality monitoring sites is insufficient to acquire the complex spatial variability of PM2.5 and fails to characterize PM2.5 distribution adequately, hence exposure, accurately. Second, satellite data can supplement PM2.5 estimates at unmonitored locations. Still, the spatial resolutions of satellite estimates of PM2.5 are at the order of kilometers, which are too coarse to capture spatial variation caused by contextual geographic factors such as buildings, roads, and wind effects. Third, studies have shown adverse health outcomes from a variety of PM measures.

The standards for PM measures, established in 1971, have continuously evolved as additional evidence became available on PM's adverse health outcomes. Current standards measure PM in terms of mass per volume. Some studies identify a positive association between adverse health outcomes and individual chemical components of PM2.5 (Peng *et al.*, 2009; Krall *et al.*, 2017), whereas other studies connect adverse health outcomes with particle size (Ibald-

Mulli *et al.*, 2002; Olsen *et al.*, 2014). Due to the limited availability of PM in terms of measures other than particle mass per volume, little evidence is available to establish new PM2.5 standards on other PM measures associated with adverse health outcomes. However, information about the spatiotemporal distribution of PM in terms of these measures may prove valuable in the near future with increasing studies on these measures. Additionally, since alternative measures focus on the properties of PM2.5 beyond particle mass, these measures may also provide additional information about PM2.5 behavior in space and time, which the current PM2.5 measures dismiss.

This dissertation has three objectives. First, it aims to evaluate the efficacy of the machine learning approach to downscale PM2.5 using satellite-based Aerosol Optical Depth (AOD) and meteorological data using the Dallas-Fort Worth region as a case study. Second, most monitoring stations are in open areas. Therefore, the effect of contextual geographic factors, especially buildings, on PM2.5 dispersion cannot be examined. To this aim, this dissertation evaluates the effects of contextual geographic factors on standard PM2.5 measures in a microenvironment on the University of Texas at Dallas campus as a study area. The third objective examines an alternative measure of PM2.5 at a local scale of city streets. To this end, a local investigation of AOD potential to downscale an alternative measure of PM2.5 and the effects of contextual geographic factors to PM2.5 distribution based on an alternative measure, Particle Number (PN), is developed based on measurements collected across the streets of San Francisco using Google streetcars.

In summary, this research focuses on the following objectives:

1. Develop and evaluate a machine learning approach to estimate PM2.5 for the Dallas-Fort Worth region.
2. Determine the effect of contextual geographic factors on the standard PM2.5 measure in mass per volume in a microenvironment of the campus of the University of Texas at Dallas.
3. Develop and evaluate a machine learning approach to estimate an alternative PM2.5 measure in particle numbers on local streets in San Francisco.

The dissertation is organized into five chapters:

Chapter 1 highlights problems on the importance of understanding PM2.5 distributions and discerns research gaps. Subsequently, it identifies three broader objectives for the dissertation research.

Chapter 2 aims to develop a machine learning approach to downscaling PM2.5 in mass per volume to a particular hour at a location. Specifically, the study models non-linear relationships between satellite-based instantaneous aerosol optical depth (AOD) data and in-situ ground-based hourly accumulated PM2.5 measures to estimate accumulative PM2.5 over a certain hour at a specific location.

Chapter 3 examines PM2.5 variations and the effects of contextual geographic factors on PM2.5 variance in mass per volume in a microenvironment.

Chapter 4 investigates an alternative measure of PM2.5, particle number (PN), develops a method to estimate PN and evaluates the effects of building morphology on PN on city streets. Chapter 5 cross-references the findings and synthesizes the characteristics of PM2.5 distributions at three spatial scales and two measurement units.

CHAPTER 2

MODELING THE PM2.5 IN AN URBAN ENVIRONMENT

2.1 Introduction

The Global Burden Disease study reported that air pollution caused 4.2 million deaths in 2015 due to particulate matter (Cohen *et al.*, 2017). Recent studies found a link between PM2.5 and several neurological disorders like dementia, Alzheimer's, and Parkinson's diseases (Chen *et al.*, 2017; Kioumourtzoglou *et al.*, 2015). Despite the identified harmful effects of PM2.5 on health, the number of ground monitoring sites providing information about PM2.5 concentration is considerably sparse and is unsuitable for spatial interpolation at a local scale. Epidemiological studies rely on the data from the nearest available monitoring site to estimate the exposure, which may not be reliable due to spatial variability present in PM2.5 (Özkaynak *et al.*, 2013). The spatial uncertainty propagates in the epidemiological findings and presents the need for models that can capture spatial variation in PM2.5.

A common approach to characterize the spatial distribution of PM2.5 is with satellite-based Atmospheric Optical Depth (AOD) as one of the predictor variables for PM2.5 estimation (Lary *et al.*, 2014; Chudnovsky *et al.*, 2014; Guo *et al.*, 2017; and Xie *et al.*, 2015). AOD measures the amount of aerosols present in the atmosphere according to the optical properties of aerosols in an atmospheric column. However, the relationship between PM2.5 and AOD is complicated. AOD is affected by the size of the particles, type of the particles, and meteorological factors. Depending on the source, the composition of the particles may vary in space and time (Bell *et al.*, 2007). Meteorological factors (such as cloud fraction, relative

humidity, temperature, boundary layer height, wind speed, and others) also affect this relationship (Lary *et al.*, 2014; Guo *et al.*, 2017). Several studies report PM2.5-AOD relationship varies with geography (Engel-Cox *et al.*, 2004), time (Guo *et al.*, 2017), the scale of regional or local studies (Chudnovisky *et al.*, 2014), and AOD data resolution (Chudnovisky *et al.*, 2014; Xie *et al.*, 2015; Guo *et al.*, 2017). Empirical models using AOD to estimate PM2.5 developed for one geographical area cannot be used in the others.

Parametric statistical frameworks, such as regression, are inappropriate for spatiotemporal modeling of PM2.5 because of the limited number of air quality stations that unlikely to capture representative variations over space. Low-cost sensors such as Purple Air have been deployed in large numbers across the United States. While these low-cost sensors help reduce the existing gap in spatial coverage of PM2.5 measurements with the standard air quality monitoring stations, the accuracy of the measurements from these sensors remains a cause of concern. A field evaluation of three Purple Air sensors carried out at Rubidoux Air Monitoring Station in California for two months indicates that in general, Purple Air sensors have shown an overall trend of PM2.5 within a day and across days but tend to overestimate PM2.5 concentration most of the times (Gupta *et al.*, 2018). Specifically, the California study highlights that the bias of Purple Air sensors increases with an increase in PM2.5 concentration. Moreover, Purple Air sensors' observations deviate widely, from 0-90% of their hourly mean values. Parametric statistical methods require specifying the functional forms of the relationship between dependent and independent variables, and proper specifications of the relationship are challenging. PM2.5 measures the ground-level concentration of particles with an aerodynamic

diameter less than 2.5 micrometers. In contrast, AOD measures the extinction of light due to aerosols in the column between ground and satellite. Both AOD and PM2.5 are individually affected by meteorological parameters, which further complicates the relationship between them. Furthermore, AOD is an instantaneous measurement from space, and PM2.5 is an hourly average measured in-situ at respective ground monitoring stations.

The literature reported several approaches to model the PM2.5-AOD relationship, like land-use regression, geographically weighted regression, back propagation artificial neural network, mixed effect models, linear regression models, and chemical transport models (Guo *et al.*, 2017). The mixed effect modeling approach appeared popular among these approaches. Some studies used AOD as the only predictor (Chudnovisky *et al.*, 2014; Xie *et al.*, 2015); other studies included additional parameters to improve model performance (Hu *et al.*, 2014; Stafoggia *et al.*, 2017). Xie *et al.* (2015) used a mixed effect model to account for spatiotemporal variations in PM2.5-AOD relationship with day-specific and site-specific parameters for AOD to estimate PM2.5. Moreover, several other studies implemented similar mixed effect models by including AOD and additional spatiotemporal parameters (Hu *et al.*, 2014; Stafoggia *et al.*, 2017). In addition to day-specific random parameters, Staffogia *et al.* (2017) introduced region-specific random parameters to account for variation in PM10-AOD relations across different regions in Italy. In the Southeastern United States, Hu *et al.* (2014) used a mixed effect model to capture PM2.5-AOD temporal variability and followed with Geographically Weighted Regression on the residuals to account for PM2.5-AOD spatial variability. Spatial and temporal parameters considered in these studies include population density, emission data, elevation, land cover, road

density, Normalized Difference Vegetation Index (NDVI), meteorological data, etc. Zheng *et al.* (2013) applied a deep learning framework to predict Air Quality Index (AQI) for Beijing at 1km resolution with region-specific parameters representative of traffic features (e.g., mean, standard deviation, and distribution of speeds on the road) and human mobility features (e.g., number of people arriving and departing a location). Such region-specific parameters may not be available or appropriate for all areas outside Beijing.

Machine learning recently gained traction on modeling PM2.5 (Lary *et al.*, 2014; Di *et al.*, 2016; Hu *et al.*, 2017; Zhan *et al.*, 2017, Li *et al.*, 2017; Park *et al.*, 2020). Several of these studies incorporated spatial dependence in the machine learning methods. Di *et al.* (2016) used an artificial neural network (ANN) for the northeastern United States to calibrate PM2.5 obtained from a chemical transport model, and Li *et al.* (2017) used the Deep Belief Network approach to estimate PM2.5 in China. They considered spatial and temporal autocorrelation using lagged spatial and temporal terms. Spatial lag was incorporated by using PM2.5 measurements from nearby stations weighted by the inverse of their distance from the monitor under consideration. An alternative way of applying weights in PM2.5 estimation was the boosting technique in machine learning. Boosting gave more weight to observations with high errors to improve model performance. Zhan *et al.* (2017) used geographically weighted gradient boosting to account for spatial non-stationarity in PM2.5 and AOD as well as meteorological factors.

Advances in deep learning opened opportunities to convolute in-situ and satellite observations for PM2.5 estimation. Park *et al.* (2020) used a convolutional neural network

(CNN) to estimate the 24-hr averaged PM2.5 across the conterminous United States using the one-year data in 2011. Hu *et al.* (2017) incorporated inverse distance weighted PM2.5 from nearby stations as input to the random forest model. Clouds or high surface brightness might obscure AOD data from MODIS. Due to the high missing rate of AOD, both studies applied the GEOS-Chem model to simulate AOD data. Hu *et al.* (2017) used GEOS-chem AOD when MODIS AOD was missing, whereas Park *et al.* (2020) used both MODIS AOD and GEOS-Chem AOD. Along with the AOD data, both studies used meteorological data, land-use variables, and National Emission Inventory (NEI) data as predictors. Several data issues were prominent in both studies. NEI database provided information about pollutant-wise emissions at annual scales. However, methods used to estimate these emissions might vary from year to year (U.S. Environmental Protection Agency, 2013; 2020). Therefore, the data from these emission inventories were unsuitable for multi-year studies. Land-use data were static in nature and could contribute very little in explaining PM2.5 which varies on an hourly basis. Li *et al.* (2017) reported that the inclusion of road networks as one of the predictors showed a minimal impact on model performance, whereas population worsened the model performance. Furthermore, in areas with sparsely distributed monitoring stations, land-use and population density around very few monitoring stations might not be representative enough to allow model generalizability for the entire study area. Xu *et al.* (2014) observed an increase in AOD values in areas with increased human activities and decreased AOD where forested areas increased and concluded that changes in land-use led to changes in AOD patterns. Therefore, this study assumes that AOD data embed the effect of land-use change on PM2.5.

Several studies assessed model performance in estimating PM2.5 through cross-validation in three different approaches for setting cross-validation data: spatially separated cross-validation (SS-CV), temporally separated cross-validation (TS-CV), and overall cross-validation (O-CV) approach (Di *et al.*, 2016; Hu *et al.*, 2017; Park *et al.*, 2020). As names suggest, SS-CV shares no common locations between the training dataset and cross-validation dataset; TS-CV uses observations for the training dataset from different days than the observations in the cross-validation dataset. In contrast, the O-CV approach imposed no restrictions in days or locations on training and cross-validation datasets. Results from studies by Di *et al.* (2016), Hu *et al.* (2017), and Park *et al.* (2020) showed that models using O-CV and TS-CV outperformed the ones using the SS-CV approach. This suggested that models developed for a set of locations did not perform well at unseen locations; the models were spatially untransferable. The performance of models using either O-CV or T-CV approach for cross-validation was comparable. Therefore, this study planned to take the O-CV approach for cross-validation.

Incorporating geographical correlations can improve model performance in PM2.5 estimation (Li *et al.* 2017), but four main challenges remain. First, many studies incorporate spatial dependence and include spatially lagged predictors and spatially lagged PM2.5 in the model (Hu *et al.*, 2017; Zhan *et al.*, 2017, Li *et al.*, 2017; Park *et al.*, 2020). For the models developed by Hu *et al.* (2017) and Park *et al.* (2020), spatially lagged PM2.5 measurements rise to the most important variable in estimating PM2.5. However, obtaining spatially lagged PM2.5 for areas with sparse distribution of monitoring stations is challenging, and the density of the

PM2.5 monitoring stations may affect model performance in areas with sparsely monitoring stations. The second challenge relates to the hindrance of real-time PM2.5 estimations without data available from nearby monitoring stations. The third challenge speaks for the mismatch between PM2.5 estimates and satellite observations. In the Dallas-Fort Worth region, for example, AOD data are instantaneous observations around 10:30 am and 1:30 pm by Terra and Aqua satellites, respectively. Although few studies such as Tian *et al.* (2010) and Xie *et al.* (2015) used PM2.5 data obtained near satellite acquisition time, most of the studies in the literature estimated the PM2.5 concentration averaged over 24 hours using instantaneous AODs. Finally, the fourth challenge relates to previous studies, which incorporated spatial dependence, used predictors from a confined spatial extent. Therefore, how the model might perform over other spatial extents is not known.

This study fills the research gaps in light of these challenges by developing a model to estimate PM2.5 in the correspondent hour when Terra or Aqua satellite overpasses the area. The model considers only spatially lagged predictors from MODIS and meteorological data but does not include PM2.5 from nearby stations. Finally, the study investigates the model performance using CNN where the size of the input image (of predictors) varies from 3×3, 5×5,….. to 19×19 with PM2.5 station at the center cells for each input image. Varying the window size in a CNN allows examining the effect of changing the spatial extent of spatial lagged predictors on the estimated value of PM2.5 at a predictor location.

2.2 Materials and methods

2.2.1 Study area

The study area is the Dallas-Fort Worth (DFW) metroplex with more than 7.5 million people. Across 2,141,104 hectares, the DFW metroplex and its surrounding area have only eight air-quality monitoring stations measuring hourly PM2.5 for 2006-2015, leaving most of the metroplex unmonitored. Figure 2.1 shows the distribution of PM2.5 monitoring stations in the DFW metroplex. Out of the eight monitoring stations, three are located in urban areas, whereas five are at the periphery of the urban areas. Information on the spatiotemporal distribution of PM2.5 at the appropriate level of detail is important because of the harmful effects of PM2.5 on health, especially for those already suffering from respiratory and cardiovascular diseases. Informed with the spatiotemporal distribution of PM2.5 at a fine resolution, people can avoid areas with high concentration and reduce the geographic context uncertainty for epidemiological studies of PM2.5 exposure. Nevertheless, a step towards estimating the spatiotemporal distribution of PM2.5 is to test how well an O-CV approach can use AOD to estimate PM2.5 at these stations corresponding to the hour of satellite overpass time. If the estimation is acceptable at these sites, the proposed model can provide the foundation for building a spatial interpolation method with AOD to estimate PM2.5 at unmonitored locations when AOD data is available.

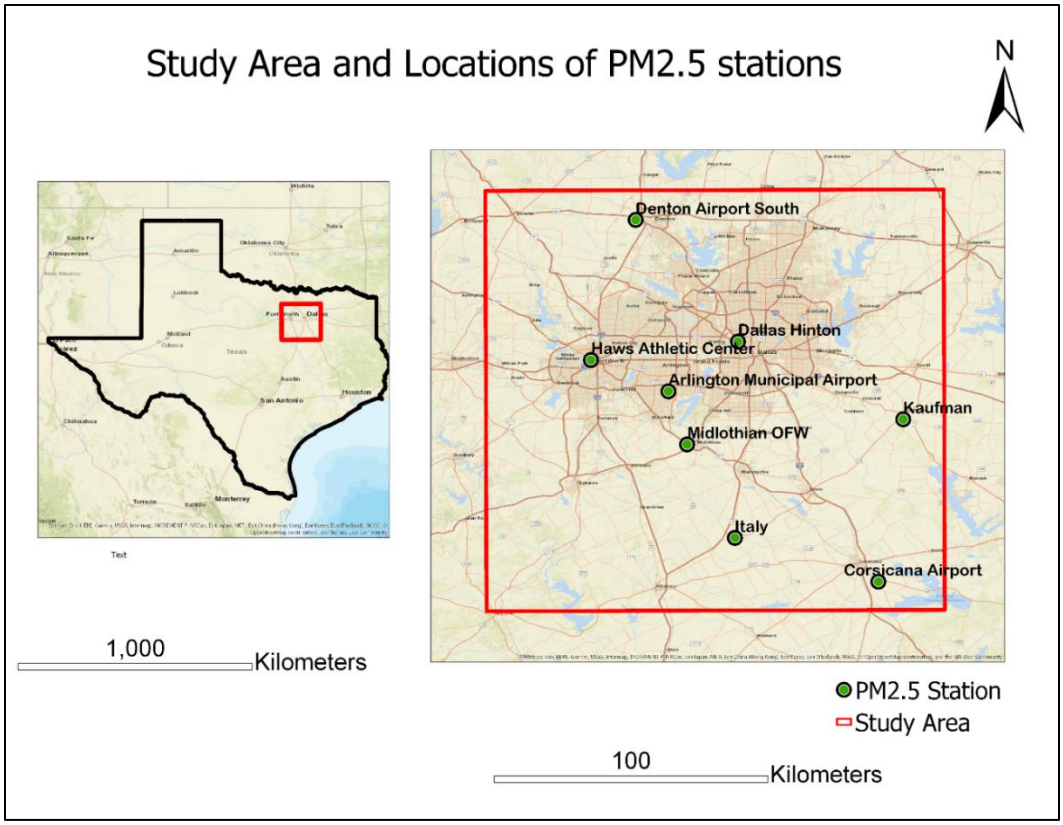


Figure 2.1: Distribution of PM2.5 stations in the study area

2.2.2 Data

The study used two sets of input data: aerosol optical depth (AOD) and AOD related variables from MODIS and meteorological data to estimate PM2.5 corresponding to the hour of MODIS overpass time.

PM2.5-

Generally, Terra and Aqua satellites overpass the study area around 10:30 am and 1:30 pm. PM2.5 data from ground monitoring stations are available at an hourly interval. The study used PM2.5 of the hour in which the MODIS overpasses the study area. For example, if MODIS

overpasses at 10:30 am, the PM2.5 measured between 10 am to 11 am was used. The data was downloaded from the Environmental Protection Agency's website (https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw) with the parameter code of the PM2.5 data 88502. A total of 10-year PM2.5 observations, from 2006-2015, were downloaded for the study area.

AOD-

AOD data from the MODIS have been available only at 10 km resolution. A recently developed algorithm, Multi-Angle Implementation of Atmospheric Correction (MAIAC) downscales AOD to 1 km resolution (Lyapustin and Wang, 2018). At 10-km resolution, two separate algorithms, Dark Target (DT) and Deep Blue (DB) retrieve aerosols from MODIS data over land surfaces for dark or vegetated surfaces and bright surfaces, respectively. In contrast, MAIAC retrieves aerosols over both dark and bright land surfaces. Besides providing AOD data at a finer resolution, MAIAC covers a greater spatial extent at a higher retrieval frequency with low bias, and high correlation with AOD from the Aerosol Robotic Network (AERONET) stations (Superczynski *et al.*, 2017; Mhawish *et al.*, 2019; Jethva *et al.*, 2019). AERONET stations are ground-based instruments that provide very accurate AOD values. Jethva *et al.* (2019) evaluate the performance of all three MODIS AOD algorithms (DT, DB, and MAIAC) over North America using 2002-2016 observations matching in space and time with AERONET stations. According to their findings, MAIAC provides more matching observations over eastern and western United States than DT or DB algorithms. Compared to DT and DB, MAIAC performs well over different surface conditions with very little bias while producing AOD at a 10-fold

finer resolution (Superczynski *et al.*, 2017; Jethva *et al.*, 2019). Furthermore, Superczynski *et al.* (2017) compare VIIRS and MAIAC on seasonal AOD coverage over North America using one-year data and conclude that MAIAC performs better over bright surfaces. The study also finds that MAIAC AOD performs almost uniformly across different zenith angles. Mhawish *et al.* (2019) confirm that MAIAC performs better than DT and DB AOD algorithms over South Asia using data from 2006-2016 regarding spatial coverage, low dependence on viewing geometry, aerosol load, and surface types.

Because of the superiority of MAIAC AOD over other AOD algorithms and its availability at a higher resolution, this study selects MCD19A2 version-6 data product for AOD estimated with MAIAC algorithm (hereafter, MAIAC AOD data). AOD is available at two wavelengths: 470 nm and 550 nm. This study uses AOD at 470 nm because AOD provided at 550 nm is derived from AOD at 470 nm, and AOD at 550 nm is marginally inferior in quality compared to AOD at 470nm (Lyapustin and Wang, 2018). MAIAC AOD data is transformed to WGS 1984 coordinate system using MODIS Reprojection Tool (MRT) and then space and time references of the MAIAC AOD are used to extract PM2.5 observations at the monitoring stations. MCD19A2 also provides quality flags for AOD, satellite retrieved water vapor content, and viewing zenith angle. This study used these variables along with MAIAC AOD. Data about zenith angle was available at 5 km resolution. Zenith angle data were resampled using nearest neighbor resampling to match the resolution of AOD data.

Meteorological data –

Meteorological data came from European Centre for Medium-range Weather Forecast (ECMWF). ECMWF provides reanalysis data worldwide, at the interval of 3,6, 9, and 12 hrs from 0:00 and 12:00 UTC (Berrisford *et al.*, 2011). Thus, it was available for the Dallas-Fort Worth metroplex four times a day, at 9 am, 12 pm, 3 pm, and 6 pm local standard time and at a spatial resolution of 0.125 degrees (~ 13 km). The reanalysis data combines weather observations with the most up-to-date weather models and provides information on different weather variables as a continuous grid (Parker, 2016). The various weather parameters obtained from ECMWF include horizontal and vertical components of the wind, wind gust, temperature, dew point temperature, clear sky surface photosynthetically active radiations, total precipitation, boundary layer height, boundary layer dissipation, total cloud cover, medium cloud cover, high cloud cover, convective precipitation, convective available potential energy, and evaporation. The study retrieved meteorological data closer (in time) to satellite acquisition time.

In total, the study used 21 predictor variables (see Table 2.1) to model PM2.5 from 8 stations. The first four predictors came from MAIAC AOD products from MODIS, and the remaining variables were from ECMWF reanalysis data. Predictors obtained from MODIS presented instantaneous observations at the time of satellite passing, whereas ECMWF reanalysis data provided four estimates per day.

