Chronic Exposure to Fine Particulate Matter and Heart Failure in New York City:
A Methodological Exploration of Environmental Justice and Health

Andrew R. Maroko

A dissertation submitted to the Graduate Faculty in Earth and Environmental Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York

2010

This manuscript has been read and accepted for the Graduate Faculty in Earth and Environmental Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date	**Dr. Juliana Maantay**, Lehman College – Chair of Examining Committee
Date	**Dr. Yehuda Klein**, Graduate Center – Executive Officer

Dr. Allan Frei, Hunter College – supervisory committee member

Dr. Ines Miyares, Hunter College – supervisory committee member

THE CITY UNIVERSITY OF NEW YORK

Dissertation Title: Chronic Exposure to Fine Particulate Matter and Heart Failure in New York City: A Methodological Exploration of Environmental Justice and Health
Author: Andrew Maroko
Date of defense: July 27th, 2010
Time of defense: 2pm

Abstract

**Chronic Exposure to Fine Particulate Matter in New York City:
A Methodological Exploration of Environmental Justice and Health**

Andrew R. Maroko

Advisor: Professor Juliana Maantay

Increased exposure to air pollution has been connected with environmentally-linked diseases (increased morbidity), decreased lifespan (increased mortality), environmental injustices (inequitable distribution of pollution based on population characteristics), reduction of quality-of-life, and increased health care costs. The main goals of this work are to analyze and quantify the potential association between chronic fine particulate matter ($PM_{2.5}$) exposure and heart failure hospitalization rates in New York City and to explore the possibility that specific populations (e.g. racial and ethnic minorities, less educated populations, lower income populations) suffer from increased chronic exposure to $PM_{2.5}$ from local stationary sources when compared to other populations in the context of environmental justice. Fine particulate matter exposure in New York City was estimated using proximity analysis, air dispersion modeling, and land use regression modeling. The characteristics, strengths, and weaknesses of each technique were compared and contrasted. A number of statistical techniques were also employed to assess and quantify these associations (odds ratios, ordinary least squares regressions, spatial autoregressive models, and geographically weighted regressions). The utility and appropriateness of each of these statistical models were examined.

The results of the analyses suggested the presence of environmental injustices, although the relationships appeared complex and non-linear. The environmental health analyses found a positive association between intra-urban chronic exposure to fine particulate matter and heart failure hospitalization rates when controlling for socio-demographics in New York City.

Acknowledgements

I would like to thank, from the bottom of my heart, many people and organizations that have made this dissertation possible. The person who contributed most to my graduate-level academic development is unquestionably **Dr. Juliana Maantay**. She provided me with guidance, knowledge, and always good advice. She connected me with funding and projects which consistently challenged and entertained me. Professor Maantay also treated me, and all students, with respect independent of academic rank or standing. Without her, I certainly would not be where I am today and for that I can never fully repay her.

I would also like to thank my committee members: the great Professors **Ines Miyares** and **Allan Frei**. They were generous enough with their time to agree to be on my committee and I will be forever thankful for their patience, kindness, and depth of knowledge. They were chosen as committee members not only due to their brilliance, but also because of their phenomenal ability to teach difficult concepts in clear and intuitive ways. To me, this shows a truly full understanding of the subject matter.

There are many other professors who I had the pleasure of learning from over the years who have earned my thanks. They include **David Harvey**, **Stuart Fotheringham**, **Charles Brunsdon**, **Martin Charlton**, and **Jochen Albrecht**. Each of these people, in their own ways, has exposed me to new concepts, ideas, or methods that have drastically influenced my approach to geography and research. I have also had many colleagues who have helped me in countless

ways during my graduate experience including **Christopher Herrmann**, **Jun Tu**, **Kristen Grady**, **Tarendra Lakhankar**, **Jennifer Brisbane**, **Rachael Weiss**, **Brian Morgan**, **Lesley Patrick**, **Grant Pezeshki**, **Holly Porter-Morgan**, **Natalia Brzezinska**, and **Dellis Stanberry**. These students have helped me with everything from spatial regression to deep fried turkey, inclusive.

The **South Bronx Environmental Justice Partnership** (SBEJP) has also been instrumental in my development. **Dr. Hal Strelnick** of the **Montefiore Medical Center**, the clinical arm of the partnership, has been invaluable to me over the past few years by acting as a PI on many fascinating SBEJP projects and studies – most of which resulted in publications. I would also like to thank the community component of SBEJP, **For a Better Bronx** (FABB). FABB, particularly **Marian Feinberg**, has helped shape my understanding of community/academic interaction and community based participatory research (CBPR) by agreeing to collaborate with me on many of the aforementioned SBEJP projects. It is also important to note the support I received from the **Lehman College department of Environmental, Geographic, and Geological Sciences**, the **Urban GISc Lab** at Lehman College, and the **Graduate Center's program in Earth and Environmental Sciences**. Regarding the latter, I would be remiss if I did not thank, by name, the two people who keep the program running smoothly and efficiently – namely **Yehuda Klein** and **Lina McClain**. They have been very helpful to me during my stay at the Graduate Center. I would also like to mention members of the **NOAA-CREST** family for their financial and academic support including **Reza Khanbilvardi**, **Shakila Merchant**, **Bruce Ramsey**, and **Ralph Ferrarro**. It has been quite enlightening to work with engineers and remote

sensing scientists as well as being given the opportunity to present my work in formats not necessarily native to geographers or health scientists.

Ultimately, none of this would have been possible without the sacrifices and understanding of my family. **Rosemary Farrell**, **Peter Farrell Maroko**, and **Sofia Rose Maroko** have shown tremendous patience and support through some difficult times. They never acted judgmentally or selfishly with respect to my studies but instead offered to help in any way they could. In the case of Peter and Sofia this generally entailed pressing buttons on my computer and asking me to explain my maps to them – an activity which really tested my knowledge of the subject. In the case of Rosemary, contributions included, but were not limited to, editing dense and nearly impenetrable papers which couldn't have been a whole lot of fun for her.

I would also like to acknowledge some of the funding agencies:

- National Institute of Environmental Health Sciences (NIEHS)
- National Oceanic and Atmospheric Administration - Cooperative Remote Sensing Science & Technology Center (NOAA-CREST)
- National Institute of Health - National Center for Minority Health and Health Disparities
- Bronx Center to Reduce & Eliminate Ethnic & Racial Health Disparities (Bronx CREED)
- Professional Staff Congress - City University of New York (PSC-CUNY)

This dissertation is dedicated to the memory of my mother, **Dr. Cleuza Maria Maroko**, and father, **Dr. Peter Richard Maroko**. They taught me the value of curiosity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 BACKGROUND

Increased exposure to air pollution has been connected with environmentally-linked diseases (increased morbidity), decreased lifespan (increased mortality), environmental injustice (inequitable distribution of pollution based on population characteristics), reduction of quality-of-life, and increased health care costs. Many major metropolitan areas, including New York City, have disease rates in excess of the national average (Maantay, 2007). In this dissertation, chronic fine particulate matter ($PM_{2.5}$) exposures in New York City (NYC) were estimated using various techniques and data sources. These estimations were then used to analyze potential environmental justice (EJ) issues as well as the pollution's influence on cardiovascular health across the city. As one of the objectives of this work is to compare methods of $PM_{2.5}$ exposure estimation, a number of techniques were utilized (proximity analysis, air dispersion modeling, and land use regression). A number of statistical techniques were also employed to assess and quantify the associations between exposure and disease (odds ratios, ordinary least squares regressions, spatial autoregressive models, and geographically weighted regressions).

The dissertation is divided into six chapters (background, methods, analysis, results, conclusions, and references), each with a number of sections and subsections. The chapters and sections are numbered hierarchically (e.g. the 5th section, 3rd subsection of chapter two would be enumerated as "2.5.3"). Figures, tables, and equations are numbered by the chapter to which they belong (e.g. the third figure in chapter two would be titles "Figure 2-3"). This background chapter serves as an introduction to the goals and hypotheses of the dissertation itself as well as to provide

information and background regarding the content of the research. It is divided into six sections including Goals (**1.1**), Hypotheses (**1.2**), Fine Particulate Matter (**1.3**), Environmental Justice (**1.4**), Environmental Health (**1.5**), and Regression Models (**1.6**).

1.1 GOALS

1. To quantify and analyze the potential association between (or the contribution of) chronic $PM_{2.5}$ exposure and heart failure hospitalization rates in New York City, using geographic information science (GISc) as the analytical framework.
2. To explore the possibility of specific populations (e.g. racial and ethnic minorities, less-educated populations, lower income populations) bearing the burden of increased chronic exposure to $PM_{2.5}$ from local stationary sources when compared to other populations in the context of environmental justice.

1.2 HYPOTHESES

1. There is a positive association between chronic exposure to $PM_{2.5}$ from local stationary sources and increased risk of hospitalization for heart failure in New York City.
2. There is a positive association between chronic exposure to $PM_{2.5}$ from major mobile sources and increased risk of hospitalization for heart failure in New York City.
3. There is a positive association between chronic exposure to ambient $PM_{2.5}$ and increased risk of hospitalization for heart failure in New York City.

4. Populations which are composed of high proportions of racial or ethnic minorities, less-educated, or those having lower incomes are more likely to be spatially co-incident with areas with high exposure to $PM_{2.5}$ from local stationary sources in NYC.

1.3 FINE PARTICULATE MATTER

Fine particulate matter, also known as $PM_{2.5}$, refers to airborne particles with a diameter smaller than 2.5 microns. It is most commonly produced by combustion (e.g. emissions from vehicles and power plants) and by chemical reactions between gases such as sulfur dioxide, nitrogen oxides, and volatile organic compounds (NYSDEC, 2010). The United States Environmental Protection Agency (EPA) has created standards for the concentration of fine particulate matter for both 24-hour ($35\mu g/m^3$) and annual averages ($15\mu g/m^3$) known as The National Ambient Air Quality Standards (NAAQS). These particulates may be either solid or liquid and are often composed of a number components including nitrates, sulfates, organic chemicals, metals, soil/dust particles, or allergens such as pollen and mold spores (EPA, 2010a). Aside from the negative health effects which will be discussed in **Section 1.5**, the EPA has noted some associations between $PM_{2.5}$ and the environment, including: reduced visibility (haze), increased acidity of lakes and streams, nutrient balance changes in coastal waters and river basins, reduced levels of nutrients in soil, damage to forests and crops, reduced diversity in ecosystems, and damage to stone and other materials (EPA, 2010a).

This dissertation uses an ecological framework, meaning that rather than focusing on individual health outcomes, the relative health of populations aggregated to geographic units is analyzed.

When using an ecological study design at a fairly fine level (e.g. census tract) to examine the affects of air pollution and any other variable (e.g. health), the optimal situation would be the existence of an air monitor representing every geographic unit of analysis (e.g. census tract). Since this is highly unlikely (NYC had only 15 monitors recording $PM_{2.5}$ concentrations in place in 2002) researchers have developed many different techniques to estimate the concentrations of, or exposure to, $PM_{2.5}$ in the absence of in situ monitors. These include proximity analysis, air dispersion modeling, and land use regression.

Proximity analysis is a commonly-used method to study the association between air pollution and environmental health as well as environmental justice (Chakraborty and Armstrong, 1997; Hodgson et al., 2007; Jerrett et al., 2005; Lin et al., 2002; Maantay, 2007; Ryan et al., 2007; van Vliet et al., 1997). This method assumes that the distance to an emission source or sources functions as an appropriate surrogate for human exposure to air pollution (Jerrett et al., 2005), and it has a great advantage over almost all other methods in that proximity analysis is very straightforward, fast, and easy to employ. However, there are disadvantages as well.

First, it assumes that air pollution disperses equally in all directions from a source only considering an absolute, fixed distance of pollutant dispersion, without taking into account physical properties of pollutants (e.g., density), source characteristics (e.g., emission rate, velocity, and temperature), local meteorological conditions (e.g., wind direction, wind speed, and temperature), topographical features (elevation changes), and effect of the surrounding built

environment (e.g. tall buildings), all of which may influence the actual pollution concentration at any given location (Maantay et al., 2009).

Secondly, the distances used (e.g. radii of the buffers) are often somewhat subjective, most commonly based on best estimates of general pollutant fate and transport as determined by environmental scientists, rather than being derived from the actual measured concentrations or specific emission data. This can result in highly varied populations considered as exposed dependent upon the distance chosen, which could affect the association found between air pollution and health.

Thirdly, the differentiation between areas, cases, and populations outside and inside a proximity buffer is often treated as binary, and as such is unable to represent the continuous spatial process of air pollution concentrations. Thus, an application of proximity analysis might not realistically reflect the complexities inherent in the relationship between air pollution exposure and human health.

A more sophisticated method, air dispersion modeling, has also been utilized to study the relationship between air pollution and human health (Bellander et al., 2001; Hodgson et al., 2007; Hruba´ et al., 2001; Maantay et al., 2009; Poulstrup and Hansen, 2004). Air dispersion modeling calculates the movement of air, or pollutants that are in the air, across a landscape using mathematical equations that consider emission quantities, meteorological and topographical factors, and describes chemical and physical processes within the atmosphere over

time and space, in order to calculate concentrations of air pollutants at different 'receptors' (researcher-defined locations for concentration calculations). If sufficient data exist and are accessible for pollution sources, emission characteristics, meteorological conditions, topographical features, and physical environment information, air dispersion modeling has the potential to provide a more accurate assessment of possible exposure without the need for extensive monitoring networks (Dent et al., 1998; Hodgson et al., 2007; Jerrett et al., 2005).

A widely-used air dispersion model is AERMOD (American Meteorological Society / Environmental Protection Agency Regulatory Model), which has been validated and frequently used to simulate air dispersion of pollutants from industrial, mining, landfill, and road sources (Cimorelli et al., 2005; Kesarkar et al., 2007; Macleod et al., 2006; Perry et al., 2005; Singh et al., 2006; Touma et al., 2007). It is an advanced steady-state model that was designed to simulate the air dispersion from sources to a distance up to 50 km. It incorporates the boundary layer theory and an understanding of turbulence and dispersion, and also considers the influence of building wakes (i.e. downwash) on plume rise and dispersion (Perry et al., 2005). AERMOD is listed as a "preferred and recommended model" by the U.S. Environmental Protection Agency (EPA) and its accuracy is well-tested and documented (EPA, 2010b). However, it is rarely used for studying the association between air pollution and environmental health or environmental justice due to the relatively costly and difficult to obtain air dispersion modeling data inputs, including source locations, source release parameters, meteorological parameters, terrain, and building locations and heights. Furthermore, a full implementation of this model for air pollution and environmental health study requires a large commitment of time and an integration of

specialized software, including a dispersion model and a geographic information system (GIS). The process of integration usually involves a certain degree of familiarity in computer programming, geographic information science, and meteorology, any one or all of which may be unfamiliar to health scientists (Jerrett et al., 2005).

The next exposure method to be discussed is land use regression (LUR). This technique sharply contrasts with both proximity and air dispersion analyses in that it is not explicitly designed to estimate the pollution from a specific source or set of sources, but rather estimates ambient concentrations based on multivariate regressions of the monitored data and the physical environment which surrounds the monitors (roadways, topography, population density, etc.). The resultant equation can then be applied to any location in the study area (e.g. census tract centroid, subject's place of residence, etc.). Utilization of these types of site-specific variables allows for the detection of small area variations that are not possible to accomplish with other interpolation methods such as kriging (Briggs et al., 1997; Gilliland et al., 2005). LUR models have been widely utilized to model a myriad of pollutants, including traffic-related pollution such as NO_2 and $PM_{2.5}$ (Briggs et al., 2000; Brauer et al., 2003; Gilbert et al., 2005; Ross et al., 2006). Although the predictor variables differ based on many factors including pollutant of interest, data availability, and study area location, there are some commonalities in terms of the types of data and their sources (**Table 1-1**).

Class	Variable used	Variable definition	Study
Road type	Road type 1	Road serving > 25000 people	Briggs et al. (1997)
	Road type 2	Road serving 5000—25000 people	Briggs et al. (1997)
	Road type 3	Road serving 1000—5000 people	Briggs et al. (1997)
	Highway	Undefined	Gilbert et al. (2005)
	Major road	Undefined	Gilbert et al. (2005)
	Major road	Average daily traffic count > 50,000	Ross et al. (2006)
	Major road	Average daily truck count > 1000	Ryan et al. (2008)
	High traffic road	Road serving > 25000 people	Brauer et al. (2003)
	Medium traffic road	Road serving 10000—25000 people	Brauer et al. (2003)
	Minor road	Undefined	Gilbert et al. (2005)
	Bus route	Public transportation route	Ryan et al. (2007)
Traffic count	Weighted traffic volume	15 * (volume < 40 m) + (volume 40–300m)	Briggs et al. (1997, 2000)
	Traffic volume	Volume (1000 vehicle km hr^{-1})	Briggs et al. (1997)
	Traffic count on nearest highway	Undefined	Gilbert et al. (2005)
	Average daily traffic count	Average number of cars traveling in both directions	Ross et al. (2006)
	Traffic intensity	Vehicles/day	Brauer et al. (2003)
	Heavy vehicle traffic intensity	Heavy traffic/day	Brauer et al. (2003)
	Average daily truck count	Average number of trucks traveling in both directions	Ryan et al. (2006)
Z	Altitude	Meters above sea level	Briggs et al. (1997, 2000)
	Elevation	Meters above sea level	Ross et al. (2006), Ryan et al. (2006)
Land cover	Land cover factor	Weighted sum of areas of industrial and high density residential	Briggs et al. (1997, 2000)
	Land cover	Area of built up land	Briggs et al. (1997)
	Industrial use land	Area of land designated for industrial use	Gilbert et al. (2005)
	Open space land	Area of land designated as open space	Gilbert et al. (2005)
	Commercial use land	Area of land designated for commercial use	Gilbert et al. (2005)
	Government/industry land	Area of land designated for government or industrial use	Gilbert et al. (2005)
	Household density	Number of houses in area	Gilbert et al. (2005), Ross et al. (2005), Brauer et al. (2003)
	Population density	Population in area	Ross et al. (2006), Brauer et al. (2003)
	Land use	Area covered by industry and multifamily housing	Ross et al. (2006), Ryan et al. (2007)
	Distance to coast	Distance to Pacific Ocean	Ross et al. (2006)

Table 1-1: Common LUR variables, adapted from Ryan and LeMasters, 2007.

There are some limitations to this model, however. For instance, a distance threshold must be determined for inclusion/exclusion of the predictor variables. This is often done by the generally accepted decay of the pollutant of interest (similar to manner in which the radii for proximity analyses are determined) or the density of the predictor variables (Ryan and LeMasters, 2007). LUR is also limited by the availability and accuracy of the data. For instance, average daily truck counts have been shown to be a very useful variable when estimating pollutants from diesel emissions; however this information is not always readily available (Ryan et al., 2006). Another

important limitation is the number and distribution of monitored locations. Interestingly, in their review and history of LUR models, Ryan and LeMasters (2007) found that there was a weak inverse relationship between the number of sampled locations and the models' R^2 values. This led them to the conclusion that the variability of land characteristics captured by the monitored locations may be more important to the functionality of a LUR than the total number of sampling sites (Ryan and LeMasters, 2007).

The final exposure estimation method that will be discussed is the use of remote sensing (RS) technologies. Satellite-based estimations of pollution exposure have been shown to be useful in certain scenarios. For instance, aerosol optical depth (AOD) data from the Moderate Resolution Imaging Spectrometer (MODIS) has been shown to be correlated with ground-based $PM_{2.5}$ measurements over certain areas of the United States. The strength of these relationships is not constant over the entire country, with the eastern portion generally showing better correlations than the western portion (Hu, 2009). The majority of health-related studies which utilized RS data, however, employed the county or similarly large geographies as the unit of analysis (Hu and Rao, 2009). As the spatial resolution increases (smaller area per estimate) the amount of error associated with the estimates increases (less certainty per estimate). There are additional problems that manifest themselves when analyzing urban environments due to high spatial and spectral variability, as well as the irregular size, shape and orientation of objects (Nichol and Wong, 2009). As such, when working with smaller study areas (e.g. New York City) the lack of spatial resolution and increased error due to the high reflectance of urban environments render RS data of limited utility. Paciorek and Liu (2009) describe many of the limitations in detail,

finding that AOD does not reflect spatial patterns in $PM_{2.5}$ well due to "systematic, spatially correlated discrepancies between AOD and PM2.5" (p. 904). They go on to suggest that RS data have little to add to models that already account for land use, emission sources, meteorology and regional variability. Paciorek and Liu go on to conclude that when $PM_{2.5}$ concentration data are available from monitors, statistical modeling tends to outperform the use of AOD (Paciorek and Liu, 2009). With that in mind, the authors have successfully utilized RS data within the context of a land use regression (Liu et al., 2009). Although there are serious limitations, this application seems the most likely to be useful in applying satellite data for the study of urban environmental health since it does not rely solely on the RS data, but rather uses them to improve a LUR model. These improvements could include the modeling of temporal as well as spatial variation in pollution concentrations.

In this dissertation, chronic exposure to fine particulate matter from local stationary sources (National Emissions Inventory facilities from the US Environmental Protection Agency), mobile sources (annual average daily traffic from the New York State Department of Transportation), and ambient $PM_{2.5}$ concentrations (Environmental Protection Agency air quality monitors) were determined using the aforementioned proximity buffers, air dispersion modeling, and land use regression.

1.4 ENVIRONMENTAL JUSTICE

An important part of this work involves examining the potential environmental justice (EJ) issues associated with the chronic fine particulate matter exposure from major point sources in NYC. Although awareness of the dangers of pollution increased greatly in the 1950's and 1960's, it wasn't until the late 1980's that the study of the biased nature of which populations groups are exposed based on race or income began in earnest (Maantay, 2002). EJ describes the scenario where certain groups of people – often communities of color, those with low socio-economic status, or populations that are otherwise marginalized due to differences in language, cultural discrimination, or geographic or social isolation – bear a disproportionately large share of the environmental burden and lack the political or social capital to affect policy or legislation. The U.S. Environmental Protection Agency (EPA) defines EJ as:

> Environmental Justice is the fair treatment and meaningful involvement of all people regardless of race, color, national origin, culture, education, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies. Fair Treatment means that no group of people, including racial, ethnic, or socioeconomic groups, should bear a disproportionate share of the negative environmental consequences resulting from industrial, municipal, and commercial operations or the execution of federal, state, local, and tribal environmental programs and policies. Meaningful Involvement means that: (1) potentially affected community residents have an appropriate opportunity to participate in decisions about a proposed activity that will affect their environment and/or health; (2) the public's contribution can influence the regulatory agency's decision; (3) the concerns of all participants involved will be

> considered in the decision-making process; and (4) the decision-makers seek out and facilitate the involvement of those potentially affected. (EPA, 2010c)

There is a large body of literature that explores environmental injustices including technological disasters, environmental catastrophes, environmental contamination, pollution events, and issues of air, water, and soil quality impacting public health (Johnston, 1994; Morello-Frosch et al., 2001; Neumann et al., 1998). It is also becoming clear that EJ communities not only carry a larger environmental burden, but are not as able to mitigate the potential effects of these exposures, often resulting in increased morbidity or mortality (Fothergill et al., 1999; Maantay and Maroko, 2008). This last point can fall under the term "environmental health justice," where even though different groups may be exposed to similar concentrations of a pollutant, certain groups may experience worse health outcomes as a result of heightened vulnerability factors.

One criticism of much of the GIS-based EJ work in the past decades is the common use of spatial coincidence (e.g. the pollution source being contained in the unit of measure such as tract or county) and proximity (e.g. distance buffers) as proxies for exposure. This limitation can be mitigated by the use of modeled pollution concentrations (dispersion modeling or land use regression) as described above. Other criticisms and questions are inherently more metaphysical, and therefore more difficult to answer. For instance, can racism truly be isolated as a factor? And, can intent to discriminate be proven? (Centner et al., 1996; Pulido, 1996). The interconnectedness and intertwining of race, ethnicity, income, and educational status often

present difficulty in identification of any potential racism. These complex social relationships are made even more convoluted when geographic differences between study areas are considered. For instance, some studies have found race to be the most important factor predicting environmental injustice (Burke, 1993; Pollock and Vittas, 1995), whereas others found income to be a better predictor (Been and Gupta, 1996; Bowen et al., 1995; Perlin et al., 1995). These differences in findings may be related to differences in the pollutant or source under investigation, the methods, the study area, or a combination of all of these. This idea is supported by Maantay in the oft-cited review paper which discussed geographic information science and environmental justice (Maantay, 2002). In the 13 papers that were reviewed, all of which used different pollution sources, methods, or study areas, environmental injustice was detected nearly unanimously, although the "type" and extent of the EJ issues varied widely.

In an expanded literature review in a research paper commissioned by the US EPA, Maantay et al. (2010) stated:

> Previous research demonstrates the existence of an uneven geographic distribution of environmental health hazards, and potentially disproportionate exposure to environmental risk in the U.S., resulting in racial/ethnic minority and lower-income communities bearing the highest burdens that, in turn, might contribute to the health disparities that have been noted extensively by public health officials and medical researchers. That these health and quality-of-life impacts are visited disproportionately on the most vulnerable populations, those least likely to be able to combat them effectively, render these impacts even more detrimental to the public's health, and the need for remedy even more urgent. The majority of reviewed studies show that both race/ethnicity and SES predicted a

> disproportionate spatial distribution of environmental burdens. When these two suites of variables were compared, SES variables pointed to more significant risks of exposure than race; however, race tended to be predictive of disproportion even when controlling for SES. (Maantay et al., 2010 p. 6).

The concept of "intent" is also difficult to show. For instance, was an environmentally burdensome land use placed in a community due to that community's inability to protect itself or did the community form around the already existing land use as a function of cheaper housing or less "vigilant" housing discrimination practices? If the latter were true the point can be made that there was no "intent" in the placement of the polluting facility. However, the argument can be made that minorities or those without the economic means may be severely limited and often constrained in their choice of where to live (Maantay, 2002). It has been argued that ultimately "intent" should not even be the issue. What is more important is that the problems are dealt with by policy-makers and related businesses as this inequity, regardless of intent, exists (Timney, 1998).

The socio-demographics which appear to mitigate or exacerbate exposure and related health outcomes must be examined with care. These racial, ethnic, economic, or educational characteristics are generally serving as proxies for other more subtle or 'unquantifiable' social phenomena. For instance, Black and Latino children are diagnosed with lead poisoning more often than White children. This is unlikely to be directly attributable to race or ethnicity, but

rather to nutritional deficiencies (e.g. iron and calcium) and exposure to lead from housing with lead-based paints or exposure to high traffic areas contaminated by lead-based gasoline (Maantay, 2002; Mahaffey, 1995). Another example is tuberculosis (TB), where racial and ethnic minorities show the highest rates, with the non-Hispanic black population having the largest proportion of cases (3,041 cases, 45.0% in 2003) (CDC, 2004). Some of the drivers of TB include overcrowded housing, residential segregation by race, and of course poverty. Classism and discrimination in housing as experienced by racial and ethnic minorities may represent many conditions that result in an increase in these communities' susceptibility to TB (Drucker et al., 1994; Stephens, 1996; WHO, 1996). As Maantay writes "… the context of race, rather than race itself, can be viewed as a risk factor" (Maantay, 2002 p. 163). As was mentioned before, the concept of race and ethnicity acting as a proxy for other factors can just as easily apply to related socio-demographic characteristics such as income and educational status.

This dissertation includes EJ analyses of chronic exposure with relation to race, ethnicity, income, and educational attainment. When the health outcome variable (heart failure hospitalization rate) is included in the analysis, foreign-born status is introduced as an independent variable. The concept behind this inclusion is that immigrants tend to be healthier than their racial and ethnic native-born counterparts, a phenomenon known as "the healthy immigrant effect". This has been shown in the United States (House et al.,1990; Stephen et al., 1994), Canada (Chen et al., 1996; Deri, 2003; McDonald, 2003; Perez, 2002), and Australia (Donovan et al., 1992). One generally accepted hypothesis as to why immigrants appear healthier is selective migration (Antecol and Bedard, 2006). This selective migration is likely due to a

number of factors. For instance, positive selection for healthier individuals either by his or her own choice or due to the manner in which screening is conducted by the U.S. immigration office is a likely explanation for some of the discrepancies between native-born and foreign-born health (Antecol and Bedard, 2006; Jasso et al., 2004; Marmot et al., 1984; McDonald, 2004). It is also possible that unhealthy immigrants are more likely to return to their countries of origin (Palloni and Arias, 2003) or that immigrants who experience economic success, which is strongly related to positive health outcomes, will be more likely to stay in the U.S. (Antecol and Bedard, 2006).