Table 2.1: List of predictors

Sr. No.	Predictor	Measurement Unit	Spatial Resolution
1	AOD	-	1 km
2	AOD QA Flag	-	1 km
3	Column Water Vapor	cm	1 km
4	Cosine of Solar Zenith Angle	-	5 km
5	2-m Temperature	K	~ 13 km
6	2-m Dew Point Temperature	K	~ 13 km
7	Clear Sky Surface Photosynthetically Active Radiations	$J\ m^{-2}$	~ 13 km
8	Photosynthetically Active Radiations at the Surface	$J\ m^{-2}$	~ 13 km
9	Total Column Water Vapor	$kg\ m^{-2}$	~ 13 km
10	Boundary Layer Dissipation	$J\ m^{-2}$	~ 13 km
11	Boundary Layer Height	m	~ 13 km
12	Total Cloud Cover	Expressed as a fraction between 0-1	~ 13 km
13	Medium Cloud Cover		
14	High Cloud Cover		
15	Convective Precipitation	m	~ 13 km
16	Convective Available Potential Energy	$J\ kg^{-2}$	~ 13 km
17	10-meter U Wind Component (Eastward)	$m\ s^{-1}$	~ 13 km
18	10-meter V Wind Component (Northward)	$m\ s^{-1}$	~ 13 km
19	10-meter Wind Gust	$m\ s^{-1}$	~ 13 km
20	Evaporation	m of water equivalent	~ 13 km
21	Total Precipitation	m	~ 13 km

2.2.3 Methodology

Figure 2.2 shows the flowchart of the data and method used in the study. The input data have been discussed in the previous subsection. The study resampled meteorological data to match the resolution of the MAIAC AOD using the nearest neighbor resampling method. This section discusses data processing, model architecture, and evaluation.

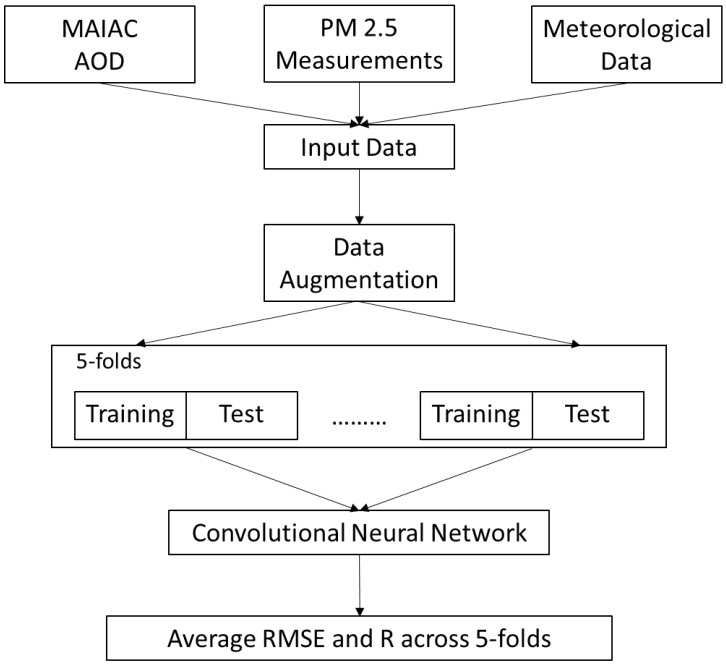


Figure 2.2: Flowchart of data and methodology

In contrast to Artificial Neural Network (ANN) approaches, CNN accounts for the influence of predictor values in the spatially adjacent locations. This is essential for the phenomenon affected by explanatory variables in the surrounding areas. As discussed in the introduction section, many studies have improved model performance after considering a correlation among variables in space. However, contrary to earlier studies, this study did not use

measurements from nearby PM2.5 monitoring stations but aimed to develop a model that uses AOD and meteorological data to estimate PM2.5 corresponding to an hour in which MODIS overpasses at specific sites. Additionally, Park *et al.* (2020) used spatially lagged predictors over a fixed distance. Instead, this study examined the influence of spatial lags over varying distances to evaluate the spatial scale effects of meteorological variables with AOD on PM2.5 estimates. The underlying grid resolution of AOD data was 1km × 1km. Expanding upon the grid size, the study constructed analysis extents ranging from 3km × 3km, 5km × 5km, 7km × 7km, …… up to 19km × 19km, centered at PM2.5 stations were extracted for this purpose from the images of the AOD, AOD quality flag, column water vapor, resampled zenith angle data and, resampled meteorological data. The process was repeated for each of the eight PM2.5 stations in the study area to form an input dataset to build a model in an O-CV approach. As such, the dependent variable in the study was PM2.5 from each of the eight stations, whereas the study's independent variables included AOD, AOD quality flag data, column water vapor data, zenith angle data, and meteorological data (as listed in Table 2.1). A total of nine CNN models, one for each grid size, were developed and compared. Each model had 21 input data grids for each grid size ranging from 3km × 3km to 19km × 19km. For example, the input data to each analysis extent was an N × N × 21 array. Figure 2.3 shows the conceptual representation of varying grid sizes of input data around the PM2.5 station.

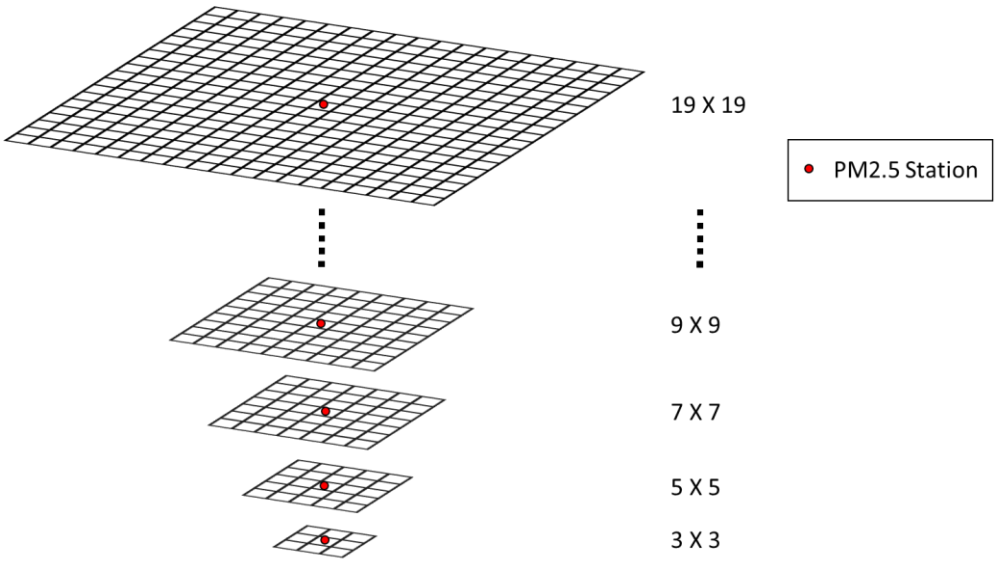


Figure 2.3: Different sized input data grids centered on PM2.5 station

2.2.4 CNN

This study built CNN models with convolutional layers and dense layers. The study used variable convolutional layers, depending on the grid size of the analysis windows for the predictor. Predictors were convoluted using filters of size 3 × 3 until the input image reduces to 1 × 1 pixel. The first and the second convolutions consisted of 24 and 16 filters, respectively, whereas each of the remaining convolutions consists of eight filters. Each of the dense or fully connected layers consists of eight neurons. Grids of size 3 × 3 and 5 × 5 required only one and two convolutions respectively whereas the remaining grid sizes required more than two convolutions. A grid of size 7 × 7 required three convolutions, whereas a grid of size 19 × 19 required seven convolutions. Figure 2.4 showed the architecture of the CNN used in the study. A blue square represents 3 × 3 filters used in all convolutions. In all layers except the last one, the study used a sigmoid activation function. The last layer, which outputs the model predictions,

used a linear activation function since sigmoid limits the output range from 0 to 1 and the linear activation regressed the predictions. The study used the Adam optimization algorithm. The learning rate of 0.01 and 200 epochs were used in the study. The learning rate of 0.01 was found to balance learning time and accuracy. The study used stride one and no padding across all convolutions. Also, batch normalization followed each convolution and dense layer prior to the ensuing activation function.

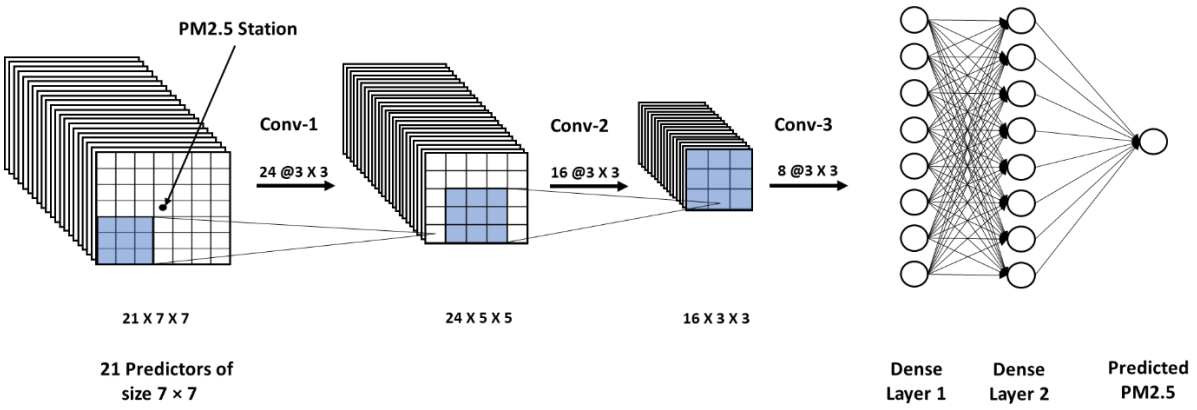


Figure 2.4: The study's CNN architecture

2.2.5 Data Augmentation

The problem of missing data in AOD was well documented. Hu *et al.* (2017) and Park *et al.* (2020) reported around 70% missing rate for 10 km AOD from MODIS in the year 2011 for the coterminous United States. Huang *et al.* (2018) found 50% of missing data for MAIAC AOD in North China during 2013-2015. Goldberg *et al.*(2020) identified the varying availability of MAIAC AOD ranging from 0.2% to 92.3 % in the eastern United States in 2008. As a larger grid

comprised of more pixels than smaller grids, the probability of missing AOD for at least one pixel was greater for larger grids. The study included only those samples with AOD data available for all pixels in each analysis extent. The problem of missing data in AOD led to decreasing the probability of AOD availability for all pixels as the spatial extent of the grid increased. Therefore, the study limited the spatial extent of the input data to the grid of size 19 × 19. Table 2.2 summarized the number of samples available across different sized input grids.

Table 2.2: AOD data availability

Input grid size	Number of samples
3 × 3	14570
5 × 5	12674
7 × 7	10686
9 × 9	8407
11 × 11	7488
13 × 13	6660
15 × 15	5703
17 × 17	5165
19 × 19	4205

Machine learning approaches, such as CNN, required a large number of samples. The relatively small study area and only 10-years of the study period resulted in small samples in the context of machine learning. Data augmentation was a common practice to generate additional samples

without introducing new information to the data, such as geometric transformation. Geometric transformations of image augmented image samples by flipping, scaling, rotating, and cropping original images (Taylor and Nitschke, 2017). Similarly, color transformation and neural style transfer are few other ways to augment the data (Shorten and Khoshgoftaar, 2019). Among these, flipping and rotating are computationally simple approaches. This study flipped image data to generate mirror copies along an axis and rotated image data to create copies of images in different orientations. Figure 2.5 illustrated a sample image of size 5 × 5 whereas
Figure 2.6 showed augmented images obtained by rotating and flipping sample images. Images of all the input variables in a particular sample were flipped or rotated in the same way to form a new sample. As a result, the process of data augmentation only repositioned the original samples without making any change to original data values or their inter-relation in spatial configuration. As the study used six different ways to augment the data (
Figure 2.6), each sample was reconfigured in 6 different ways, resulting in a 6-fold increase in the number of training samples.

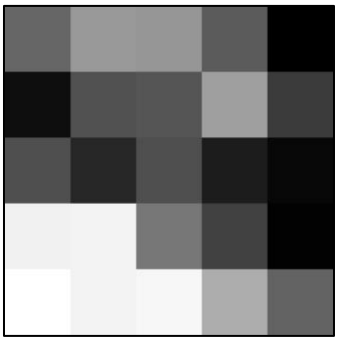


Figure 2.5: Sample image

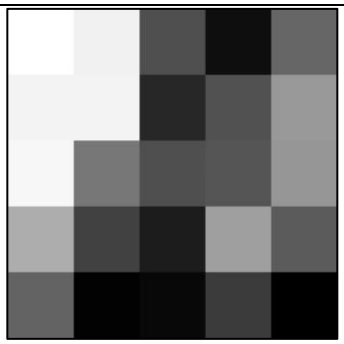
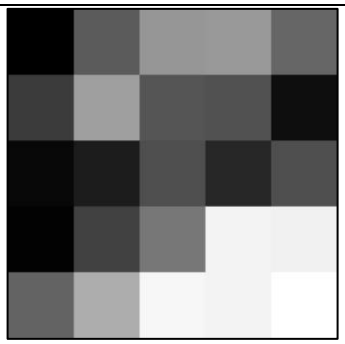
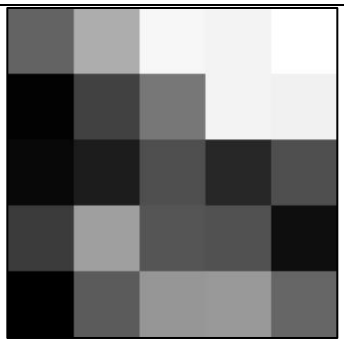
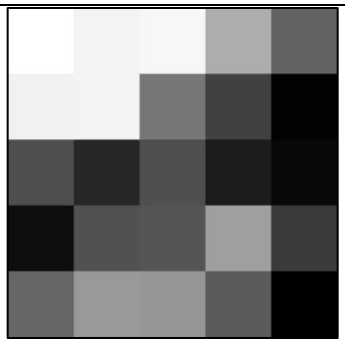
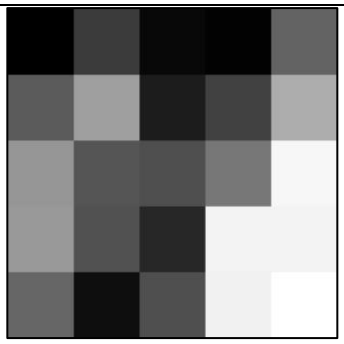
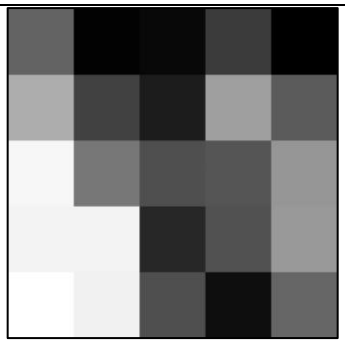
Rotated Images	Flipped Images
 (a) 90^0 rotation	 (d) Original image flipped about the vertical axis
 (b) 180^0 rotation	 (e) Original image flipped about the horizontal axis
 (c) 270^0 rotation	 (f) 270^0 rotated image flipped about the vertical axis

Figure 2.6: Augmented images

2.2.6 Cross-validation/test dataset

As already mentioned in the introduction, this study used the O-CV approach for evaluating the model performance. Specifically, the study adopted the 5-fold O-CV approach. The data is split into 5 groups, each group is iteratively used for testing the model performance and the remaining 4 groups for training the model. The average correlation coefficient (R) and root mean squared error (RMSE) across all 5 groups was used to compare model performance.

2.3 Results and Discussion

Figure 2.7 shows the results for CNN across different sized input grids for PM2.5 estimation. A larger grid size represents spatially lagged predictors from a wider area around a PM2.5 station. Increasing grid size resulted in increased R between observed and estimated values and decreased RMSE. This showed that predictor values in the surrounding area influenced PM2.5 concentration. Out of all the grid sizes, the model with input grid size 19×19 performed the best with R of 0.87 (or R^2 of 0.76) and RMSE of 2.57 $\mu g/m^3$. Unlike other studies in the literature, this study achieved reasonably good performance without including PM2.5 from nearby stations as a covariate. In the study carried out for the conterminous United States, Di *et al.* (2014) achieved R^2 of 0.84 whereas Park *et al.* (2020) reported R^2 of 0.84 and RMSE of 2.55 $\mu g/m^3$. Similarly, the study performed in China for daily PM2.5 estimation (Zhan *et al.*, 2017), reported R^2 and RMSE of 0.76 and 13 $\mu g/m^3$ respectively. All these studies estimated PM2.5 aggregated over 24-hours. Aggregation of PM2.5 over 24-hrs masked variation throughout the day. This study developed the model to estimate PM2.5 at two different times a day, corresponding to the

hours of MODIS overpass time. The MODIS AOD was an instantaneous observation whereas each PM2.5 observation from the ground station was a measurement averaged over a respective hour. The CNN model developed in the study did not rely on measurements from nearby PM2.5 stations and still achieved R^2 comparable to other studies.

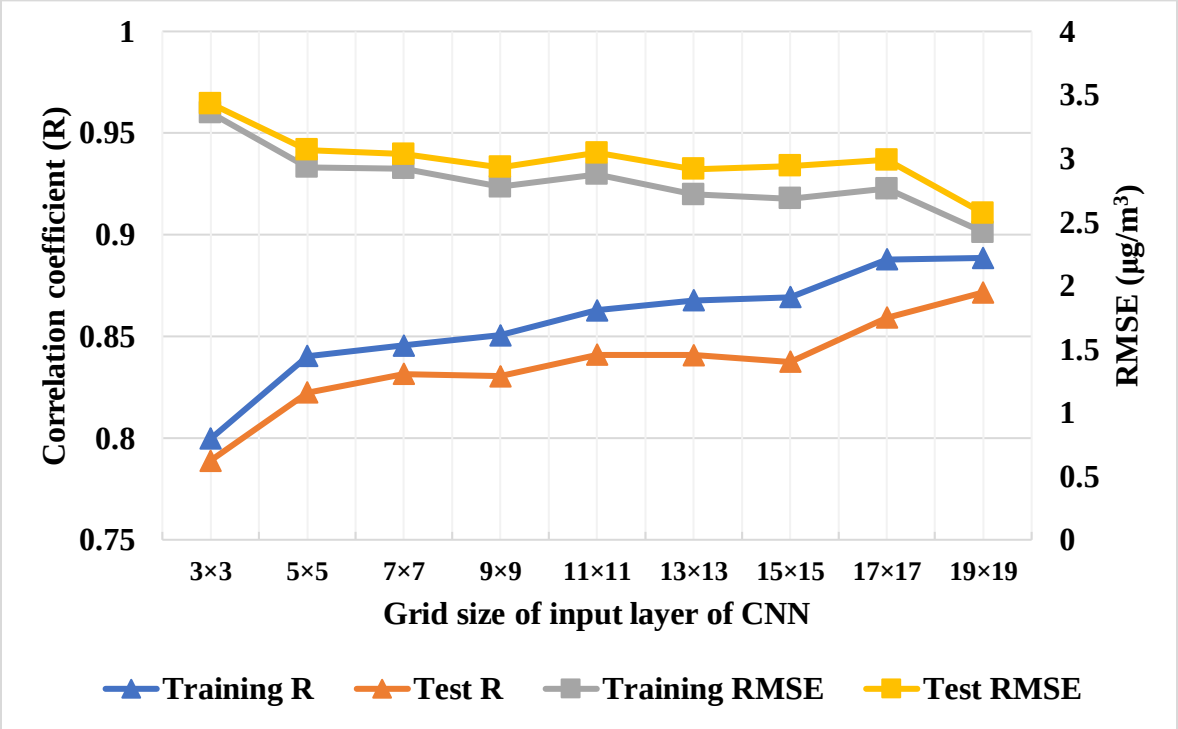


Figure 2.7: Correlation coefficient and RMSE for CNN with varying grid size

Machine learning methods require a large amount of data to train the model. AOD data are often susceptible to data gaps due to cloud cover or bright surfaces. There was a limited number of samples for the study area over 10 years of study period because of the missing data problem in AOD. Also with the larger grid size, chances of missing AOD data in the grid increased. That further reduced the number of samples available for the analysis. This study used a data augmentation technique to artificially increase the number of samples. The data

augmentation technique involved flipping and rotating original images to increase the sample size. Thus the data augmentation technique helped to increase the number of samples without introducing any new information to the data. For the training dataset, a correlation between estimated PM2.5 from MAIAC AOD and observed PM2.5 at monitoring stations increased with the grid size, but for the test dataset, the model performance was substantially degraded (Figure 2.8). Similarly, the model did not perform as well on the test dataset as the training dataset in terms of RMSE. This suggested that the model performed well on the training dataset with a smaller number of observations, but it failed to perform equally well over unseen data, a case of overfitting. Figure 2.9 compared the performance of models with and without data augmentation on the test dataset in terms of % change in R and RMSE. Data augmentation improved R and decreased RMSE in models of all grid sizes.