Other potential explanations include differences in diet and environmental exposures in the immigrants home countries versus the United States. For instance, it has been shown that immigrants tend to arrive in the U.S. with lower body mass index scores (BMI), but over time tend to adopt American behaviors (e.g. diet and physical activity) when they are exposed to the U.S. environment. This adoption often leads to increased BMI, obesity, and the subsequent inevitable health risks (Antecol and Bedard, 2006; Kasl and Berkman, 1983; Marmot and Syme, 1976; McDonald, 2004; Stephen et al., 1994). Selective migration and differences in behavioral risk factors and chronic environmental exposures successfully explain some of the reason for the "healthy immigrant effect," and as such appear important to include in regression analyses.

1.5 ENVIRONMENTAL HEALTH

There has been a significant amount of research that attempts to analyze the associations between air pollution and health outcomes (Ciccone et al., 1998; Dent et al., 1998; Dockery et al., 1993; Guo et al., 1999; Hruba et al., 2001; Hu et al., 2008; Ihrig et al., 1998; Jalaludin et al., 2006; Janssen et al., 2001; Kim, 2004; Kunzli et al., 2005; Nitta et al., 1993; Nyberg et al., 2000; Oosterlee et al., 1996; Poulstrup and Hansen, 2004; Ryan et al., 2006; Schwartz, 2004; Van Vliet et al., 1997; Venn et al., 2001; Wyst et al., 1993). These works have shown that even though there are many potential triggers, exacerbators, and causes for environmentally-linked diseases, such as behavioral risk factors (e.g. smoking, poor diet, lack of exercise) and genetic predisposition, air quality can play an important role. Although some diseases, such as asthma, have been relatively well studied using an ecological design, meaning that the study focuses on populations rather than individuals (Delfino et al., 2003; Dockery et al., 1993; Edwards et al., 1994; English et al., 1997; Friedman et al., 2001; Kunzli et al., 2005; Lin et al., 2002; Neutra, 1999; Schwartz et al., 1993; Wilkinson et al., 1999), other health outcomes such as cardiovascular disease and its relationship to chronic exposure to fine particulate matter have not been as fully examined with an intra-urban ecological framework.

It has been shown that there is a correlation between vascular inflammation and certain types of air pollution, most notably coarse (PM_{10}) and fine ($PM_{2.5}$) particulate matter (Brook et al., 2004; Chen et al., 2005; Henneberger et al., 2005; Miller et al., 2007; Pope et al., 2004; Ruckerl et al., 2007). The New York State Department of Environmental Conservation claims that breathing

air with high $PM_{2.5}$ concentrations can lead to premature death as well as many illnesses including increased respiratory symptoms and disease, chronic bronchitis, and decreased lung function (NYSDEC, 2010). The United States Environmental Protection Agency (EPA) states that inhalable particles, particularly $PM_{2.5}$, "have the greatest demonstrated impact on human health" (EPA, 2010a). Their small diameters allow them to penetrate deeply into the lungs and subsequently cause inflammation in the lungs, blood vessels, or heart (and possibly other organs as well). The EPA continues to state that studies have shown a significant association between $PM_{2.5}$ exposure and premature death from heart or lung disease. Fine particulate matter is linked to the aggravation of heart and lung diseases and related health effects (e.g. cardiovascular symptoms; cardiac arrhythmias; heart attacks; respiratory symptoms; asthma attacks; and bronchitis). The outcomes of these effects include increased hospital admissions and emergency room visits, economic loss due to absences from school or work, and a compromised quality of life (EPA, 2008).

Although relatively few intra-urban ecological study designs have been implemented to examine the association between heart failure and chronic $PM_{2.5}$ exposure, there have been many recent studies which confirm the general relationship (**Table 1-2**). Most of the recent work has focused on individual-level data, most often a patient, with exposure being estimated from nearby monitors. Of the ecological studies (population-based rather than individual based) most of the units of analyses are relatively large (e.g. counties or cities) focusing on general regional phenomena rather than small-area intra-urban variations. Notable exceptions were carried out by

Maheswaran et al. (2005a and 2005b) in Sheffield, England. Maheswaran found quantifiable associations between coarse particulate matter (PM_{10}) and coronary heart disease mortality, stroke mortality, coronary heart disease emergency hospital admission, and stroke emergency hospital admission using census enumeration districts (n= 1030) as the unit of analysis. For both studies, the pollution exposure was classified using quintiles, and the highest (mean concentration = 23.3 $\mu g/m^3$) and lowest (mean concentration = 16.0 $\mu g/m^3$) groups were compared. As of the writing of this dissertation, there are few other intra-urban ecological studies examining the relationship between heart disease and chronic exposure to fine particulate matter, particularly in U.S. cities.

Author	Date	Description	Location	Unit of Analysis	Findings
Zanobetti et al.	2010	Used EKG to measure HR variability vs. $PM_{2.5}$ measurements from monitors	Boston, MA	Individual	Exposure to PM decreases HR variability
Ostro et al.	2010	Tested which elements of $PM_{2.5}$ were most harmful using monitored data	California	Individual	Long-term exposure to $PM_{2.5}$ and some of its elements are positively associated with increased mortality
Puett et al.	2009	GIS-based general additive mixed model to estimate monthly exposure, longitudinal study	Northeast and Midwest U.S.	Individual	Increased mortality with $PM_{2.5}$ exposure
Hu	2009	Used satellite AOD to estimate $PM_{2.5}$ exposure	Eastern U.S.	County	Positive association with AOD and chronic coronary heart disease
Hu and Rao	2009	Used satellite AOD to estimate $PM_{2.5}$ exposure	Eastern U.S.	County	Positive association with AOD and chronic ischemic heart disease
Rosenthal et al.	2008	Compared EMS data on out-of-hospital cardiac arrest with monitored $PM_{2.5}$ concentrations	Indianapolis, IN	Individual	Positive association between short-term $PM_{2.5}$ exposure and cardiac arrest
Elliot et al.	2007	Compared association of black smoke and SO_2 to mortality (all-cause, cardiovascular, and respiratory) in Great Britain	Great Britain	Electoral wards	Black smoke and SO_2 exposure is positively associated with all-cause, cardiovascular, and respiratory mortality
Ruckerl et al.	2007	Compared MI patients inflammation levels with monitored pollution data	6 European Cities	Individual	Exposure to PM is positively associated with inflammation
Jalaludin et al.	2006	Compared monitored air pollution with cardiovascular disease emergency room visits	Sydney, Australia	Individual	Positive association between cardiovascular disease and increased short-term pollution exposure
Maheswaran et al.	2005a	Compared highest and lowest quintile of PM_{10} to mortality and hospitalization rates of coronary heart disease	Sheffield, UK	Census enumeration district	PM_{10} associated with greater risk of coronary heart disease (mortality and hospitalization)
Maheswaran et al.	2005b	Compared highest and lowest quintile of PM_{10} to mortality and hospitalization rates of stroke	Sheffield, UK	Census enumeration district	PM_{10} associated with greater risk of stroke (mortality and hospitalization)
Peters et al.	2004	Compared MI patients with traffic exposure	Augsburg, Germany	Individual	Transient exposure to traffic increases the risk of myocardial infarction
Clancy et. al.	2002	Compared mortality rates before and after pollution control measures	Dublin, Ireland	City	Pollution control (decrease in PM) associated with decreased mortality rates

Table 1-2: Some recent peer-reviewed studies analyzing the relationship between particulate matter and heart disease.

1.6 REGRESSION MODELS

Although it is not necessarily traditional to incorporate a discussion of statistics in the background section of a dissertation such as this, it seems reasonable to include an overview of the basic regression models that will be used throughout the paper as methodological considerations are one of the central foci of this work. This section contains a brief review of the three regression models (ordinary least squares, spatial autoregressive model, and geographically weighted regression) followed by a worked hypothetical example designed to illustrate some of the main differences among the three models.

1.6.1 ORDINARY LEAST SQUARES

Ordinary least squares regression (OLS), put simply, is a technique used to describe a linear relationship between a dependent variable (e.g., disease rate) and one or more independent variables (e.g., pollution exposure and socio-demographic characteristics). The OLS mathematically determines the line that best fits the data (i.e. explains the variance in the dependent variable by minimizing the square of the errors – hence "least squares") and provides diagnostics explaining the strength and nature of the relationships and the model (e.g. R^2 represents the percent of the variance explained). The OLS assigns a constant and parameter coefficients (β) to create a general linear equation (**Equation 1-1**). The constant represents where the regression line would intercept the y-axis if the values for all the independent variables were zero. The parameter coefficients, when left un-standardized, show the amount that the dependent variable would change for each 1 unit change of the corresponding independent variable (Hamilton, 1990).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \qquad \text{Eq. 1-1}$$

where:
Y = the dependent variable
β_0 = the constant (intercept)
β_n = the parameter coefficient for each independent variable (1 to n)
X_n = the independent variable (1 to n)
ε = the error term

OLS regressions require certain assumptions, which include that the relationship is linear, there are no outliers overly influencing the model, the model is properly specified (e.g. no important independent variables are excluded), observations are independent, and that the independent variables are not excessively correlated with one another. There are also assumptions regarding the error terms in the regression including that they have a mean of zero, are normally distributed, have a constant variance (homoscedastic), and not autocorrelated (including spatial autocorrelation), (Hamilton, 1990). In this dissertation, OLS regressions were most often performed in the SPSS Statistics 17.0 software package.

1.6.2 SPATIAL AUTOREGRESSIVE MODELS

The assumptions regarding independence of the error terms are often not met when working with ecological data such as neighborhood socio-demographics. When values in one location are related to those at neighboring locations, there is a spatial dependence. The two commonly cited causes of spatial dependence are (1) measurement error (e.g. the boundaries of the geographic

units of analysis don't match the underlying process) and (2) the phenomenon or process demonstrates truly spatial characteristics (e.g. diffusion, hierarchies, etc) (Anselin, 2004).

There are two primary types of spatial dependence which present problems in regression analysis: (1) spatial error - spatially correlated error terms where the errors from an OLS are correlated in space often due to an omitted, spatially correlated covariate, and (2) spatial lag – where the dependent variable is influenced by independent variables in neighboring geographic units which often represents a truly spatial process (e.g. diffusion). The spatial error scenario violates the OLS assumption of uncorrelated error terms, whereas the spatial lag scenario violates both the uncorrelated error term and the independent observations assumptions (Anselin, 2004; Anselin and Bera, 1998).

A spatial autoregressive model (SAR) is similar to an OLS, except that it is specifically designed to control for spatially correlated data by introducing a new variable (lag or error) and uses a Maximum Likelihood approach. Within the OpenGeoda software package designed by Luc Anselin, both the lag and error regression models can be performed after a spatial weights matrix is defined. A weights matrix imposes a geographic neighborhood on the data to enable the quantification of potential spatial autocorrelation. In OpenGeoda there are two basic categories of weight matrices: contiguity and distance. Contiguity-based weights can be rook (border is shared) or queen (border or vertex is shared) with the option to choose the number of "orders" to include (1st order means that only one ring of "neighbors" will be used). Distance-based methods include the simple distance threshold (similar to a buffer) and k-nearest neighbors

(KNN). With KNN, the researcher can define the number of neighbors (k) for a location. This insures that all locations will have the same number of neighbors – although it also means that each location will have a different amount of area associated with its "neighborhood" (GeodaCenter, 2010).

When an OLS is run in OpenGeoda, a number of tests are conducted to assess spatial dependence, including the simple LaGrange Multiplier (LM) test for a missing spatially lagged dependent variable; the simple LM test for error dependence; and the robust versions of the LM(lag) and LM(error) tests. These robust LM tests examine dependence in one scenario while accounting for the possible presence of spatial dependence in the other scenario (e.g., robust LM(lag) tests for significance of a lagged variable while accounting for spatially correlated error terms). If one of the simple LM tests is significant and the other is not, then the choice of SAR model is simply the result of the significant LM test. However, if both simple LM tests are significant, then the robust LM tests are used to determine which SAR to use (Anselin, 2003; Anselin; 2005, Anselin et al., 2004; Gibson and Olivia, 2010).

The spatial lag model uses a lagged version of the independent variable (WY) on the right-hand side of the equation. This lagged term is essentially the average of the 'Y' values in the neighboring units of analysis based on the defined weights matrix. The equation is written below (**Equation 1-2**)

$$\mathbf{Y = \beta_0 + \rho WY + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon} \qquad \textbf{Eq. 1-2}$$

where:
Y = the dependent variable
β_0 = the intercept
W = an n x n spatial weights matrix
ρ = the spatial autoregressive coefficient
X = the independent variable (1 to n)
β_n = the parameter coefficient (1 to n)
ε = the error term

The inclusion of the lagged variable (WY) and its parameter coefficient rho (ρ) improve the reliability of regressions where the phenomena meet the criteria for this type of spatial dependence (Anselin and Bera, 1998; Gibson and Olivia, 2010).

The spatial error model is expressed by a different equation that uses a spatially weighted error term and can be defined as **Equation 1-3** below. Note that the first line is a standard OLS equation, but the re-definition for the error term (ε) involves a spatially weighted version of the OLS residuals as well as its coefficient (λ) and a new random error term (μ).

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon} \qquad \textbf{Eq. 1-3}$$
$$\mathbf{\varepsilon = \lambda W \varepsilon + \mu}$$

where:
ε = the error term
λ = the spatial autoregressive coefficient
W = an n x n spatial weights matrix
μ = a vector of errors that are assumed to be independently and identically distributed

In the spatial error model, the error term for one observation is dependent upon the weighted average of the error terms for neighboring observations. The inclusion of this spatial error term,

along with its parameter coefficient lambda (λ), is appropriate for relationships that meet the criteria for this type of spatial dependence (Anselin and Bera, 1998; Anselin et al., 2004).

1.6.3 GEOGRAPHICALLY WEIGHTED REGRESSION

Geographically weighted regression (GWR) is the last regression technique that will be used in this dissertation. GWR, a technique developed by Stuart Fotheringham, Chris Brunsdon, and Martin Charlton, quantifies locally varying relationships among data, rather than computing a global relationship as OLS does. Instead of calculating one set of parameter estimates based on one regression and resulting in one set of summary statistics (i.e. global), GWR performs many local regressions, each of which is influenced by the surrounding data, resulting in a set of summary statistics for each regression point. Each local regression utilizes surrounding data, with nearby data being weighted more heavily than distant data (e.g., distance decay). In this way, GWR is able to show local variations in the relationships and is able to account for potential spatial non-stationarity (local variation in parameter estimates). Locally varying relationships may suggest a number of things, including possible model misspecification, sampling variation, or simply a relationship that intrinsically varies over space (Fotheringham et al., 2002).

Similar to the assignment of a spatial weights matrix in SAR modeling, a kernel bandwidth must be specified for the model. The two most common choices are fixed or adaptive kernels. The fixed kernel incorporates sample values within the distance threshold, with distal samples being weighted less heavily than proximal ones. The adaptive kernel works in a similar way to k-

nearest neighbors, selecting a certain number of samples per local regression, still weighting the near samples more heavily than distant ones. As such, the adaptive kernel is able to "grow" when the samples are sparse, and "shrink" when there is a high density of sample points. Unlike SAR regressions in OpenGeoda, when GWR regressions are run in the GWR3 software package, the number of nearest neighbors or the value of the bandwidth can be determined empirically using an iterative process which is designed to minimize the Akaike Information Criterion (AIC) – a diagnostic statistic that describes the performance (fit) of the model (Fotheringham et al., 2002). The generic geographically weighted regression equation is written below (**Equation 1-4**)

$$\mathbf{Y}_{(u,v)} = \boldsymbol{\beta}_{\mathbf{0}(u,v)} + \boldsymbol{\beta}_{\mathbf{1}(u,v)}\mathbf{X}_{\mathbf{1}(u,v)} + \ldots + \boldsymbol{\beta}_{\mathbf{n}(u,v)}\mathbf{X}_{\mathbf{n}(u,v)} + \boldsymbol{\varepsilon}_{(u,v)} \qquad \textbf{Eq. 1-4}$$

where:
$\beta_{0\,(u,v)}$ = the intercept at location (u, v)
$Y_{(u,v)}$ = the dependent variable at location (u, v)
X = the independent variable (1 to n) at location (u, v)
$\beta_{(u,v)}$ = the parameter coefficient (1 to n) at location (u, v)
ε = the error term at location (u, v)

1.6.4 WORKED HYPOTHETICAL EXAMPLE FOR REGRESSION COMPARISON

It may be easiest to see the difference, and utility, of different regression models by using a hypothetical example with fabricated data. To that end, a grid of 900 polygons was created. Each cell, which can be thought of as a census tract, was given a value for "pollution exposure" and "social vulnerability" – the two independent variables that will be used to create the "disease rate" dependent variable. To accomplish this, the polygons were divided into two equal groups, with the northern group being designated as "socially vulnerable" and the southern group designated as "socially non-vulnerable". This variable is meant to represent socio-demographic

characteristics that are positively associated with increased disease rates such as low educational attainment, poverty, and racial/ethnic minority status. The polygons in the "socially vulnerable" area were calculated as having a social vulnerability score (SV) with a mean of 80 and standard deviation of 5 (from a normal distribution). The "non-vulnerable population" polygons were assigned SV scores with a mean of 20 and a standard deviation of 5 (again, normally distributed). This results in a random distribution of high SV scores in the northern section, and a random distribution of low SV scores in the southern section. Pollution exposure was calculated as a distance from a point (representing a pollution source such as a smoke stack), then split into five "distance zones" (D) given scores of one through five (**Figure 1-1**). The concept is that as distance from the point increases, exposure will decrease.

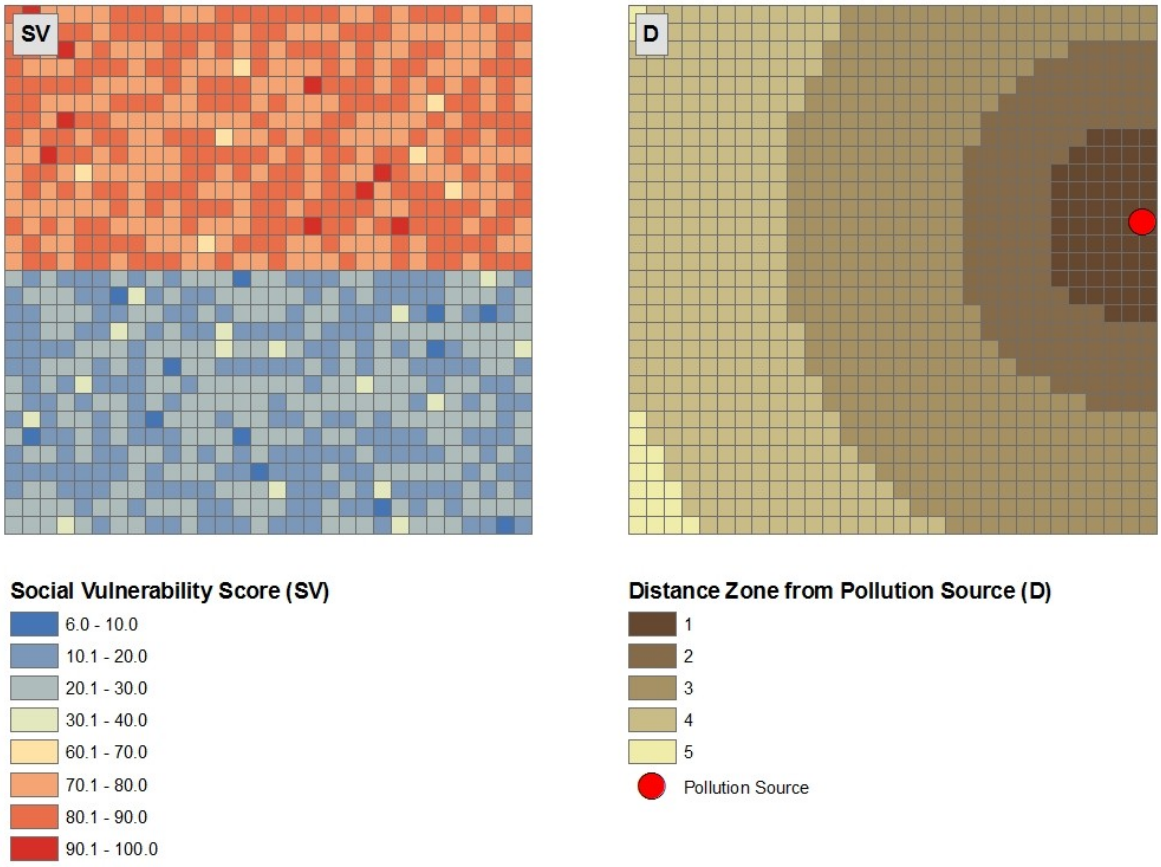


Figure 1-1: Social vulnerability score (SV) and distance zone categories (D) of hypothetical data.

To calculate the disease rate, which will serve as the dependent variable in the regressions, the following general equation was used (**Equation 1-5)**. This health outcome variable, named "Disease Rate 1" (DR_1) is defined by two variables: social vulnerability (SV), and the inverse distance squared of the pollution zone (IDS). They were given coefficients that would give SV more influence than the pollution zone with a disease rate range of approximately 10 through 25. An error term, with a mean of 0 and standard deviation of 1 was also added to increase the variability of the DR_1 term (**Equation 1-6**, **Figure 1-2)**.

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon} \qquad \textbf{Eq. 1-5}$$

where:
Y = the dependent variable, DR_1
β_0 = the intercept
β_n = the parameter coefficient of social vulnerability (β_1) or pollution exposure (β_2)
X_n = the variables for social vulnerability (X_1) or pollution exposure (X_2)
ε = the error term (mean of 0 and standard deviation of 1 from a normal distribution)

$$\mathbf{DR_1 = 10 + .1*SV + 5*\ IDS + \varepsilon} \qquad \textbf{Eq. 1-6}$$

where:
DR_1 = disease rate 1
SV = social vulnerability
$IDS = 1/d^2$, where d = the impact zone distance designation for pollution exposure

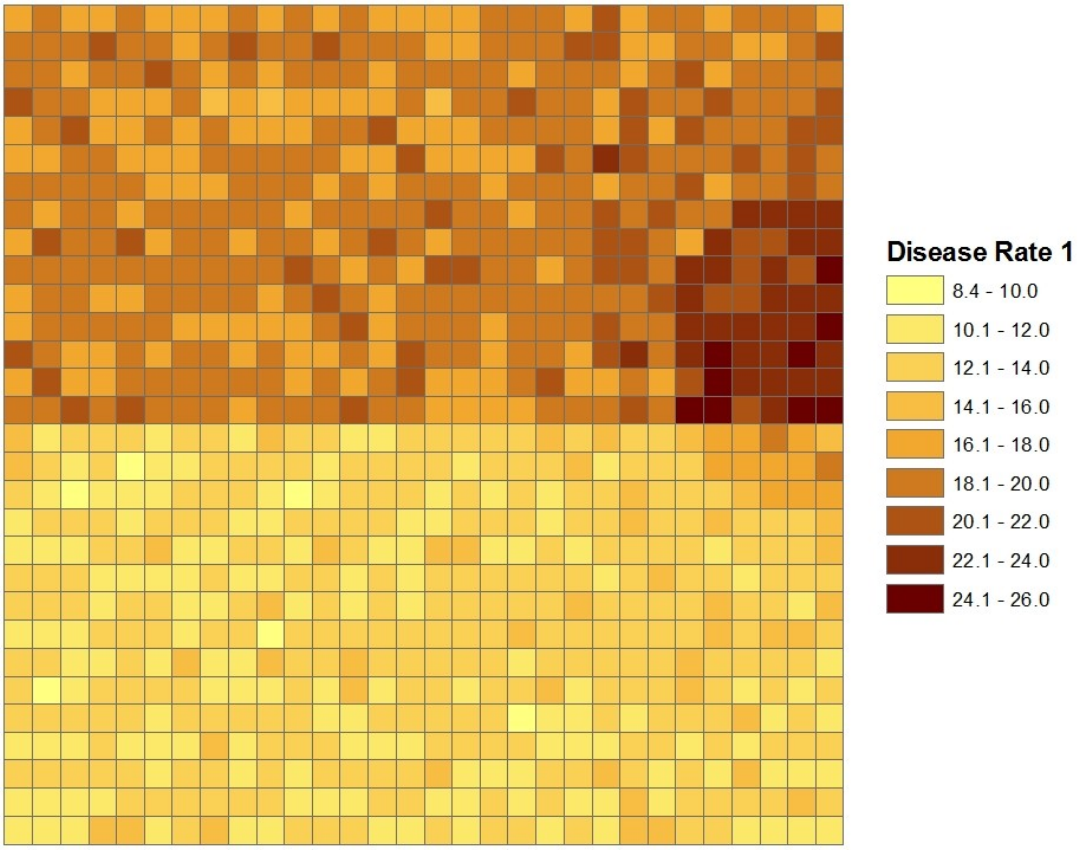


Figure 1-2: Disease Rate 1 (DR_1) of hypothetical dataset. DR1 = 10 + .1*SV + 5*IDS + ε

As expected, when these data are analyzed with an ordinary least squares regression, the relationships between social vulnerability and the inverse squared distance to the exposure source (IDS) are well captured (R^2 = .92, AIC = 2573). The parameter estimates were quite close to those used in the creation of DR_1, with values of .102 and 4.94 for SV and IDS, respectively (by definition, these values are .1 and 5).

When the data were put into a spatial autoregressive model, there was no significance in the Lagrange Multiplier diagnostics. This suggests, correctly, that a SAR will not improve the model since there is no significant spatial autocorrelation. When a GWR was used, the bandwidth converged at a local sample size of 876 (out of 900). This means that the Akaike Information Criterion (AIC) was minimized at 876 samples. Since it is so close to the full dataset of 900, the GWR suggests a 'global' relationship. In other words, the GWR is trying to make itself behave like an OLS. This is supported by the lack of statistically significant spatial variability in the GWR-derived parameter estimates using a Monte Carlo test. The R^2 and AIC values are nearly identical for both models. By looking at the bandwidth selection, model diagnostics, and parameter estimates, it is clear that an OLS is the proper model for these data (**Table 1-3**). It can be useful to look at the spatial distribution of the OLS and GWR regression residuals as well as the t-values and local R^2 of the GWR model (**Figure 1-3**). The residuals for both OLS and GWR models appear randomly distributed. Note that the lack of spatial patterns in the t-values and local R^2 in the GWR model are due to the 'global' nature of the model (i.e. the GWR is trying turn itself into an OLS).

	R^2	AIC	CONSTANT	SV	IDS
OLS	0.92197	2573.85	9.91	0.10165	4.93668
SAR	--	--	--	--	--
GWR	0.92215	2576.65	9.87 to 9.96	0.101 to 0.103	4.78 to 5.01

Table 1-3: Model comparison where the dependent variable = Disease Rate 1 (DR_1) and independent variables = social vulnerability (SV) and the inverse square of the distance zone (IDS). All variables are significant ($p<.001$) in the OLS model. The SAR model was not run due to insignificant Lagrange Multiplier diagnostics. The GWR model did not show significant spatial variability in any of the parameter estimates and converged at a local sample size of 876 out of 900.