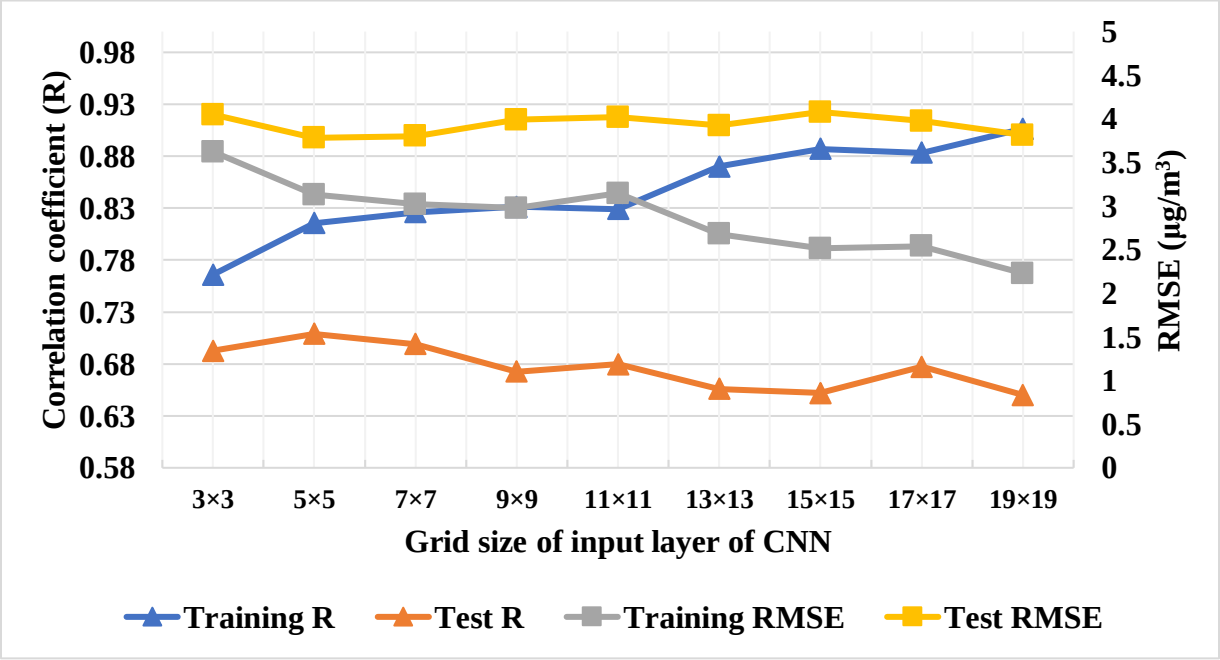


Figure 2.8: Correlation coefficient and RMSE for CNN with varying grid size without data augmentation

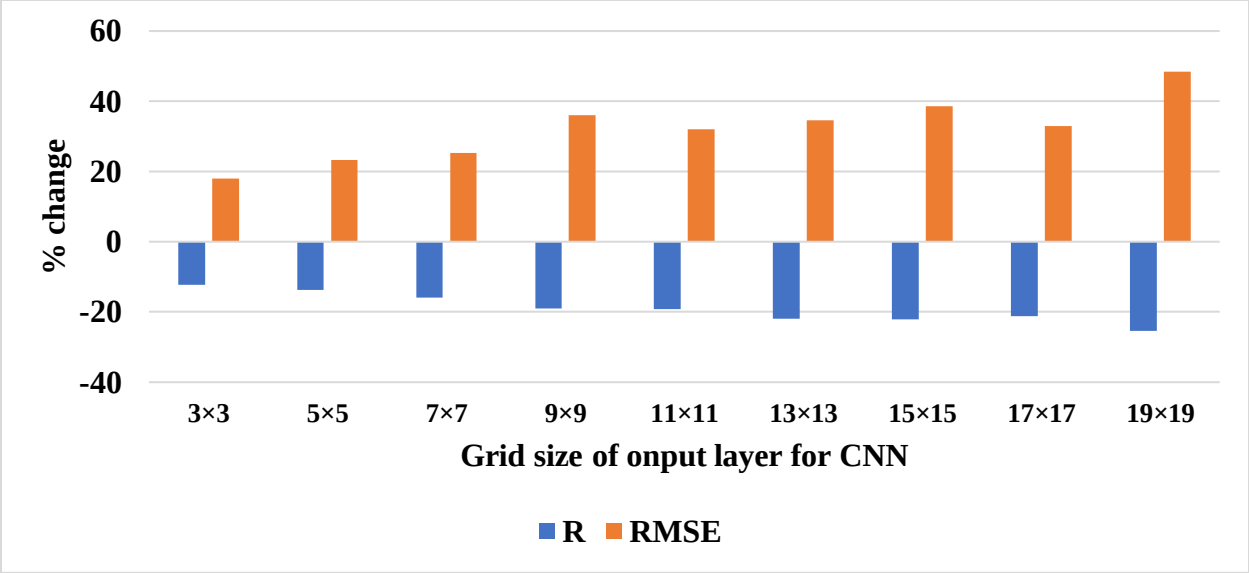


Figure 2.9: Percent change in R and RMSE in models without data augmentation

2.4 Conclusion

This chapter outlined a CNN architecture to estimate PM2.5 corresponding to an hour of MODIS overpass time using MAIAC AOD and meteorological variables. The CNN accounted for the effect of predictors from spatially adjacent locations. The study evaluated the performance of the CNN over grids (of predictors) of different sizes. A larger grid size incorporated the influence of predictors from a greater number of spatially adjacent pixels. The results of the study showed that the larger grid size improves model performance. Highest R and lowest RMSE were achieved with 19 × 19 grid size. That showed that the inclusion of spatially lagged predictors over a wider area can improve model performance. Although this study did not find optimum grid size for PM2.5 estimation, the inclusion of spatially lagged predictors within 10km provided satisfactory model performance with R of 0.87 and RMSE of 2.57 $\mu g/m^3$. Furthermore, the model developed in the study allowed estimation of PM2.5 corresponding to the hour of

MODIS overpass time. Also, PM2.5 is estimated at a resolution of 1 km resolution. The counterpart studies estimated PM2.5 at a coarser spatial and temporal resolution. Previous studies stressed the importance of incorporating spatial dependence, evidenced by the improved model performance with the inclusion of spatially lagged PM2.5 from nearby stations. However, PM2.5 stations are often sparse and the density of these stations may affect the model performance. The model developed in the study achieved comparable performance without including spatially lagged PM2.5. Therefore, the CNN architecture developed in the study can be used to improve the availability of PM2.5 at finer spatial and temporal scales at unmonitored locations.

The study suffered from a limited number of samples due to missing AOD. The study did not conduct any analysis to see if missing AOD values created any systematic bias in the model. Regardless of the missing AOD data, the data augmentation technique proved helpful to train the model. However, alternatively, this limitation can be overcome in the future by using imputation techniques to fill missing AOD values. The CNN architecture used a fixed number of filters and neurons in convolution and dense layers, respectively. Hyperparameter tuning can help identify an optimal number of filters and neurons in these layers. Furthermore, this study did not account for temporal autocorrelation in the data as machine learning methods do not require temporal independence in the data. Also, the high correlation between observed and estimated PM2.5 values suggested that covariates used in the model accounted for the temporal variability in the data. However, future studies can explore the inclusion of temporal autocorrelation in machine learning methods. Lastly, the sparse configuration of the air quality stations did not provide

enough spatial variability to estimate PM2.5 values at the unmonitored locations using interpolation techniques such as cokriging. However, recently low-cost sensors from Purple Air are deployed in large numbers across the United States. Though the accuracy of these sensors remains a cause of concern, Barkjohn *et al.* (2020) developed a model to correct bias in the low-cost sensor measurements using temperature and relative humidity. Therefore, with appropriate correction, these sensors may provide spatially dense measurements across the region to use interpolation techniques to estimate PM2.5 at an unmonitored location. A comparative analysis of estimates from PM2.5-AOD modeling and estimates provided by cokriging will be a good future study to gain insights about the other factors which are responsible for the errors.

CHAPTER 3

MICROSCALE DYNAMICS OF PM2.5: A CASE STUDY AT THE UNIVERSITY OF TEXAS AT DALLAS

3.1 Introduction

Information about PM2.5 concentration from the nearest air quality monitoring station may not represent PM2.5 in a microenvironment area. MODIS provides AOD data at 10km, 3km, and 1km resolutions. These resolutions are insufficient to estimate PM2.5 variations within neighborhoods or local urban environments. A finer scale estimation requires covariates at a finer scale. Moreover, most monitoring stations are in open areas to avoid the effect of buildings and trees. Because of their fixed locations, PM2.5 observations from these stations provide limited opportunities to study spatial variabilities of PM2.5 in places where most people live, and the health effects are most impactful. Recently, several studies applied mobile monitoring platforms to measure pollutant concentration rather than relying on data from ground monitoring stations (Harrison *et al.*, 2015a; Shi *et al.*, 2016). Mobile monitoring platforms provide flexibility to gather data at high spatial resolutions and temporal frequencies, in contrast to data from ground monitoring stations which are often sparsely located. Harrison *et al.* (2015a) and Shi *et al.* (2016) have collected PM2.5 data across roads in urban areas using vehicles as a platform for PM2.5 monitoring. Few studies used mobile aerial or ground vehicles to study the horizontal and vertical profiles of PM2.5 (Harrison *et al.*, 2015b; Peng *et al.*, 2015). These platforms offer flexibility to measure PM2.5 concentration in varied environments.

Spatial obstructions like buildings and trees affect the dispersion and concentration of PM2.5 (Vos *et al.*, 2013, Ginzburg *et al.*, 2015; Shi *et al.*, 2016; Brown *et al.*, 2019). Shi *et al.* (2016) included building morphological characteristics in the land-use regression model to estimate PM2.5 and PM10 during stable meteorological conditions and improved model R^2 by 10% in the high-density area of Hong Kong downtown. Other studies examined how urban morphology influences the flow of air pollutants in the high-density environment of Hong Kong. Some of the key variables affecting the air pollution in these studies include site coverage, average building height, distance between buildings, and degree of enclosure (Edussuriya *et al.*, 2014; Yang *et al.*, 2020).

In addition to physical structures or trees that might influence dispersion, PM2.5 distributions have complex relationships to geographic features. PM2.5 refers to all particles with an aerodynamic diameter smaller than 2.5 μm. The chemical components of PM2.5 vary depending on the sources of emission in a region. Vehicle exhaustion is one of the major sources of black carbon, and black carbon correlated better with traffic patterns than with PM2.5 distributions (Wang *et al.*, 2018). PM2.5 observations from near-road sites appeared strongly correlated with nearby sites on background pollution (Brown *et al.*, 2019), but PM2.5 averages at near-road sites appeared higher than other nearby sites (Ginzburg *et al.*, 2015; Brown *et al.*, 2019). This indicates that on average, near road areas are susceptible to relatively greater PM2.5 exposure. However, studies conducted at near-road sites across the United States have shown that the traffic characteristics like traffic volume and traffic speed do not correlate well with PM2.5, but meteorological factors and their interaction with site characteristics have profound

impacts on PM2.5 (Ginzburg *et al.*, 2015; Brown *et al.*, 2019). In both prior studies, a stable meteorological condition at night showed elevated PM2.5 concentration, whereas higher wind speed reduced the PM2.5 level. A source apportionment analysis at a roadside location in Maryland showed that on-road traffic contributes only 12.5-17% of PM2.5 (Ginzburg *et al.*, 2015). According to Brown *et al.* (2019), high PM2.5 is associated with wind near perpendicular to the road since the wind sweeps the maximum surface area of the road towards the monitor than when it blows precisely perpendicular to the road segment. Moreover, the monitoring station with buildings on the windward side of the wind displayed consistently high PM2.5, suggesting that the presence of buildings trapped the pollutant coming from the roads.

Thus, besides emission sources, geographic features, spatial forms, and their interactions with wind play a role in determining PM2.5 concentration. Mobile monitoring platforms allowed studying PM2.5 distributions in different urban microenvironments, such as the effect of trees and building near emission sources (e.g., roads) on the PM2.5 concentration (Shi *et al.*, 2016; Deshmukh *et al.*, 2019;). However, PM2.5 research in low density and non-near road sites appeared lacking but can offer insights into the dynamics of PM2.5 concentration in microenvironments typical of pedestrian areas or shopping plazas. This study aimed to investigate PM2.5 variations in such a built environment.

3.2 Data and Methods

3.2.1 Study area

This study chose the University of Texas at Dallas (UTD) campus, located in the city of Richardson, Texas, USA, as an example of pedestrian-dominant urban settings where city dwellers are most likely to experience direct exposure to air pollution than indoor or in-vehicle in other urban settings. Four roads surrounded the University: Campbell road, Waterview Parkway, Floyd road, and Synergy Parkway, with peak hour traffic counts of 1912, 1698, 569, and 625 respectively (Traffic Count Program, Annual report, Dec 2019, City of Richardson). Specifically, the study focused on five paths in the interior of the University, away from all the major roads mentioned above. Buildings with various characteristics surround each of the five paths. Data collection followed these paths when traffic was light. Figure 3.1 shows the locations of the selected five paths: A, B, C, D, and E on the campus.

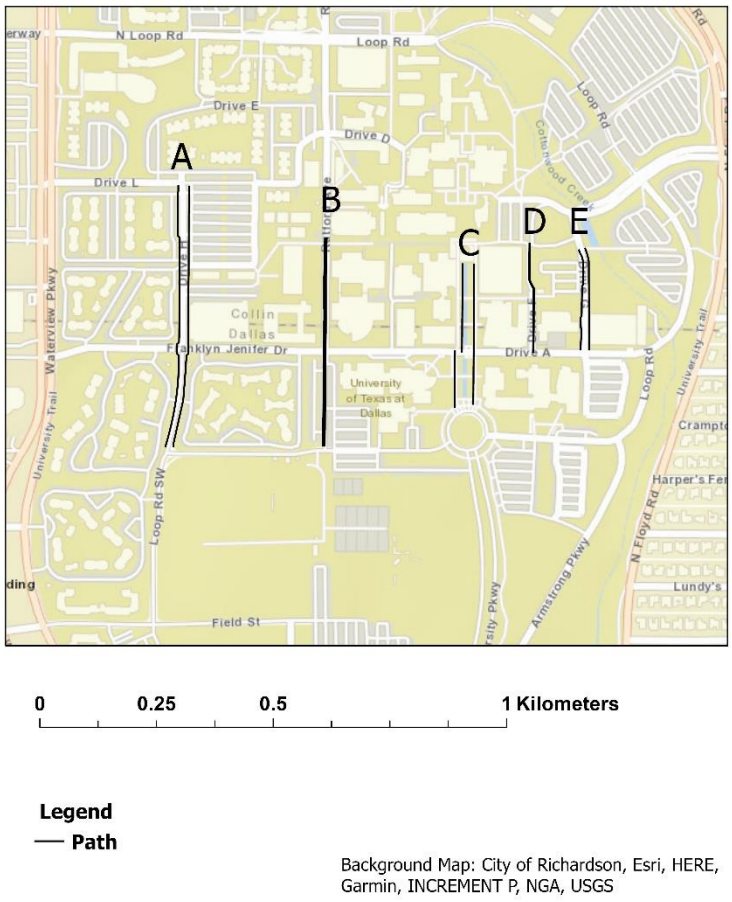


Figure 3.1: UTD Campus with Data Collection Paths A, B, C, D and E

3.2.2 Data Source

PM2.5 data

The study used DR1000, a flying laboratory from Scentroid, to collect the PM2.5 data. Before data collection, DR1000 was calibrated using co-located measurements from the fine dust measurement device, Fidas® Frog. Fidas® Frog measured the mass of particles in different size bins. DR1000 records PM2.5 measurements at an interval of 3-4 seconds, whereas Fidas® Frog

for every 5 seconds. After the calibration, both instruments collected data simultaneously for about an hour. The correlation between the one-minute average of PM data collected from both the instruments was 0.9. Each collection of PM2.5 observations ran through all five paths with the DR1000 mounted on a bicycle at a height of about 1.2m.

A total of nine data-collection runs were completed: eight runs in Dec 2019 and one run in Feb 2020. Out of these nine runs, three runs collected data in the mornings, three in afternoons, and three in evenings. Each data collection run took about an hour. Each run collected data from the north to south direction, repeated from the south to north and repeated three times in each direction.

Weather data

Variables such as wind direction, wind speed, temperature, and relative humidity could affect the PM2.5 observations. The weather station on the roof of Residential Hall West on campus supplied weather data at the observation frequency of one minute. The weather station was at about 320m distance from path A in the northwest. The study retrieved weather data closest to the PM2.5 timestamps.

Building data

The study used building footprints released by Microsoft (https://github.com/Microsoft/USBuildingFootprints). However, footprints for new construction on campus were missing in the data. Such footprints were digitized manually.

Digital Surface Model (DSM)

A high-resolution LiDAR point cloud for the study area was downloaded from https://nationalmap.gov/3DEP/. A DSM with the 50-cm resolution was derived from the LiDAR point cloud using ArcGIS Pro.

3.2.3 Data Preparation

Figure 3.2 depicts the flowchart for data preparation. The following subsection describes the steps involved in preparing the final dataset for the analysis.

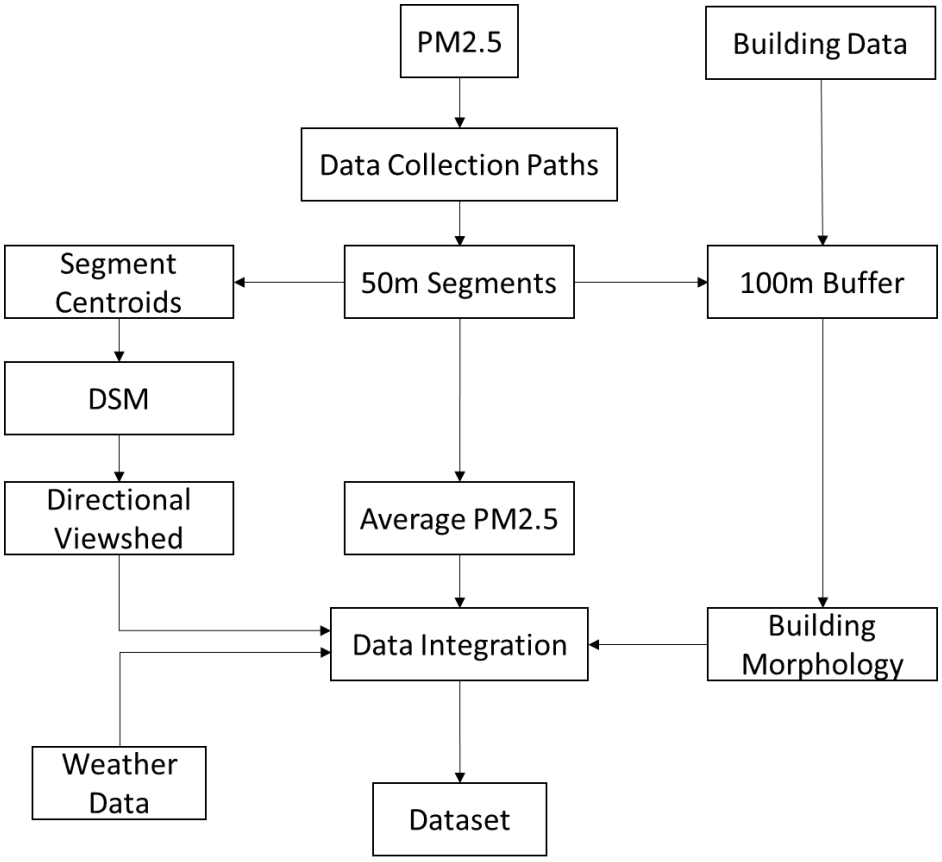


Figure 3.2: Flowchart of data preparation

Segment-wise PM2.5 Data

The PM2.5 data came from data collection paths shown in Figure 3.1. Data processing included steps to divide each path into segments of 50m and average PM2.5 observations along each 50m segment as a representative measure. The divide-and-average process smoothened the measurement uncertainty from the instrument and variations in movement speed during the collection, while sufficient to capture environmental variabilities within and between all paths.

Building morphology

Building morphological parameters included building coverage ratio (ratio of the area occupied by buildings to total buffer area), mean building area, number of buildings, the average distance between nearest-neighbor buildings and mean building height. All these parameters could affect pollution dispersion (Edussuriya *et al.*, 2014; Yang *et al.*, 2020). In addition to building morphological characteristics, Shi *et al.* (2016) considered parameters: frontal area index and sky view factor to model PM2.5 and PM10. Both these factors were related to openness. The authors found that the frontal area index plays an important role in determining PM2.5 concentration. A frontal area index represented building area in the direction of the wind. Since this study already considered openness in the direction of the wind (see directional viewsheds in the next section), the frontal area index was excluded from the analysis.

This study assumes that the 100m buffer around each 50m segment is sufficient to study the effect of building morphology on PM2.5 at the 50m segment. Therefore, buffers of 100m around each 50m segment delineated the proximity around roads to extract building characteristics for calculating morphological parameters.

Depending on the spatial relation of buildings with the 100m buffer, we categorize buildings into three classes:

1) Buildings, with their centroid falling in the buffer
2) Part of the building/buildings intersecting with the buffer
3) Buildings, within or touching the buffer

Figure 3.3 depicts buildings in all three classes highlighted in orange color. Buildings, with their centroid falling in buffer (class 1) are used to compute mean building height, mean building area, and the number of buildings per 1000 sqm. This is to avoid considering the buildings of which only a small portion fall in the buffer. In order to compute the building coverage ratio, this study considered the portions of the building intersecting with the buffer (class 2). Furthermore, to determine the average nearest neighbor distance between buildings in the buffer, this study considers all the buildings which were either within the buffer or touching the buffer (class 3) as a partial account of buildings would not reflect the true distance between them.

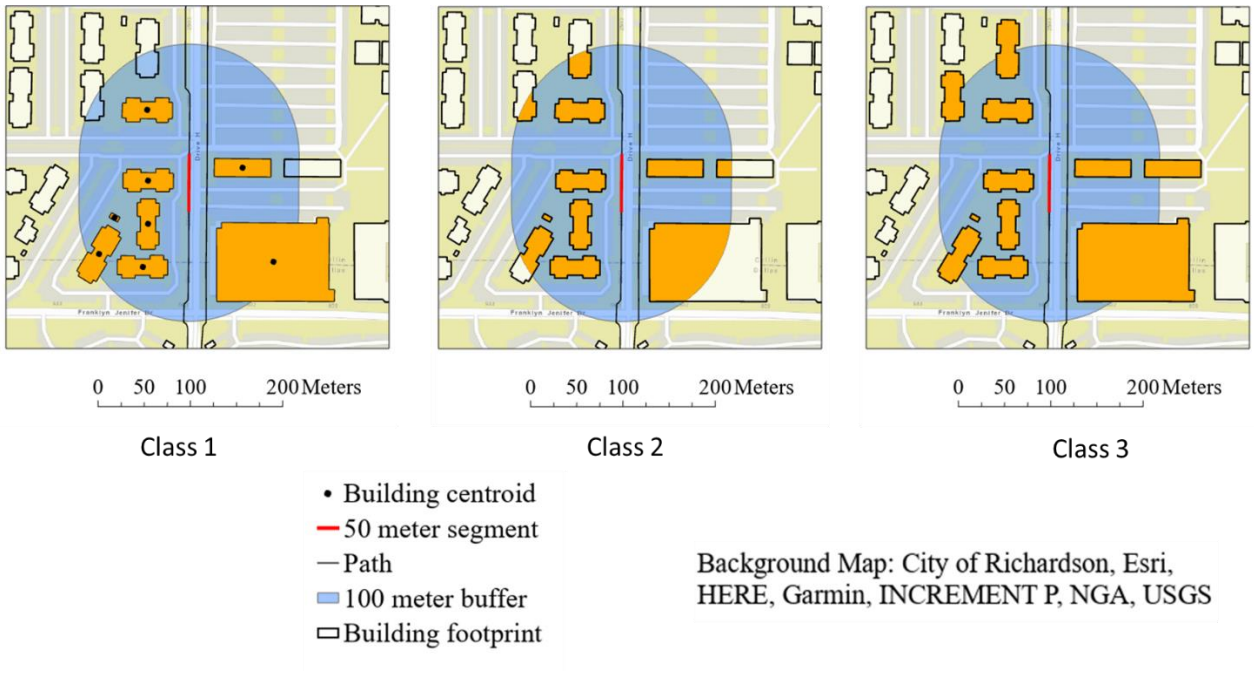


Figure 3.3: Buffer and buildings used in building morphology calculation

The discussion below clarifies the relevance and computation of each of these parameters:

1. Mean building area.

Mean building area measured the average size of buildings in the area around the street segment. Larger and taller buildings provided a greater total enclosure and hence hampered the PM2.5 dispersion and vice-versa.

Mean building area = (Sum of areas of all building footprints in 100m buffer)/(Number of buildings in 100m buffer)

2. Number of buildings

The number of buildings along with the mean building area gave an idea about the amount of built-up area. A large number of large-sized buildings occupied more space and hence left less

room for dispersion and vice versa. Segments at the end of a road could have a length smaller than 50m. The buffer drawn around these segments would have a smaller area than buffers around the other segments. Therefore, the number of buildings is standardized to the number of buildings per 1000 square meter buffer area.

*Number of buildings per 1000 square meters = (Number of buildings in 100m buffer)*1000/Area of buffer in square meters.*

3. Mean building height

Mean building height = (Sum of heights of all buildings in 100m buffer)/ Number of buildings in 100m buffer.

4. Average nearest neighbor distance between buildings

This study used the generate near table tool from ArcGIS Pro to calculate the distance between a building and its nearest neighbor building and based on which calculated the average nearest neighbor distances corresponding to all the buildings in the buffer. Tightly packed buildings led to pollutant accumulation, whereas a greater distance between buildings left more room for dispersion.

5. Building coverage ratio

The area covered by building footprints was also one of the features that impacted pollution dispersion. The larger the area covered by buildings, the lesser was the space available for pollution dispersion. The building coverage ratio measured the areal ratio of the 100m buffer

covered of building footprints to the total area of 100m buffer. The greater the building coverage ratio was, the less room available for dispersion and vice-versa.

Building coverage ratio = Area of 100m buffer occupied by buildings/Area of 100m buffer

Figure 3.4 to Figure 3.8 depict the building morphological characteristics within 100 m buffer around segments across all five paths in the study area.