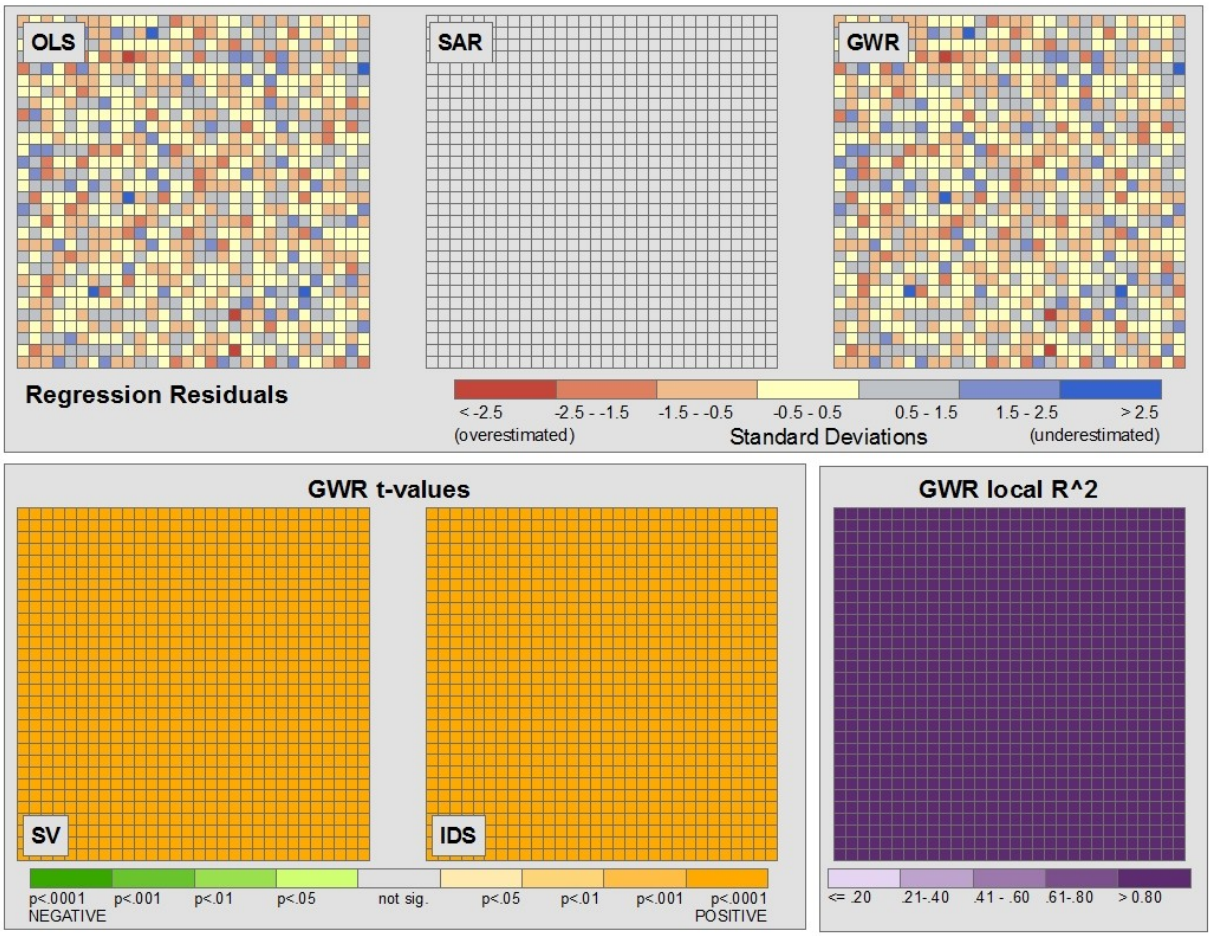


Figure 1-3: Regression model comparison of Disease Rate 1 (DR_1) vs. social vulnerability (SV) and the inverse square of the distance zone (IDS). No SAR model was run due to the lack of significance of the LM diagnostics. The lack of significant spatial variability in the GWR model outputs, and 'global' nature of the relationship, can be seen in the bottom three maps.

The input data can be modified in order to show a more complex scenario. For instance, what if there was a very active and strong community within the northern section of the study area that enabled the residents to not suffer disproportionately from pollution exposure due to "social vulnerability" when compared with the "non-vulnerable" populations? How would OLS, SAR,

and GWR respond to this somewhat misspecified model? To examine this, an area within the northern section was identified as a "strong community" and a new variable representing this state (C) was created and assigned a value of "0" if it is "strong" and "1" for all other locations (**Figure 1-4**). A new disease rate term, DR_2, was defined as **Equation 1-7**.

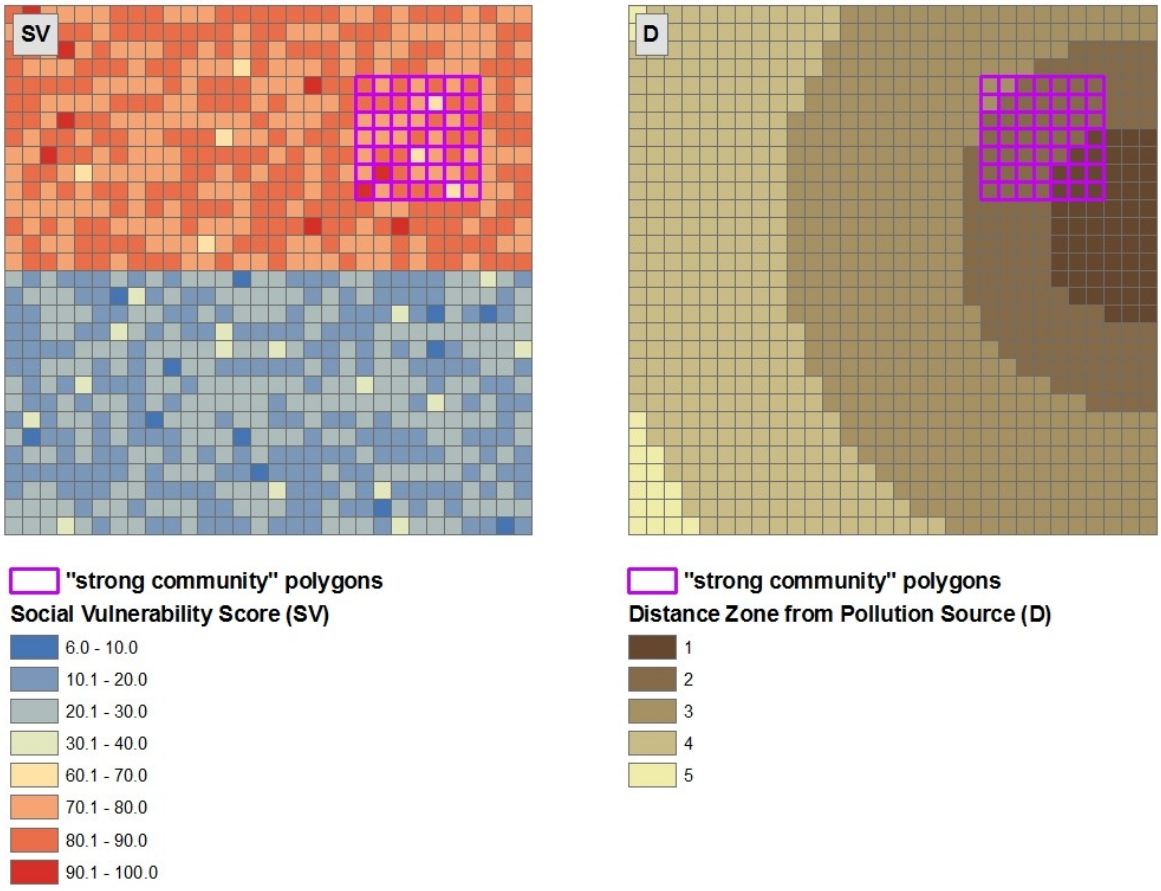


Figure 1-4: "Strong Community" polygons superimposed above social vulnerability scores (SV) and distance zone from pollution source scores (D).

$$\mathbf{DR_2 = 10 + C^{*}.1^{*}SV + 5^{*}\ IDS + \varepsilon} \qquad \textbf{Eq. 1-7}$$

Where:

DR2 = Disease Rate 2

C = a dummy/interaction variable representing a "strong community" where "0" = a strong community and "1" = all other communities

The "C" dummy variable will change the SV score to zero in the Disease Rate 2 (DR_2) calculations. This is meant to simulate the effect of a phenomenon which is either omitted from a regression (misspecification) or data which is, by its very nature, unquantifiable (**Figure 1-5**). The regression equations, however, do not include this interaction term.

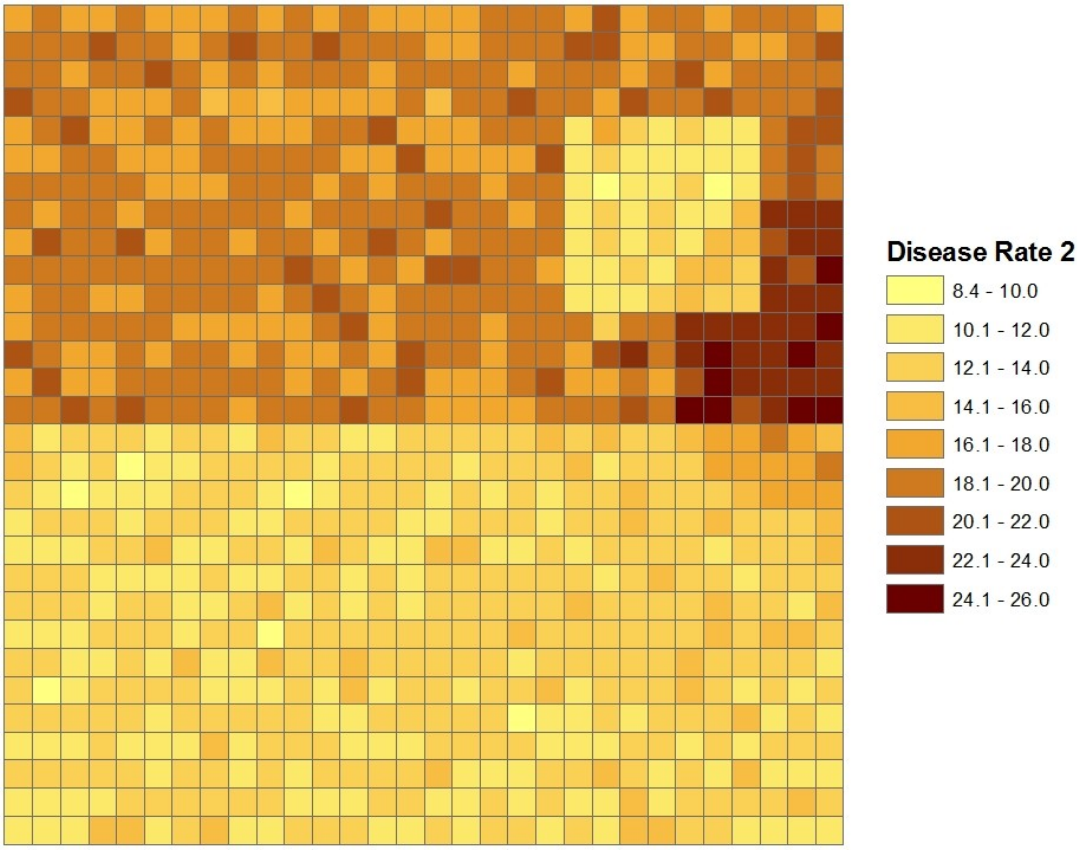


Figure 1-5: Disease Rate 2 (DR_2) of hypothetical dataset. $DR_2 = 10 + C*.1*SV + 5*IDS + \varepsilon$.

When an OLS is used to assess the association of DR_2 with SV and IDS, the model has an R^2 of .68 and an AIC of 3818 – both of which suggest a worse-fitting model than the DR_1 model. The constant's coefficient is slightly overestimated (10.36 rather than 10) while the coefficients of SV and IDS are underestimated (.089 and 3.47 rather than .1 and 5, respectively). Lagrange Multiplier diagnostics based on first order queen's contiguity pointed to using the spatial error model SAR rather than the spatial lag model. The SAR performed quite well, with a high R^2 (.853), low AIC (3243), and comparatively accurate parameter estimates. It also includes a spatial term (Lambda), which accounts for some of the effect of the model misspecification (there are highly autocorrelated error terms around the "strong community" polygons). The GWR is very similar to the SAR in terms of model diagnostics, although the R^2 is slightly lower (.845) and AIC slightly higher (3308). This suggests that the GWR does not fit the data quite as well as the SAR, but it does offer information that the global estimates do not. For this dataset, the GWR converged at a local sample size of only 84 out of 900, indicating an extremely local model that uses less than 10% of the data in each local regression. The local nature of the GWR allows the parameter estimates (and t-values, errors, etc.), all of which demonstrated significant spatial variation ($p<.001$), to change over space. The most striking effect of this is the wide range and inverted directionality (negative to positive) of parameter estimates associated with the IDS score (**Table 1-4**).

	R^2	AIC	CONSTANT	SV	IDS	Lambda
OLS	0.68334	3818.07	10.36	0.08920	3.46808	--
SAR	0.85338	3243.06	10.02	0.09399	3.92835	0.81013
GWR	0.84482	3308.16	-11.7 to 23.8	-0.05 to 0.29	-102.2 to 74.9	--

Table 1-4: Model comparison for dependent variable = Disease Rate 2 (DR_2) and independent variables = social vulnerability (SV) and the inverse square of the distance zone (IDS). All variables are significant ($p<.001$) in the OLS and SAR models. The GWR model showed significant spatial variability in all of the parameter estimates and converged at a local sample size of 84 out of 900.

Again, the spatial distribution of the residuals for the three models and the parameter estimates and local R^2 for the GWR can be viewed cartographically (**Figure 1-6**). The OLS and SAR residuals are distributed nearly identically, although the magnitude of the error is lower in the SAR. The GWR residuals, while still clustered, have a different pattern as the model varies locally to attempt to account for the 'odd' behavior of the "strong community" polygons. It is interesting to observe the patterns of significance, as represented by each parameter's t-value, in the GWR maps. The social vulnerability variable maintains high level of significance through the middle of the study area from east to west, where there are large amounts of variation between the higher SV values to the north and lower SV values to the south. Since the GWR has become quite local (sample size 84), it is unable to detect the same strong relationships when the number of local samples does not extend from the "socially vulnerable" area to the "non-vulnerable" area (or vice-a-versa). Similarly, the pollution variable (IDS) is only significant in the expected direction (positive) where its influence is the strongest (nearest to the pollution source), and loses significance as the distance increases and local variation decreases. Interestingly, around the "strong community" polygons, the GWR changes the directionality of the IDS variable rather

than the SV variable to account for the unexpected DR_2 values. Although these results do not match "reality" as has been defined in this example, they do clearly bring our attention to these polygons and alert us to the spatial heterogeneity of the relationship. It is also illuminating, and not unexpected, to note that the highest local R^2 values are either where the pollution (IDS) is highest or where there are high amounts of variation in the social vulnerability score (SV). Since the GWR is detecting very local relationships, the associations in areas with small variation have demonstrated weaker correlations (i.e., lower local R^2).

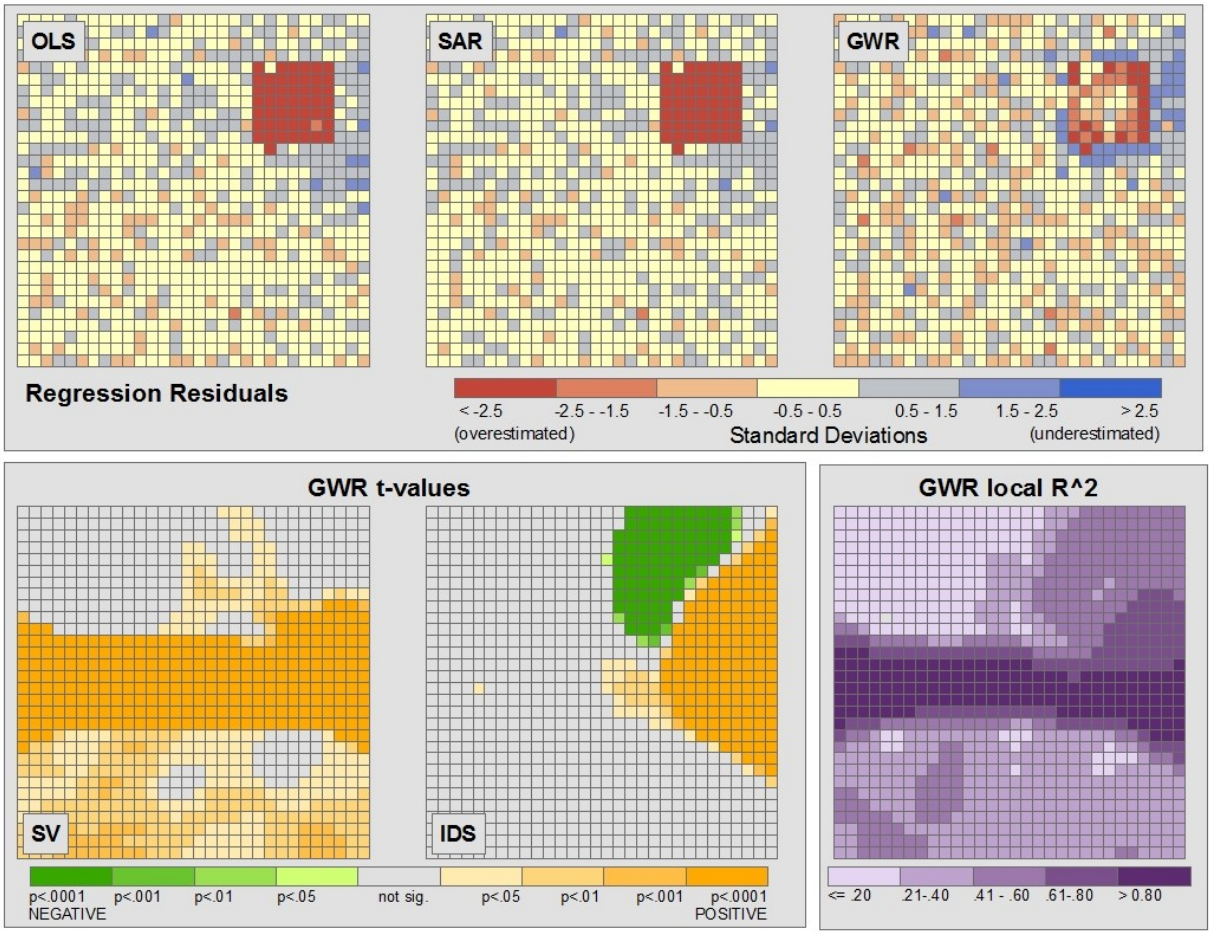


Figure 1-6: Regression residuals of OLS, SAR, and GWR predicting DR_2 vs. SV (social vulnerability) and IDS (inverse squared distance zone from pollution source) (top left, top center, and top right, respectively). GWR t-values shown with levels of significance, where green areas represent negative associations, orange areas represent positive associations, and gray areas represent no significant association (bottom left and bottom center). GWR local R^2 values show areas of better performance in darker shades (bottom right).

The last hypothetical scenario that will be explored is the forced inclusion of spatial autocorrelation. This was done by creating a third Disease Rate outcome variable (DR_3) defined as the first disease rate (DR_1) plus a spatial lag term (**Equation 1-8**). The lagged term (DR_{LAG}) was calculated using first order queen's contiguity on the DR_1 variable, then dividing by 10. This

term, when added to DR_1, results in a modified version (DR_3) whose value at each polygon is influenced by the values of the surrounding polygons (**Figure 1-7**).

$$\mathbf{DR_3 = DR_1 + .1 * DR_{LAG}} \qquad \textbf{Eq. 1-8}$$

Where:

DR_3 = Disease Rate 3

DR_{LAG} = the lagged version of DR_1 using first order queen's contiguity

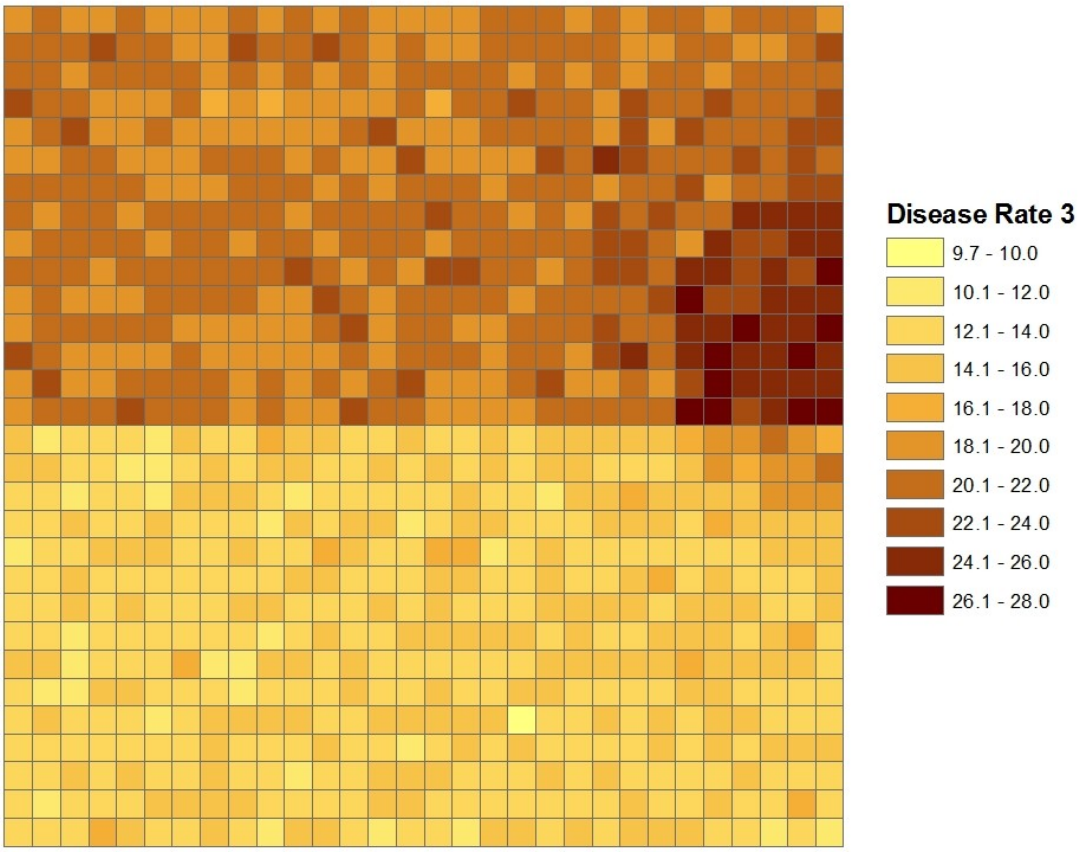


Figure 1-7: Disease Rate 3 (DR_3) of hypothetical dataset. $DR_3 = DR_1 + .1 * DR_{LAG}$

When the spatially lagged dependent variable (DR_3) is used in an OLS, the results are predictably similar to the DR_1 model. The R^2 actually increased from .92 for the DR_1 regression to .93 in the DR_3 regression. This slight increase in R^2, however, is foiled by an increase in the AIC as well. The coefficients in the DR_3 OLS model tend to be overestimated for both the variables and the constant. Lagrange multiplier diagnostics appropriately suggest the use of the spatial lag model for this scenario. This adds the W_DR_3 variable, which is the spatially lagged (weighted) version of the dependent variable (DR_3). The coefficients determined by the SAR are very close to "reality", and the R^2 and AIC support the use of SAR over OLS. The GWR selected a local sample size of 559 out of 900 which suggests a relatively global relationship similar, although not as extreme, as the DR_1 model (**Table 1-5**). Spatial variability was not significant for the intercept, but was for SV ($p<.05$) and IDS ($p<.001$). Again, these data can be looked at cartographically in order to fully understand the distributions of residuals (which all appear uncorrelated) and GWR outputs (**Figure 1-8**).

	R^2	AIC	CONSTANT	SV	IDS	W_DR_3
OLS	0.93316	2583.54	10.95	0.11108	5.39748	--
SAR	0.93407	2574.55	9.74	0.10010	4.86134	0.10600
GWR	0.93380	2585.49	10.83 to 11.44	.105 to 0.115	2.911 to 5.978	--

Table 1-5: Model comparison for dependent variable = Disease Rate 3 (DR3) and independent variables = social vulnerability (SV) and the inverse square of the distance zone (IDS). All variables are significant ($p<.001$) in the OLS and SAR models. The GWR model showed significant spatial variability in SV and IDS but not in the constant and converged at a local sample size of 559 out of 900.

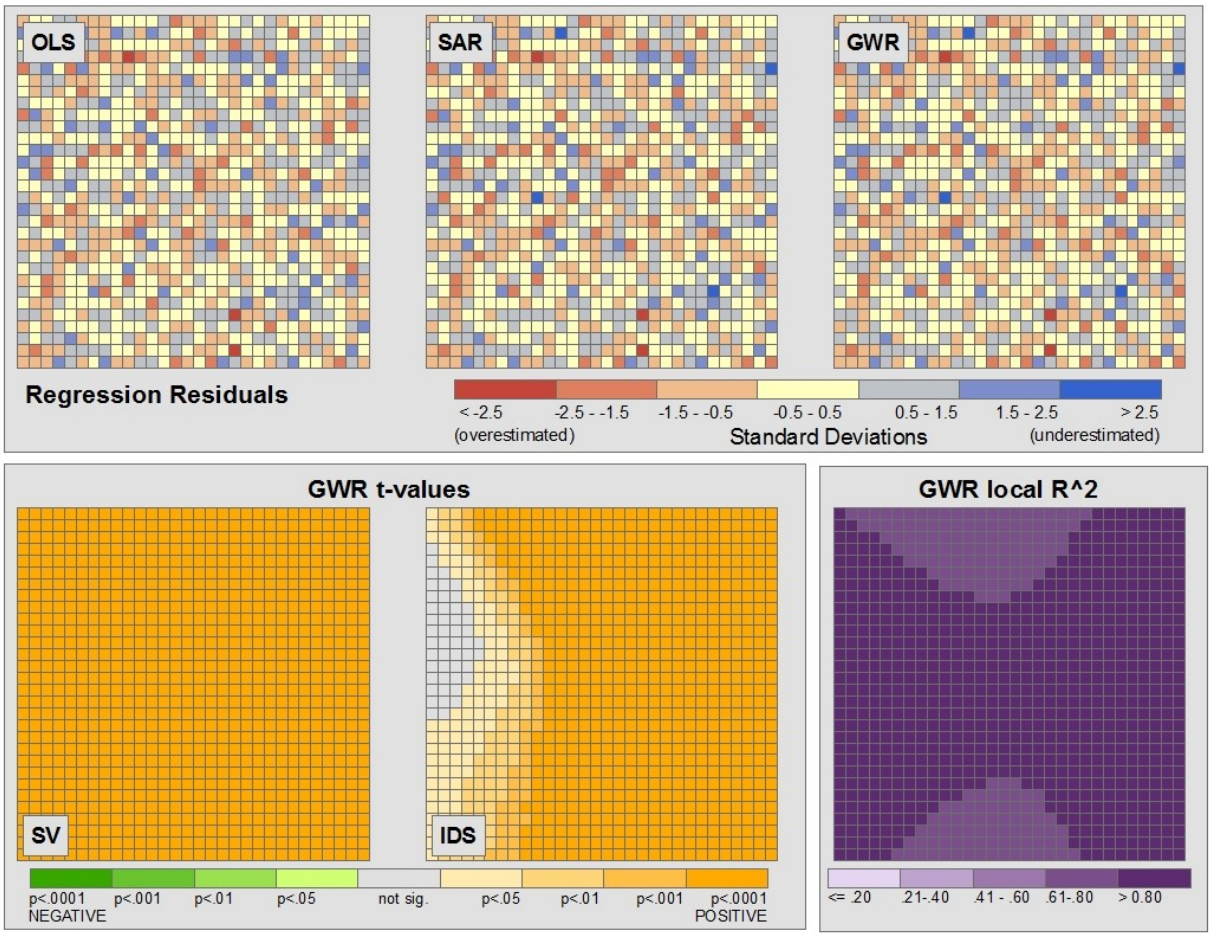


Figure 1-8: Regression residuals of OLS, SAR, and GWR predicting DR_3 vs. SV (social vulnerability) and IDS (inverse squared distance zone from pollution source) (top left, top center, and top right, respectively). GWR t-values shown with levels of significance, where green areas represent negative associations, orange areas represent positive associations, and gray areas represent no significant association (bottom left and bottom center). GWR local R^2 values show areas of better performance in darker shades (bottom right).

The three models can be looked at simultaneously to aid in understanding some of the different effects of misspecification and spatial autocorrelation on the three regression types (**Table 1-6**).

DEPENDENT VARIABLE	REGRESSION TYPE	DIAGNOSTICS		COEFFICIENTS				Sample Size[##]
		R^2	AIC	CONSTANT	SV	IDS	Spatial Term[#]	
DR_1	OLS	0.922	2574	9.91*	0.102*	4.94*	--	900
	SAR	--	--	--	--	--	--	--
	GWR	0.922	2577	9.9 to 10.0	.101 to .103	4.8 to 5.0	--	876
DR_2	OLS	0.683	3818	10.36*	0.089*	3.47*	--	900
	SAR	0.853	3243	10.02*	0.094*	3.93*	0.810*	900
	GWR	0.845	3308	-11.7 to 23.8'''	-.050 to .293'''	-102.2 to 74.9'''	--	84
DR_3	OLS	0.933	2584	10.95*	0.111*	5.40*	--	900
	SAR	0.934	2575	9.74*	0.100*	4.86*	0.106*	900
	GWR	0.934	2585	10.8 to 11.4	.105 to 0.115'	2.9 to 6.0'''	--	559

Table 1-6: Comparison of three dependent variables with OLS, SAR, and GWR. The coefficients, by definition, should be 10 for the intercept (constant), .1 for SV, and 5 for IDS.
*significant at p<.001;'sig. spatial variability (p<.05); '''sig. spatial variability (p<.001);
[#] spatial term in the DR_2 spatial error SAR is Lambda, in the DR_3 spatial lag SAR it is W_DR_3
[##]sample size represents "n" for OLS and SAR and represents the local sample size for GWR

It seems apparent that to gain a full understanding of the associations between or among variables when the unit of analysis is geographic by nature (e.g., an ecological study), it can be useful and productive to perform various types of regressions. In this way, it is possible to observe how the data behave when looked at globally (OLS or SAR), spatially (SAR or GWR), or locally (GWR). By comparing the various regression outputs to one another, characteristics and idiosyncrasies of the correlations that would otherwise be obscured may become apparent. For instance, the anomaly of the "strong community" in DR_2 would not have been apparent without using GWR, or at least mapping the residuals of the OLS/SAR. By using an assortment of regression techniques, it is not only possible to estimate more accurate parameter coefficients,

but also to have more confidence in one's results if the outputs support one another such as the DR_1 and DR_2 models. It must be remembered, however, that the three hypothetical examples above by no means cover all possible uses of these statistical techniques nor do they comprehensively describe potential outcomes or interpretations, but rather show discrete possibilities for their utilization in the context of this dissertation.