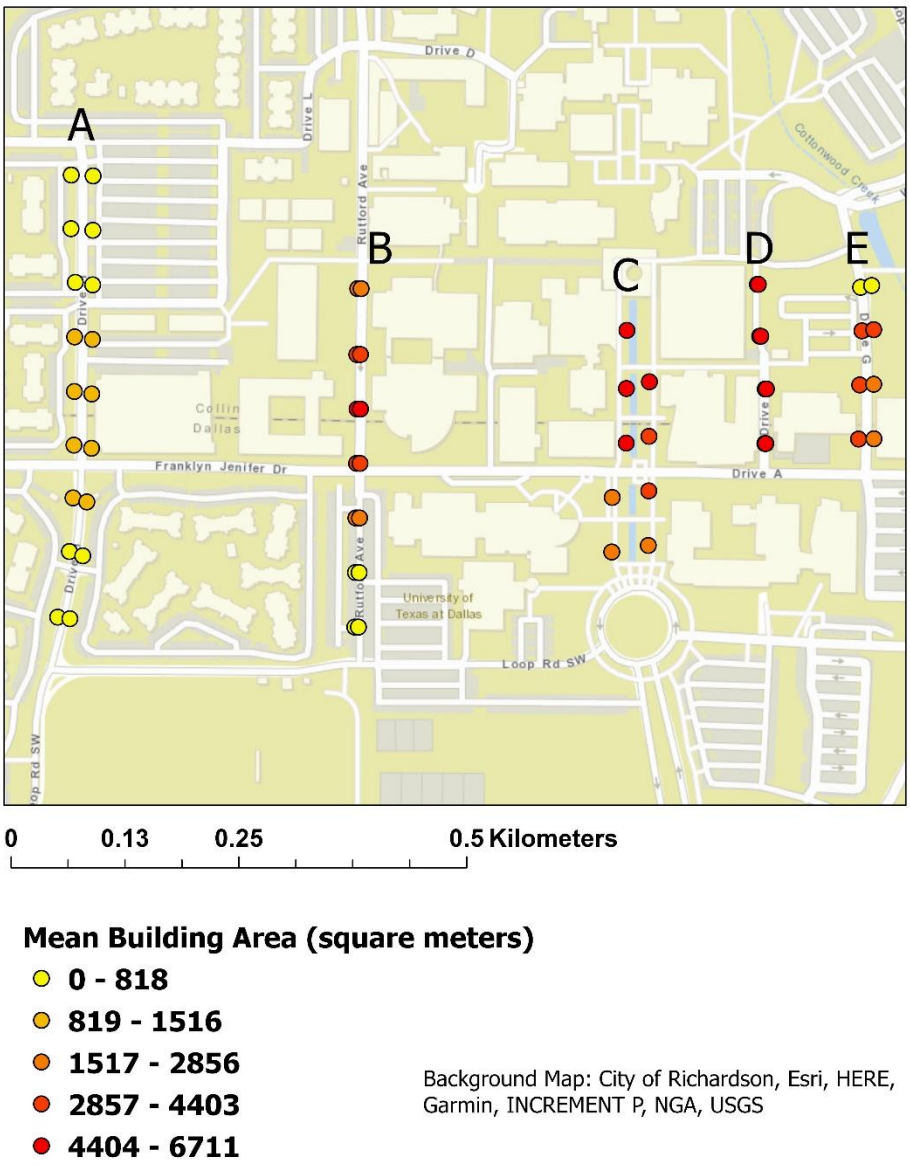


Figure 3.4: Mean building area

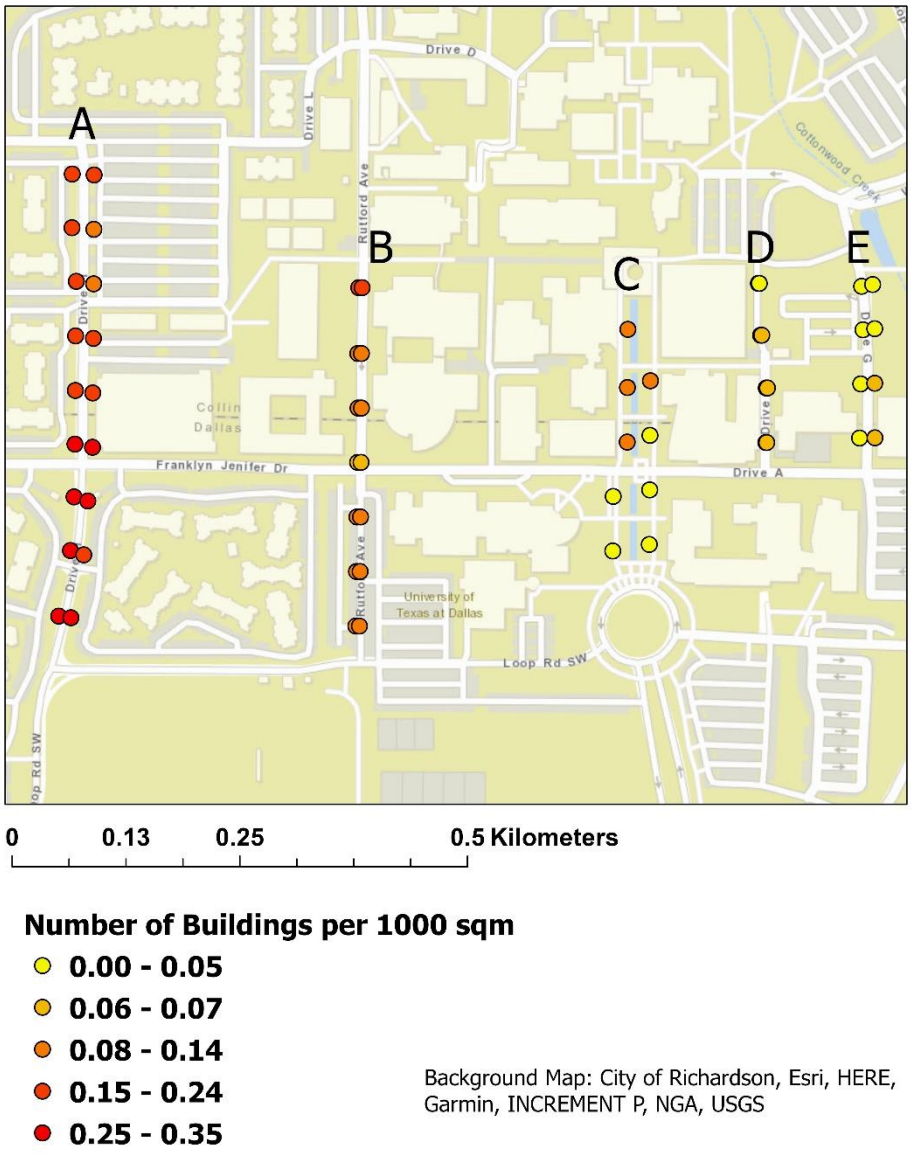


Figure 3.5: Number of buildings per 1000 m^2

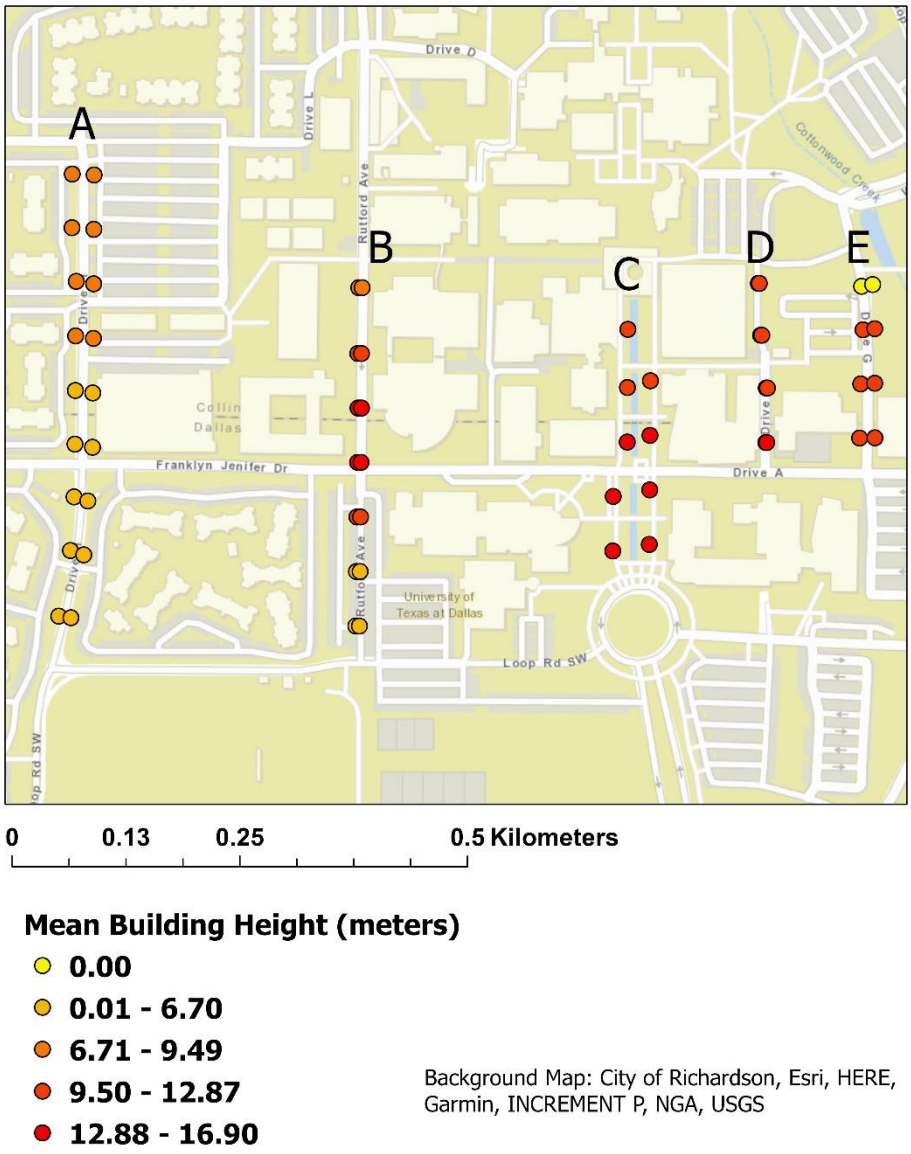


Figure 3.6: Mean building height

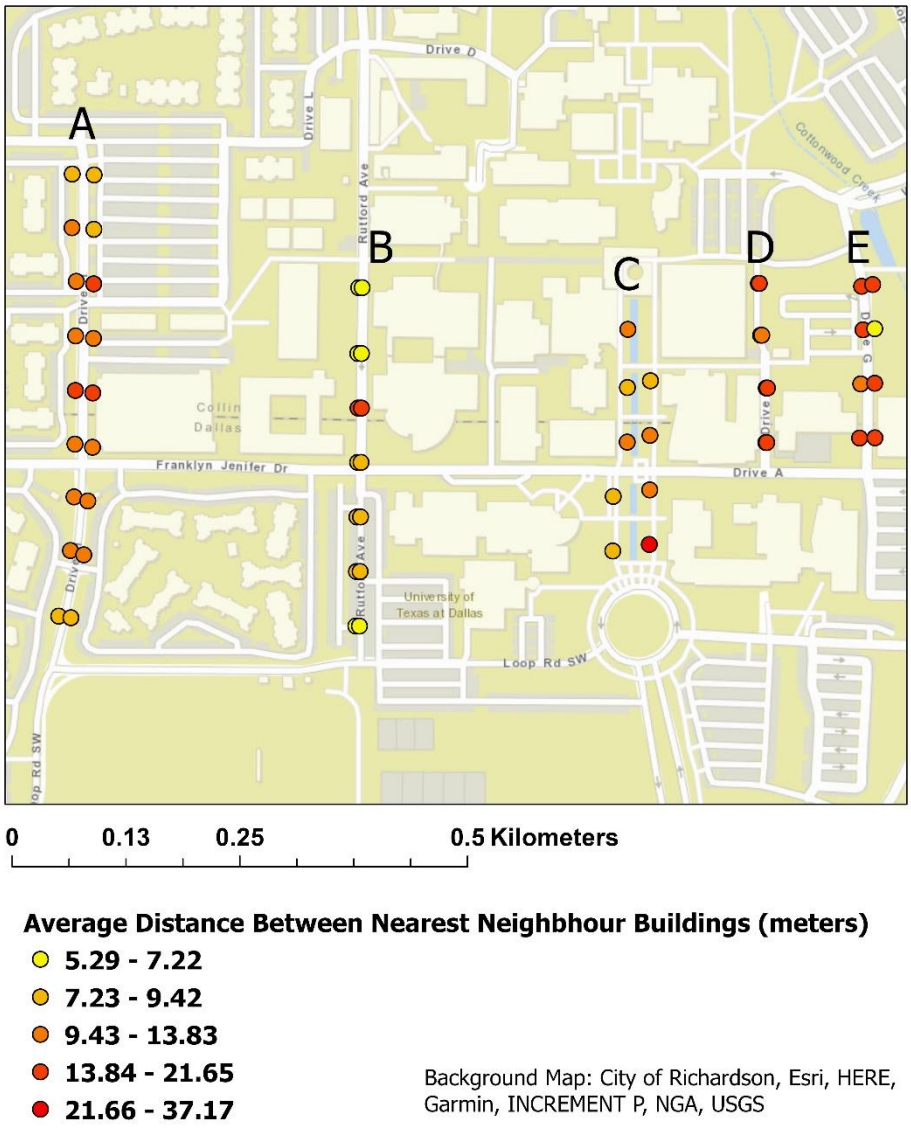


Figure 3.7: Average distance between nearest neighbhor buildings

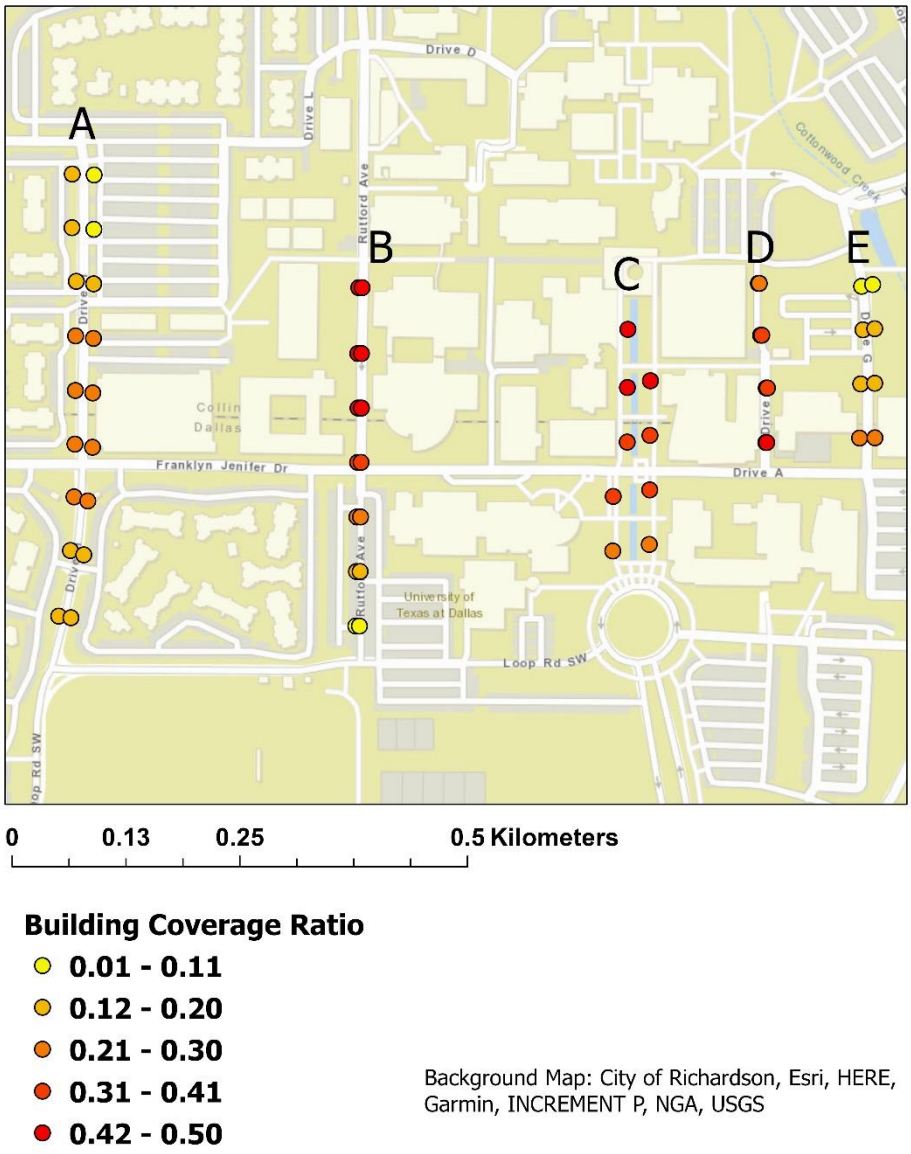


Figure 3.8: Building coverage ratio

Table 3.1 summarized building morphological characteristics along data collection paths. Among all the paths, the building coverage ratio of paths A and E were smaller. Many small buildings

surrounded path A, whereas a few bigger and taller buildings surrounded path E. Path C had the largest site coverage ratio, followed by path D, and followed by path B. In contrast, few bigger and taller buildings surrounded path E. Path C had the largest site coverage ratio, followed by path D, and followed by path B. Still, on average, buildings around paths C and D were bigger in size, taller in height, and fewer in number compared to path B. In short, path A constituted smaller, shorter, and denser buildings with intermediate building coverage. Path B consisted of moderately dense medium-sized buildings with median height and median building coverage. Path C was composed of relatively low density large-sized, and taller buildings but with the highest building coverage. Path D contained largest-sized low density and distantly placed buildings with medium height and relatively high building coverage. Finally, path E had small-sized sparsely placed buildings surrounded by short buildings. Path E also had the lowest building coverage among all paths.

Table 3.1: Summary of building morphological characteristics around data collection paths

	Building coverage ratio	**Mean building area (meter^2)**	**Number of buildings per 1000 meter^2**	**Average distance between nearest neighbor buildings (meter)**	**Mean building height (meter)**
Path A	0.19	886	0.23	12	6.70
Path B	0.33	2837	0.12	9	9.70
Path C	0.38	4132	0.07	13	14.80
Path D	0.36	5661	0.06	14	12.00
Path E	0.16	2520	0.04	16	8.40

Directional viewshed

The amount of open area around a location could impact the PM2.5 value observed at a location. Brown *et al.* (2019) provided evidence for the effect of interaction between wind direction and site characteristics on PM2.5. Thus, this study calculated open area in the wind direction for each segment centroid. The study calculated open area in the direction of wind for each segment centroid at varying distances of 100m, 200m, 400m, 800m, and 1500m and applies viewshed analysis to determine the amount of area visible from any given location in all directions. The viewshed analysis considered 12 different wind directions, starting with 0^0 in an increment of 30^0, resulting in viewsheds at 15^0, 45^0, 75^0….,345^0. These directional viewsheds served as the basis to analyze the interactions of PM2.5 and openness based on the corresponding wind direction.

Data integration

In each data collection run, mean PM2.5 for each 50 m segment was obtained by averaging all the PM2.5 data points associated with the respective segment. Based on the timestamp of the segment-wise PM2.5, the weather data closest in time is combined with the segment-wise PM2.5 data. Further, based on the wind direction, directional viewshed data was integrated with PM2.5 and weather data. In combination, the study datasets consist of the segment-wise PM2.5, weather-related variables, and directional viewshed. The resulting dataset was further combined with the corresponding building morphological data.

3.2.4 Modeling

This study collected spatial panel data because the data consisted of the same set of locations with observations at multiple times. The study included time-variant variables to explain temporal variations in the PM2.5 observations. Location-specific variables, such as building morphological characteristics, would not change over time, and they would have fixed effects on the response variable. The study built a fixed effect panel model to assess the effect of time-varying variables on PM2.5 and extract the individual or location-specific fixed effects of time-invariant variables.

The equation for the fixed effect panel model was expressed as follows:

$$Y_{it} = \alpha_i + \beta * X_{it} + e_{it}$$

Here, Y was a dependent variable, and *i* referred to a fixed location; *t* referred to the time at which an observation was collected; and e_{it} was the error term. Intercept α_i represented a fixed effect which was time-invariant but varied across locations. Following fixed effects models, the study estimated coefficient β by "demeaning" that removed the average over time from each observation:

$$(Y_{it} - \overline{Y_\iota}) = (\alpha_i - \overline{\alpha_\iota}) + \beta * (X_{it} - \overline{X_\iota}) + (e_{it} - \overline{e_\iota})$$

$$\text{Where } \overline{Y_\iota} = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \text{ , } \overline{X_\iota} = \frac{1}{T}\sum_{t=1}^{T} X_{it} \text{ and } \overline{e_\iota} = \frac{1}{T}\sum_{t=1}^{T} e_{it}$$

The variables with bars (i.e. $(\overline{Y_\iota}, \overline{\alpha_\iota}, ..)$ represented temporal means. Demeaning removed the fixed effects α_i since a constant equated its mean over time ($\overline{\alpha_\iota}$). Thus, only time-variant effects remained in the equation. The study was subject to the drawback of the fixed effects

model that the model could not include individual time-invariant covariates in the model because demeaning the estimated α_i (i.e., the sum of all fixed effects) effectively removed the time-invariant observations

Specifically, this study used the fixed effects model to assess the effect of weather-related variables and the openness in the wind direction on PM2.5. Although openness in the wind direction was not a time-variant variable, this variable might not exhibit full time-invariant nature since its value changed according to the wind direction. Wind direction was a circular variable, and, therefore, it was used in terms of sine and cosine components in the model. Another wind-related variable resulted from the interaction of wind direction and the inlet of the DR 1000 instrument. When the wind blew into the instrument's inlet, the DR1000 recorded an elevated PM2.5 compared to the situation when the wind blew away from the inlet. Therefore, in addition to weather-related variables and directional viewshed, this fixed effect model included the angle between instrument travel direction and wind direction. When the wind blew into the inlet, this angle was 180^0, whereas when it blew away from the inlet, it was 0^0. In other cases, the angle varied between 0^0 to 180^0. Consequently, PM2.5 maxima coincided with the angle at 180^0 and minima at 0^0, confirming the cosine of this angle as the appropriate measure to include in the model.

Considering all the variables explained above, this study specified the fixed effect panel data models as follows:

$$PM2.5_{it} = \alpha_i + Wind\ Speed_{it} + Cos\ Wind\ Dir_{it} + Sin\ Wind\ Dir_{it} + Temperature_{it} + Relative\ Humidity_{it} + Cosine\ of\ the\ angle\ between\ travel\ direction\ and\ wind\ direction_{it} + Directional\ Viewshed_{it}$$

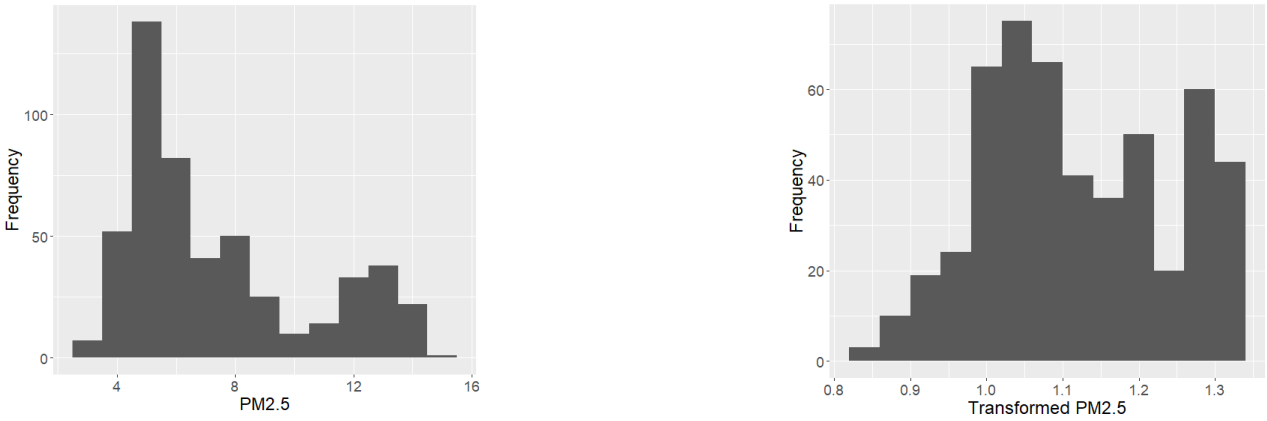
Further, this study also tested spatial autocorrelation in the errors for each data collection run.