1.7 BACKGROUND CHAPTER CONCLUSORY STATEMENT

This background chapter has introduced the goals and hypotheses of this dissertation. The nature of fine particulate matter and its relationship to environmental justice and environmental health have also been described in general terms. The regression models (ordinary least squares, spatial autoregressive models, and geographically weighted regressions) that are used in this work were also explained and demonstrated using a worked hypothetical example. The next chapter, Methods, delves into the source data, data preparation, and exposure estimation techniques.

2 METHODS

As one of the objectives of this dissertation is to explore ways to estimate exposure to fine particulate matter and analyze those estimates vis-à-vis environmental justice and health outcomes, the methods section is central to this study. There are two sections in this chapter: Data (**2.1**) and Exposure Estimation (**2.2**). The data section is subdivided into each main data type, which include Population Data (**2.1.1**), Health Data (**2.1.2**), and Pollution Data (**2.1.3**)

The exposure estimation section is broken down into three parts: Proximity Analysis (**2.2.1**), Air Dispersion Modeling (**2.2.2**), and Land Use Regression Modeling (**2.2.3**). Each of these sections works through the estimation of $PM_{2.5}$ exposure or concentration in New York City. It is important to note that even though each of these techniques is designed to examine the same pollutant, they are in essence measuring different aspects of fine particulate matter. Proximity analysis, for instance, is best used to estimate exposure to pollutants due to physical closeness to specific facilities or locations. As such, it is more of a proxy for exposure that is calculated by estimating the number and socio-demographics of populations proximal to environmentally burdensome land uses. Conversely, air dispersion modeling is designed to get more precise estimates of actual pollution concentrations from known sources. The results of this modeling are influenced by the meteorological and physical environments and enable a more "true" analysis of exposure to, or effects of, a specific pollutant originating from specific sources. The last method, land use regression, differs from the previous two in that it is designed to estimate total ambient pollution rather than the pollutant(s) emanating from any specific source. Therefore this

technique is best suited for the analysis of the effects of total pollution burden on health outcomes as it does not intrinsically rely on specific facilities or emissions.

2.1 DATA

This section of the dissertation contains information and some exploration of the source datasets that drive the overall study. The quality and understanding of these data are extremely important. Without a full knowledge of the information in terms of its content and spatial distributions, it would be very difficult to understand how to design hypotheses, create models, or assign any implications to the model outputs or study results.

This section is arranged in three parts. The first describes the population data used. These include socio-demographics from the U.S. Census Bureau (2000) as well as the disaggregation technique known as the Cadastral-based Expert Dasymetric System (CEDS). The second part describes the health outcome data utilized in the study. It includes heart failure hospitalization data, originally compiled by the Statewide Planning and Research Cooperative System (SPARCS), but supplied to me via two sources: Emerging Health Information Technologies and Infoshare.org. The third part describes the pollution data. These include information regarding $PM_{2.5}$ production from major stationary point sources via the National Emissions Inventory (NEI), annual average daily traffic data from the New York State Department of Transportation (NYSDOT), information

about the US EPA's monitoring stations in NYC, and aerosol optical depth from the Moderate Resolution Imaging Spectroradiometer (MODIS) space-borne sensor.

2.1.1 POPULATION DATA

Population and demographic data is the foundation for the vast majority of ecological studies. These data comprise the denominators of disease rates, underlying "exposed" or "at-risk" populations, as well as specific sub-populations (e.g. race, ethnicity, income, and education) which can be used for environmental justice analyses and predictors in health outcome regressions. As such, their accuracy and precision are of paramount importance.

In the United States, the most common source for this type of demographic data is the U.S. Bureau of the Census. The decennial census data are provided at "census aggregations" – geographic units used to enumerate the population(s). The three smallest units (i.e. units with the highest spatial resolutions) are hierarchically arranged, or nested, and include the census block, census block group, and census tract (from smallest area to large). In NYC, the census block (n = 36,592) is often similar to a "city block" or "street block," but data at this spatial resolution are limited to information for redistricting for the protection of confidentiality of respondents. More detailed information from both the 100% count (short form) and the 20% sample (long form), including more detailed socioeconomic data are released at the block group level (n=5734 in New York City). Unfortunately, many health studies are forced to use coarser aggregates due to the manner in which the health outcome data are aggregated. This results in much work being done at the census tract level (n = 2216), zip code tabulation area (n = 180+, dependant on

selection method for partial zip codes and building-specific zip codes), United Hospital Fund neighborhood (n = 34+, depending on availability), or county (n = 5).

2.1.1.1 CENSUS DATA

The census-derived socio-demographic data used in this study include percent non-Hispanic White, percent non-Hispanic Black, percent Hispanic/Latino, percent below poverty, median household income, percent of adults greater than 25 who do not have a high school diploma, and percent foreign born. All of the rates were derived by dividing the number of people in the group of interest by the appropriate denominator (e.g., number of non-Hispanic Blacks divided by total population equals percent non-Hispanic Black). These data are traditional environmental justice and environmental health variables meant not only to stand on their own, but also to serve as a proxy for a number of unmeasured factors including possible institutionalized racism, classism, behavioral characteristics (for health risk), cultural differences, social deprivation, social cohesion, and political capital (e.g., political "clout"). Although these secondary factors are not measured directly, the complexity and nuance of the social space is much larger than simple socio-demographic categories and it should be remembered that a variable such as "percent Hispanic / Latino" often represents much more than simply an ethnicity. Additionally, the measured variables are quite broad in scope; treating potentially heterogeneous groups of individuals as homogeneous. For instance, an ethnic label of "Hispanic / Latino" does not take into account ancestry or country of origin (e.g., Puerto Rican, Dominican, Mexican, etc.). It should be noted that the racial/ethnic categories are arranged by non-Hispanic White, non-

Hispanic Black, and Hispanic/Latino in order for the classes to be mutually exclusive as white/black is considered "race" whereas Hispanic/non-Hispanic is considered ethnicity (i.e., in this dataset Latinos who have identified themselves racially as white or black are in the same group). Lastly, other racial and ethnic groups are not explicitly included in this study due to comparatively small proportions city-wide.

Demographically, New York City has an extremely diverse population. This diversity, however, is not homogeneously distributed throughout the city by race/ethnicity, income, education, or foreign-born status when aggregated by borough (**Table 2-1**). This heterogeneity can be more easily seen when it is mapped (**Figure 2-1**). Since pollution is also unlikely to be homogeneously distributed across NYC, it is possible that the environmental burden from exposure to $PM_{2.5}$ is not equitably shared among all the socio-demographic groups in the city.

	Total Population	Percent non-Hispanic White	Percent non-Hispanic Black	Percent Hispanic / Latino	Percent without High Sch. Degree	Median Household Income	Percent Below Poverty
Bronx	1,332,650	14.6	31.2	48.4	37.7	27,611	30.7
Brooklyn	2,465,326	34.7	34.3	19.8	31.2	32,135	25.1
Manhattan	1,537,195	45.8	15.2	27.2	21.3	47,030	20.0
Queens	2,229,379	32.9	18.8	25.0	25.6	42,439	14.6
Staten Island	443,728	71.4	9.0	12.1	17.4	55,039	10.0
NYC	*8,008,278*	*35.0*	*24.4*	*27.0*	*27.7*	*38,293*	*21.2*

Table 2-1: Socio-demographics NYC-wide and by borough. Data source: US Census, 2000.

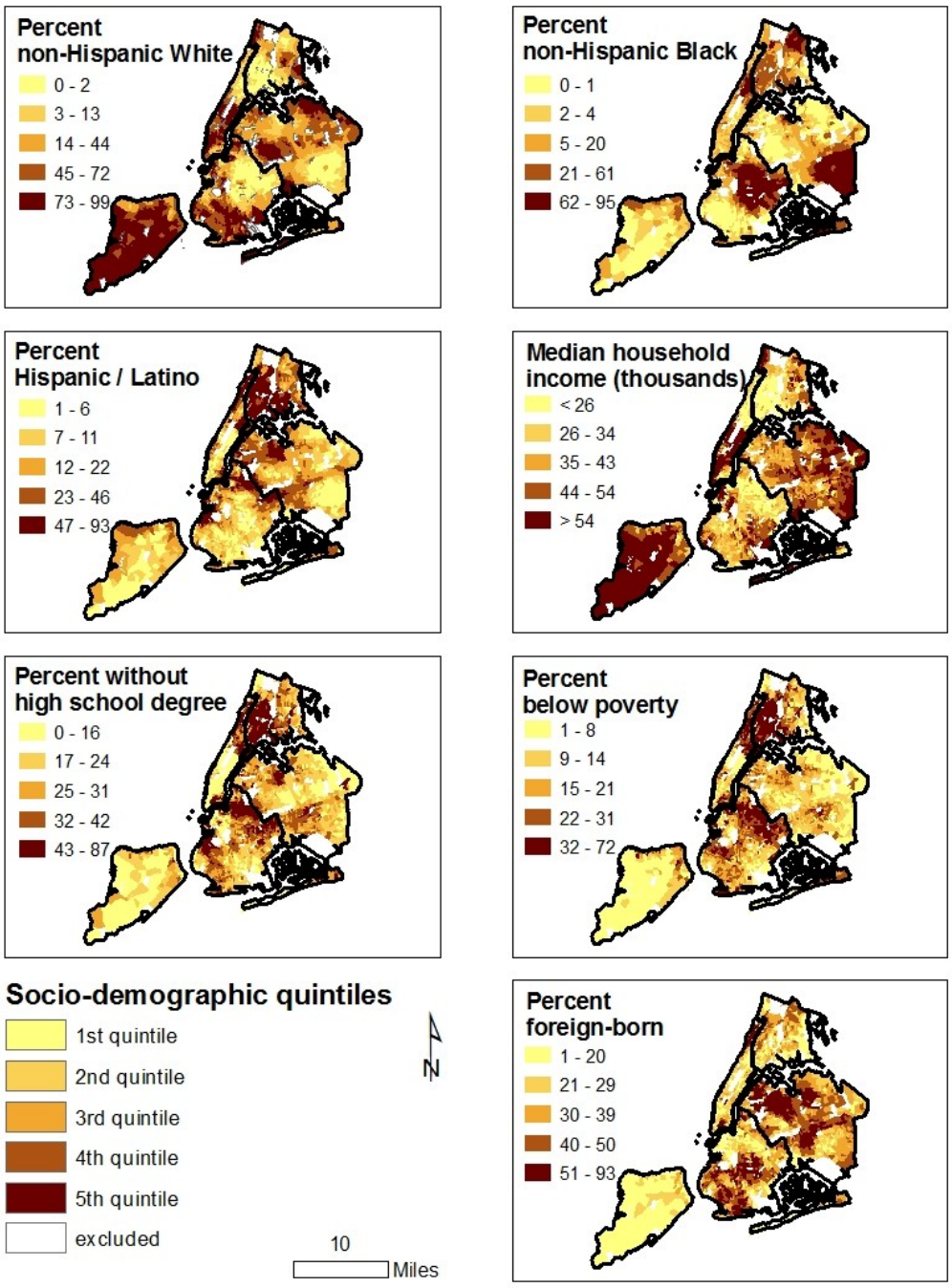


Figure 2-1: NYC socio-demographics arranged in quintiles (5 groups with equal representation on each individual map) by census tract in NYC. Tracts with populations less than 415 (bottom 5%) were excluded to stabilize rates. Data source: U.S. Census, 2000.

It can be useful to examine the spatial distributions of the socio-demographic characteristics using statistical techniques to attempt to confirm the visual assessment that there are clusters of high or low values. This can be done using Moran's I for a global measure and local indicators of spatial autocorrelation (LISA) for a local measure.

Spatial autocorrelation is a central concept in spatial analysis and exploratory spatial data analysis. If the value in one geographic unit is dependent upon the values in its neighboring units, it can be considered spatially autocorrelated (Cliff and Ord, 1973). Moran's I is a global measure (global in the sense that there is one value for the entire study area) for autocorrelation which ranges from -1 to 1. When values approach 0, there is no spatial autocorrelation; as the Moran's I approaches 1 or -1, there is positive (clustering) or negative (dispersion) autocorrelation, respectively. A standardized Z score can be used to assess significance (with the null hypothesis representing a random spatial distribution). In order to calculate a Moran's I, however, a spatial weights matrix must first be defined. This matrix defines the spatial relationship among the samples (e.g., census tracts) based on the chosen parameter (e.g., percent non-Hispanic White). There are many options regarding the definition of the spatial weights matrix, the most common of which include polygon contiguity, simple distance threshold, distance decay (e.g. inverse distance weighting), and k-nearest neighbors, similar to what was discussed earlier. A local indicator of spatial autocorrelation (LISA) can be used to quantify spatial autocorrelation locally by calculating a Moran's I and an associated significance level for each spatial unit (local). The sum of all of the LISAs will be proportional to the global measure of spatial autocorrelation (Anselin, 1995).

Spatial autocorrelation has been assessed at the census tract level in order to match the unit of aggregation for the health data (**Section 2.1.2**) using first order queen contiguity. Globally, all of the socio-demographic variables demonstrate statistically significant spatial autocorrelation at the .01 level (**Table 2-2**). This suggests highly clustered data (as Moran's I values approach '1', the spatial distribution approaches 'perfect' clustering).

Socio-demographic	Moran's I	Z Score	sig.
Percent non-Hispanic White	0.80	61.60	p<.01
Percent non-Hispanic Black	0.87	66.42	p<.01
Percent Hispanic/Latino	0.78	60.06	p<.01
Percent with no High School Degree	0.67	49.86	p<.01
Percent Below Poverty	0.65	48.81	p<.01
Median Household Income	0.59	44.35	p<.01
Percent Foreign-born	0.75	53.75	p<.01

Table 2-2: Spatial autocorrelation (clustering) of socio-demographics in NYC using Moran's I (first order queen contiguity). Data source: U.S. Census, 2000.

When examined locally using a LISA (local Moran's I), the clusters of high and low values, as well as the outliers, can be mapped (**Figure 2-2**). Notice the statistically significant clusters of high values for percent Hispanic/Latino, percent below poverty, and percent without a high school degree in the South Bronx. Also note that this area shows statistically significant clusters of low values for median household income and percent non-Hispanic White. Other interesting collocations of high and low clusters of the different variables can be seen across NYC (e.g., most of Staten Island represents a high cluster of percent non-Hispanic White, and median

household income while also having low clusters of percent without a high school degree, percent below poverty, and percent foreign-born.)

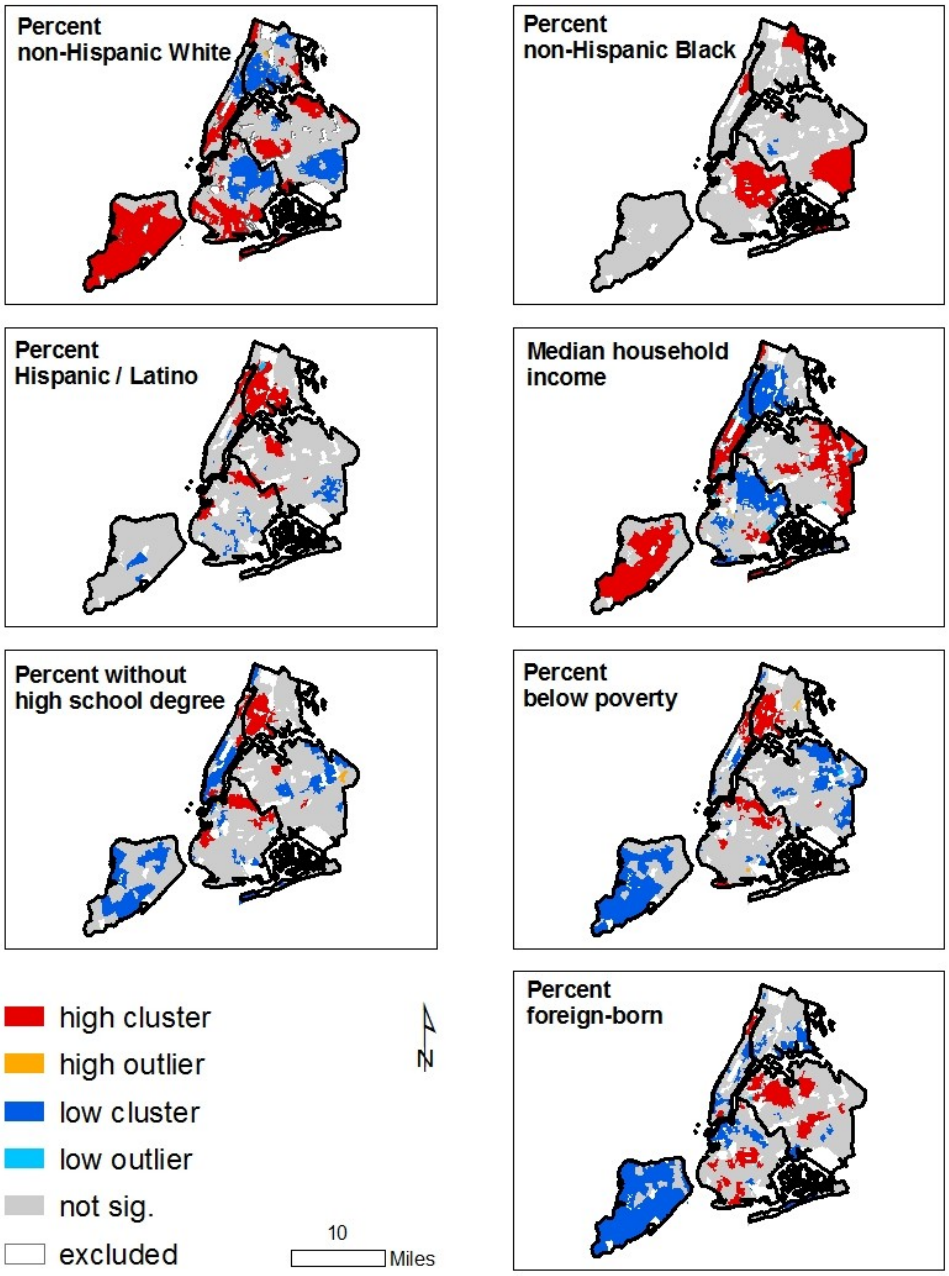


Figure 2-2: Local indicator of spatial autocorrelation (local Moran's I) for socio-demographic variables in NYC. 'High cluster' = high values surrounded by other high values. 'High outlier' = a high value surrounded by low values. 'Low cluster' = low values surrounded by other low values. 'Low outlier' = low value surrounded by high values. Not sig. = tracts without statistically significant local autocorrelation. First order queens contiguity was used

to define the spatial weights matrix. Tracts with fewer than 415 persons (bottom 5%) were excluded to stabilize rates. Data source: U.S. Census, 2000.

2.1.1.2THE CADASTRAL-BASED EXPERT DASYMETRIC SYSTEM (CEDS)

The heterogeneity and clustering present at the city-wide level are mirrored at the micro-level in terms of land-use, housing availability, and population distribution. As the spatial resolution of exposure estimation to an environmental hazard improves, so must the estimation of the potentially affected population(s) in order to fully exploit the increased geographic accuracy of the pollution data. It was mainly for this reason that the Cadastral-based Expert Dasymetric System (CEDS) was developed. It has been shown to be effective in estimating affected residential populations in a number of scenarios including proximity analysis (Maantay, 2002), pollution plume analysis (Maantay et al., 2009), flood risk (Maantay and Maroko, 2008), and crime analyses (Herrmann and Maroko, 2006).

CEDS estimates total population and specific sub-population distribution for urban areas, or any geographies with sufficient cadastral (tax lot) data, in order to develop an improved "denominator," allowing for more correct rates in GIS analyses as well as more accurate estimations of exposure due to residential location. Rather than using data aggregated by arbitrary administrative boundaries such as census tracts, dasymetric mapping, a disaggregation method using ancillary information to delineate areas of homogeneous values, can be used.

Dasymetric mapping techniques have been utilized extensively in various forms (Bielecka, 2005; Eicher et al., 2001; Forster, 1985; Holt et al., 2004; Holloway et al., 1997); however remotely sensed data and land cover are often used as the ancillary datasets (Langford et al., 1991; Mennis, 2003; Sleeter , 2004; Wu and Murray, 2007; Wu et al., 2005). This contrasts starkly with CEDS, which exploits cadastral data (tax lot information) to redistribute the populations. This method is more applicable than its remote-sensing counterparts to hyper-heterogeneous urban environments, such as that of New York City (**Figure 2-3**). This heterogeneity can be particularly troubling when quantification of an exposed population (e.g., people living near major pollution sources) is heavily biased within the geographic unit (e.g., the people are not evenly distributed in the census aggregate, **Figure 2-4***).* Additionally, CEDS uses an expert system routine and validation against various census enumeration units in order to further refine the estimates.

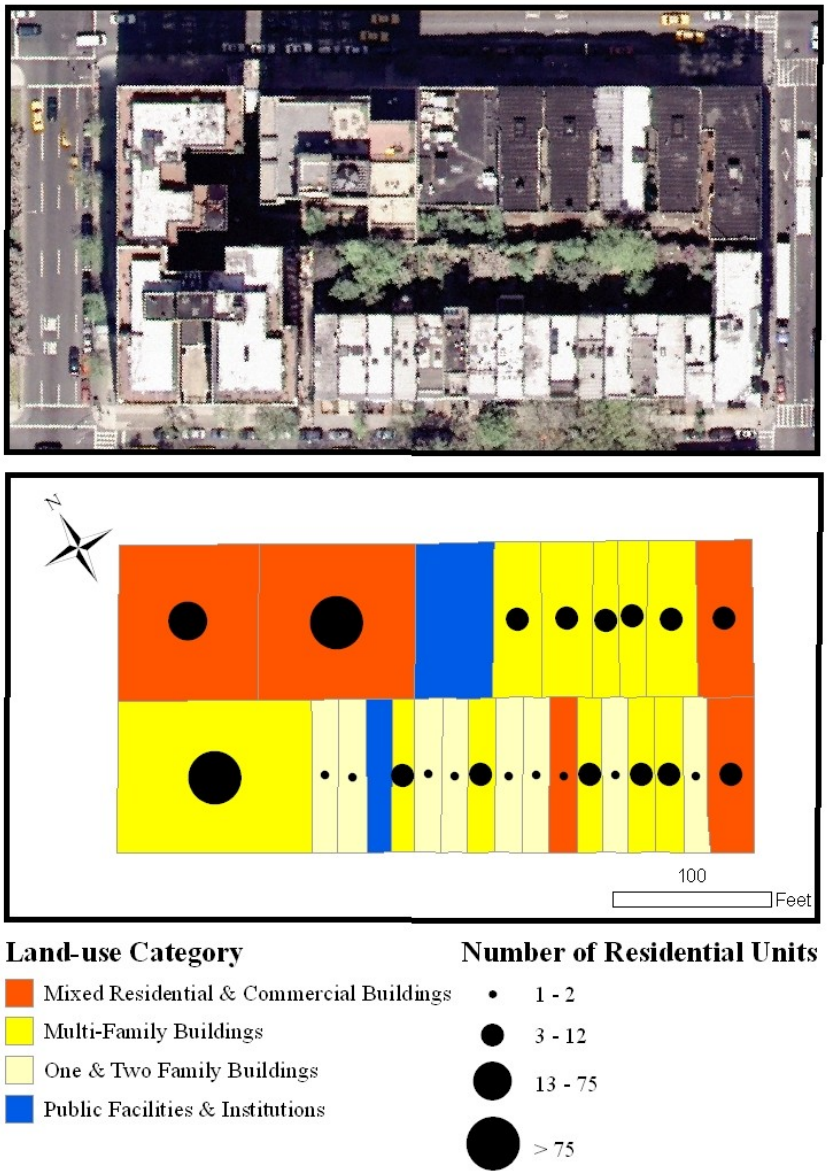


Figure 2-3: Heterogeneity of a Manhattan city block. The orthophoto (above) and the cadastral map show the uneven distribution of land use categories and residential units at the tax-lot level even when examining only one city block. (There are, on average, more than 16 city blocks in a New York City census tract.) Data source: NYCMap, 2004; LotInfo 2001.

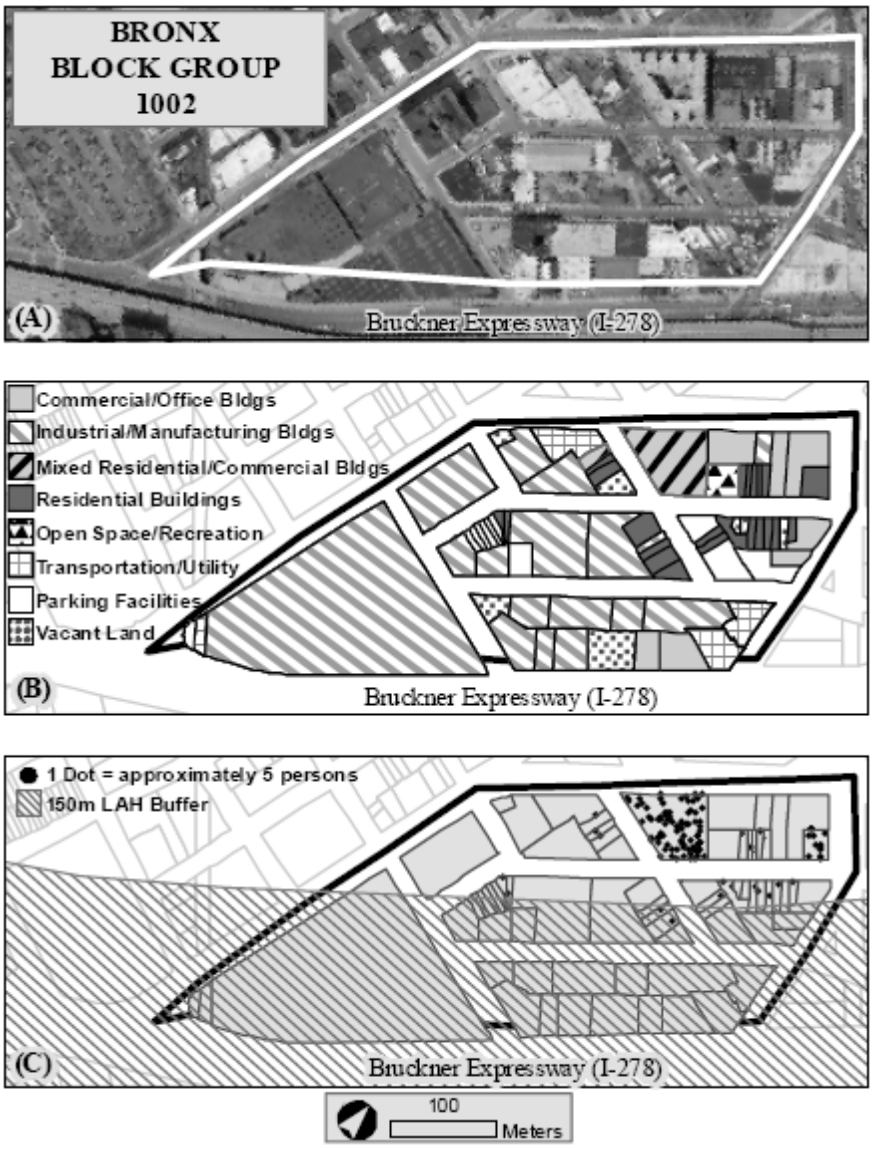


Figure 2-4: Sample heterogeneous block group (a) aerial photo; (b) land use map; (c) Population estimate and 150m pollution buffer from a major highway in the Bronx. Data Sources: LotInfo, 2003; NYCMap, 2002.

In urban environments, the most commonly used forms of data disaggregation are areal weighting (AW) and filtered areal weighting (FAW). These methods are very simple dasymetric techniques which utilize area as the ancillary dataset for redistribution of the data. For instance, if there is a source zone (e.g., census tract) that has 25% of its area exposed to unsafe levels of

pollution (target zone), then areal weighting would estimate that 25% of the residential population of that census tract are exposed (**Equation 2-1**).

$$\mathbf{POP_{AW} = POP_S * AREA_t / AREA_S} \qquad \textbf{Eq. 2-1}$$

where:
POP_{AW} = estimated population in target zone from areal weighting;
POP_S = source zone population (known quantity from census tract, block group, etc.);
$AREA_t$ = area of target zone (e.g. area exposed to pollution)
$AREA_S$ = area of source zone (e.g. census tract, block group, etc.).

Filtered areal weighting is refinement of simple AW in that it removes areas known to not contain any population (e.g., parks, open spaces, and water bodies) (**equation 2-2**).

$$\mathbf{POP_{FAW} = POP_S * M\_AREA_t / M\_AREA_S} \qquad \textbf{Eq. 2-2}$$

where:
POP_{FAW} = estimated population in target zone from filtered areal weighting;
POP_S = source zone population (known quantity from census tract, block group, etc.);
M_AREA_t = modified area of target zone (open spaces excluded); and
M_AREA_S = modified area of source zone (e.g. census tract area with open spaces excluded).

CEDS, on the other hand, employs tax-lot level information regarding amount of residential area (RA) and number of residential units (RU) with which to redistribute population. The CEDS technique, although simple in theory, can be complex in practice and does require sufficient cadastral data in order to function. To calculate CEDS-derived population estimates, one must first calculate the total number of residential units (RU) and residential area (RA) in the source zone (e.g. census block group). Then the RU and RA are calculated for the target zones (e.g., tax lots). A ratio is established for both ancillary datasets and that ratio is multiplied by the

population in the source zone. The results are the estimated populations in the target zone (one based on RA and one based on RU, **Equation 2-3**). To determine which ancillary dataset to use, CEDS employs an expert system which disaggregates the data from a larger source zone (e.g., census tract) to a smaller, but known, target zone (e.g., census block group). Since the target zone's 'true' data are known, the expert system compares RU- and RA-based estimates to these known quantities and selects the better performing dataset (**Equation 2-4**).

$$\mathbf{POP_{CEDS} = POP_S * U_t / U_S} \qquad \textbf{Eq. 2-3}$$

where:
POP_{CEDS} = CEDS-derived lot-level population;
POP_S = source zone population (block group or tract);
U_t = the number of proxy units (RU or RA) in the target zone (e.g. tax lot); and
U_S = the number of proxy units (RU or RA) in the source zone (e.g. census tract or block group).

$$\mathbf{POP_{diff} = | POP_{BG} - POP_{est} |} \qquad \textbf{Eq. 2-4}$$

where:
POP_{diff} = the difference between census and estimated populations per block group;
POP_{BG} = census block group population; and
POP_{est} = estimated population (*RU* or *RA*) derived from the census tract (not block group).

By comparing the estimated population to the census population for both the *RU*- and *RA*-based techniques, it can be assumed that the process that resulted in estimates more similar to the census block group values (i.e., smaller POP_{diff} values) more accurately redistributed the data. After re-joining the POP_{diff} data with the LotInfo data, the expert system would then select the superior proxy unit as the disaggregation technique for each block group. This can be described as follows (**Equation 2-5**):

IF RU_POP_{diff} <= RA_POP_{diff}, THEN $POP_l = POP_{RU\_BG}$, ELSE $POP_l = POP_{RA\_BG}$ **Eq. 2-5**

where:
RU_POP_{diff} = the absolute difference between the census block group population and the estimated block group population derived from the census tract population based upon number of residential units;
RA_POP_{diff} = the absolute difference between the census block group population and the estimated block group population derived from the census tract population based upon residential area;
POP_l = the final estimated tax lot population dasymetrically derived from the census block group population (not the census tract);
POP_{RU_BG} = the estimated tax lot population dasymetrically derived from the census block group population (not the census tract) based on number of residential units; and
POP_{RA_BG} = the estimated tax lot population dasymetrically derived from the census block group population (not the census tract) based on the adjusted residential area.

In essence, it is the performance of the tract-level disaggregation that defines the proxy units used for each block group disaggregation, ultimately resulting in a final dasymetrically derived value individually tailored for each block group. It is important to note that CEDS is a pycnophylactic, or mass preserving, technique. This means that if BG population data are disaggregated to the tax lot level, the estimated population data of the tax lots, when summed, will equal that of the BG from which they were derived.

The differences in these three techniques – areal weighting, filtered areal weighting, and the cadastral based expert dasymetric system – can be easily visualized diagrammatically (**Figure 2-5**).

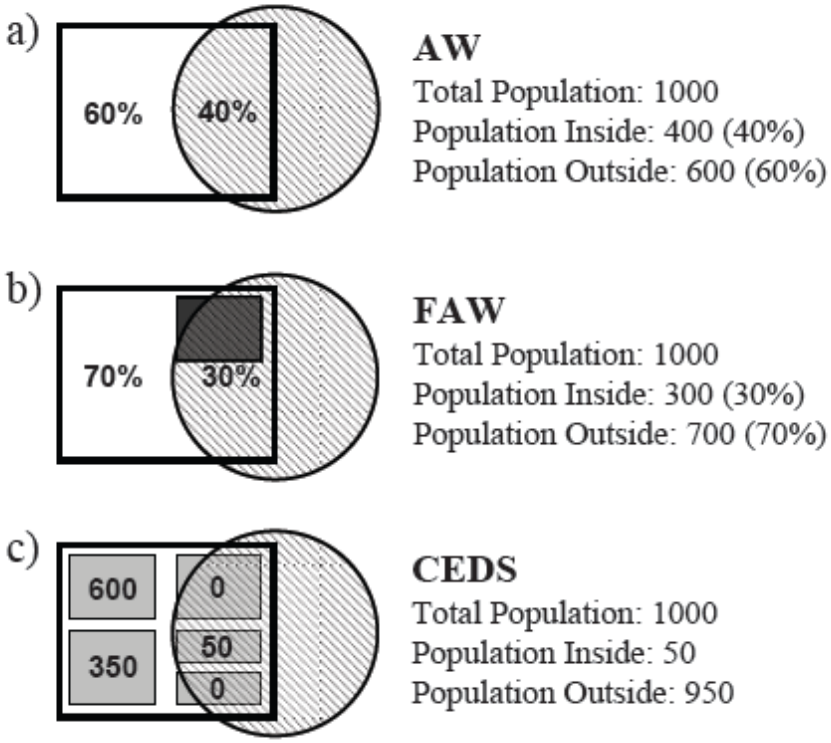


Figure 2-5: Diagrammatic comparison of population disaggregation methods. (a) Areal Weighting (AW): Block group intersected by an impact buffer. (b) Filtered Areal Weighting (FAW): block group intersected by an impact buffer, and showing an uninhabited area (dark rectangle). (c) CEDS: Block group showing tax lot boundaries.

The CEDS method has been validated in a very similar way to how the expert system is employed. Census tract (TR) data are disaggregated to census block groups (BG) using CEDS, ratio of residential area, ratio of residential units, and FAW. Note that the residential area and residential units ratios are intermediate steps for CEDS and are included to show the improvement of the use of the expert system. The estimates of each method are compared to the 'observed' BG populations as reported by the census. Previous work has validated the data using bivariate regressions and percent deviation. The CEDS method's performance appears to be

superior to that of filtered areal weighting, residential area, and residential units when estimating BG population based on TR populations in NYC (**Figure 2-6**).

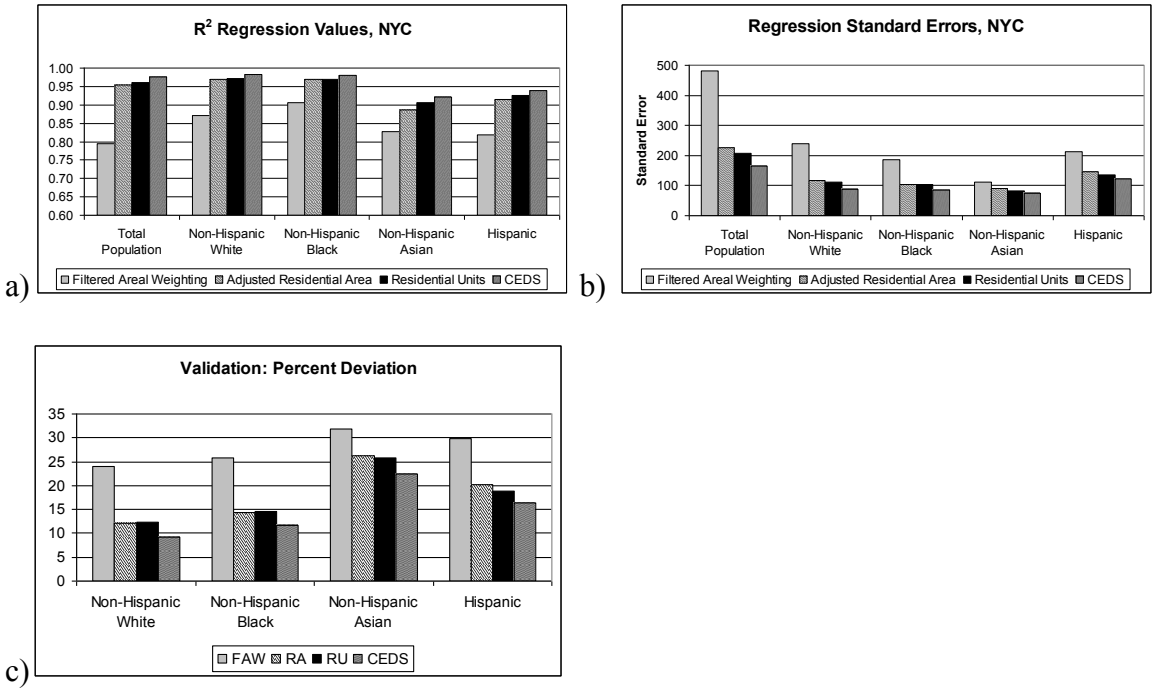


Figure 2-6: R2 values (a) and standard error values (b) from simple linear regressions of selected populations for filtered areal weighting, residential area, residential units, and CEDS estimated block group populations vs. Census-reported block group population. Figure (c) shows percent deviation of derived data as compared to Census-reported data.

The relationship between estimated and observed population values can be observed graphically using scatter plots. They clearly suggest that CEDS estimates are "closer" to observed census values than the filtered areal weighting estimates at the block group level (**Figure 2-7**).

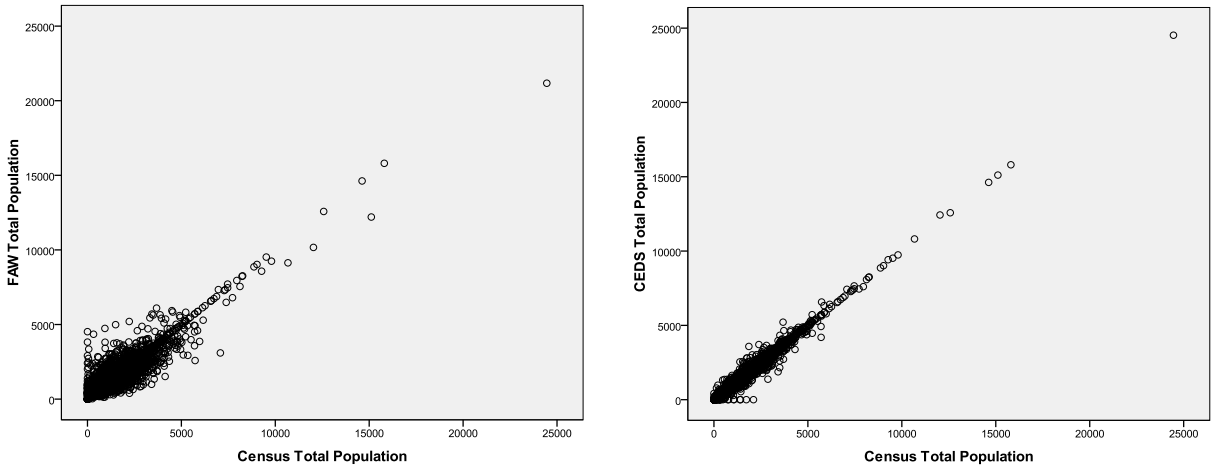


Figure 2-7: Scatter plots of FAW-derived (left) and CEDS-derived (right) block group estimates of total population vs. census-reported block group total population.

To confirm and support these validations, three added measures were employed on the total population and racial/ethnic demographic categories. These were bias, distance, and correlation. Bias was measured by simply comparing the means of the estimated block group data (filtered areal weighting, residential area, residential units, and CEDS) and the "observed" data (Census-reported block group populations) (**Equation 2-6**).

Mean error = (Σε)/N **Eq. 2-6**

where:
ε = error
N = number of observations

Distance was measured using root-mean squared error (RMSE). RMSE quantifies how close the estimated data are to the observed data by calculating the "distance" from each estimate to the observed value, squaring this value (to prevent negative numbers from cancelling out positive

ones), calculating the mean, and finally taking the square root. The smaller the RMSE, the closer the fit is to the data. Put simply, the RMSE is the average distance of estimated data from the observed data (**Equation 2-7**).

$$\mathbf{RMSE = [(1/N)\ \Sigma\varepsilon^2]^{0.5}} \qquad \textbf{Eq. 2-7}$$

where:
RMSE = root mean square error
ε = error
N = number of observations

Correlation is calculated using Pearson and Spearman correlation tests, which result in "goodness-of-fit" measures either parametrically (Pearson) or non-parametrically (Spearman). The results of these diagnostics suggest that there is slightly more bias with CEDS when compared to FAW, and FAW tends to overestimate, whereas the cadastral data and CEDS tend to underestimate. In terms of distance and correlation, CEDS outperforms the other methods with consistently lower RMSE values and higher Pearson's and Spearman's correlations (**Table 2-3**).

Population Group	Disaggregation Method	BIAS	CORRELATION		DISTANCE
		Mean of Estimate - Mean of Census	Pearson Correlation	Spearman's Rho	RMSE
Total Population	Filtered Areal Weighting	1.334	.891	.789	482.72
	Residential Area	-2.336	.977	.954	229.39
	Residential Units	-2.297	.980	.960	211.08
	CEDS	-5.540	.988	.975	164.96
non-Hispanic White	Filtered Areal Weighting	1.180	.934	.947	239.58
	Residential Area	-0.907	.984	.975	118.75
	Residential Units	-0.859	.986	.974	112.72
	CEDS	-0.792	.991	.979	87.85
non-Hispanic Black	Filtered Areal Weighting	-0.003	.950	.937	186.39
	Residential Area	-0.718	.985	.959	103.95
	Residential Units	-0.771	.985	.959	104.08
	CEDS	-1.727	.990	.964	84.51
non-Hispanic Asian	Filtered Areal Weighting	0.036	.910	.924	112.03
	Residential Area	-0.144	.941	.942	91.82
	Residential Units	-0.129	.952	.942	82.89
	CEDS	-0.459	.960	.948	75.52
Hispanic / Latino	Filtered Areal Weighting	0.099	.904	.899	214.60
	Residential Area	-0.512	.956	.944	145.96
	Residential Units	-0.487	.962	.949	135.21
	CEDS	-0.713	.969	.956	122.22

Table 2-3: Validation diagnostics for filtered areal weighting, residential area-based disaggregation, residential unit-based disaggregation, and CEDS.

Although these diagnostics, and those that preceded them, do suggest that CEDS is a better estimator of population distribution than filtered-areal weighting, some limitations and caveats remain with the method. It appears that the bias of underestimation may be at least partially due to an incomplete cadastral dataset. If there are block groups wherein none of the tax lots have information regarding residential area or residential units, then the CEDS method will fail (assuming that there is population present). This failure not only potentially leads to an

underestimation bias, but also a loss of the pycnophylactic nature of CEDS. This phenomenon can be seen in the scatter plots above (**Figure 2-7**) with the "line" of points that have zero CEDS-estimated population but existing Census-reported population. Although this only appears to be an issue with less than 2% of the CEDS data, there are a number of ways that this limitation can be dealt with, the most enticing of which is the use of a third ancillary data set to be used when the other two (residential area and residential units) fail. This "fail-safe" could be total lot area (independent of building class), total land area (independent of lot size), or some combination of other variables (e.g. total lot area minus commercial/industrial lot area). This analysis is not included in this dissertation, but it will be conducted in future research. Another limitation is the use of CEDS in regression analysis. It is extremely important to note that although the absolute numbers of estimated populations and sub-populations seem reliable, the rates within each tax lot (e.g., percent non-Hispanic Black) are not independent from the parent block group. In other words, if the block group contained a population that is 50% non-Hispanic Black, then all the populated tax lots within that block group would have very similar rates – resulting in data that are not independent or uncorrelated. As such, CEDS is most useful when working with absolute numbers, or for the purpose of re-aggregating the data in non-census boundaries (e.g., buffer zones, high pollution areas, etc.).

In this study, CEDS has been used (where appropriate) to estimate human exposure to $PM_{2.5}$. Due to the physical and social heterogeneity of New York City, and the high spatial resolution of the modeled pollution surface, CEDS seems like the best option for estimating exposure of populations and selected sub-populations (environmental justice analyses). It was not used for

the health outcome portion of this study. The heart failure hospitalization data are aggregated to the census tract, and cannot be disaggregated reliably with any known method. As such, CEDS is of limited utility and standard census tract socio-demographic data were used in the environmental health outcome modeling.

2.1.2 HEALTH DATA

The need for reliable health outcome data is of clear importance to this study, as it will act as the dependant variable in subsequent models and analyses. This type of data, however, is notoriously difficult to acquire. This section is divided into three sub-sections. In these sub-sections the sources and limitations of the raw data will be discussed (SPARCS), manipulation of the data in terms of age-adjustment will be described, and some exploratory spatial data analysis will be utilized in order to present what is arguably the most critical dataset in the dissertation.

2.1.2.1SPARCS

The original plan was to use record-level hospitalization data, geocoded to the patients' home addresses, that were being collected for a grant-funded project of the South Bronx Environmental Justice Partnership (SPEJP) – a collaboration among researchers from Lehman College, Montefiore Medical Center / Albert Einstein College of Medicine, and the For a Better Bronx (FABB) community group. After a significant monetary investment from the SBEJP and a number of years of waiting, the data were finally given from the Statewide Planning and

Research Cooperative System (SPARCS) (original compilers) to Emerging Health Information Technologies (processors) to us. After working with these data, however, it became clear that there were some serious quality issues. The most important of these shortcomings was an apparent unnatural grouping of hospitalizations in Queens (**Figure 2-8**). Notice that when the data are aggregated from points to the census tract level, and compared to the same SPARCS data acquired from Infoshare.org, there are marked differences. The tracts with extremely high numbers and rates of hospitalizations in the Emerging Health Information Technologies dataset are near the centroids of the zip codes, leading me to believe that much of the data were not geocoded to the actual home address, but rather to the zip code (when data is aggregated to the census tract level, and then displayed as a dot density, the dots are placed randomly within the aggregate unit). After numerous dialogues with the lead investigator at Emerging Health Information Technologies, it became clear that the geocoding success rate for all the boroughs, particularly Queens (23%), was far below what can be considered acceptable (**Table 2-4**).

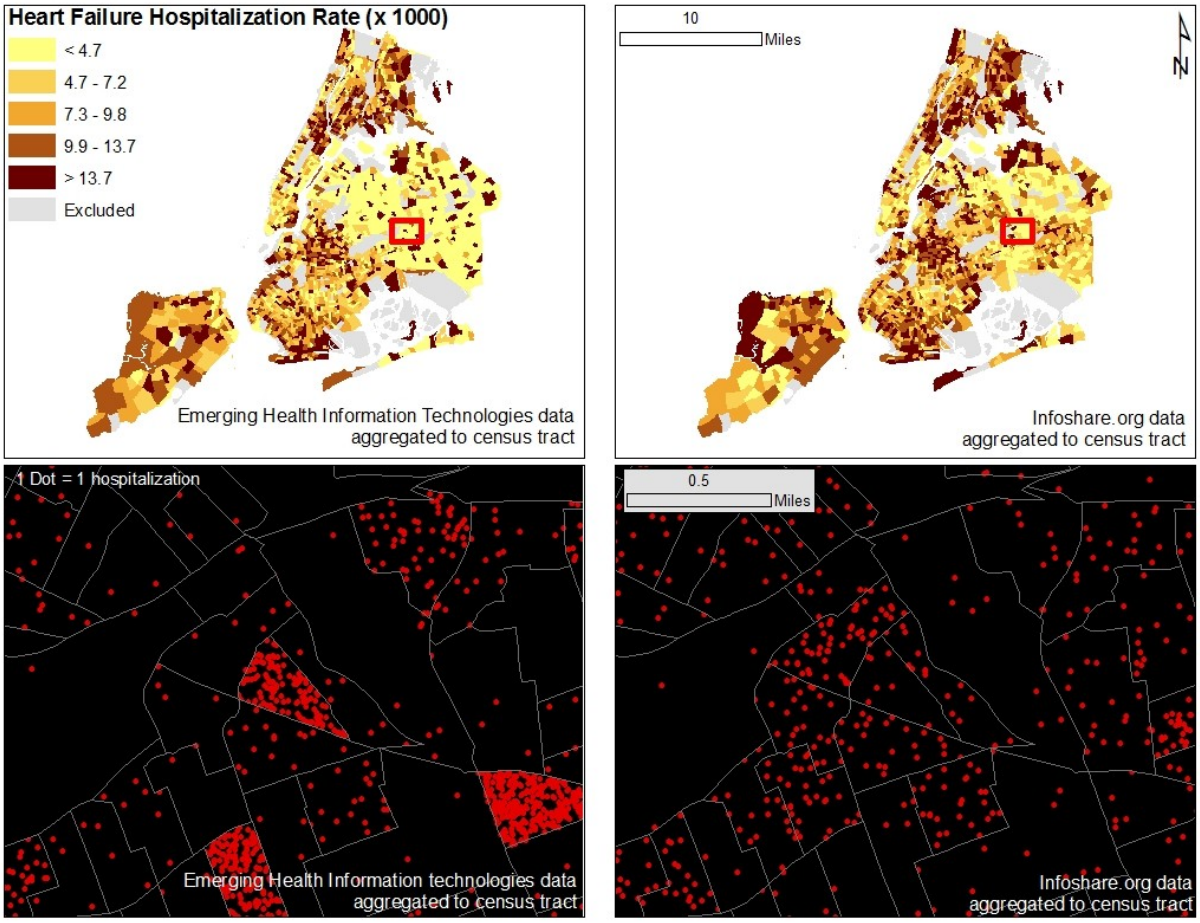


Figure 2-8: Heart Failure rates using Emerging Health Information Technologies SPARCS data aggregated to the census tract level versus Infoshare.org SPARCS data aggregated to the census tract level. Top left and top right show the raw heart failure rates for the Emerging Health Information Technologies and Infoshare.org data, respectively. The bottom left and bottom right show a "zoom-in" of the number of hospitalizations shown as dot densities. The "unnatural" groupings in the left maps are clearly visible.

County	Percent Matched to Address	Percent Matched to Zip Code Centroid	Percent Failed
Bronx	65.9%	28.0%	6.1%
Brooklyn	75.8%	21.1%	3.1%
Manhattan	70.8%	23.7%	5.5%
Queens	23.3%	74.8%	1.9%
Staten Island	78.3%	20.2%	1.6%

Table 2-4: Geocoding success rate for Emerging Health Information Technologies SPARCS data. Data source: Emerging Health Information Technologies.

These geocoding issues rendered the data completely useless for spatial analysis as even the zip code level is too coarse for the type of analyses employed in this study. Although I was clearly unable to use these data for spatial analyses, it proved to be a good exercise in the importance of exploratory spatial data analysis, since without the latter it would not have been clear that much of the data was spatially distributed in a highly biased and unrealistic fashion (at the zip code centroid rather than home address location).

Thankfully, SPARCS data for NYC is freely available aggregated to the census tract level from Infoshare.org / Community Studies of New York, Inc., a non-profit corporation founded by Leonard Rodberg and John Seley of Queens College, CUNY. Although the patient-level data may have resulted in a much more nuanced and novel set of analyses, the aggregate data would suffice. It is important to note that "hospitalizations" are not equivalent to incidence, but rather a proxy for prevalence, incidence, and severity (as usually only the most severe health events result in hospital admission). There is also potential bias in the data as a result of how disease(s) is managed, access to primary care physicians, health education, and insurance status. These biases, although unavoidable, must be considered thoughtfully.

The SPARCS data are arranged by the International Classification of Diseases, 9th Revision (ICD-9) codes. These codes identify the diagnoses used by the hospital at the time of the patient's admission. All persons who were hospitalized for heart failure (ICD-9: 428) between 2001 and 2003, inclusive, were queried from the database. Note that these data represent persons

hospitalized rather than all hospitalizations (if the same individual was hospitalized multiple times in one year, only the first instance is included).

2.1.2.2 AGE ADJUSTMENT

As heart failure hospitalization rates are closely related to age, age-adjusted rates were calculated. Indirect-age adjustment was used rather than direct adjustment since the age of the individual patients being hospitalized was not available via Infoshare. Indirect adjustment averages the age-specific rates in the standard population, weighted by the distribution of the study population. Although used less frequently than direct adjustment, it can be useful for a number of situations, including when age-specific numbers are not available in the study population (Curtin and Klein, 1995). In this case, the estimate is of the occurrence of hospitalizations per census tract relative to what might be expected if the population had the same hospitalization rates as NYC, which is designated as the "standard" population. Age data were downloaded from the U.S. Census Bureau and processed into 5-year cohorts at the census tract level, hospitalization data are from SPARCS via Infoshare.org aggregated by census tract, and the NYC-wide hospitalization data are from SPARCS via Emerging Health Information Technologies, since the aspatial information (e.g., age) associated with this dataset was not problematic (**Equation 2-8**).

$$\mathbf{AAHR_{CT} = CHR_{NYC} * H_{CT}/[\Sigma_{age\ groups}\ (HR_{asNYC} * P_{ageCT})]} \qquad \textbf{Eq. 2-8}$$

where:
$AAHR_{CT}$ is the age-standardized hospitalization rate in the census tract
CHR_{NYC} is the crude hospitalization rate for NYC
H_{CT} is the number of hospitalization in the census tract.
HR_{asNYC} is the age-specific hospitalization rate for NYC.
P_{ageCT} is the number of people in the age group in the census tract's population.

The results of the heart failure hospitalizations rates age adjustment are quite striking when mapped. Distinct differences in the concentrations of hospitalization rates can be seen in various areas of NYC (e.g., rates in the South Bronx appear more severe after age adjustment, whereas rates in southern Brooklyn appear to be reduced) (**Figure 2-9**). Tracts with either low population (lowest 5 percentile) or no hospitalizations over the three-year period were excluded in order to stabilize rates.