The fixed effects in the model represented location-specific characteristics that were not included in the model. The study related these fixed effects with building morphological characteristics- building coverage ratio, mean building area, number of buildings, average distance between nearest-neighbor buildings, and mean building height. The study assumed that the key time-invariant factor was building morphology and correlated the estimated fixed effects with time-invariant building morphological parameters. In order to assess the contribution of building morphological characteristics, the study developed a regression model relating building morphological characteristics with fixed effects.

3.3 Results

3.3.1 PM2.5 and weather data

Figure 3.9 (a) showed the distribution of PM2.5 across all segments and all data collection runs. The PM2.5 data were positively skewed and adjusted with Box-Cox transformation (Figure 3.9 (a)) in order to use it in the proposed fixed effect panel model. The Box-Cox transformation parameter, lambda, was -0.6057.



(a) (b)

Figure 3.9: Distribution of PM2.5 (a) raw PM2.5 observations (b) Transformed PM2.5 data.

The details of data collection times and average PM2.5, and weather data values were provided in Table 3.2. During these data collection runs, maximum and minimum observed PM2.5 values were 13.10 µg/m3 and 4.21 µg/m3, respectively. The difference between maximum and minimum PM2.5 recorded in each run varied only slightly from 1.2 µg/m3 to 4.5 µg/m3.

Table 3.2: Details of date and time of data collection rounds

Run	Date and Time	Mean PM2.5 ($\mu g/m^3$)	PM2.5 Range ($\mu g/m^3$)	Wind Speed (mph)	Wind Direction	Temperature (Degrees Fahrenheit)	Relative Humidity (%)
1	10 Dec 2019 13:32:16	4.21	4.00	3.85	90	44.50	39
2	10 Dec 2019 17:42:52	5.20	2.00	4.42	58	50.00	42
3	12 Dec 2019 11:02:42	6.20	2.00	7.00	174	51.00	45
4	12 Dec 2019 14:23:48	7.40	2.75	3.40	187	42.50	53
5	13 Dec 2019 11:21:26	13.10	3.00	2.00	296	75.60	51
6	13 Dec 2019 17:22:52	11.50	4.50	4.40	270	65.00	58
7	14 Dec 2019 11:26:38	5.00	1.20	7.80	87	49.00	54
8	15 Dec 2019 16:35:22	8.40	3.40	5.60	297	40.00	70
9	21 Feb 2020 15:24:03	5.00	2.00	2.60	250	20.00	47

3.3.2 Variation in PM2.5 across runs and paths

Figure 3.10 shows the PM2.5 distribution across the study area. All data-collection runs encountered low variations in PM2.5 measurements. Runs five and six exhibited comparably higher variations. The PM2.5 distributions at individual data collection paths during all data collection runs are shown in Figure 3.11. When the wind was from the south, PM2.5 values at all paths were highly similar (run 3 and 4). All the paths oriented in the north-south direction with open passages in the same orientation could facilitate dispersion and explain the similar PM2.5 observations.

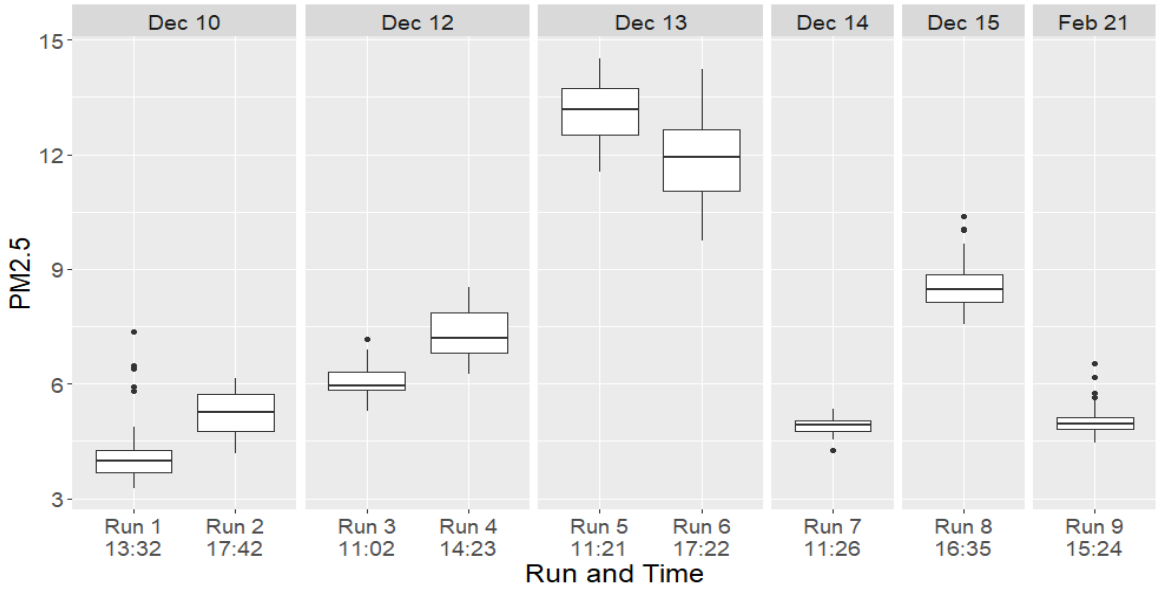


Figure 3.10: Distribution of PM2.5 across all paths during each data collection run

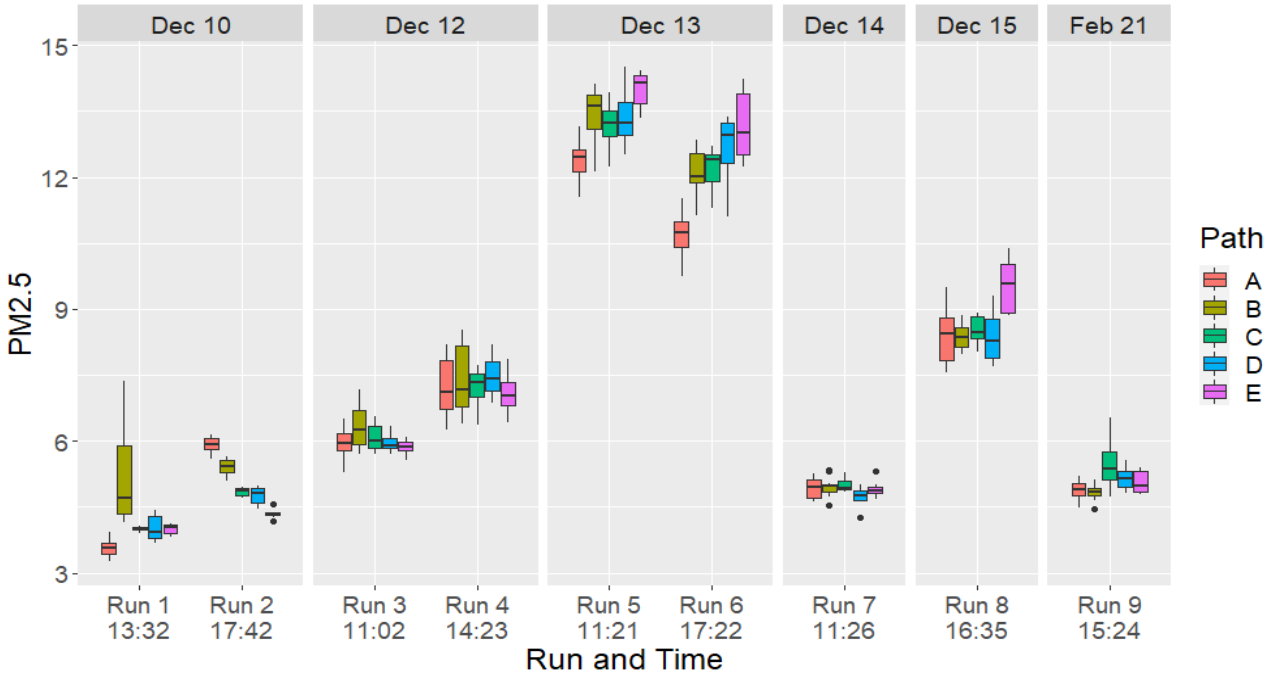


Figure 3.11: PM2.5 at each path during data-collection runs

While variations in PM2.5 observations across all paths appeared low, there were few exceptions. During northwest winds (run 5, 6, and 8), path E collected higher PM2.5 values compared to other paths. A possible explanation was that the open area to the north of path E allowed efficient dispersion of pollutants from surrounding areas, and the pollutants subsequently were trapped by large buildings along this path. In other cases, when the wind came from different directions, these buildings shielded the incoming airflow. In general, path A had low PM2.5 values, except when the wind was from the northeast (run 2) across a sizable parking lot. Higher wind speed led to better dispersion. At the wind speed of 7.8 mph (run 7), the highest wind speed observed among all data collection runs, PM2.5 values were uniform across all paths.

3.3.3 Results from the Fixed effect model

This study calculated directional viewshed at multiple distances to account for the unknown effective distance to which openness in a viewshed would influence PM2.5 dispersion. An experiment of five fixed effect models, each with an effective distance, examined the distance effect on model performance. All other weather-related variables remained same in all the five models. Results showed comparable model performance (R^2) at varying viewshed distances (Table 3.3).

Table 3.3: Fixed effect model results with varying viewshed distance

Directional viewshed distance	R^2	Is directional viewshed significant?
100m	0.82496	No
200m	0.82516	No
400m	0.82658	Yes
800m	0.82812	Yes
1500m	0.82778	Yes

Nevertheless, small viewshed distances up to 200m were not statistically significant in the model. The model R^2 value increased up to 800m, and then it started decreasing. Besides the directional viewshed, all other weather-related variables used in the model appeared statistically significant. Table 3.4 summarized the detailed results of this model built with a viewshed of 800m.

Table 3.4: Fixed effect model results

Variable	Coefficient	p-value
Wind speed	-0.0003894	$2.358* 10^{-6}$
Cosine of wind direction	-0.014296	0.0001239
Sine of wind direction	-0.062830	$< 2.2*10^{-16}$
Temperature	0.006214	$< 2.2*10^{-16}$
Relative humidity	0.0040963	$< 2.2*10^{-16}$
Cosine of angle between travel direction and wind direction	-0.01895	0.0024263
Directional viewshed (800m)	$3.3831*10^{-7}$	0.0013893

The small transformed PM2.5 values (0.84-1.32) in the dependent variable contributed to the small coefficients for all explanatory variables. Wind speed had a negative impact on PM2.5 as the model asserted a negative coefficient. Both the components of the wind direction, sine and cosine, had negative coefficients in the model, whereas both temperature and humidity had positive coefficients. This study used a digital surface model with 50-cm resolution to calculate the viewshed of a location as the number of 50×50 cm^2 cells visible in the given direction and up to a given distance. Depending on the location's visibility, the number of visible cells varied from hundreds to thousands, which explained the small coefficient for directional viewsheds. The amount of open area in the wind direction had a small but positive impact on PM2.5.

3.3.4 Error analysis

Moran's I was calculated to evaluate the spatial autocorrelation in the errors for each data collection run. All but run 3 had a significant moderate to strong spatial autocorrelation in the errors., suggesting the possible omission of variables that were responsible for the PM2.5 variations across the study area. In order to account for spatial dependence, this study considered both spatial lag models and spatial error models. Spatial lag models examine the existing autocorrelation in the response variable (Anselin, 2003). As the study intended to mediate the model bias due to spatially autocorrelated errors and investigate the effect of predictors on PM2.5 estimates, the study opted for the spatial error modeling approach. Spatial autocorrelation in the errors increased with spatial specification in each panel. Table 3.5 presented Moran's I, p-value, and significance for the errors for each data collection round in the panel model (aspatial specification) and the spatial panel model. The results suggested that aspatial models suffered

less than spatial models from spatial dependence in the errors. Therefore, the study selected the fixed effect panel model over the fixed effect spatial panel data model for further analysis.

Table 3.5: Spatial autocorrelation in errors of the fixed effect models

Data collection run	Number of observations	Moran's I (Panel Data)	Significance	Moran's I (Spatial Panel Data Model)	Significance
1	57	0.47	Significant	0.68	Significant
2	57	0.87	Significant	0.90	Significant
3	57	0.14	Insignificant	0.30	Significant
4	57	0.31	Significant	0.26	Significant
5	57	0.31	Significant	0.55	Significant
6	57	0.48	Significant	0.80	Significant
7	57	0.43	Significant	0.75	Significant
8	57	0.40	Significant	0.68	Significant
9	57	0.48	Significant	0.59	Significant

3.3.5 Relation between fixed effects and building morphology

Correlation between fixed effects and building morphological characteristics varied from 0.79 to -0.51. Fixed effects were weak to moderately correlated with all building morphological characteristics, except for the average nearest-neighbor distance between buildings (Figure 3.12). Mean building area, mean building height, and building coverage ratio were strongly positively correlated with each other. On the other hand, while the average nearest-neighbor distance between buildings was only weakly correlated with all the other building morphological characteristics, the number of buildings was moderately negatively correlated with mean building height.

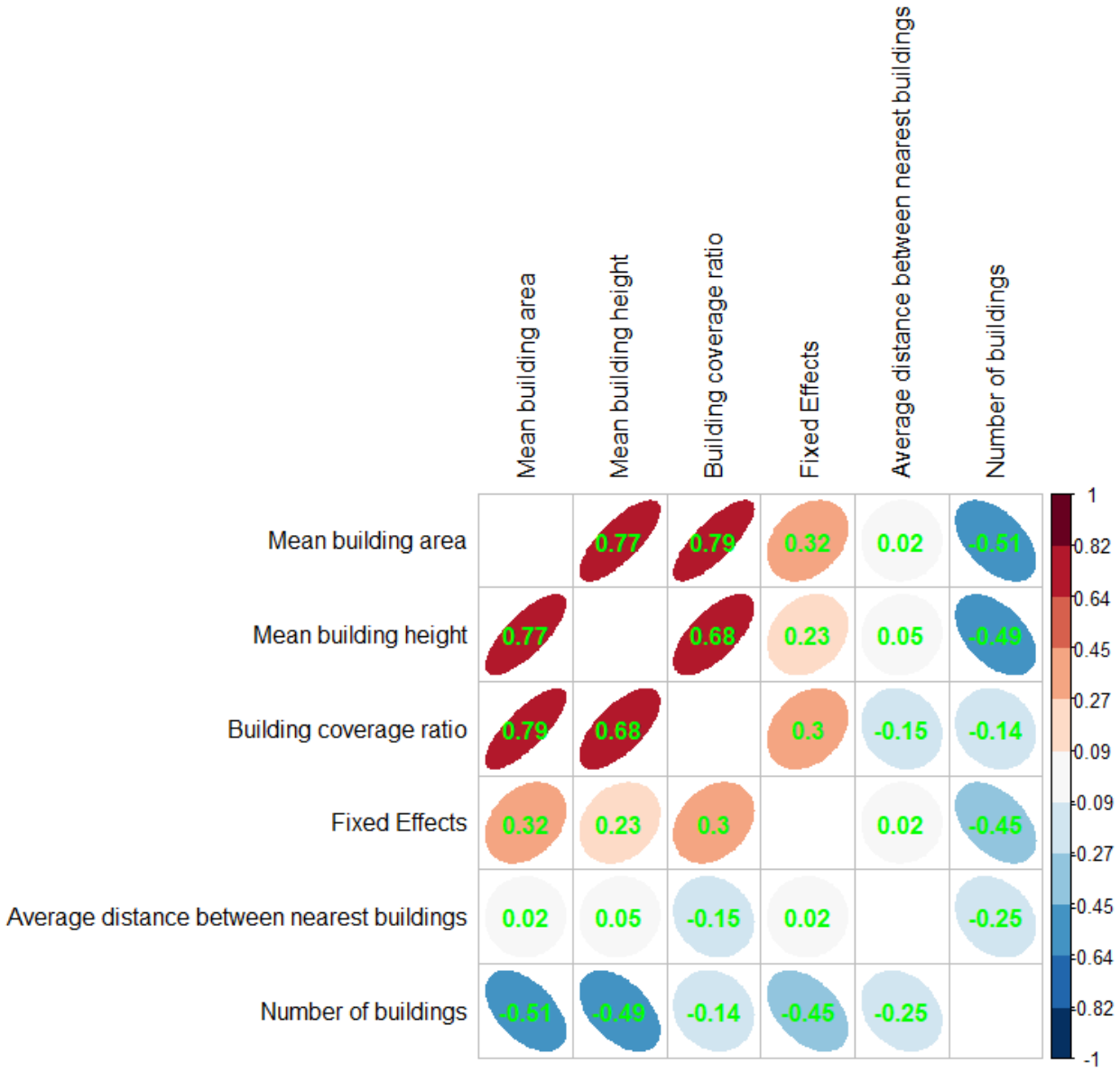


Figure 3.12: Correlation between fixed effects and building morphological variables

As such, the study excluded the average nearest-neighbor distance between building in regression modeling as it was almost uncorrelated with the fixed effects. Table 3.6 showed the regression results. Out of the considered building morphological characteristics, the study found only the building coverage ratio and the number of buildings as significant variables that

explained variations in fixed effects. Overall, building morphological characteristics explained 33.22% variation in the fixed effects.

Table 3.6: Results of regression between fixed effects and building morphological characteristics

Variable	Coefficient	p-value
Mean building area	$-2.33 * 10^{-6}$	0.1939
Building coverage ratio	0.0731	0.0046**
Mean building height	-0.00162	0.1329
Number of buildings per 1000 sqm	-0.1099	0.0001***
Model R^2: 0.3322		

3.4 Discussion

Despite the moderate variation in building morphology, the study observed very low spatial variability in PM2.5 in the area of less than 1 square kilometers across all data collection runs. The small spatial variation in PM2.5 observed in this study conforms with the findings of Harrison *et al.* (2015a). Harrison *et al.* (2015a) collected PM2.5 observations for a month in a 100km^2 area encompassing the University of Texas at Dallas. Their study found that depending on the weather condition the spatial scale of PM2.5 variation in the area varied from 0.8 km to 5.2 km. The building morphology accounted for 33.22% of the variations in the fixed effects, suggesting that the built environment potentially affects PM2.5. Regression results from fixed effects showed that the building coverage ratio had a positive impact on PM2.5. Thus, the more

built-up area likely hindered the dispersion of the pollutants. The same inference followed on the negative impact of the number of buildings per 1000 m^2 on PM2.5. Nevertheless, the study could not assess each variable in isolation as an inference about it could change based on its relationships with other variables in the model. Still, a negative correlation between the number of buildings per 1000 m^2 and mean building area, mean building height, and building coverage ratio in the study area denotes the contrast between areas with many smaller buildings and areas with fewer larger buildings. Areas with smaller lower buildings experienced better dispersion than other areas. The insignificant impact of mean building height and mean building area could be attributed to limited variability among building morphological characteristics in the study area.

The negative effect of wind speed on PM2.5 in the model reaffirmed findings from the previous studies that increasing wind speed improved dispersion and hence reduced the PM2.5 concentration (Ginzburg *et al.*, 2015; Brown *et al.*, 2019). Increased temperature promoted air circulation and was negatively related to PM2.5 measures (Ginzburg *et al.*, 2015). Therefore, the temperature was expected to have a negative effect on PM2.5 contrary to the positive effect found in the study. A possible explanation was that temperature contributed to chemical reactions that might form new particles, which could include PM2.5, in the atmosphere (Wang *et al.*, 2015). As for the relative humidity, depending on its value, it could increase or decrease in PM2.5 mass. Increased humidity promoted particle size growth, but after a certain threshold, very high humidity could lead to particle deposition due to heavy particle growth (Wang *et al.*, 2015). Moreover, Lou *et al.* (2017), in their 3-year study in the Yangtze river delta, described the

relationship between relative humidity and PM2.5 as an inverted U-shape, where hygroscopic growth of particles continued with an increase in relative humidity until relative humidity reached 70% and then it started decreasing. Relative humidity observed in the study ranged from 39-70%, not high enough to cause particle deposition, and resulted in a positive correlation with PM2.5. The positive coefficient associated with directional viewshed suggested that a more open area in the direction of the wind is consistent with additional incoming pollution from the other areas as confirmed by Brown *et al.* (2019). Out of different viewshed distances used in the model, model performance consistently increased up to 800m distance, and then it started decreasing, suggesting openness up to 800 m distance affect PM2.5. However, increasing distance led to only marginal improvements in the model, and openness within 400 m also gave satisfactory results.

3.5 Conclusion

The study collected the PM2.5 data using mobile monitoring platform at a non-near road site with different building morphological characteristics. Building morphological characteristics varied with high density, small-sized, shorter buildings to low density, and medium to large-sized taller buildings. The data collection strategy included multiple runs across paths representing varied building morphology. The fixed effect panel model was used to investigate the effect of weather-related variables and building morphological characteristics on PM2.5. Unlike meteorological variables, building morphological characteristics did not change with time. A regression model was developed to find the contribution of building morphological characteristics to fixed effects extracted from the panel data model. While the weather-related

variables explained variations in PM2.5, building morphological characteristics also showed positive effects on PM2.5. Furthermore, openness in the direction of wind allowed pollutants from other areas and raised PM2.5 concentration in the area.

The model presented in the study carried spatially autocorrelated errors even with spatial specifications, possibly due to the small study area with low PM 2.5 variability. A larger study extent and coarser unit of spatial analysis might mediate the issues of spatially autocorrelated errors. Also, the fixed effect model applied the same spatial weight matrix to all panels. Depending on the weather conditions and its interactions with the surrounding built environment, the nature of spatial dependence between errors could change. Therefore dynamic spatial weight matrix would be more appropriate to account for this spatial dependence. Alternatively, space-time convolution could also address the complex space-time dependence in the data overlooked in the current model specification. Nevertheless, this study showed small spatial variation in PM2.5 in a small area ($< 1km^2$) typical of non-near road sites with moderate variation in building morphology. The study considered interactions between wind and openness in the wind direction and overall building morphological characteristics within a 100m buffer. Future studies can incorporate interactions between wind and building heights and measures related to spatial arrangement of buildings to further understand the effects of buildings on PM2.5.

CHAPTER 4

STREET-LEVEL QUANTIFICATION OF ALTERNATIVE PM2.5 MEASURE

4.1 Introduction

Current regulatory standards for Particulate Matter (PM), both PM2.5 and PM10, are based on measurements of particles in terms of mass per volume. Geographical and seasonal variations in health risks associated with PM2.5 (Dominici *et al.*, 2006; Bell *et al.*, 2008) lead to postulate the possible role of PM2.5 chemical composition behind these geographical differences in the risk. Bell *et al.* (2009) test this hypothesis and find the association between the greater content of PM elements like nickel, vanadium, and elemental carbon and hospital admissions related to cardiovascular and respiratory diseases in an elderly population. Like chemical composition, particles also differ in size and surface area. Although all particles with a size less than or equal to 2.5 μm are referred to as PM2.5, smaller particles are greater in number and have larger surface area for the same amount of mass than large-sized particles (Valavanidis *et al.*, 2008; Kwon *et al.*, 2020). Finer particles have greater reach and deposition frequency in deeper parts of the lungs like bronchiole and alveoli, crucial elements of the human respiratory system (Salma *et al.*, 2002). However, since these smaller particles contribute very little to the total PM2.5 mass (Kwon *et al.*, 2020), the impact of these particles on human health remains inconspicuous with the current standards of air quality.