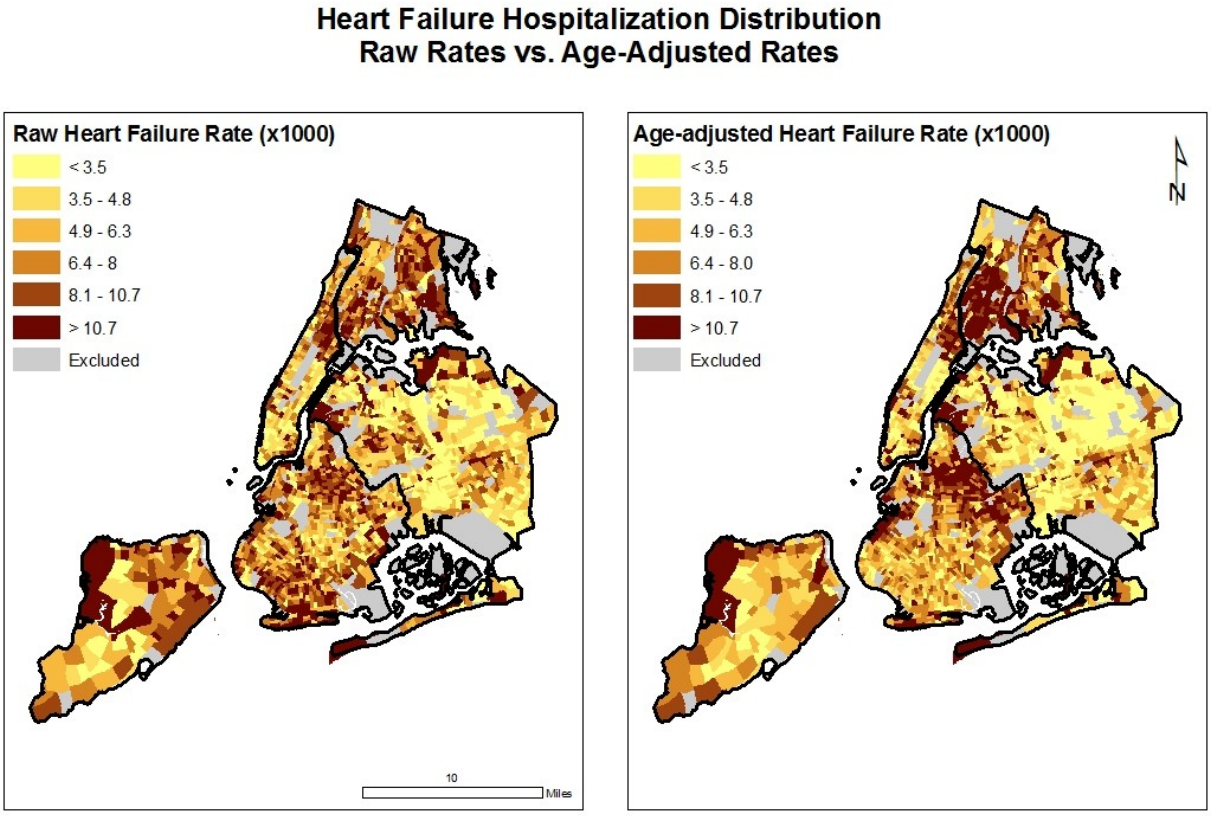


Figure 2-9: Raw hospitalization rate vs. age-adjusted hospitalization rate for heart failure in NYC (2001-2003, inclusive).

It is notable that the aspatial distribution of the data is quite skewed with some extremely high values. The histogram, which excludes census tracts with low populations (lowest 5 percentile which equates to tracts with fewer than 415 residents), are shown with a logarithmic scale for the vertical axis in order to be able to discern the positive tail (**Figure 2-10**). The NYC-wide rate of heart failure hospitalizations is 8.8 per 1000 over 3 years. The two extremely high values represent census tracts near the Broad Channel (346 per 1000 people over 3 years) and Maspeth (317 per 1000 people over 3 years) neighborhoods. The next highest hospitalization rate is near the Hunters Point neighborhood, just south of the previously mentioned tract near Maspeth, with 105 heart failure hospitalizations per 1000 over 3 years (**Table 2-5**). If the census tracts with the top 3% of hospitalization rates (n = 61) are trimmed, the remaining values range from 0.2 to 21.7 per 1000 over 3 years (excluding tracts with low populations or no hospitalizations, **Figure 2-11**). The affect of this distribution is explored later in this chapter and the next (**Chapter 3, Analysis**).

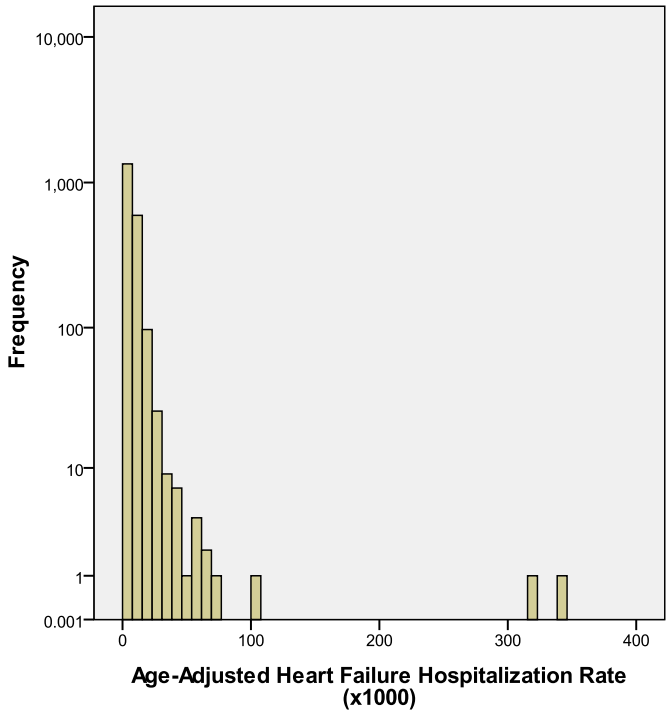


Figure 2-10: Histogram for age-adjusted heart failure rates (2001-2003, inclusive). Vertical axis is logarithmic.

	Neighborhood		
	Broad Channel, Queens	***Maspeth, Queens***	***Hunters Point, Queens***
Census Tract FIPS	36081107201	36081001900	36081000100
Total Population	2630	794	1370
Percent Non-Hispanic White	93.7	37.8	50.8
Percent Non-Hispanic Black	0.0	3.1	22.9
Percent Hispanic/Latino	4.9	46.3	10.8
Percent Without High Sch. Degree	21.0	16.7	22.8
Percent Below Poverty	12.1	14.8	20.1
Percent Foreign-born	4.5	33.3	26.4
Hospitalization Rate	345.76	316.78	105.39

Table 2-5: Socio-demographics for three census tracts with the highest heart failure hospitalization rates per 1000 over 3 years. NYC-wide rate is 8.8.

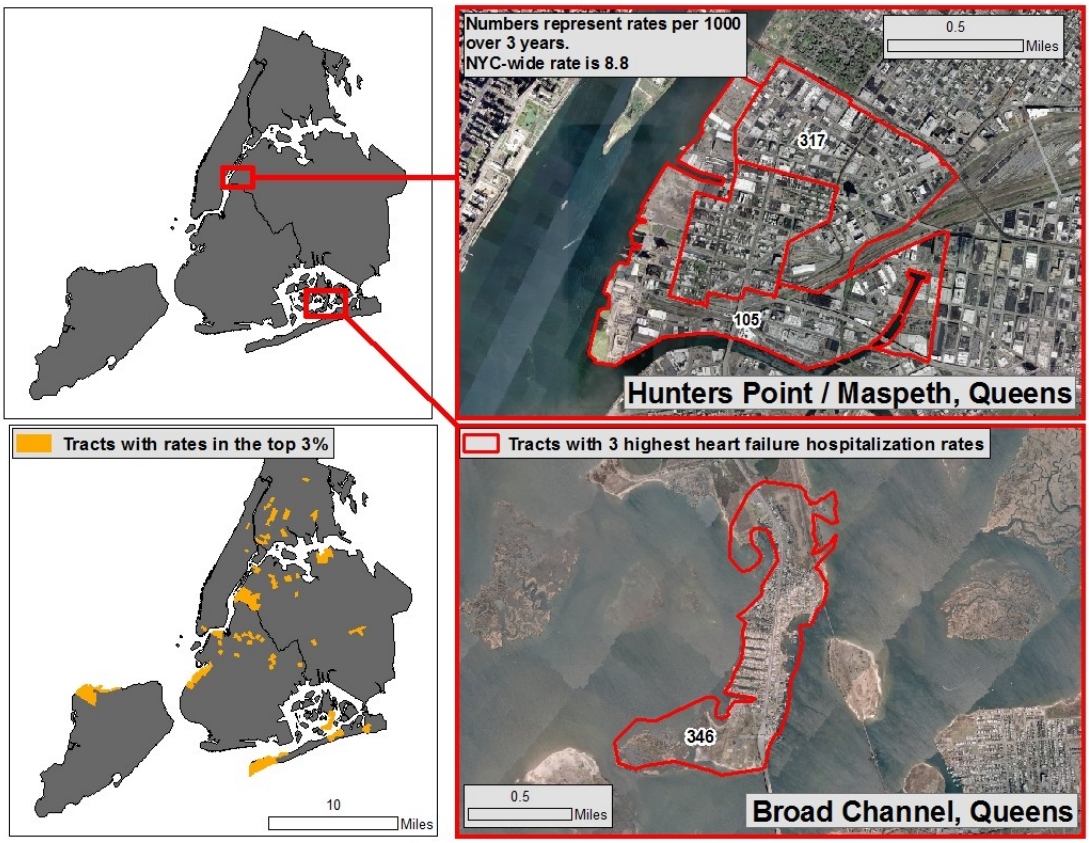


Figure 2-11: Areas of high heart failure hospitalization rates in NYC. Data source: SPARCS, US Census, 2000, NYCMap, 2002.

2.1.2.3 EXPLORATORY SPATIAL DATA ANALYSIS

Aside from the comparison of adjusted and unadjusted hospitalization rates, it can be very useful to continue the spatial exploration of the data by conducting an analysis of the clustering or dispersion of the hospitalization rates. This can be done both globally via Moran's I and locally via local indicators of spatial autocorrelation (LISA).

The Moran's I for heart failure hospitalization rate (low population tracts and tracts with no reported hospitalizations excluded) suggests a statistically significant clustering of hospitalizations ($p<.01$) with a Moran's I value of 0.14 using 1st order polygon contiguity (Z score of 11.82). When the positive tail of the data is trimmed, this clustering becomes stronger. For instance, when the top 2 outliers are removed, the Moran's I increases to 22.06 (Z score of 22.06, $p<.01$). When the census tracts with the highest 3% of hospitalization rates are trimmed, the Moran's I becomes 0.42 (Z score of 29.19, $p<.01$). This clustering can be further explored using local indicators of spatial autocorrelation (LISA). When the Z-scores of the local Moran's I are looked at in this fashion, once again using first order contiguity, the areas with local clusters of high hospitalization rates can be clearly seen (**Figure 2-12**). It is important to note that what is being seen in this map are the local clusters of similar or dissimilar rates (neighboring geographic units in this case), rather than the rates themselves. As can be seen on the map, there are clusters of high values near and around: Hunter's Point/Maspeth (Queens), College Point / Whitestone (Queens), Sunset Park (Brooklyn), Bushwick / Stuyvesant Heights (Brooklyn), central part of the Rockaways, and areas within the South Bronx when the untrimmed dataset is used.

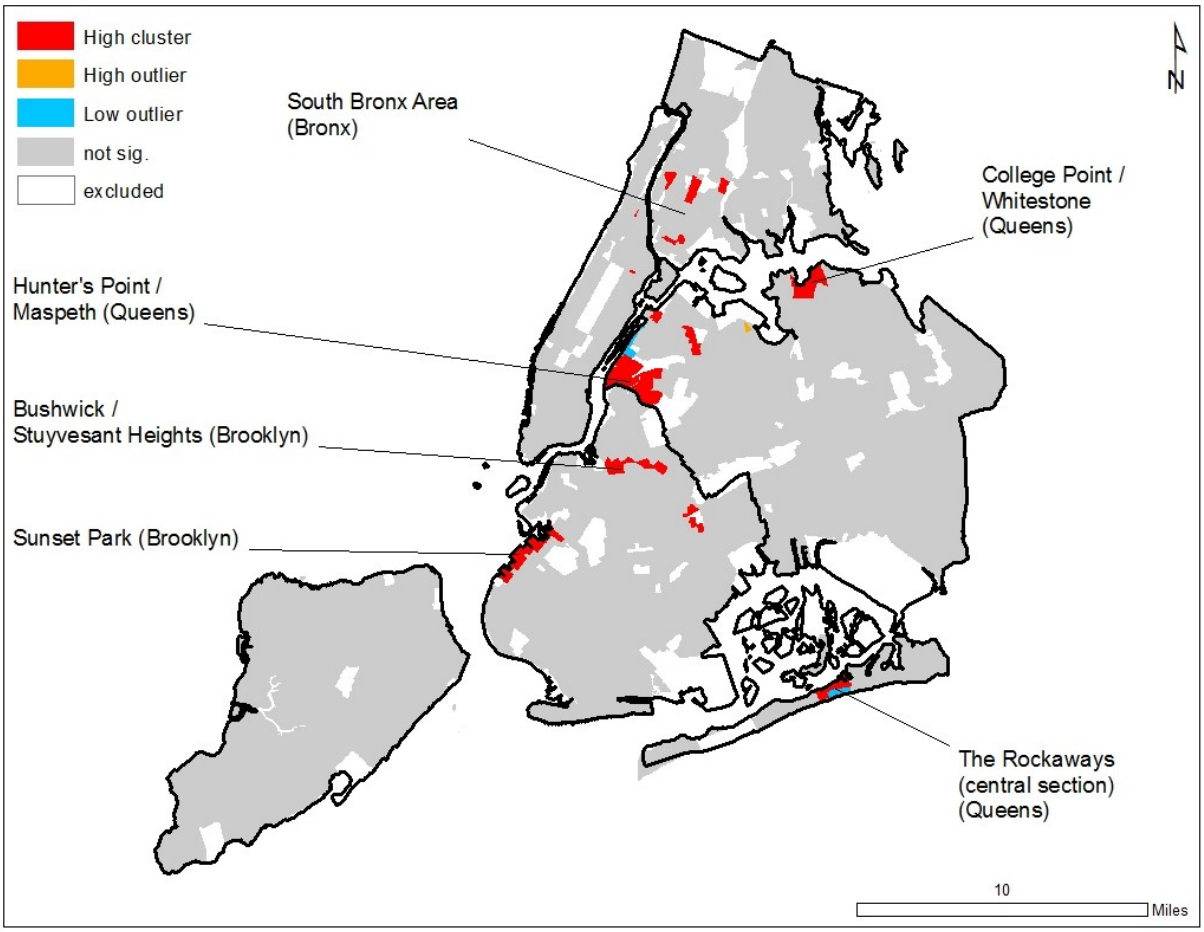


Figure 2-12: Local Moran's I of age-adjusted heart failure hospitalization rate clusters and outliers. 'High cluster' = high values surrounded by other high values. 'High outlier' = a high value surrounded by low values. 'Low outlier' = low value surrounded by high values. Not sig. = tracts without statistically significant local autocorrelation. There are no statistically significant 'low clusters' (low values surrounded by other low values). Note that the Broad Channel census tract has no significance due to the use of first order queen's contiguity being used to define the spatial weights matrix. Tracts with fewer than 415 persons (bottom 5%) were excluded to stabilize rates.

It can be interesting to explore this data further by systematically trimming some of the positive tail that was exposed in the histogram (**Figure 2-10**). Two more LISA maps were created; one with the top two outliers removed and one with the tracts exhibiting the highest 3% of the heart

failure hospitalization rates removed. Note how the intricacy and heterogeneity of the spatial autocorrelations are increase as the positive tail is trimmed (**Figure 2-13**).

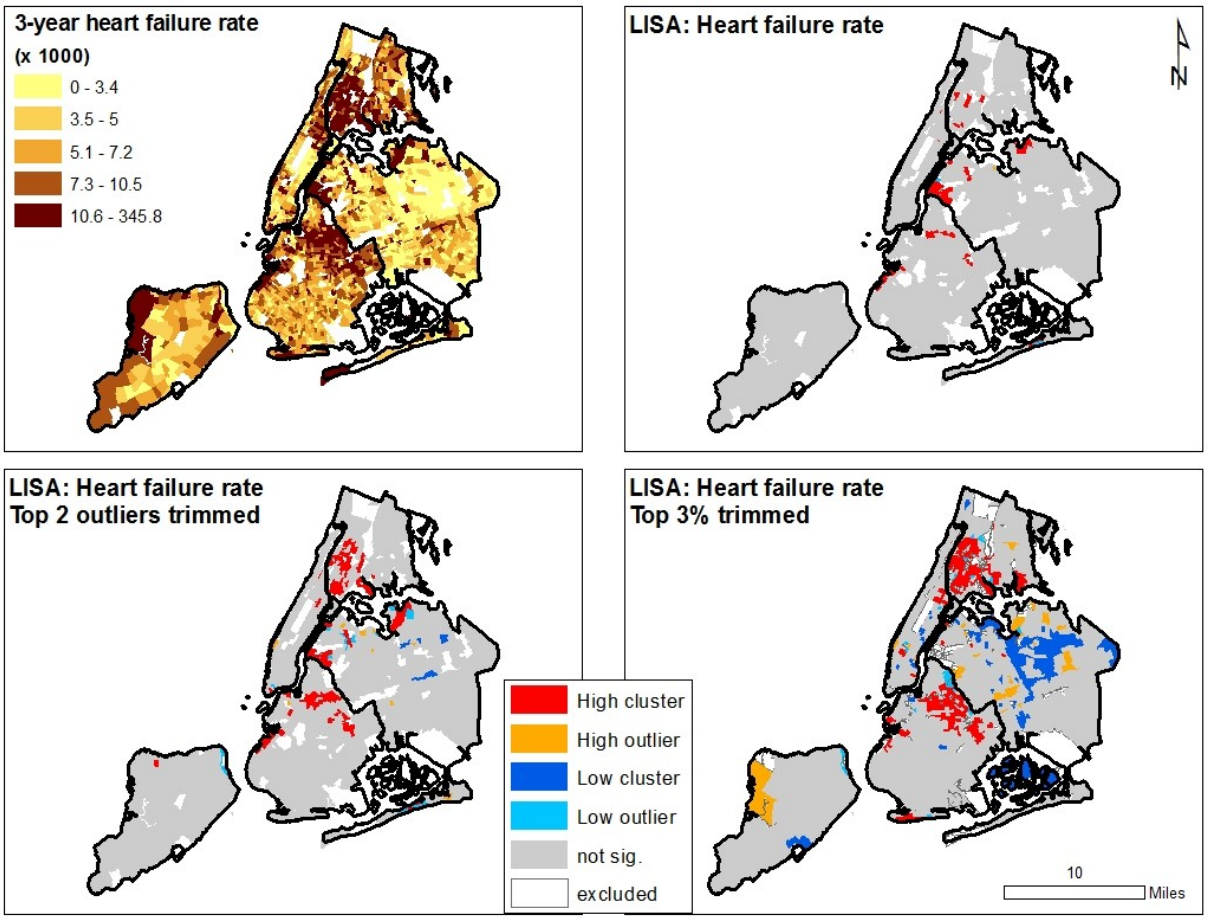


Figure 2-13: Age-adjusted heart failure hospitalization rate by quintiles (top left), LISA with untrimmed data (top right), LISA with top two outliers trimmed (bottom left), and LISA with the highest 3% removed (bottom right). All maps exclude census tracts containing fewer than 415 persons (lowest 5%).

2.1.3 POLLUTION DATA

In this study, a number of different sources for $PM_{2.5}$ data were utilized. These data were then used to calculate proximity (**2.2.1**), processed in an air dispersion model (**2.2.2**), or incorporated into a land use regression model (**2.2.3**). Proximity analyses and air dispersion modeling of $PM_{2.5}$ originating from stationary sources relied upon National Emission Inventory (NEI) data from the USEPA. Air dispersion modeling of mobile sources was based upon annual average daily traffic data from the New York State Department of Transportation (NYSDOT). The land use regression models used data acquired from EPA air quality monitors in NYC as well as results from the air dispersion modeling. Remotely sensed aerosol optical depth data from MODIS were also explored for use as a contributing variable in land use regressions.

2.1.3.1STATIONARY SOURCES / NATIONAL EMISSIONS INVENTORY (NEI)

The NEI data (2002) are the foundation for both the proximity analysis and the air dispersion modeling of $PM_{2.5}$ from major stationary point sources (EPA, 2006a; EPA, 2006b). Created by the USEPA's Emission Inventory Analysis Group (EIAG), they are designed to be a comprehensive inventory of criteria air pollutants (CAPs) and hazardous air pollutants (HAPs) emanating from point sources across the entire country to enable air quality modeling. CAPs, which include SO_2, VOCs, NOx, CO, Pb, PM_{10}, $PM_{2.5}$ and NH_3, are reported in two general groups: Type A (large sources reported annually) and Type B (smaller sources which report every 3 years). The 2002 NEI data used in this study include both Type A and Type B. The main asset of NEI is the inclusion of stack information (diameter, height, gas temperature, etc.), as these data are a necessity when modeling air dispersion.

The emission release points (stacks) were identified by first geocoding the facilities based on their addresses. Then points representing the individual stacks (one or more per facility) were manually placed using high resolution orthophotos as a guide. Google 'street view' and LotInfo (a spatial dataset containing tax lot information which includes land use) were used to confirm the stack placement, or aid in identifying proper locations. When stacks were not able to be visually recognized using these data, they were placed in 'likely' locations within the facility's property lot. In NYC, there were 306 release points (stacks) in the 2002 NEI data (**Figure 2-14**).

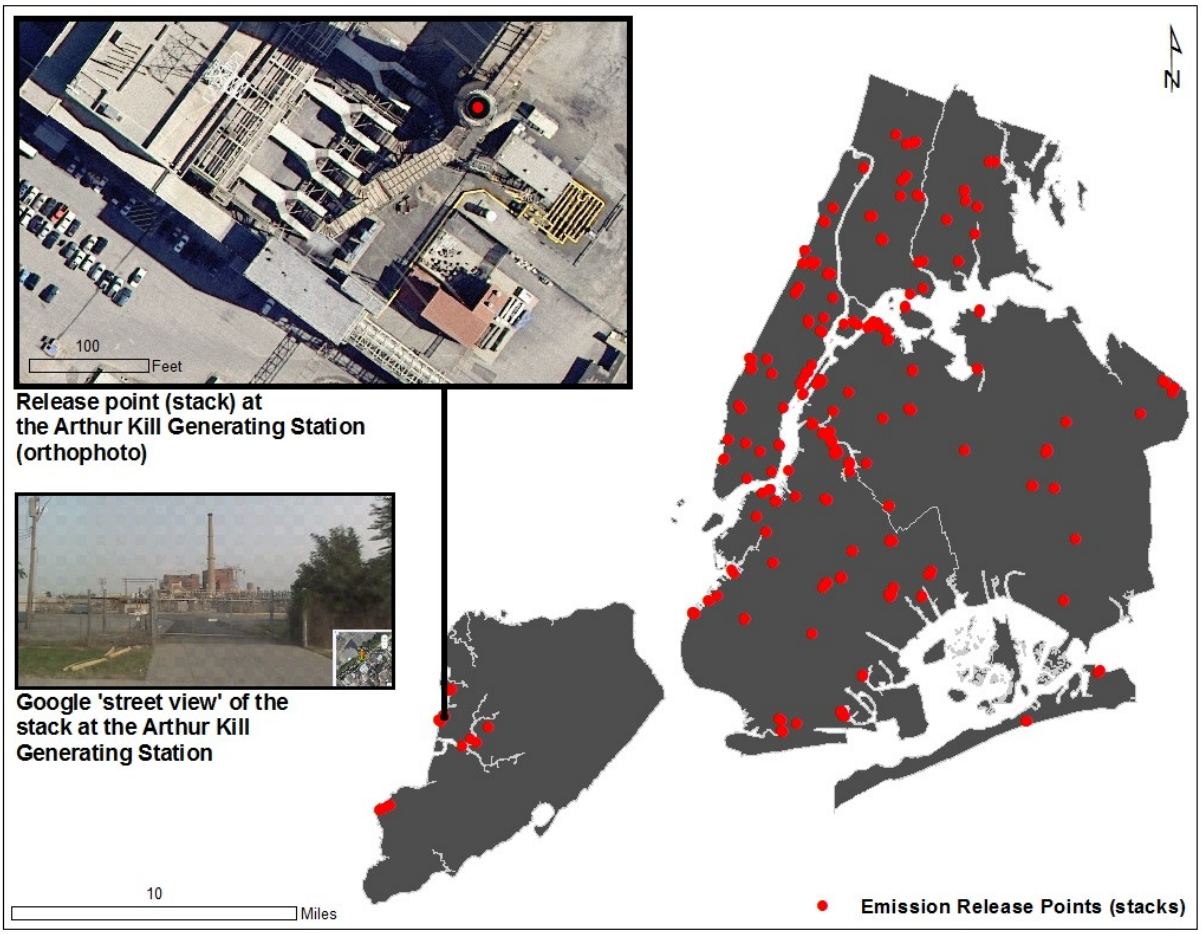


Figure 2-14: Emission release points (stacks) of $PM_{2.5}$ in NYC. The inset maps show how the specific locations of the stacks were identified.

Of the 306 emission release points, 233 stacks were identified as emitting $PM_{2.5}$. The range of average annual emissions rates (in grams per second) is from $7.0*10^{-6}$ g/s to 4.56g/s with a mean of 0.16 and standard deviation of 0.58. The amount of $PM_{2.5}$ emitted per year (or average amount emitted per second) is not distributed evenly across the city (**Table 2-6**).

BOROUGH	$PM_{2.5}$ (g/s)
Bronx	0.87
Brooklyn	3.21
Manhattan	8.91
Queens	22.42
Staten Island	2.44
NYC	*37.85*

Table 2-6: Total grams of $PM_{2.5}$ per second emitted per borough (2002). Data source: USEPA, 2002.

Notice the extremely high value for Queens which is the borough that houses the largest number of major producers of $PM_{2.5}$. This information, however, can be deceiving as a large number of emission points fall very near the border of other boroughs. Therefore, an examination of the spatial distribution of release rates can be informative by using proportional symbols to represent the average annual release rate of $PM_{2.5}$ (**Figure 2-15**). The NEI data suggests that the majority of the large producers of $PM_{2.5}$ tend to be power generating facilities (e.g., Consolidated Edison generating stations), with the smaller producers showing a range in land use including: large institutions (e.g., Bronx Zoo, City College, and St. Mary's Hospital), housing complexes (e.g., Amalgamated Housing, Tracey Towers, and Co-op City), and industrial uses (e.g., Poly Plastic Packaging, Grace Asphalt, and Acme Steel Company).

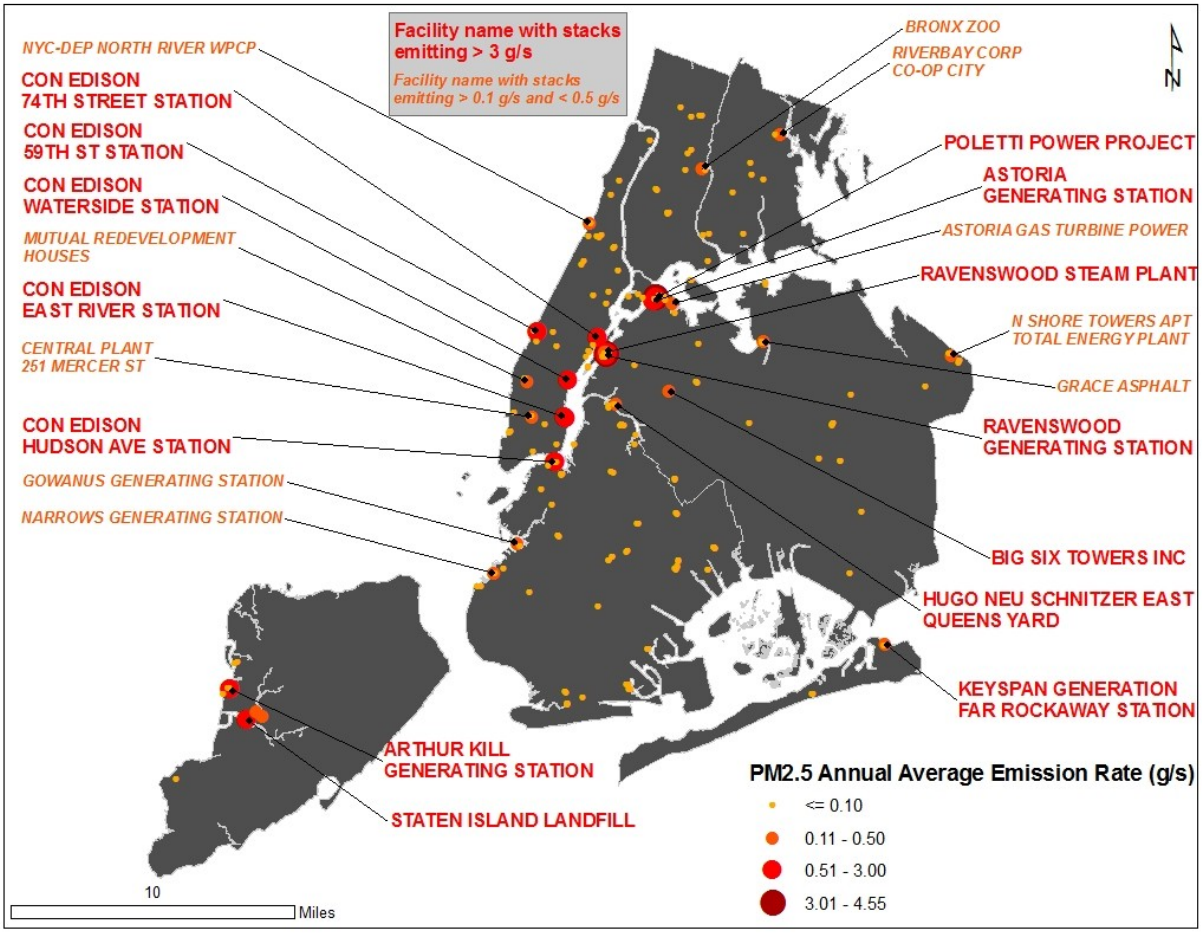


Figure 2-15: Annual average emission rate of $PM_{2.5}$ in NYC (2002). The facilities possessing the largest emission points (>3.0 g/s) are labeled in red. Those that release > 0.10 g/s are labeled in orange. If the same facility possesses multiple stacks that emit >0.10 g/s, only the largest amount is labeled. There are no labels for stacks releasing between 0.51 and 3.0 g/s since they are already accounted for in the >3.0 g/s category in this dataset. Data source: USEPA, 2002.