Studies focusing on the causal mechanism that leads to harmful health effects call attention to alternate measures of PM2.5. Currently, PM2.5 measures lack information on chemical constituents, surface area, or particle-size distributions of PM2.5 particles, which may

have detrimental health effects and may serve as better measures to evaluate the air quality than the existing air quality measure. Peters *et al.* (2015) find the association between particle numbers (PN) in size 0.1-1 μm and metrics of heart rate variability in individuals with underlying conditions like type-2 diabetes and impaired glucose tolerance. An analysis of particles retained in human lungs of 10 elderly residents of Vancouver shows that only 5% of particles have an aerodynamic diameter below 0.1 μm, whereas 96% of particles are PM2.5 (Churg and Brauer, 1997). When particles are measured in PN, smaller-sized particles contribute greatly to PN than larger-sized particles (Kwon *et al.*, 2020). Variable proportions of mass of submicron particles at emission and non-emission sites show that size segregated PN distribution may be an indicator of potential sources (Tsai *et al.*, 2005). PN can be measured for the entire size range of 0 – 2.5 μm particles, but specific PN bins of distinct particle sizes can be most helpful to uncover a potential association between particle size and adverse health effects.

Depending on the size, genesis, lifespan in the atmosphere, particles are categorized as Aitken mode (0.01-0.1 μm), accumulation mode (0.1-1 μm), and coarse mode (> 1 μm) particles (Alfarra, 2004). Accumulation-mode particles grow in number in high humidity as condensation nucleates ultra-fine particles (< 0.1 μm). Meanwhile, coarse-mode particles decrease in number due to wet deposition (Hussein *et al.*, 2018). Unlike humidity, the temperature negatively relates to particles in the accumulation mode (Dinoi *et al.*, 2020).

Distributions of PM and PN vary with distance from highways: PN in the range 6-220nm exhibit greater variation compared to PM (Zhu *et al.*, 2002). This suggests that spatial variation

in PN is different than that in PM. PN concentration may be 20-80% more in traffic affected areas than urban sites unaffected by traffic in Augsburg, Germany (Cyrys *et al.*, 2008). Between the spatial distribution of mean annual PM10 and PN concentration (of particles greater than 7nm**)** in Stockholm, PM10 has a smoother spatial gradient across the city, whereas PN concentration at the center of the city is five times higher compared to background PN concentration (Johansson *et al.*, 2007). However, PN measures are uncommon in most cities, and consequently, there is inadequate epidemiological evidence on the effects of PN counts of different particle sizes on health outcomes (Atkinson *et al.*, 2015; Baldauf *et al.*, 2016). Particles in various modes differ in their interaction with the human respiratory system. Therefore it is important to consider these differences while assessing the effects of particles on health outcomes (Alfarra, 2004). More studies on smaller-sized particles and information on the spatial distribution of particles in terms of alternative measures will help epidemiological studies seeking the effect of particle size on human health (Kwon *et al.*, 2020).

The purpose of this study is to develop a machine learning approach to PN estimations and investigate the effect of building morphology on street-level PN, using the city of San Francisco as a case study. As traffic predominantly contributes to PN concentration, many studies on PN compare PN concentrations at traffic and non-traffic sites. Besides traffic, how contextual factors may affect the PN dispersion is unclear. Publicly available traffic-count data are limited to Annual Average Traffic Count (AADT), too coarse for PN estimation at a fine temporal resolution as traffic counts tend to vary throughout the day. Aerosol Optical Depth (AOD) serves as a proxy to estimate the PM2.5 and PM10 in many studies. However, its use for

PN estimation has not been investigated. The study uses AOD and meteorological variables as covariates along with building morphological variables. AOD from satellite data can serve as a surrogate for particles in the atmosphere, whereas finer scale variables such as building morphology may help explain within pixel (of AOD) variability in PN if it has any impact on particle dispersion. The study employs a neural network to downscale PN of different particle sizes at street level using the data collected by Google streetcars during 2016-17 (Google, 2017). Data collected in this campaign has a particle size between 0.3-2.5 μm, which are further divided into a total of five bins. This study aims to investigate the role of AOD as well as building morphological parameters to downscale street-level instantaneous PN concentration in each of these size ranges.

4.2 Data and methods

PN Data

The study used Air Quality Data from Google. Google streetcars collected high-resolution data about NO, NO_2, O_3, BC (black carbon), and PN at five distinct sizes at a sampling frequency of 1Hz during May-September 2016 and April-June 2017 in the city of San Francisco. Google streetcars collected data between 9 am-5 pm during weekdays. This study focused only on PN data. The different sized particles were noted as PN1 to PN5, with PN1 referring to the smallest particles whereas PN5, the largest (Table 4.1).

Table 4.1: Particle size table

PN	Size (in microns)
PN1	0.3-0.5
PN2	0.5-0.7
PN3	0.7-1.0
PN4	1.0-1.5
PN5	1.5-2.5

The study followed the PN collection strategy and considered street segments as the basic unit of analysis for calculating PN values. A street segment was the street section between two intersections. For each street segment in the study area, the study calculated the mean PN value for each PN bin with PN data collected on that street segment. Short street segments might not have enough PN measurements to calculate representative mean PN value, and therefore this study considered only segments greater than 30m. While the PN measurements came with GPS data, GPS related positional errors resulted in some points at some distance off streets, a well-known "map matching problem" (Newsman and Krumm, 2009). Another map-matching issue arises from GPS points around intersections, which are challenging to identify corresponding street segments. In order to overcome the map matching problem, each street segment took on all PN measurements within a 2.5m buffer from the segment. PN measurements around the last 2.5m on both ends of a street segment were assigned to the intersections (Figure 4.1).

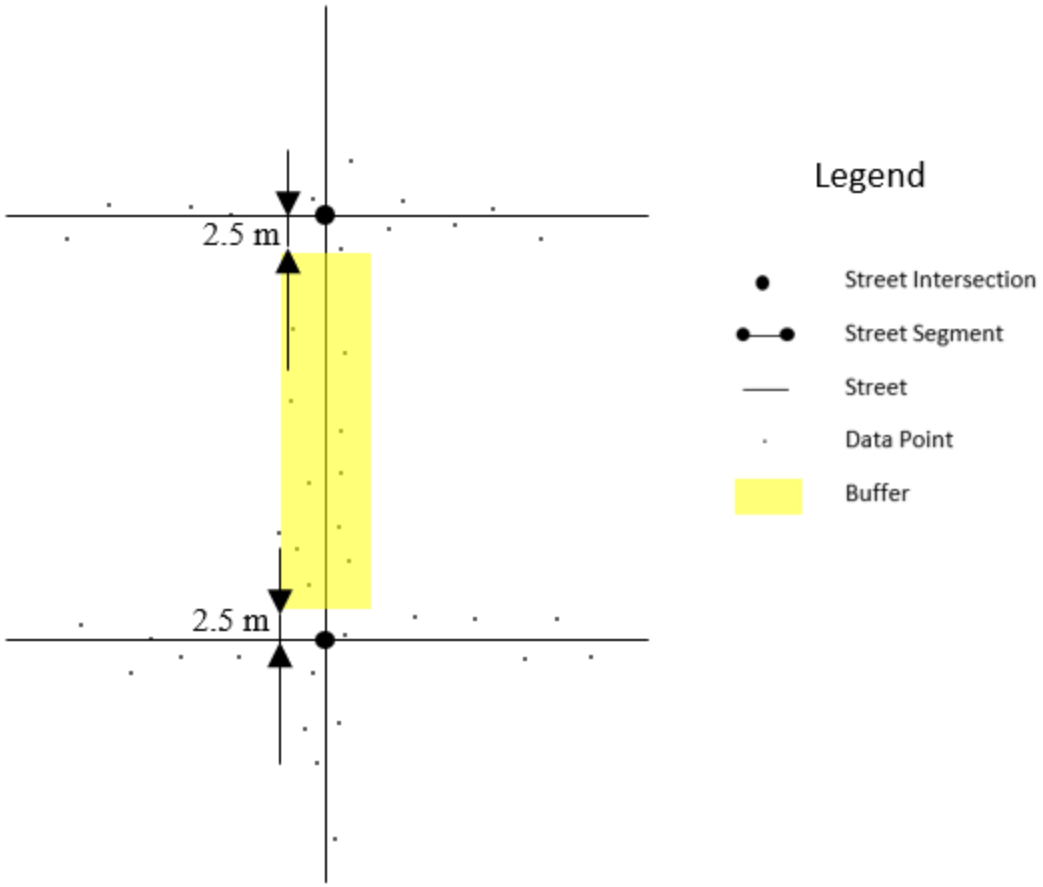


Figure 4.1: street segments and PN assignments

Aerosol Optical Depth (AOD) data

The study uses MAIAC AOD from MODIS at 1 km resolution. Chapter 2 details this dataset. In sun-synchronous orbits, the Terra and Aqua MODIS satellites pass over any given location on the Earth at the same local time every day. AOD data from MODIS are available twice a day at about 10:30 am (Terra) and 1:30 pm (Aqua). The study extracts AOD data covering street segment midpoints and joins the AOD data with street-segment mean PN data. The time of AOD acquisition and PN measurements may not coincide, so the study selects PN observations within

±30 minutes of AOD acquisition time assuming that PN values would not change substantially in that period. Besides AOD, the study includes MAIAC column vapor content, quality assurance flag, and column water vapor which may impact the AOD.

PN data collection frequency across San Francisco

Google streetcars carried out the data collection exercise with the aim to collect data across each road street at least once. However, Google streetcars collected more data for some street segments than others. Figure 4.2 shows the frequency of data collection across all street segments (with lengths greater than 30 meters) across San Francisco during May-Sept 2016 and April-June 2017. Most road segments are mapped one to ten times. Although Google streetcars covered each road segment at least once, very few street segments had data acquisition time within ± 30 minutes of MODIS overpass time (Figure 4.3).

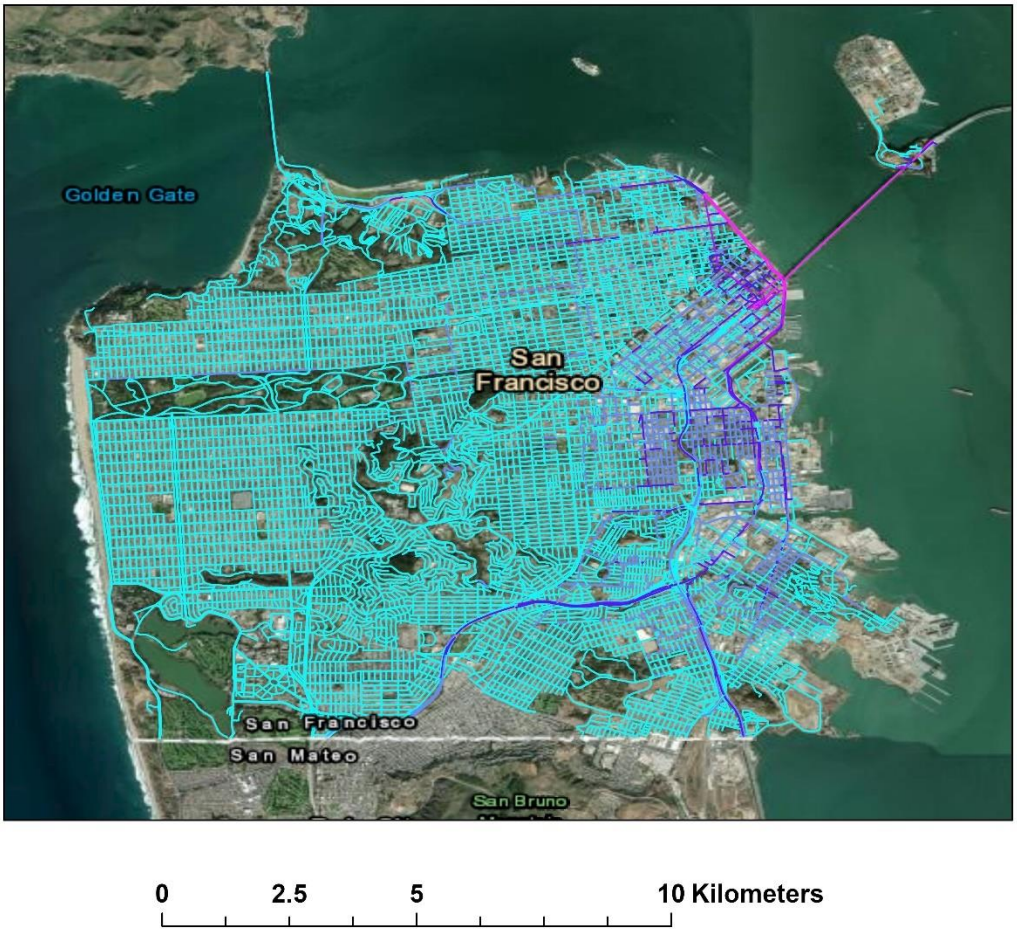


Figure 4.2: Data collection frequency across street segments in San Francisco

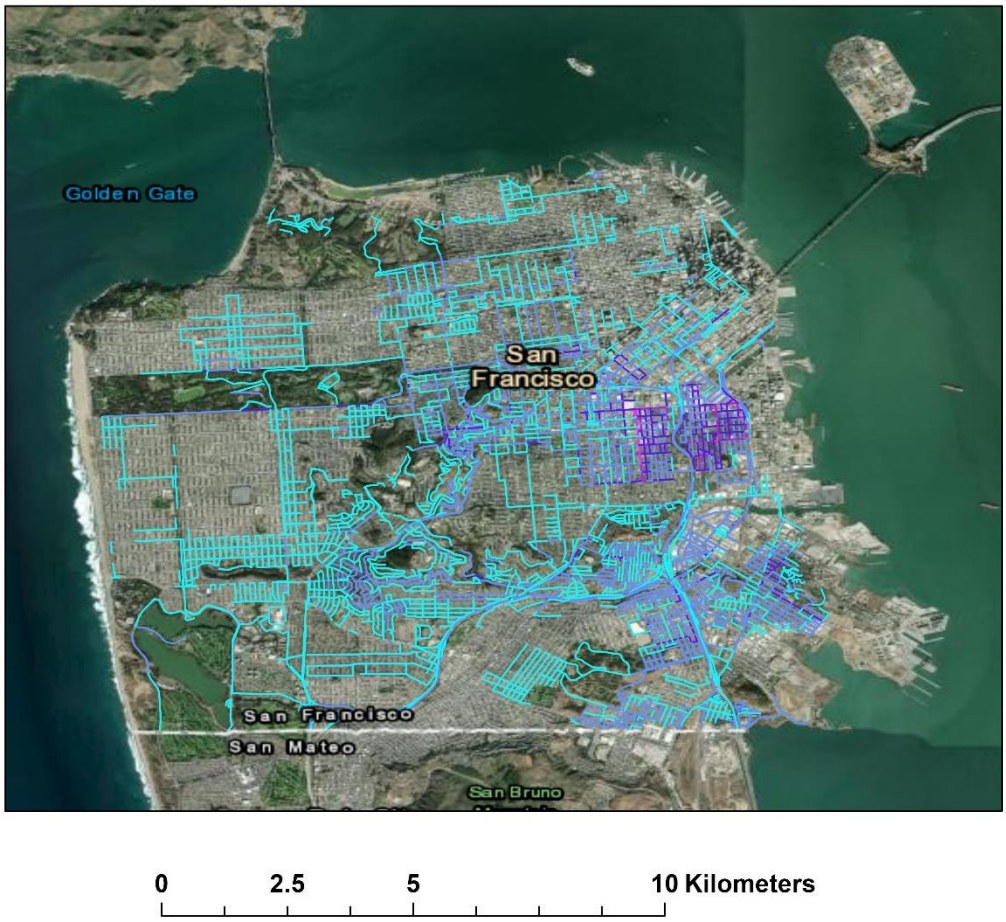


Figure 4.3: Street segments with overlapping data acquisition time within ± 30 minutes of MODIS overpass time

Weather data

Weather data from European Centre for Medium-Range Weather Forecasts (ECMWF). ECMWF provides a reanalysis dataset for several weather parameters. This dataset is available at 0.125-degree (~13km) resolution and at intervals of 3, 6, 9, and 12 hrs from 0:00 and 12:00 UTC (Berrisford *et al.*, 2011). Therefore this dataset was available at 7:00 am, 10:00 am, 1:00 pm, and 4:00 pm at PST during the day. ECMWF data closest in time with PN data provided the weather parameters for the analysis. Weather parameters considered in the study included temperature, dew point temperature, wind speed, and wind direction.

Building data

Building data, including footprints and heights, support calculations of building morphological parameters. Building data for the city of San Francisco came from the City and County of San Francisco under Open Data Commons (https://data.sfgov.org/Geographic-Locations-and-Boundaries/Building-Footprints/ynuv-fyni).

Building morphology

Building morphological parameters are important measures of the spatial complexity of a city. Building data support the calculation of five building parameters: mean building area, number of buildings per 1000 m^2, mean building height, average nearest-neighbor distance between buildings, and building coverage ratio. These parameters are calculated as discussed in chapter 3.

Summary statistics of building morphological characteristics

Table 4.2 shows the summary statistics of building morphological characteristics within 100m buffer of road segments. The summary statistics include mean, minimum, 25th percentile, median, 75th percentile, and maximum value of each building morphological characteristic.

Table 4.2: Summary statistics of building morphological characteristics

	Mean building area	**Building coverage ratio**	**Average nearest neighbor distance between buildings**	**Building count per 1000 sqm**	**Mean building height**
Mean	317	0.31	3.15	1.73	7.92
Std Dev	443	0.12	15.70	0.97	3.84
Min	0	0.00	0.00	0.00	0.00
25%	123	0.24	0.40	0.98	6.01
Median	160	0.32	0.74	1.76	7.18
75%	280	0.39	1.71	2.45	8.95
Max	5252	0.66	208.74	4.81	58.87

Neural network modeling

The study took a neural network approach to estimate PN in each particle-size bin using AOD, weather, and building morphological data. A separate neural network model was developed for PN of each size bin. The study used neural networks with a single hidden layer (Figure 4.4). The first layer consisted of inputs, the hidden layer consisted of 24 neurons, and finally the output layer. The study evaluated the performance of the neural network at 8, 16, and 24 neurons to

determine the optimal number of neurons in the hidden layer. In general, the result showed that an increase in neurons improved R and RMSE values and hence the model performance (Refer Table A.1 in Appendix). The results also showed only small differences between performance metrics for training and test data in all networks with 8, 16, and 24 neurons in the hidden layer. The average R on test data was above 0.82 for all PN sizes for a neural network with 24 neurons in the hidden layer. A leaky relu activation function with alpha 0.2 was used to transform input parameters, and a linear activation function was used to transform neurons in the hidden layer to the final output. A total of 500 epochs and a batch size of 512 were used in training neural networks for all particle sizes.

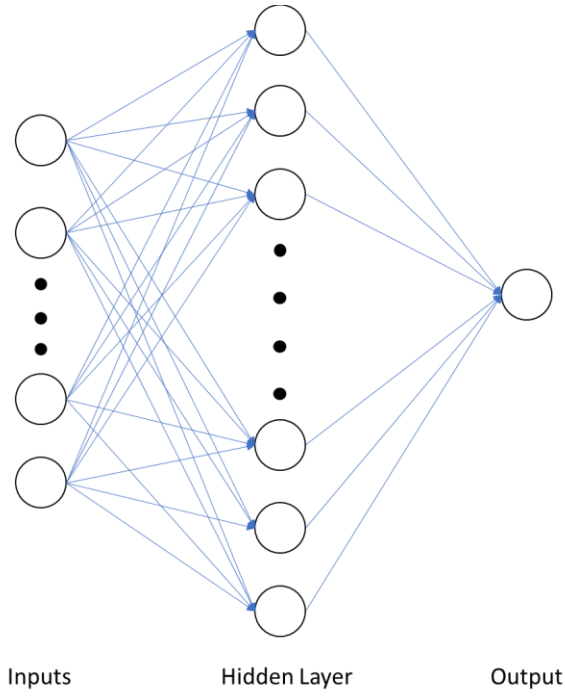


Figure 4.4: Neural network architecture

An RMSProp optimization algorithm was used to minimize the loss function, and 5-fold cross-validation was used to evaluate the model performance. PN values beyond 3 standard deviations were considered as outliers and were removed before running the neural network. Table 4.3 summarized the range of particle number concentrations across different sizes.

Table 4.3: Range of particle number concentration across different particle sizes

PN	Range for San Francisco Data	Range after removing outliers
PN1	2000-86000	2000-46000
PN2	120-60000	120-13000
PN3	20-32000	30-6000
PN4	0-17000	0-2700
PN5	0-28000	0-2900

The model considered covariates in three classes: meteorological variables, AOD measures, and building morphological parameters. A base model included the covariates from all the classes. In order to evaluate the role of building morphology in PN estimation, a neural network was run by excluding covariates corresponding to building morphology. The study then compared the model performance with the base model. The same strategy was adopted to investigate the impact of AOD and AOD related variables on model performance.

SHapley Additive exPlanation (SHAP):

Despite the success of machine learning approaches in numerous prediction and classification problems, these approaches are referred to as black box because how the algorithms reach decisions remains unclear. Several methods have been proposed recently to explain the output from machine learning. SHAP, proposed by Lundberg and Lee (2017), is one of those methods that help gain insights into the contribution and impacts of each input feature in the model. SHAP calculates the marginal contribution of each feature in the model by considering all possible combinations of features used in the model. In each of these combinations, the difference in the model output with and without the feature of interest provides the estimated contribution of that feature in the model. The average contribution of the feature of interest across all possible combinations provides a SHAP value for that feature. This study applied Kernel SHAP, a model-agnostic approximation method to calculate SHAP values.

4.3 Results and Discussion

While the proposed neural network architecture estimated the PN concentration across different sized particles with correlation coefficients above 0.82, some predictors appear more important than the others. Table 4.4 shows the number of observations used for PN in each size bin, predictors, correlation coefficient (R), and root mean squared error (RMSE) for training and test data.

Table 4.4: PN estimation results for the base model and with the exclusion of AOD and building morphological characteristics

Dependent variable	Independent variables	N	Train R	Test R	Train RMSE	Test RMSE
PN1	Weather, AOD, building morphology	10011	0.8869	0.8835	4253	4281
PN1	Weather and building morphology		0.7997	0.7964	8079	8046
PN1	Weather and AOD		0.8613	0.8616	5513	5462
PN2	Weather, AOD, building morphology	10144	0.8793	0.8768	1325	1339
PN2	Weather and building morphology		0.7638	0.7570	2603	2609
PN2	Weather and AOD		0.8140	0.8168	1705	1683
PN3	Weather, AOD, building morphology	10153	0.8751	0.8722	751	754
PN3	Weather and building morphology		0.7871	0.7852	1184	1186
PN3	Weather and AOD		0.8598	0.8556	940	943
PN4	Weather, AOD, building morphology	10150	0.8724	0.8639	293	302
PN4	Weather and building morphology		0.7745	0.7738	457	457
PN4	Weather and AOD		0.8291	0.8301	323	323
PN5	Weather, AOD, building morphology	10159	0.8399	0.8374	337	340
PN5	Weather and building morphology		0.7504	0.7461	435	440
PN5	Weather and AOD		0.8178	0.8109	352	360

Figure 4.5 and Figure 4.6 summarize the results in Table 4.4. Figure 4.5 shows correlation coefficients between observed and predicted values of PN across different size-bins and RMSE on test data. It also shows how the correlation coefficient is impacted after removing AOD and building morphology-related parameters from the base model. The more the decrease in

correlation coefficients due to the absence of a variable in the model, the more important that variable is in estimating the PM2.5. Likewise, an increase in RMSE due to the absence of a variable in the model indicates the importance of that variable. The greater the increase in RMSE due to the absence of a variable, the more important that variable is in predicting PN and vice versa.

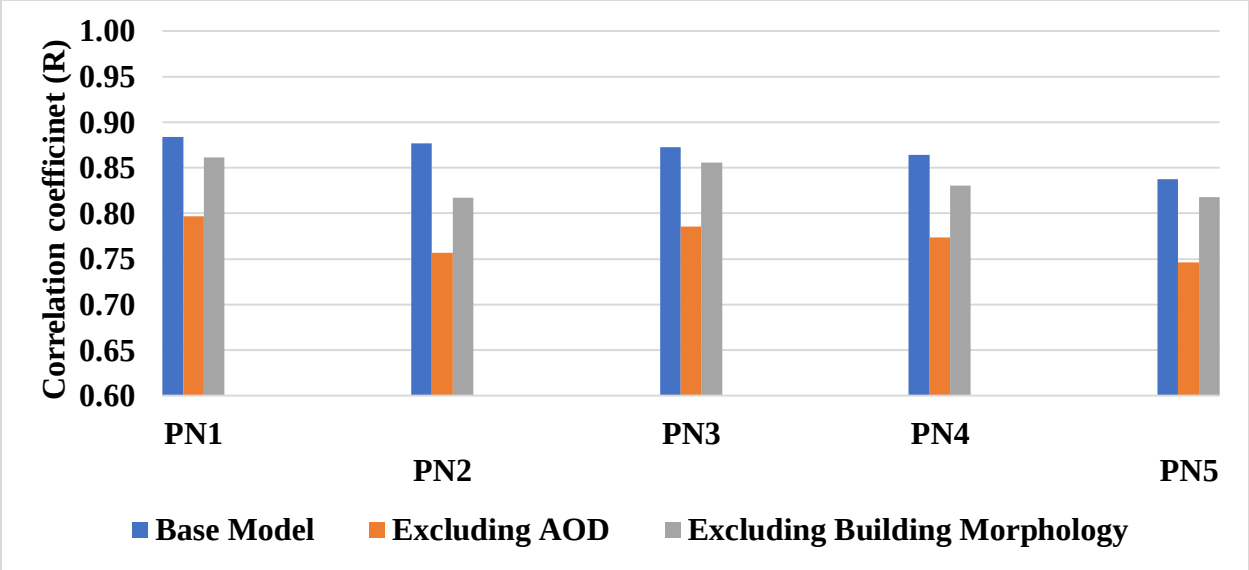


Figure 4.5: Correlation coefficient for different sized PNs across different models

For all PN sizes, exclusion of AOD and AOD related variables have a greater impact on correlation coefficient than exclusion of building morphology-related variables (Figure 4.5). Similarly, removing AOD and AOD related parameters leads to an increase in RMSE for all PN sizes (Figure 4.6). Removal of building morphology parameters results in greater RMSE for all PN sizes. However, the impact of AOD and AOD related parameters on RMSE is greater than that of the building morphological parameters.

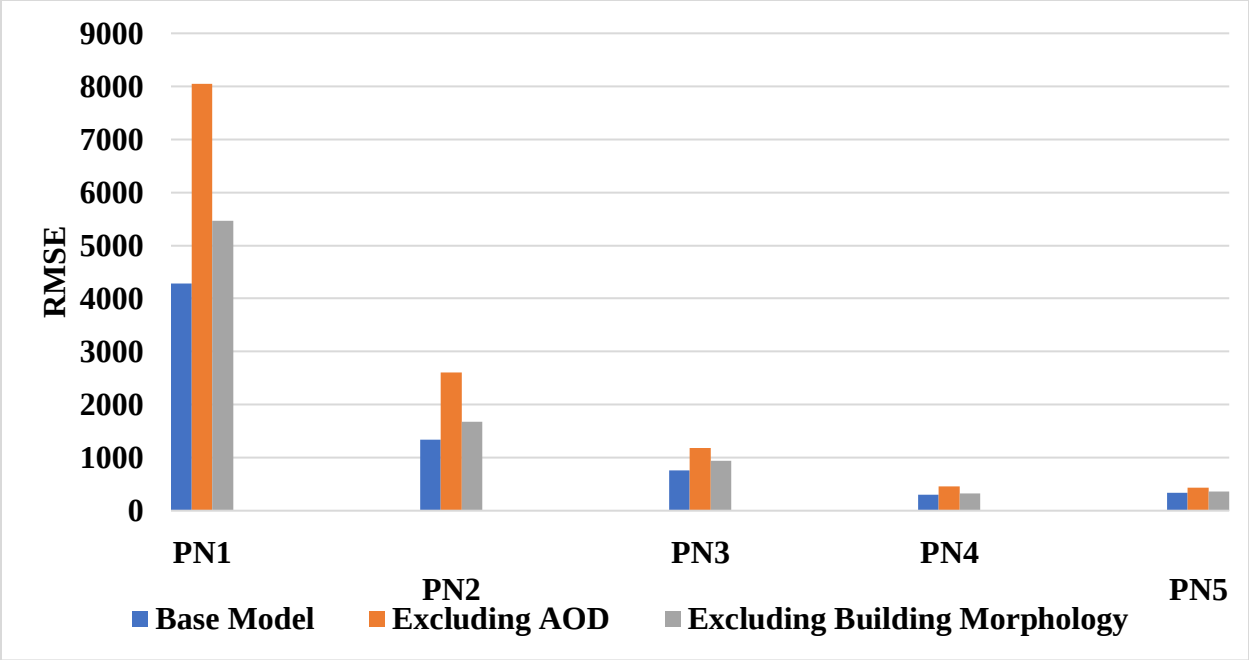


Figure 4.6: RMSE for different sized PNs across different models

Figures, from Figure 4.7-Figure 4.11 show SHAP summary plots for all particle size models in the order of their importance from top to bottom. Variables at the top have a greater impact on the model output than those at the bottom. The plots also show variable's values and impacts on the model. Blue color indicates low variable values, and red color is for high variable values. Across all particle sizes, weather-related variables temperature, dew point temperature, eastward and northward wind have a larger impact on the PN concentration across all particle sizes. AOD plays an important role in the PN concentration estimation after weather-related variables, whereas building morphology-related variables have the least impact on the model. Also, high values of temperature and AOD have a positive influence on the PN concentration. The eastward wind has a negative effect on the PN concentration, suggesting that cleaner air from the ocean reduces PN concentration. Except for PN1, the negative influence of dew point

temperature on remaining all particle sizes implies that increased moisture content in the air reduces PN concentration in the San Francisco area.

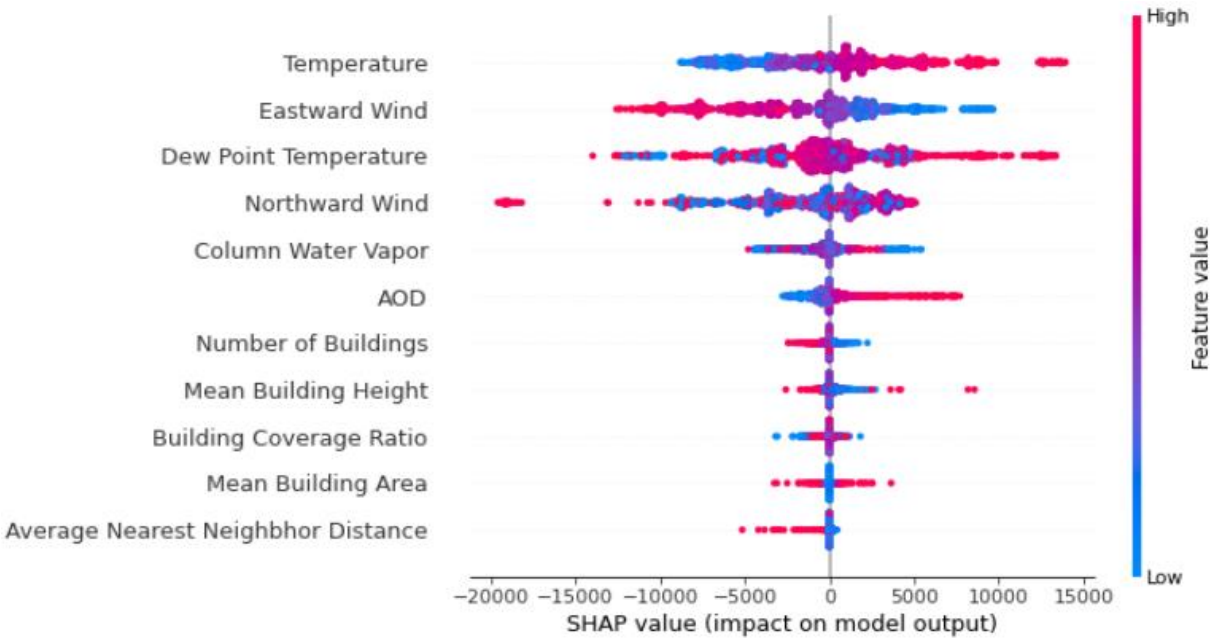


Figure 4.7: SHAP summary plot for PN1

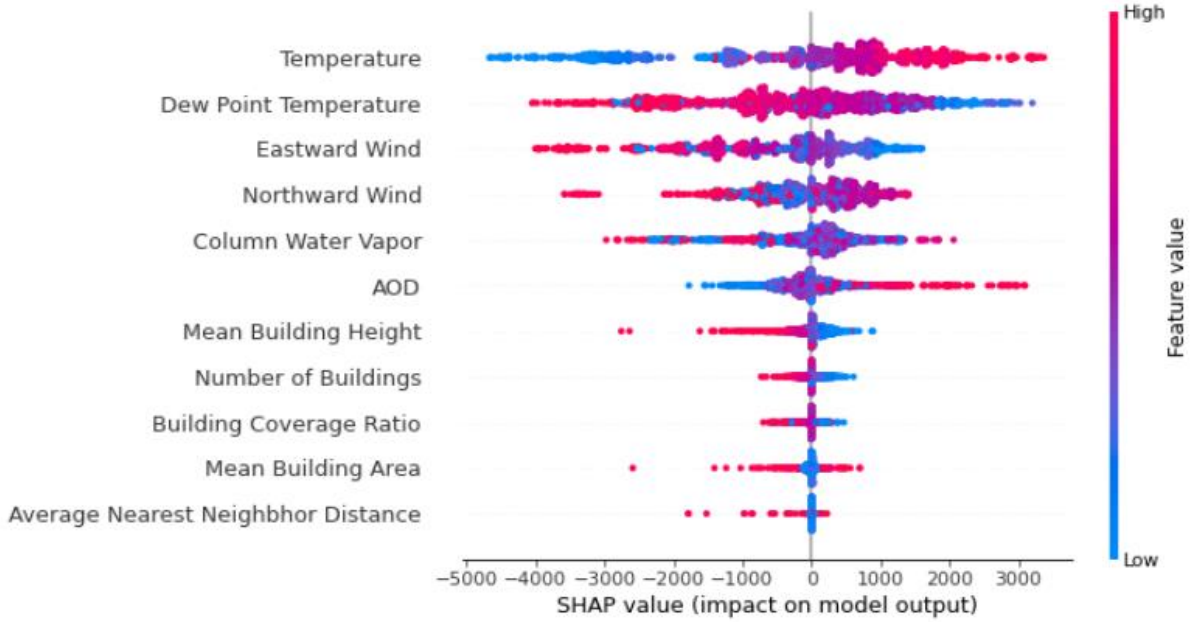


Figure 4.8: SHAP summary plot for PN2

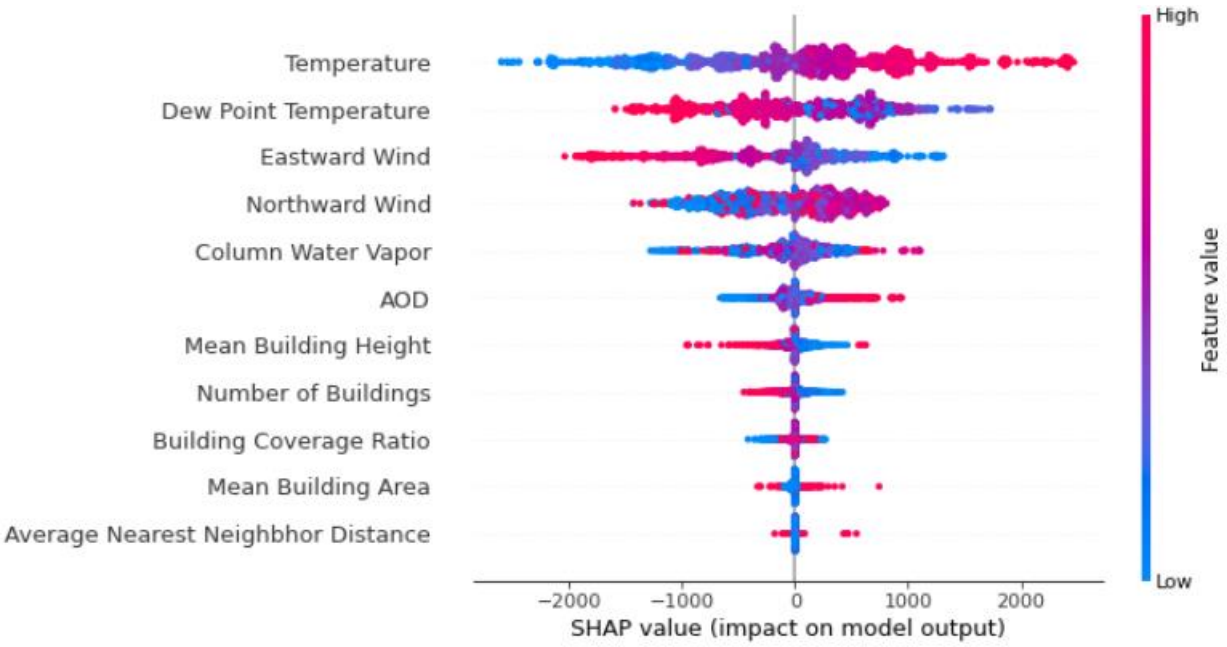


Figure 4.9: SHAP summary plot for PN3

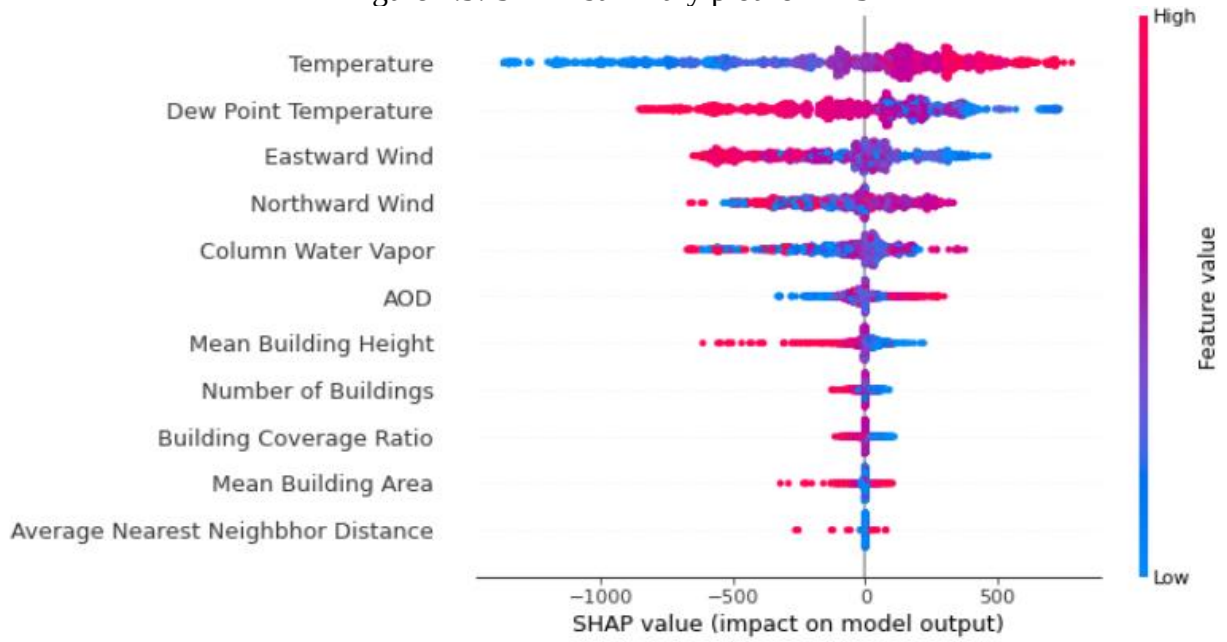


Figure 4.10: SHAP summary plot for PN4

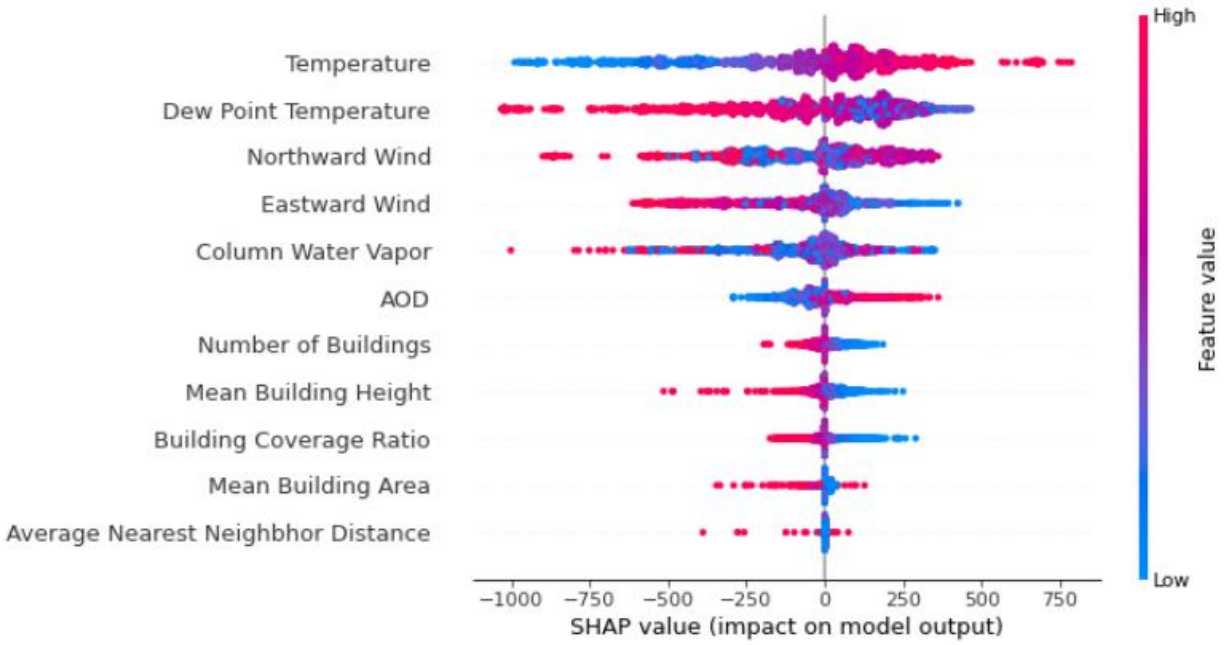


Figure 4.11: SHAP summary plot for PN5

4.4 Conclusion

The study applied a neural network approach to quantify the street-level PN concentration across five different particle sizes in the 0.3-2.5 µm range. The covariates included meteorological data, AOD, and parameters related to AOD quality and building morphology. The neural network architecture employed in the study successfully estimated the PN for different sized particles with correlation coefficients above 0.82. While the previous studies used the AOD to estimate PM2.5, AOD's potential to estimate PN was unknown. Furthermore, the study investigated the effects of AOD and building morphology on model performance. Compared to building morphology, AOD proved to be a more important covariate in estimating PN in all sizes. Considering the importance of alternate PM measures and the lack of availability of the data in terms of these measures, the study developed the model to illustrate the possibility

of using readily available meteorological data and AOD to estimate PN in the hour of AOD acquisition at the street level in San Francisco. As the model is trained for the city of San Francisco, the same model may not apply to other cities due to differences in meteorological and physical characteristics. However, the performance of a similar approach can be investigated in the future for other cities and regions. Future research may also explore what boundary conditions will allow for transfer learning.

CHAPTER 5

CONCLUSION

PM2.5 is one of the major air pollutants associated with various health problems related to breathing and lung functions. Epidemiological studies on the adverse health impact of PM2.5 often rely on PM2.5 measurements from the nearest air quality station to estimate exposure. However, these air quality stations are few and insufficient to characterize PM2.5 variability. There are three challenges associated with PM2.5 estimation. First, although as an alternative to in-situ PM2.5 measurements, AOD data from satellites have been used to develop models to estimate PM2.5, the relationship between PM2.5 and AOD is complicated. A model developed for one area is seldom applicable to other areas. Second, the resolution of AOD data is often coarse, from 1km to 10km, and therefore PM2.5 variability within the single pixel of MODIS imagery needs to be studied. Third, current standards measure PM2.5 in mass per volume, but findings from recent studies stress the importance of alternative PM2.5 measures in particle numbers. This research addresses the three challenges in chapters 2, 3, and 4, respectively.

In chapter 2, the study proposed a CNN-based approach to estimate PM2.5 averaged over an hour in which MODIS overpassed the study area, Dallas-Fort Worth with MAIAC AOD and meteorological data. The proposed model produced good PM2.5 estimates with a correlation coefficient (R) of 0.87 and RMSE of 2.57 $\mu g/m^3$. Although previous studies showed similar success in predicting PM2.5 using satellite AOD, the temporal and spatial resolution of the predicted PM2.5 in these studies were daily average at 1-10 km^2 resolutions. This study used

MAIAC AOD at 1km resolution and estimated PM2.5 averaged over an hour in which MODIS overpassed the study area. Moreover, the study systematically investigated the impacts of predictor variables from spatially adjacent areas to predictor's location (i.e., in-situ stations) on the PM2.5 estimation. The study found that predictors from spatially adjacent locations were helpful in estimating PM2.5 and improving the model performance. Model performance improved with predictors from a wider area around the PM2.5 stations. Unlike the previous studies, the proposed model did not rely on PM2.5 measurements from the nearby stations, and hence it can be used in near-real time settings to estimate PM2.5 concentration. Unavailability of AOD data due to clouds or retrieval quality affected the number of samples available for training the model. This study showed that data augmentation overcame this problem. The proposed model developed using augmented data gave a comparable performance on training and test datasets.

The resolution of the PM2.5 estimated from satellite AOD is subject to the resolution of the AOD data used for estimation. Therefore, to study the variability in PM2.5 in an area with a smaller spatial extent (less than 1 km^2) and the effect of built-up area on PM2.5, in chapter 3, the study collected PM2.5 data in a small area on the University of Texas at Dallas campus using a mobile sensor. The study observed small spatial variability within individual data collection runs. But variability in PM2.5 concentration from one run to the other appeared prominent. The findings suggested that for a small study area with a spatial extent less than 1 km^2, temporal variables such as weather were major driving factors for observed PM2.5 variability. The study investigated the effects of meteorological variables, building morphology, and openness in the

wind direction on PM2.5 concentration. The positive effect of temperature and relative humidity and the negative effect of wind speed on PM2.5 was consistent with findings in the literature. Despite low spatial variability in PM2.5, building morphological characteristics explain approximately one-third of the variation in the model's fixed effects. In addition to building characteristics, openness in the wind direction also impact the PM2.5 but with a positive correlation.