2.1.3.2 MOBILE SOURCES / ANNUAL AVERAGE DAILY TRAFFIC (AADT)

Mobile sources of $PM_{2.5}$ were estimates using annual average daily traffic (AADT) data for NYC from the New York State Department of Transportation (NYSDOT). These data contain the

number of vehicles per day, averaged over the year, for 322 street segments in the city (**Figure 2-16**).

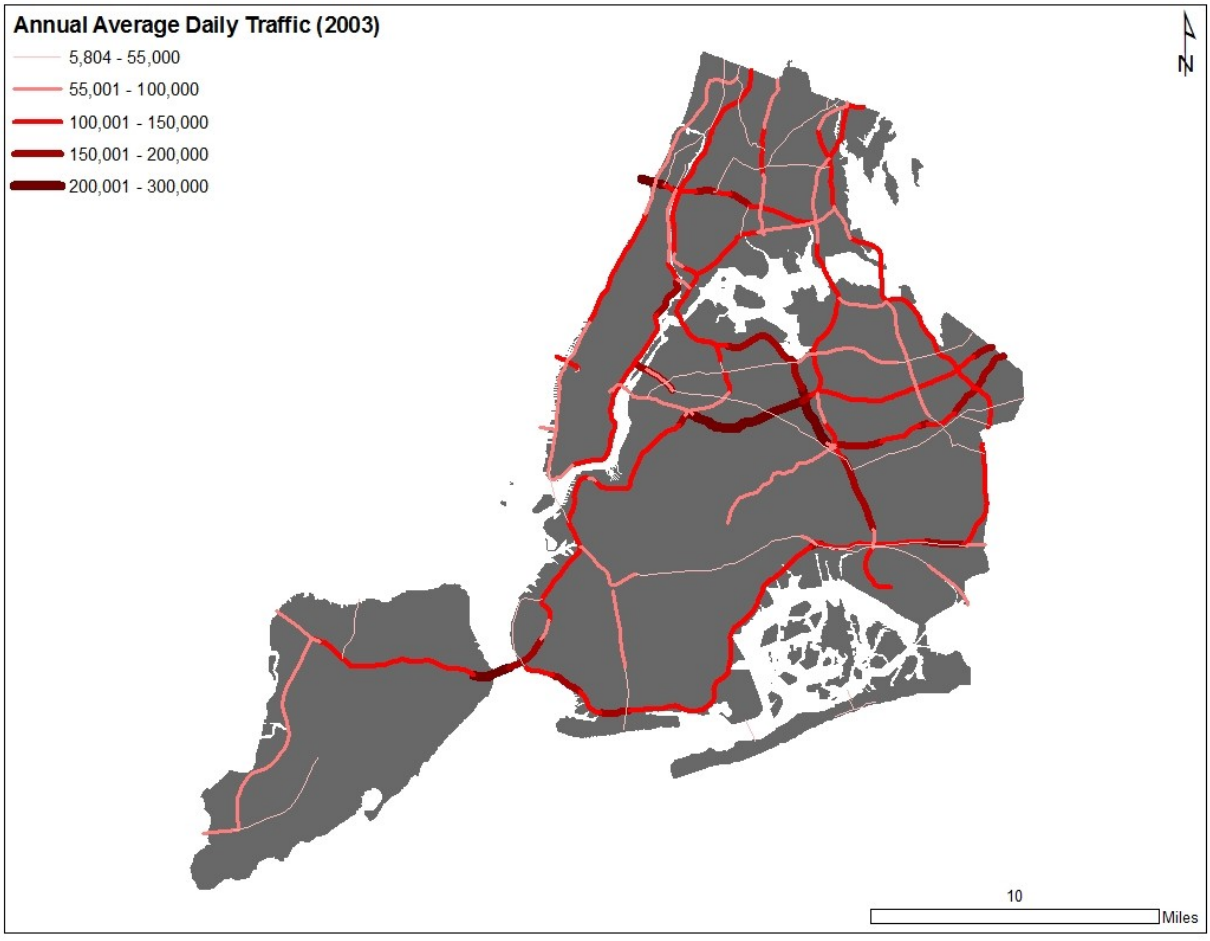


Figure 2-16: Annual average daily traffic in NYC, 2003. Data source: NYSDOT, 2003

The AADT data can then be joined with outputs from the MOBILE6.2 model provided by NYSDOT as well (NYSDOT, 2008). These outputs contain "emission factors," which are information regarding the amount of grams per mile that would be created based on default ratios of types of non-idling vehicles based on road type. For instance, in the Bronx during 2003 on a class 14 road, of all the vehicles present, approximately 59.3% would be light-duty gasoline

vehicles (LDGV) and 20.8% would be light-duty gasoline trucks less than 6000 lbs with a loaded vehicle weight between 3,751 and 5,750 lbs (LDGT2). There are 27 vehicle classes in all, which when combined, provide the information needed to calculate $PM_{2.5}$ emission in grams per day by multiplying the emission factor with the AADT and length for each street segment (**Equation 2-9**). This can then be converted to grams per second by dividing by 24*60*60 (hours, minutes, and seconds). Again, mapping the data can be informative (**Figure 2-17**).

$$\mathbf{AADPM_{2.5} = AADT * EF * L_s} \qquad \textbf{Eq. 2-9}$$

Where:
$AAPM_{2.5}$ = annual average daily $PM_{2.5}$ (grams) for the street segment
AADT = annual average daily traffic of the street segment
EF = $PM_{2.5}$ emission factor (g/mile) for the street segment
L_s = length of the street segment

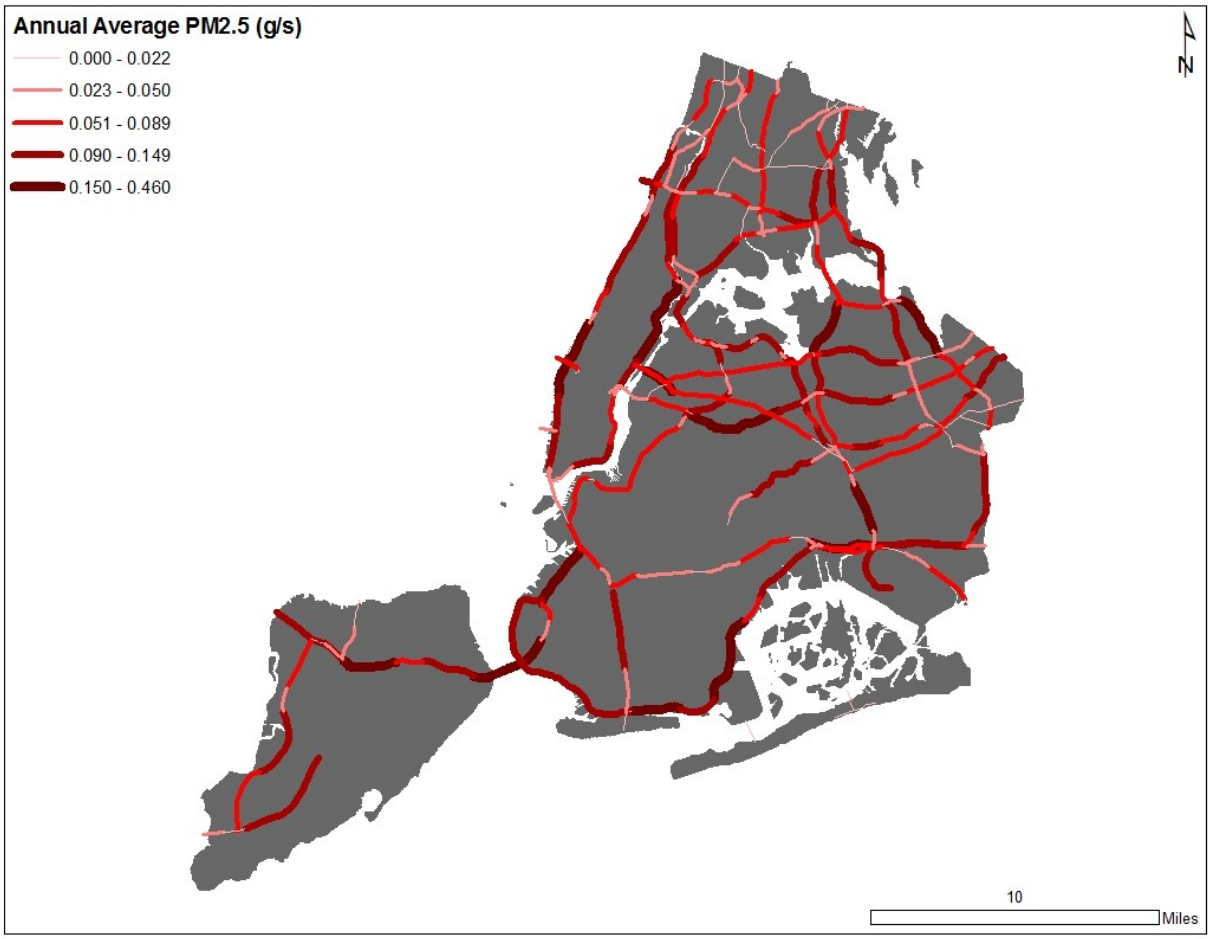


Figure 2-17: $PM_{2.5}$ emission rate in grams per second based on AADT and emission factors from MOBILE6.2

These data can now be used as inputs to an air dispersion model for processing. The main limitation that makes this information somewhat suspect is the fact that the majority of roads, many of them quite busy with heavy truck traffic, are not included in the dataset. It seems that most of the roadways in the AADT data are limited access highways (LAHs), and that most major truck routes (MTRs) are not considered (**Figure 2-18**).

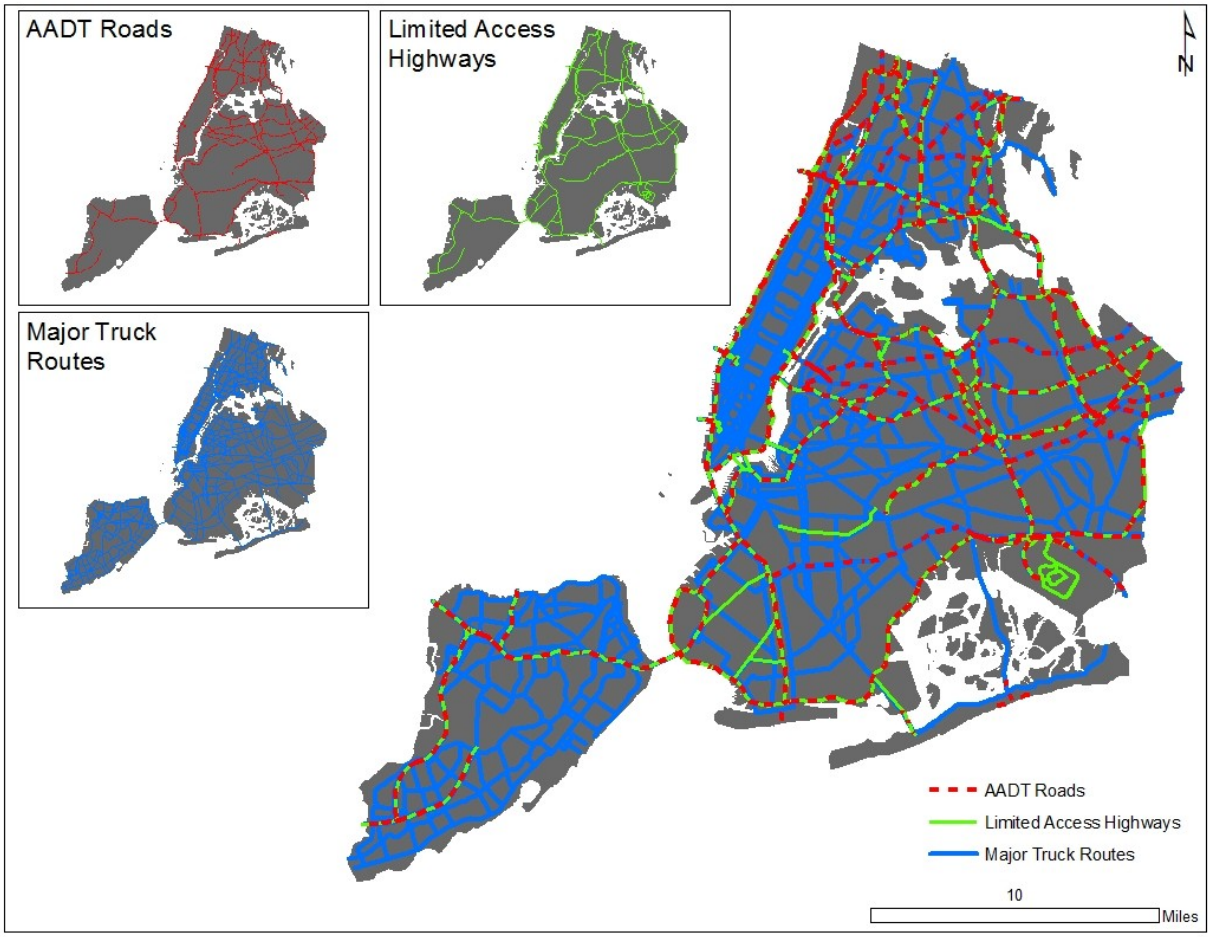


Figure 2-18: Comparison of AADT roads, limited access highways, and major truck routes in NYC.

2.1.3.3 AIR QUALITY MONITORS

The final source of $PM_{2.5}$ data used in this study is the annual average concentration (2002) from the EPA air quality monitors in NYC. These data are queried from the Air Quality System (AQS), which contains ambient air pollution data collected by EPA, state, local, and tribal air pollution control agencies around the country. The monitors in NYC are considered SLAMS (State and Local Air Monitoring Stations), as opposed to being associated with national or other

programs. During 2002, there were only 15 discrete monitors in NYC collecting $PM_{2.5}$ data (excluding monitors with incomplete data or whose purpose is to quality control collocated monitors). They have a range of monitored $PM_{2.5}$ concentrations between 11.5 and 15.9, with a mean of 14.0 and standard deviation of 1.15. As there are so few monitors, they can be easily described in a table (**Table 2-7**). Note the high values (> 15 $\mu g/m^3$) in Manhattan (Site IDs 56, 62, and 128) and the Bronx (Site ID 80).

Borough	Site ID	Address	Land Use	Latitude	Longitude	$PM_{2.5}$
Bronx	80	MORRISANIA CENTER, 1225-57 GERARD AVE	residential	40.83606	-73.92009	15.34
Bronx	83	HARDING LAB, 200TH ST AND SOUTHERN BLVD	commercial	40.86585	-73.88083	13.45
Bronx	110	IS 52, 681 KELLY ST	residential	40.81618	-73.90200	14.25
Brooklyn	52	PS 314, 330 59TH ST	residential	40.64182	-74.01871	13.95
Brooklyn	76	PS 321, 180 7TH AV	residential	40.67185	-73.97824	13.28
Brooklyn	122	JHS 126 424 LEONARD ST	industrial	40.71961	-73.94771	14.05
Manhattan	56	PS 59, 228 E. 57TH STREET, MANHATTAN	commercial	40.75912	-73.96661	15.88
Manhattan	62	POST OFFICE,350 CANAL STREET	commercial	40.72052	-74.00409	15.42
Manhattan	79	IS 45, 2351 1ST AVENUE	residential	40.79970	-73.93432	14.12
Manhattan	128	PS 19, 185 1ST AVENUE	commercial	40.73000	-73.98446	15.63
Queens	94	PS29, 125-10 23RD AVE	residential	40.77798	-73.84318	13.31
Queens	96	3115 140TH STREET	commercial	40.77039	-73.82841	13.12
Queens	124	QUEENS COLLEGE, 65-30 KISSENA BLVD, LOT#6	residential	40.73614	-73.82153	12.77
Staten Island	55	POST OFFICE, 364 PORT RICHMOND AVE.	residential	40.63307	-74.13719	13.79
Staten Island	67	SUSAN WAGNER HS, 1200 MANOR RD	residential	40.59664	-74.12525	11.52

Table 2-7: EPA $PM_{2.5}$ monitors in NYC, 2002. Data source: USEPA, 2002.

The spatial distribution of the monitors themselves is extremely important, as a lack of coverage and low number of "samples" can create problems in exposure assessment or land use regression modeling. The monitor locations do not appear to have an unbiased distribution, with sparsely distributed monitors across the "hinterlands" of NYC. When the monitored $PM_{2.5}$ concentrations ($\mu g/m^3$) are mapped it can be seen that there are higher levels in Manhattan and the South Bronx (values > 15$\mu g/m^3$) than in the remainder of the city (**Figure 2-19**). Although this is interesting,

by themselves the monitored values do not provide enough information to conduct intra-urban spatial analyses regarding the relationship of $PM_{2.5}$ concentration with either socio-demographics or health outcomes.

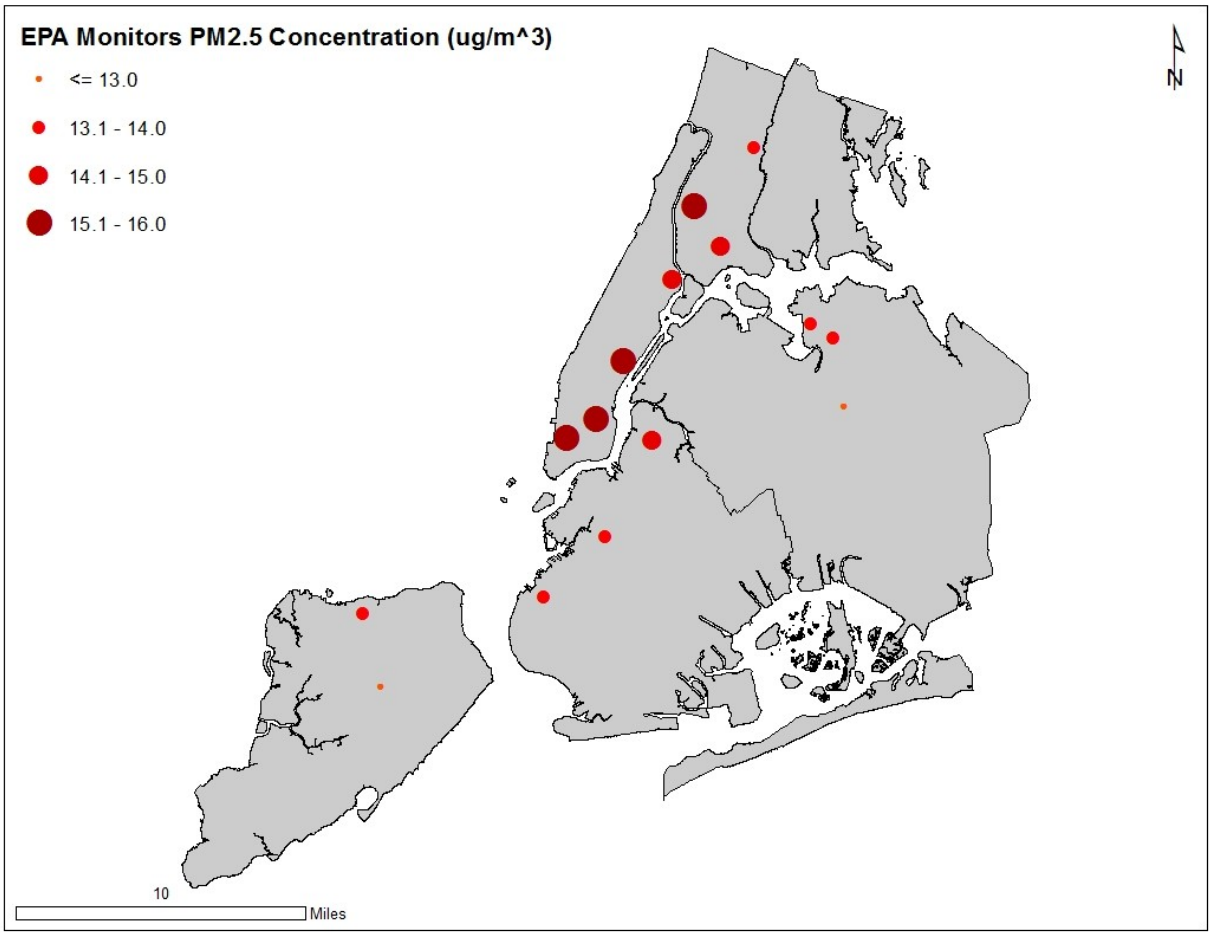


Figure 2-19: EPA monitor locations and measured $PM_{2.5}$ concentrations (ug/m3) in NYC, 2002. Data source: USEPA, 2002.

2.1.3.4 REMOTE SENSING

Remote sensing (RS) could be a tremendous asset to health geography as remotely sensed images are updated frequently and cover large, often not easily accessible, areas. It would hypothetically be possible to use RS data to estimate exposure by incorporating GIS population and health data. Unfortunately, RS data are often unreliable over urban areas due to high spatial and spectral variability, as well as the irregular size, shape and orientation of objects (Nichol and Wong, 2009). Additionally, there is a scalar mismatch in the definition of "high spatial resolution" between remote sensing scientists and health geographers. An area like NYC may be viewed as very small and homogeneous by remote sensing standards, whereas it is often viewed extremely large and heterogeneous by the standards of health geographers.

In order to explore the potential utility of RS data for the spatial analysis of $PM_{2.5}$ exposure, Aerosol Optical Depth (AOD) from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument was processed and examined. AOD, or τ, quantifies the amount of light that is prevented from being transmitted through the atmosphere due to aerosols absorbing or scattering the solar energy (NASA, 2009). This reveals another major limitation to RS data for health research, namely that AOD estimates the particulate matter in the entire atmospheric column, rather than just the aerosols near the ground that have the greatest potential to influence human health.

The majority of the processing was done by Dr. Tarrendra Lankhashar at City College of New York, CUNY following procedures outlines in Hu (2009) and Hu and Rao (2009). One year of

AOD data (2002, AOD at 0.55 µm, "collection 5") was queried and re-sampled to a .04 degree resolution covering a study area between 40^0 to 42^0 latitude and -71^0 to -75^0 longitude, which includes NYC. The MODIS instrument is aboard two spacecraft, Terra and Aqua; both were used in order to gather the greatest amount of information possible. It is important to note that not all measurements were recorded at the same time of day. Some days have multiple measurements and some have no measurements. When the data are examined in a times series (averaged over all points within the study area), the typical seasonal variations can be seen with higher values in the summer (**Figure 2-20**).

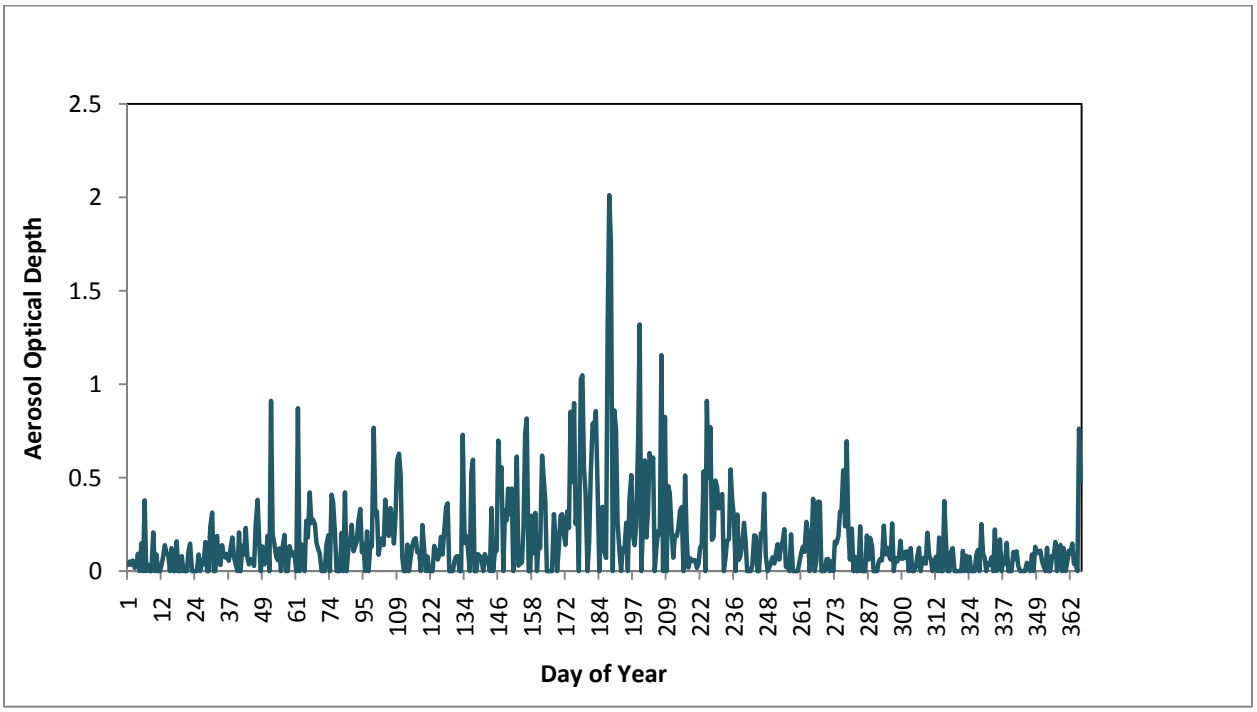


Figure 2-20: MODIS Aerosol optical depth over the greater NYC area showing higher values over the summer season.

When the data are aggregated to the year, they can be mapped and examined spatially. Although believable patterns of AOD can be seen over the greater metropolitan region (**Figure 2-21**), it can also be noted that there is poor spatial resolution over NYC-proper, and the data does not correlate particularly well with the ground-based $PM_{2.5}$ concentrations as measured by the EPA monitors (**Figure 2-22**). This correlation trouble can be seen more starkly with the lack of significance with bivariate Pearson correlation (-.057, p=.839), Spearman correlation (-.125, p=.657), and scatter plot (**Figure 2-23**). Note that these correlations did not improve when an attempt was made to add more variability to the AOD data by interpolating values between raster cell centroids.

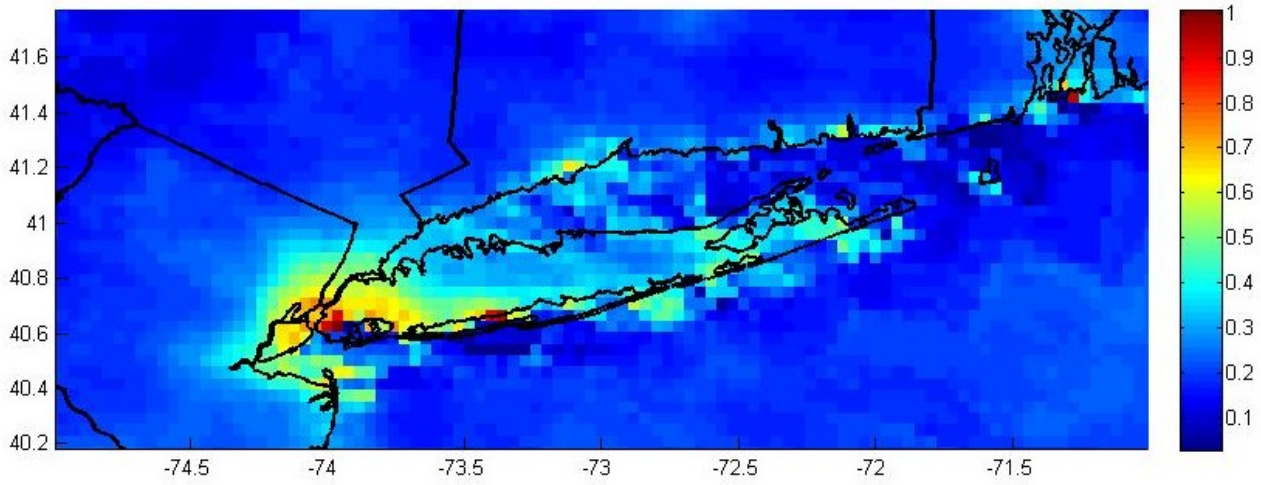


Figure 2-21: MODIS AOD (2002) over the greater NYC area (40: 42 latitude, -75: -71 longitude). Higher AOD values indicate more aerosols in the atmosphere.