Chapter 4 focused on PN, an alternative measure of PM2.5. The study estimated the average PN values at five particle-size bins over street segments at the two hours MODIS satellite passed the study area. The proposed model used AOD, meteorological data, and building morphological characteristics as covariates for PM2.5 PN in each bin in San Francisco. Good estimation with correlation coefficient (R) values above 0.82 suggested that the AOD, commonly used to estimate standard measures of PM2.5, showed the ability to estimate PN. A comparative analysis of variable importance denoted that after weather-related variables, AOD, and AOD-related variables were more important covariates than the building morphological characteristics in estimating PN.

The three studies together examined the standard and alternate measures of PM2.5 at three different scales. The studies showed the usefulness of AOD and meteorological factors to estimate standard and alternate PM2.5 measures. MAIAC AOD data are available on an average twice a day at a given location. Future studies may explore similar approaches using the newly available AOD products from GOES-16 satellite providing AOD data at 5 min frequency and

spatial resolution of 2km. Prevalence of temporal variability in PM2.5 over spatial variability on the campus of University of Texas at Dallas and variable importance in the San Francisco study signified the important role of weather-related variables in modeling PM2.5 and PN concentrations. Moreover, AOD and meteorological variables alone achieved the correlation coefficient (R) of 0.87 between observed and estimated PM2.5 in Dallas-Fort Worth region. The study also demonstrated the appreciable influences of contextual factors, such as built-up area and openness in the wind direction on PM2.5.

The study has several limitations. The models in the study were developed for Dallas-Fort Worth and San Francisco. The models lack direct applicability to other cities. However, the modeling approaches are transferable since the key variables from MODIS data and weather reanalysis data are readily available worldwide. The study explored the effects of a limited number of building morphological characteristics on PM2.5 and PN concentrations. Other building morphological characteristics or indices that account for spatial configuration of building arrangements could be more useful in estimating different PM measures. The microenvironment study proceeded at a single site on a university campus with limited variations in building morphology. Variation in PM2.5 depends on many factors like weather parameters and emission sources that vary across different parts of the year. All the scenarios cannot be explored at a single site with a limited number of runs. More sites with different urban settings need to be studied for an extended period of time to improve our understanding of microscale dynamics of PM2.5. The study used a CNN-based approach to model PM2.5 because of its ability to account for the influence of spatially adjacent locations. However, the study does not

explore the mechanism by which spatially adjacent locations affect the PM2.5 concentration. Using explainable AI techniques, future studies can investigate how predictors from spatially adjacent locations contribute to improved model performance, especially analyzing scenarios that lead to elevated PM2.5 concentration. This will help gain insights into spatial processes responsible for PM2.5 dynamics.

APPENDIX

SUPPLEMENTARY TABLE

Table A.1: Results across the neural networks with 8, 16 and 24 neurons for five particle bins

PN	Number of neurons	Correlation coefficients (R)		Root Mean Squared Error (RMSE)	
		Train	Test	Train	Test
PN1	8	0.8577	0.8533	5754	5823
	16	0.8704	0.8652	5332	5362
	24	0.8869	0.8835	4253	4281
PN2	8	0.8282	0.8252	1685	1694
	16	0.8521	0.8510	1696	1714
	24	0.8793	0.8768	1325	1339
PN3	8	0.8398	0.8357	756	760
	16	0.8527	0.8440	757	774
	24	0.8751	0.8722	751	754
PN4	8	0.8360	0.8307	348	351
	16	0.8589	0.8570	306	307
	24	0.8724	0.8639	293	302
PN5	8	0.7773	0.7733	386	389
	16	0.8393	0.8366	345	344
	24	0.8399	0.8374	337	340

REFERENCES

1. Alfarra, M. (2004). Insights into atmospheric organic aerosols using an aerosol mass spectrometer (Doctoral dissertation, University of Manchester).
2. Anselin, L. (2001). Spatial Econometrics. In Baltagi, B., H. (Ed.), A companion to theoretical econometrics (Vol. 1). (pp. 310-330). Blackwell
3. Atkinson, R. W., Mills, I. C., Walton, H. A., & Anderson, H. R. (2015). Fine particle components and health—a systematic review and meta-analysis of epidemiological time series studies of daily mortality and hospital admissions. Journal of exposure science & environmental epidemiology, 25(2), 208-214.
4. Baldauf, R. W., Devlin, R. B., Gehr, P., Giannelli, R., Hassett-Sipple, B., Jung, H., Martini G., McDonald J., Sacks J. D., & Walker, K. (2016). Ultrafine particle metrics and research considerations: review of the 2015 UFP workshop. International journal of environmental research and public health, 13(11), 1054.
5. Barkjohn, K. K., Gantt, B., & Clements, A. L. (2020). Development and Application of a United States wide correction for PM 2.5 data collected with the PurpleAir sensor. Atmospheric Measurement Techniques Discussions, 1-34.
6. Bell, M. L., Dominici, F., Ebisu, K., Zeger, S. L., & Samet, J. M. (2007). Spatial and temporal variation in PM2. 5 chemical composition in the united states for health effects studies. Environmental Health Perspectives,115(7), 989-995.
7. Bell, M. L., Ebisu, K., Peng, R. D., Samet, J. M., & Dominici, F. (2009). Hospital admissions and chemical composition of fine particle air pollution. American journal of respiratory and critical care medicine, 179(12), 1115-1120.

8. Bell, M. L., Ebisu, K., Peng, R. D., Walker, J., Samet, J. M., Zeger, S. L., & Dominici, F. (2008). Seasonal and regional short-term effects of fine particles on hospital admissions in 202 US counties, 1999–2005. American journal of epidemiology, 168(11), 1301-1310.
9. Berrisford, P., Dee, D., Poli, P., Brugge, R., Fielding, K., Fuentes, M., Kallberg, P., Kobayashi, S., Uppala, S. and Simmons, A. (2011) The ERA-Interim archive, version 2.0. ERA report series. 1. Technical Report. ECMWF pp23.
10. Brown, S. G., Penfold, B., Mukherjee, A., Landsberg, K., & Eisinger, D. S. (2019). Conditions Leading to Elevated PM2. 5 at Near-Road Monitoring Sites: Case Studies in Denver and Indianapolis. International journal of environmental research and public health, 16(9), 1634.
11. Chen, H., Kwong, J. C., Copes, R., Hystad, P., van Donkelaar, A., Tu, K., Brook J. R., Goldberg M. S., Martin R. V., Murray B. J., Wilton A. S., Kopp A. & Burnett R. T. (2017). Exposure to ambient air pollution and the incidence of dementia: a population-based cohort study. Environment international, 108, 271-277.
12. Chudnovsky, A., Lyapustin, A., Wang, Y., Tang, C., Schwartz, J., & Koutrakis, P. (2014). High resolution aerosol data from MODIS satellite for urban air quality studies. Open Geosciences, 6(1), 17-26
13. Churg, A., & Brauer, M. (1997). Human lung parenchyma retains PM2. 5. American journal of respiratory and critical care medicine, 155(6), 2109-2111.
14. Cohen A. J., Brauer .M, Burnett R, Anderson H. R., Frostad J., Estep K., Balakrishnan K., Brunekreef B., Dandona L., Dandona R., Feigin V., Freedman G., Hubbell B., Jobling A., Kan H., Knibbs L., Liu Y., Martin R., Morawska L., Pope C. A., Shin H., Straif K.,

Shaddick G., Thomas M., van Dingenen R., van Donkelaar A., Vos T., Murray C. J. L., Forouzanfar M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 389(10082), 1907-1918.

15. Cyrys, J., Pitz, M., Heinrich, J., Wichmann, H. E., & Peters, A. (2008). Spatial and temporal variation of particle number concentration in Augsburg, Germany. Science of the Total Environment, 401(1-3), 168-175.
16. Deshmukh, P., Isakov, V., Venkatram, A., Yang, B., Zhang, K. M., Logan, R., & Baldauf, R. (2019). The effects of roadside vegetation characteristics on local, near-road air quality. Air Quality, Atmosphere & Health, 12(3), 259-270.
17. Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016). Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States. Environmental science & technology, 50(9), 4712-4721.
18. Dinoi, A., Conte, M., Grasso, F. M., & Contini, D. (2020). Long-term characterization of submicron atmospheric particles in an urban background site in Southern Italy. Atmosphere, 11(4), 334.
19. Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., & Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. Jama, 295(10), 1127-1134.
20. Edussuriya, P., Chan, A., & Malvin, A. (2014). Urban morphology and air quality in dense residential environments: Correlations between morphological parameters and air pollution at street-level. Journal of Engineering Science and Technology, 9(1), 64-80.

21. Engel-Cox, J.A., Holloman, C.H., Coutant, B.W. and Hoff, R.M. (2004). Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. Atmospheric Environment, 38(16), pp.2495-2509.

22. Ginzburg, H., Liu, X., Baker, M., Shreeve, R., Jayanty, R. K. M., Campbell, D., & Zielinska, B. (2015). Monitoring study of the near-road PM2. 5 concentrations in Maryland. Journal of the Air & Waste Management Association, 65(9), 1062-1071.

23. Goodchild, M. F. (2011). Scale in GIS: An overview. Geomorphology, 130(1-2), 5-9.

24. Google (2017), "California_201604_201709_GoogleAclimaAQ"

25. Guo J., Xia F., Zhang Y., Liu H., Li J., Lou M., He J., Yan Y., Wang F., Min M. & Zhai P. (2017). Impact of diurnal variability and meteorological factors on the PM 2.5-AOD relationship: Implications for PM 2.5 remote sensing. Environmental Pollution, 221, 94-104.

26. Gupta, P., Doraiswamy, P., Levy, R., Pikelnaya, O., Maibach, J., Feenstra, B., Polidori A., Kiros F. & Mills, K. C. (2018). Impact of California fires on local and regional air quality: The role of a low-cost sensor network and satellite observations. GeoHealth, 2(6), 172-181.

27. Harrison, W. A., Lary, D., Nathan, B., & Moore, A. G. (2015) a. The neighborhood scale variability of airborne particulates. Journal of Environmental Protection,6(05), 464

28. Harrison, W. A., Lary, D. J., Nathan, B. J., & Moore, A. G. (2015) b. Using remote control aerial vehicles to study variability of airborne particulates. Air, Soil and Water Research,8, ASWR-S30774.

29. Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017). Estimating PM2. 5 concentrations in the conterminous United States using the random forest approach. Environmental science & technology, 51(12), 6936-6944.
30. Hu, X., Waller, L. A., Lyapustin, A., Wang, Y., Al-Hamdan, M. Z., Crosson, W. L., Estes, M. G. Jr., Estes, S. M., Quattrochi, D. A., Puttaswamy, S. J., & Liu, Y. (2014). Estimating ground-level PM2. 5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sensing of Environment, 140, 220-232.
31. Hussein, T., Sogacheva, L., & Petäjä, T. (2018). Accumulation and Coarse Modes Particle Concentrations during Dew Formation and Precipitation. Aerosol and Air Quality Research, 18(12), 2929-2938.
32. Ibald-Mulli, A., Wichmann, H. E., Kreyling, W., & Peters, A. (2002). Epidemiological evidence on health effects of ultrafine particles. Journal of Aerosol Medicine, 15(2), 189-201.
33. Johansson, C., Norman, M., & Gidhagen, L. (2007). Spatial & temporal variations of PM10 and particle number concentrations in urban air. Environmental monitoring and assessment, 127(1-3), 477-487.
34. Johansson, C., Norman, M., & Gidhagen, L. (2007). Spatial & temporal variations of PM10 and particle number concentrations in urban air. Environmental monitoring and assessment, 127(1), 477-487.
35. Kioumourtzoglou, M. A., Schwartz, J. D., Weisskopf, M. G., Melly, S. J., Wang, Y., Dominici, F., & Zanobetti, A. (2015). Long-term PM2. 5 exposure and neurological

hospital admissions in the northeastern United States. Environmental health perspectives, 124(1), 23-29.

36. Krall, J. R., Mulholland, J. A., Russell, A. G., Balachandran, S., Winquist, A., Tolbert, P. E., Waller L. A. & Sarnat, S. E. (2017). Associations between source-specific fine particulate matter and emergency department visits for respiratory disease in four US cities. Environmental health perspectives, 125(1), 97-103.
37. Kwon, H. S., Ryu, M. H., & Carlsten, C. (2020). Ultrafine particles: unique physicochemical properties relevant to health and disease. Experimental & Molecular Medicine, 1-11.
38. Lary, D. J., Faruque, F. S., Malakar, N., Moore, A., Roscoe, B., Adams, Z. L., & Eggelston, Y. (2014). Estimating the global abundance of ground level presence of particulate matter (PM2. 5). Geospatial health, 8(3), 611-630
39. Li, T., Shen, H., Yuan, Q., Zhang, X., & Zhang, L. (2017). Estimating ground-level PM2. 5 by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophysical Research Letters, 44(23), 11-985.
40. Lou, C., Liu, H., Li, Y., Peng, Y., Wang, J., & Dai, L. (2017). Relationships of relative humidity with PM 2.5 and PM 10 in the Yangtze River Delta, China. Environmental monitoring and assessment, 189(11), 1-16.
41. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).

42. Lyapustin, A., & Wang, Y. (2018). MODIS Multi-Angle Implementation of Atmospheric Correction (MAIAC) Data User's Guide. NASA: Greenbelt, MD, USA.

43. McGill, B. J. (2010). Matters of scale. Science, 328(5978), 575-576.

44. Newson, P., & Krumm, J. (2009, November). Hidden Markov map matching through noise and sparseness. In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 336-343).

45. Olsen, Y., Karottki, D. G., Jensen, D. M., Bekö, G., Kjeldsen, B. U., Clausen, G., Hersoug, L.-G., Holst, G. J., Wierzbicka, A., Sigsgaard, T., Linneberg, A., Møller, P., & Loft, S. (2014). Vascular and lung function related to ultrafine and fine particles exposure assessed by personal and indoor monitoring: a cross-sectional study. Environmental Health, 13(1), 1-10.

46. Özkaynak, H., Baxter, L. K., Dionisio, K. L., & Burke, J. (2013). Air pollution exposure prediction approaches used in air pollution epidemiology studies. Journal of exposure science & environmental epidemiology, 23(6), 566-572.

47. Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., & Moon, T. (2020). Estimating PM2. 5 concentration of the conterminous United States via interpretable convolutional neural networks. Environmental Pollution, 256, 113395.

48. Parker, W. S. (2016). Reanalyses and observations: What's the difference? Bulletin of the American Meteorological Society, 97(9), 1565-1572.

49. Peng, R. D., Bell, M. L., Geyh, A. S., McDermott, A., Zeger, S. L., Samet, J. M., & Dominici, F. (2009). Emergency admissions for cardiovascular and respiratory diseases

and the chemical composition of fine particle air pollution. Environmental health perspectives, 117(6), 957-963.

50. Peng, Z. R., Wang, D., Wang, Z., Gao, Y., & Lu, S. (2015). A study of vertical distribution patterns of PM2. 5 concentrations based on ambient monitoring with unmanned aerial vehicles: A case in Hangzhou, China.Atmospheric Environment,123, 357-369.
51. Peters, A., Hampel, R., Cyrys, J., Breitner, S., Geruschkat, U., Kraus, U., Zareba, W. & Schneider, A. (2015). Elevated particle number concentrations induce immediate changes in heart rate variability: a panel study in individuals with impaired glucose metabolism or diabetes. Particle and fibre toxicology, 12(1), 7.
52. Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama, 287(9), 1132-1141.
53. Quiros, D. C., Lee, E. S., Wang, R., & Zhu, Y. (2013). Ultrafine particle exposures while walking, cycling, and driving along an urban residential roadway. Atmospheric Environment, 73, 185-194.
54. Salma, I., Balásházy, I., Winkler-Heil, R., Hofmann, W., & Záray, G. (2002). Effect of particle mass size distribution on the deposition of aerosols in the human respiratory system. Journal of Aerosol Science, 33(1), 119-132.
55. Shi, Y., Lau, K. K. L., & Ng, E. (2016). Developing street-level PM2. 5 and PM10 land use regression models in high-density Hong Kong with urban morphological factors. Environmental science & technology,50(15), 8178-8187

56. Shi, Y., Lau, K. K. L., & Ng, E. (2016). Developing street-level PM2. 5 and PM10 land use regression models in high-density Hong Kong with urban morphological factors. Environmental science & technology,50(15), 8178-8187

57. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48.

58. Stafoggia, M., Schwartz, J., Badaloni, C., Bellander, T., Alessandrini, E., Cattani, G., Donato, F., Gaeta, A., Leone, G., Lyapustin, A., Sorek-Hamer, A., Hoogh, K., Di, Q., Forastiere, F., & Kloog, I. (2017). Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. Environment international, 99, 234-244.

59. Superczynski, S. D., Kondragunta, S., & Lyapustin, A. I. (2017). Evaluation of the multi-angle implementation of atmospheric correction (MAIAC) aerosol algorithm through intercomparison with VIIRS aerosol products and AERONET. Journal of Geophysical Research: Atmospheres, 122(5), 3005-3022.

60. Tate, N., & Atkinson, P. M. (Eds.). (2001). Modelling scale in geographical information science. John Wiley & Sons.

61. Taylor, L., & Nitschke, G. (2018, November). Improving deep learning with generic data augmentation. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1542-1547). IEEE.

62. Tian, J., & Chen, D. (2010). A semi-empirical model for predicting hourly ground-level fine particulate matter (PM2.5) concentration in southern Ontario from satellite remote

sensing and ground-based meteorological measurements doi://doi.org/10.1016/j.rse.2009.09.011

63. Traffic Count Program, Annual Report, Dec 2019, City of Richardson available at https://www.cor.net/home/showdocument?id=28259 last accessed on April 7, 2021.
64. Tsai, J. H., Chang, K. L., Lin, J. J., Lin, Y. H., & Chiang, H. L. (2005). Mass-size distributions of particulate sulfate, nitrate, and ammonium in a particulate matter nonattainment region in southern Taiwan. Journal of the air & waste management association, 55(4), 502-509.
65. U.S. Environmental Protection (April 2020), 2017 National Emission Inventory, Technical Support Document accessed at https://www.epa.gov/sites/production/files/202004/documents/nei2017_tsd_full_30apr2020.pdf on 08/31/2020
66. U.S. Environmental Protection Agency (September 2013), 2008 National Emission Inventory, version 3, Technical Support Document DRAFT accessed at https://www.epa.gov/sites/production/files/2015-07/documents/2008_neiv3_tsd_draft.pdf
67. United Nations. (2018). World Urbanization Prospects: The 2018 Revision, Key Facts. Technical report. Available at https://population.un.org/wup/Publications/Files/WUP2018-KeyFacts.pdf
68. Valavanidis, A., Fiotakis, K., & Vlachogianni, T. (2008). Airborne particulate matter and human health: toxicological assessment and importance of size and composition of particles for oxidative damage and carcinogenic mechanisms. Journal of Environmental Science and Health, Part C, 26(4), 339-362.

69. von Storch, H., & Zorita, E. (2019). The history of ideas of downscaling—from synoptic dynamics and spatial interpolation. Frontiers in Environmental Science, 7, 21.

70. Vos, P. E., Maiheu, B., Vankerkom, J., & Janssen, S. (2013). Improving local air quality in cities: to tree or not to tree?. Environmental pollution, 183, 113-122.

71. Wang, J., & Ogawa, S. (2015). Effects of meteorological conditions on PM2. 5 concentrations in Nagasaki, Japan. International journal of environmental research and public health, 12(8), 9089-9101.

72. Wang, Z., Zhong, S., Peng, Z. R., & Cai, M. (2018). Fine-scale variations in PM2. 5 and black carbon concentrations and corresponding influential factors at an urban road intersection. Building and Environment, 141, 215-225.

73. Xie, Y., Wang, Y., Zhang, K., Dong, W., Lv, B., & Bai, Y. (2015). Daily estimation of ground-level PM2. 5 concentrations over Beijing using 3 km resolution MODIS AOD. Environmental science & technology, 49(20), 12280-12288.

74. Xu, J., Jiang, H., Zhang, X., Lu, X., & Peng, W. (2014). Study on spatial-temporal variation of aerosol optical depth over the Yangtze Delta and the impact of land-use/cover. International Journal of Remote Sensing, 35(5), 1741-1755.

75. Yang, J., Shi, B., Zheng, Y., Shi, Y., & Xia, G. (2020). Urban form and air pollution disperse: Key indexes and mitigation strategies. Sustainable Cities and Society, 57, 101955.

76. Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., ... & Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM2. 5 concentrations across

China using a spatially explicit machine learning algorithm. Atmospheric environment, 155, 129-139.

77. Zheng, Y., Liu, F., & Hsieh, H. P. (2013, August). U-air: When urban air quality inference meets big data. InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining(pp. 1436-1444). ACM.

78. Zhu, Y., Hinds, W. C., Kim, S., & Sioutas, C. (2002). Concentration and size distribution of ultrafine particles near a major highway. Journal of the air & waste management association, 52(9), 1032-1042.

BIOGRAPHICAL SKETCH

Yogita Karale completed her Bachelor of Technology in Civil Engineering in 2011 from Shivaji University Kolhapur. While pursuing her bachelor's degree in Civil Engineering, she developed interest in Remote Sensing and GIS through a distance learning course offered by Indian Institute of Remote Sensing, Dehradun. To follow her interest in the same field, she pursued a Master of Technology in Geo-informatics and Natural Resources Engineering at Indian Institute of Technology Bombay where she received the Institute Silver Medal for the year 2012-2013. She worked as a researcher at TCS Innovation Labs Mumbai. Later she started her doctoral studies in Geospatial Information Sciences at The University of Texas at Dallas in 2016. She presented her research work at annual meetings of the American Association of Geographers in 2019 and 2021. Her research interests include remote sensing and machine learning for geospatial data. While pursuing her doctoral studies, she worked as a teaching assistant for several GIS courses. She was also an instructor for GIS courses Principles of Geospatial Information Sciences and Spatial Data Science during Summer 2020 and Fall 2020.

CURRICULUM VITAE

Yogita Karale

Education:

- PhD in Geospatial Information Sciences (Expected Summer 2021)

  University of Texas at Dallas
- M.Tech in Geo-informatics and Natural Resources Engineering 2011-2013

  Indian Institute of Technology Bombay
- B.Tech in Civil Engineering 2007-2011

  Shivaji University Kolhapur (Walchand College of Engineering, Sangli)

Professional Experience:

- Instructor, University of Texas at Dallas Summer 2020, Fall 2020
- Teaching Assistant, University of Texas at Dallas Fall 2016-Spring 2020, Spring 2021
- Researcher, TCS Innovation Labs Mumbai July 2013- March 2016

Awards and Scholarships:

- Pioneer Student Research Scholarship 2018, 2019, 2020

  University of Texas at Dallas
- Institute Silver Medal 2013

  Indian Institute of Technology Bombay

Professional Memberships:

- American Association of Geographers
- Gamma Theta Epsilon Honor Society

Journal Articles:

- Cummings, A. R., Karale, Y., Cummings, G. R., Hamer, E., Moses, P., Norman, Z., & Captain, V. (2017). UAV-derived data for mapping change on a swidden agriculture plot: Preliminary results from a pilot study. International Journal of Remote Sensing, 38(8-10), 2066-2082.

Conference Proceedings:

- Karale, Y. Y., Mohite, J., & Jagyasi, B. (2014, November). Crop classification based on multi-temporal satellite remote sensing data for agro-advisory services. In Land Surface Remote Sensing II (Vol. 9260, p. 926004). International Society for Optics and Photonics.

Conference Presentations:

- Karale Y. and Yuan M., 2019, Downscaling PM2.5 through data fusion in Dallas-Fort Worth, Annual Meeting of American Association of Geographers, Washington DC
- Karale Y. and Yuan M., 2021, How does convolutions in neural network improve hourly PM2.5 estimates, Annual Meeting of American Association of Geographers, Virtual