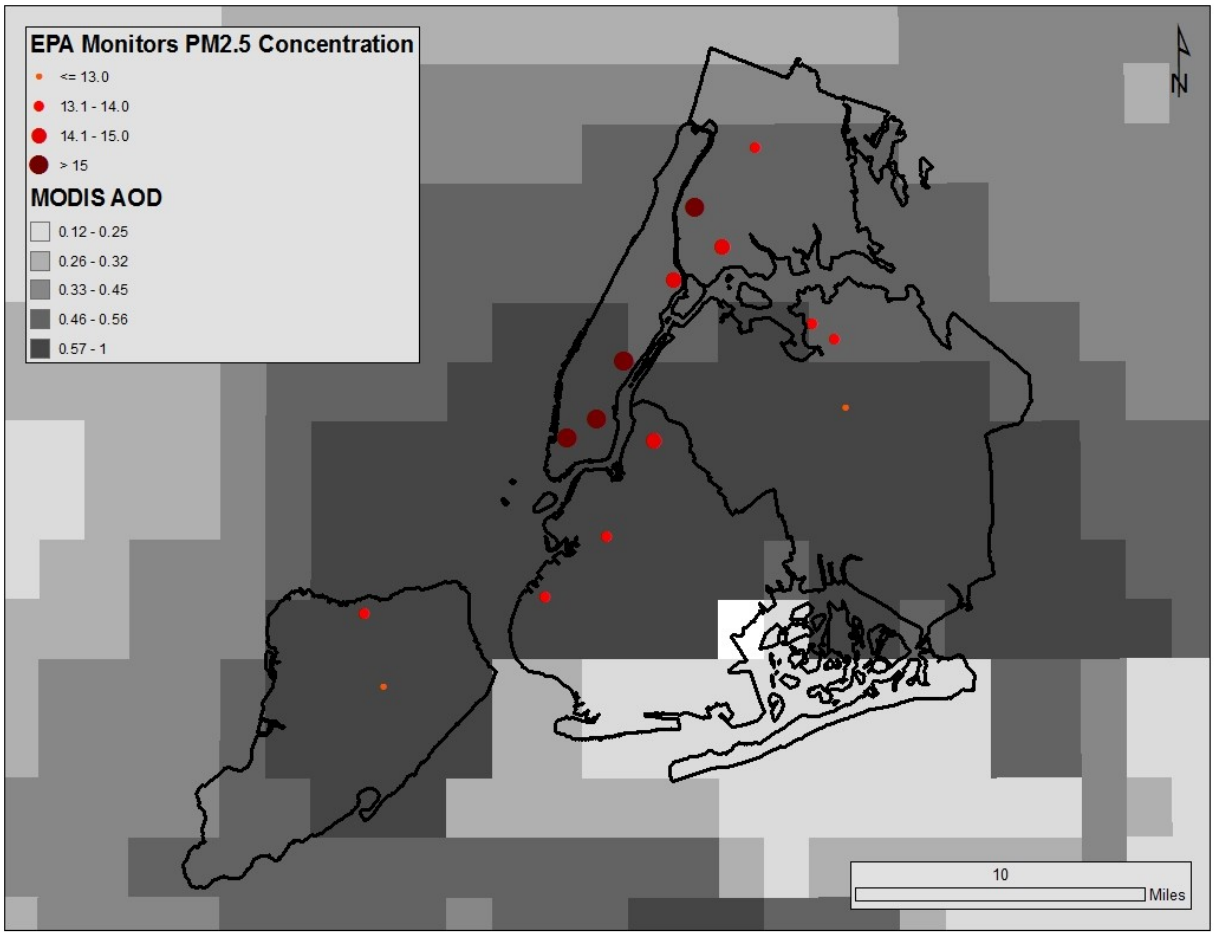


Figure 2-22: MODIS AOD (2002) vs. EPA Monitors in NYC.

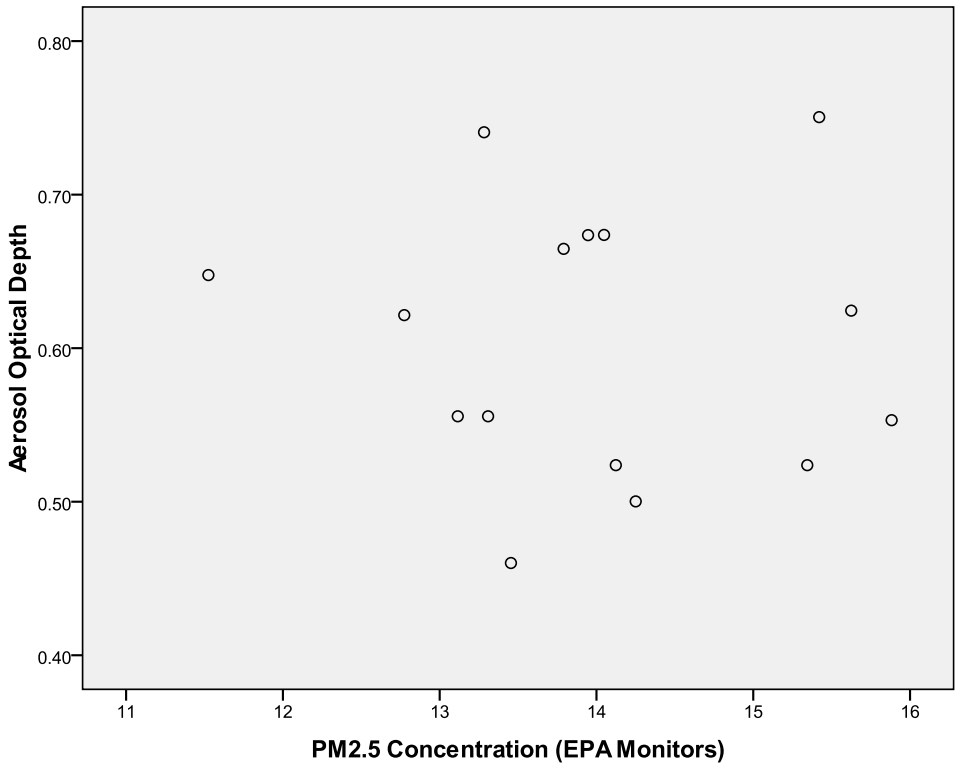


Figure 2-23: Scatter plot of MODIS AOD (2002) vs. EPA Monitors (2002) in NYC.

These limitations, and the lack of correlation in the study area between MODIS aerosol optical depth and monitored near-ground-level $PM_{2.5}$ values, unfortunately rendered the remotely sensed data unusable in this study. There is promise for the utilization of this type of data for both longitudinal exploration or when examining a larger study area that could use larger units of aggregation where small-area fluctuations in $PM_{2.5}$ estimates would not severely affect the study outcomes (e.g. a state-wide or regional analysis).

2.2 EXPOSURE ESTIMATION

One of the main foci of this work is to compare three different methods for estimating pollution exposure. These are (1) proximity, (2) air dispersion modeling, and (3) land use regression. This part of the overall study is extremely important, as the estimation of the effects of human interaction with $PM_{2.5}$, in terms of either environmental justice or environmental health, are the goals of the study. These three methods, and variations of them, range from very simple and quick (e.g. proximity) to extremely complex, time consuming, data-hungry, and processor intensive (e.g. air dispersion modeling). The outputs also represent different metrics, with proximity analysis employing a simple distance threshold from known pollution sources, air dispersion modeling estimating $PM_{2.5}$ concentrations from specific sources, and land use regression estimating ambient $PM_{2.5}$ concentrations as a function of selected land uses. As each of these methods quantifies somewhat different aspects of pollution exposure, the results of analyses utilizing different methods should be examined with care.

2.2.1 PROXIMITY ANALYSIS

Proximity analysis in this study, as was discussed in the background section (**Chapter 1**), utilizes a fixed distance buffer from known sources of pollution in order to assess human exposure. In this case, the known pollution sources that were analyzed are the release points (stacks) of the National Emission Inventory facilities that release $PM_{2.5}$ (**Section 2.1.3.1**). Note that generally, proximity analysis is done from either the center or the boundary of the facility site, rather than the actual release point location. Since the release point information was already prepared for

other analyses, it seemed prudent to use these more specific data rather than the more generalized facility location. The goal of proximity analysis for this study is to determine the number of people living within ¼ mile from a stack and their socio-demographic characteristics. As was already discussed, proximity analysis can be a very attractive option when conducting this type of study due to its simplicity, speed of processing, and comparatively undemanding data requirements.

A circular buffer with a ¼ mile radius was drawn around each of the stacks, and each tax-lot was identified as either being 'exposed' (i.e. within ¼ mile) or 'unexposed' (i.e. greater than ¼ mile from a stack) using basic geoprocessing functions within ArcGIS. When a tax-lot was transected by a pollution buffer, it was considered 'exposed' only if the geographic center of the lot was within the buffer – otherwise it was defined as 'unexposed' (**Figure 2-24**). This is unlikely to add excessive error or uncertainty since there are over 730,000 populated tax lots in NYC with an average area of 3,870 square feet.

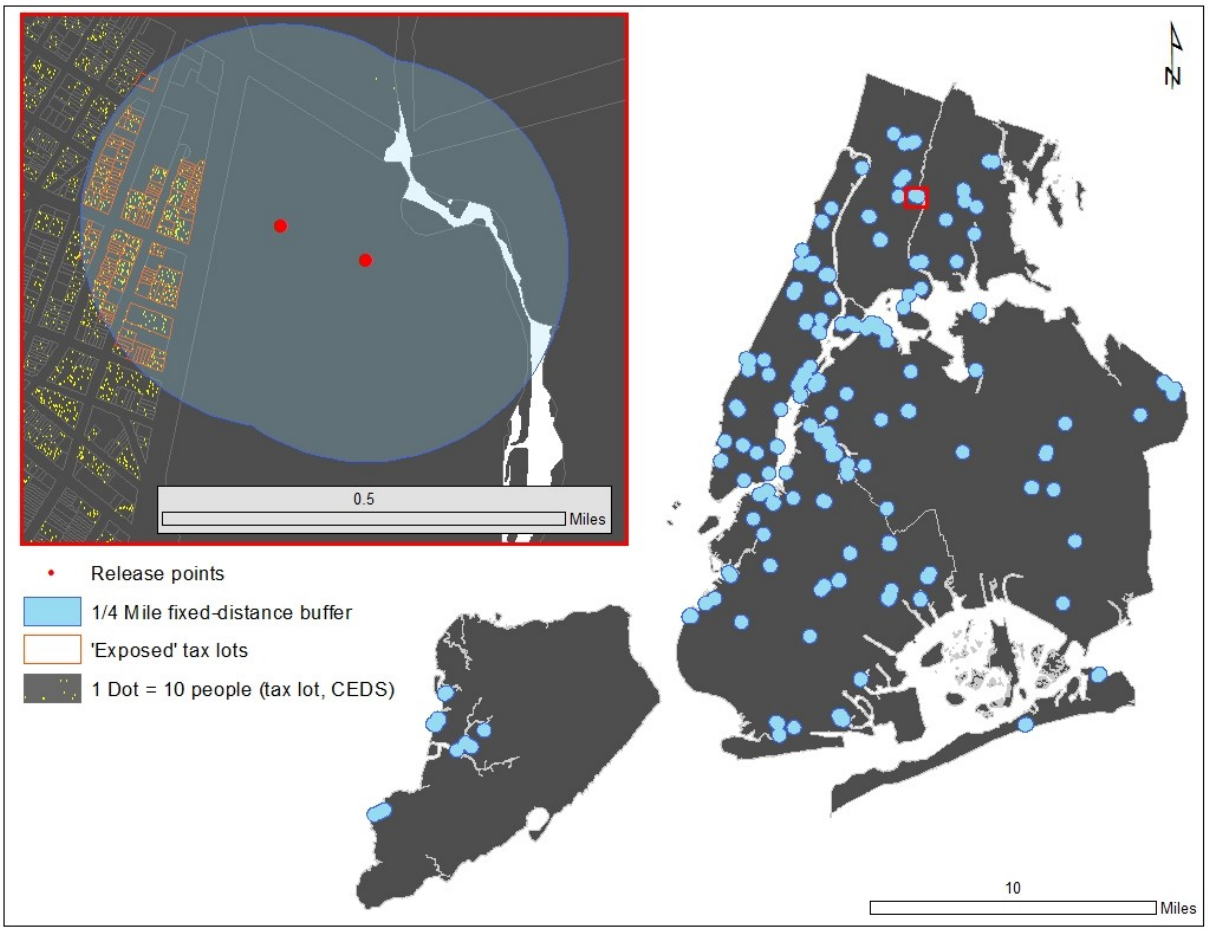


Figure 2-24: Proximity analysis of $PM_{2.5}$ release points and CEDS-derived population. 'Exposed' tax lots that are populated are indicated in orange.

The CEDS-derived socio-demographic characteristics were then attached to the lot-level 'exposure' estimates. The result of this procedure is to have data at the lot level that describes the population characteristics (total population, non-Hispanic white, non-Hispanic black, Hispanic/Latino, poverty status, and adults lacking a high school diploma) and the exposure state (exposed or unexposed). Note that the source population (denominator) of those below poverty and those without a high school diploma are not the same as the total population (a subset of the total population was 'tested' for poverty, and the education variable is for adults older than 25).

This socio-demographic information can then be aggregated (summed) as the total exposed and unexposed populations and sub-populations (**Table 2-9**) that can then be used for environmental justice analyses.

	Exposure	Total Population	non-Hispanic White	non-Hispanic Black	Hispanic / Latino	Below Poverty	Poverty Denominator	No High School Degree	High School Denominator
NYC	unexposed	6,975,182	2,468,173	1,702,838	1,833,589	1,429,917	6,873,500	1,266,710	4,583,169
	exposed	1,006,490	323,262	245,433	318,632	235,920	973,400	187,569	676,563
Brooklyn	unexposed	2,254,609	777,410	774,346	445,637	550,229	2,231,599	441,686	1,419,606
	exposed	208,520	75,341	73,156	41,798	59,932	202,509	41,852	131,839
Bronx	unexposed	1,142,196	168,072	343,322	563,988	344,762	1,117,912	259,453	679,208
	exposed	175,867	23,976	64,098	76,134	49,859	169,076	34,922	105,660
Manhattan	unexposed	1,049,049	513,385	154,203	257,376	194,989	1,023,937	152,315	778,867
	exposed	481,523	186,455	78,448	158,355	101,796	464,353	86,251	343,251
Queens	unexposed	2,087,253	694,737	391,471	513,460	296,156	2,064,903	362,209	1,412,769
	exposed	140,053	37,119	29,716	42,291	24,298	136,938	24,486	95,413
Staten Island	unexposed	442,075	314,569	39,497	53,127	43,780	435,149	51,047	292,719
	exposed	526	370	15	54	36	524	58	400

Table 2-9: Proximity analysis of socio-demographics as 'exposed' (< ¼ mile of an NEI $PM_{2.5}$ release point) and 'unexposed' (>1/4 mile of an NEI $PM_{2.5}$ release point). "Poverty Denominator" and "High School Denominator" fields are the "universe" from which the numbers of those below poverty or without a high school degree are subsets.

More discussion pertaining to the environmental justice ramifications suggested by proximity analysis can be found in the analysis chapter (**Section 3.1.1**).

2.2.2 AIR DISPERSION MODELING

In stark contrast to the simplicity and ease of proximity analysis, air dispersion modeling is a labor intensive, time consuming, and data-heavy endeavor. AERMOD (American Meteorological Society/Environmental Protection Agency Regulatory Model) was used to model the dispersion of $PM_{2.5}$ both the stationary NEI point sources and the mobile annual average daily traffic (AADT). This model operates by taking various environmental variables into

account such as meteorological data, topography and building downwash (structures close enough to the stack to influence the dispersion, **Figure 2-25**). Earlier work has shown the utility of the loose-coupling of AERMOD and ArcGIS in the examination of the relationship between local sources of criteria air pollutants (CAPS) and asthma hospitalizations in the Bronx, NY (Maantay et al., 2009).

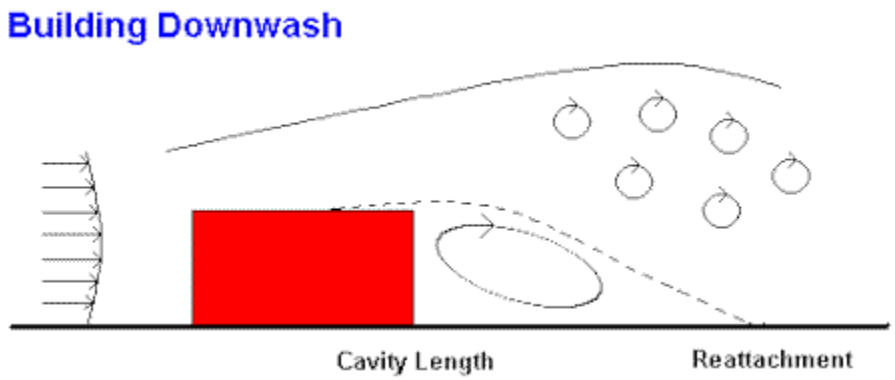


Figure 2-25: source: http://support.lakes-environmental.com/FAQ/definitions.html

2.2.2.1 NATIONAL EMISSIONS INVENTORY (NEI) POINT SOURCES

Stacks were defined and plotted for the NEI sources and the emission parameters (e.g. emission rate and gas temperature) assigned to their geographic locations as described in **Section 2.1.3.1** and imported to the AERMOD software (**Table 2-10**). Note that some of the data for AERMOD were not directly from NEI, but had to be derived. For instance, NEI only provides tons/year of $PM_{2.5}$ emission, whereas AERMOD requires grams/second. This presents one of the major limitations of the data: the assumption that the pollution is being emitted at a constant rate over the entire year. Although this assumption is clearly untrue, the state of the data does not allow a more precise temporal measurement.

Srcid	Source ID (up to 8 characters)
Srctyp	Source type (point for NEI data).
Xs	X coordinate of the source location in meters
Ys	Y coordinate of the source location in meters
Zs	Source elevation location in meters or feet.
Ptemis	Point emission rate in g/s.
Stkhgt	Release height above ground in meters.
Stktmp	Stack gas exit temperature in degrees K.
Stkvel	Stack gas exit velocity in m/s.
Stkdia	Stack inside diameter in meters.

Table 2-10: Emission inputs for NEI point sources in AERMOD

Hourly meteorological data for one full year (1999) was acquired from the National Climatic Data Center (NCDC) of the National Oceanic and Atmospheric Administration (NOAA). Surface air information from New York LaGuardia Airport station (ID 147321, 40°47'N / 73°53'W) and upper air data from New York City Station (ID 94703, 40°52'N / 72°52'W) were utilized within AERMOD. The surface meteorological variables include hourly ceiling height, wind direction, wind speed, pressure, temperature, relative humidity, and cloud cover. Upper air data consist of wind direction, wind speed and temperature. The meteorological data were preprocessed to in order to be usable within ISC-AERMOD by Meteorological Solutions Inc.

Shuttle Radar Topography Mission (SRTM) elevation data, at 30m resolution, were imported directly into AERMOD and processed within the software. The elevation data are utilized not only for the values themselves, but also to modify base elevations of buildings, emission points, and receptors by using the AERMAP module within AERMOD. A receptor grid, which identifies the points where AERMOD will calculate the pollution concentrations, was then created. Although it is relatively simple to create a regular point grid, this proved to be inefficient since as the number of receptors increases, the speed of the model's processing decreases.

Additionally, there is no functional reason to calculate concentrations in areas where there is no population since this analysis is based on locations of residential population vis-à-vis estimated pollution concentrations. This issue was resolved in ArcGIS by first creating a regular grid over the NYC area with one point every 200m (approximately 19,000 points within the NYC land boundary). Next, tax lots within the city that had CEDS-derived populations greater than zero were spatially intersected with the 200m grid. Only those points that intersected the populated lots were saved (6,590 points). As this resulted in large areas with no receptors, a 600m grid was created to 'fill in' the blank regions within the NYC land boundary in order to make graphical representations of the pollution more understandable. This resulted in an irregular point grid consisting of 7,447 points, which is vastly preferred to the original 19,000 points of the regular 200m grid over NYC land area (**Figure 2-26**).

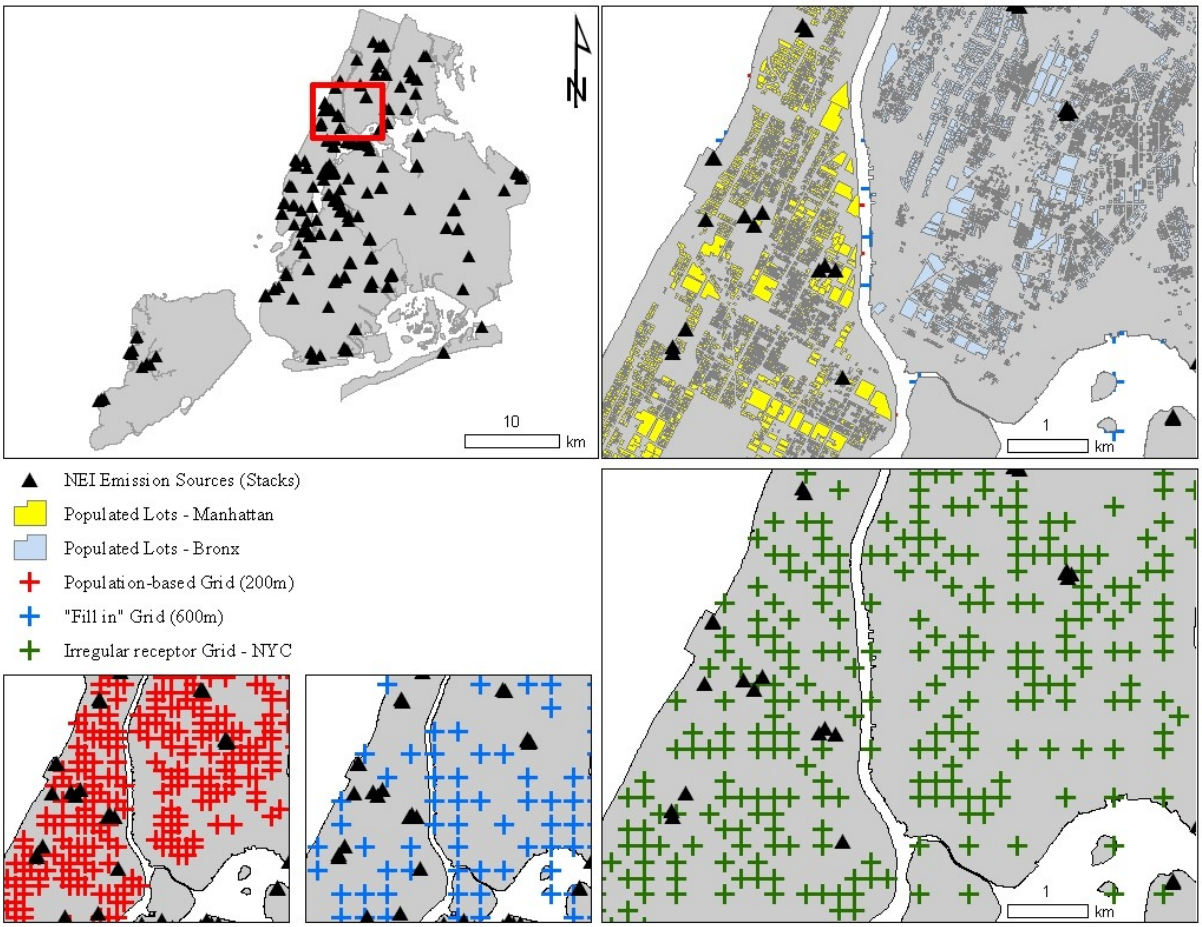
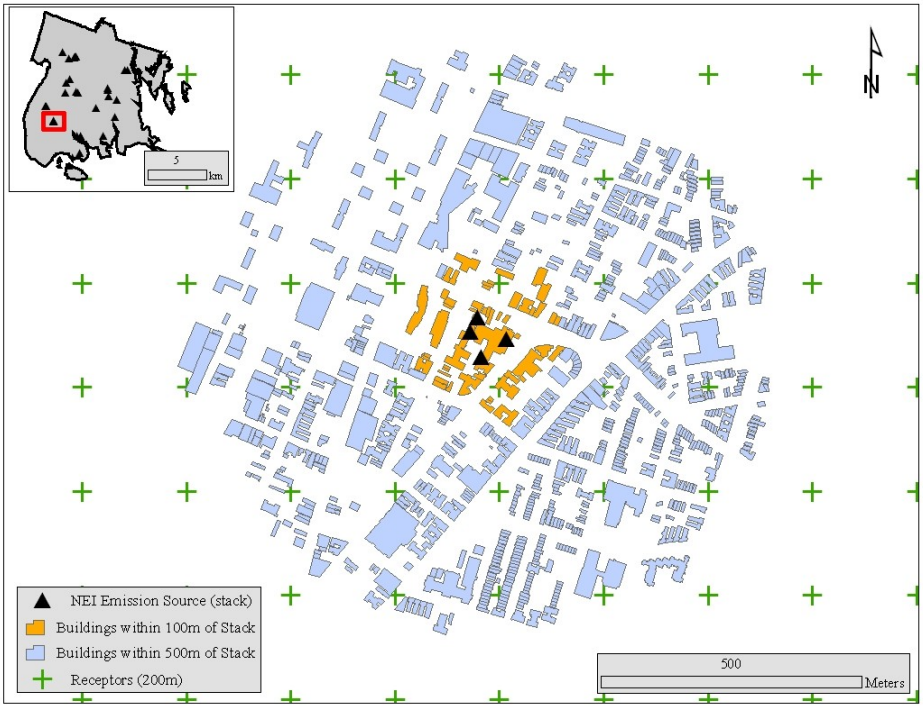


Figure 2-26: NEI stationary point sources (stacks), populated tax lots, and the irregular receptor grid in NYC.

Building footprints and heights were acquired from the New York City Department of Information Technology and Telecommunications (DOITT). The height data were provided as points and were assigned to their respective building footprint via a spatial join in ArcGIS. There was a small amount of error in this process due to points not consistently demonstrating spatial collocation with the building footprints (approximately 2% of buildings were not properly assigned height data). As only buildings in relatively close proximity to emission points are modeled as affecting pollution distribution, it is not necessary to import all structures into

AERMOD. Different distances from stacks were tested to determine the software's ability to process large amounts of data. Both 500m and 100m buffers were created around the stacks and buildings within these proximities were tested (**Figure 2-27a**). Using the 500m buffer drastically slowed, or altogether crashed, the downwash processing. Ultimately, buildings which are within 5*L of the stacks were used, where L = the height of the buildings. This means that if a building is closer than 5x its own height from a stack, it will be included in the model (EPA, 2004) (**Figure 2-27b**).

(a)



(b)

Figure 2-27: Buildings potentially affecting pollution dispersion from NEI stacks (a) within 100m and 500m of stacks; (b) within 500m and 5L of stacks (L = height of building). Data source: USEPA, 2002; DOITT.

After the heights were assigned, the data were then converted to a DXF R12 file (an export format native to AutoCAD required by AERMOD). Unfortunately, ArcGIS is not able to export as an R12 file (DXF R14 is the most similar format it is able to export), and does not retain the attribute data once it is imported into AERMOD as an R14. In the previous Maantay, Tu, and Maroko study, a program called ArcV2CAD 5.0 was used to convert the ArcGIS shapefile to a DXF R12 (Maantay et al., 2009). Although this does successfully retain the height data, it produces unnecessarily complex spatial data that causes AERMOD to bog and crash when working with larger datasets. As such, rather than converting via ArcV2CAD, AutoCAD 2007 was used to convert the R14 data to R12, which resulted in a preservation of the height data and

clean, simple spatial information. Data necessary for calculating building downwash during the final model run was then processed using the Building Profile Input Program (BPIP) analysis module within AERMOD.

As this is very computationally intensive modeling, the study area (NYC) had to be broken down into 13 different sub-areas in order to prevent the computers from crashing (**Figure 2-28**). However, it was necessary to include all the release points city-wide when doing these analyses so that pollution from stacks outside the sub-area boundaries could still be included in the estimates. This method allowed the models to run without crashing, and still take into account all the NYC NEI sources for $PM_{2.5}$. Ultimately, computer processing alone took more than two weeks, excluding data collection and preparation.

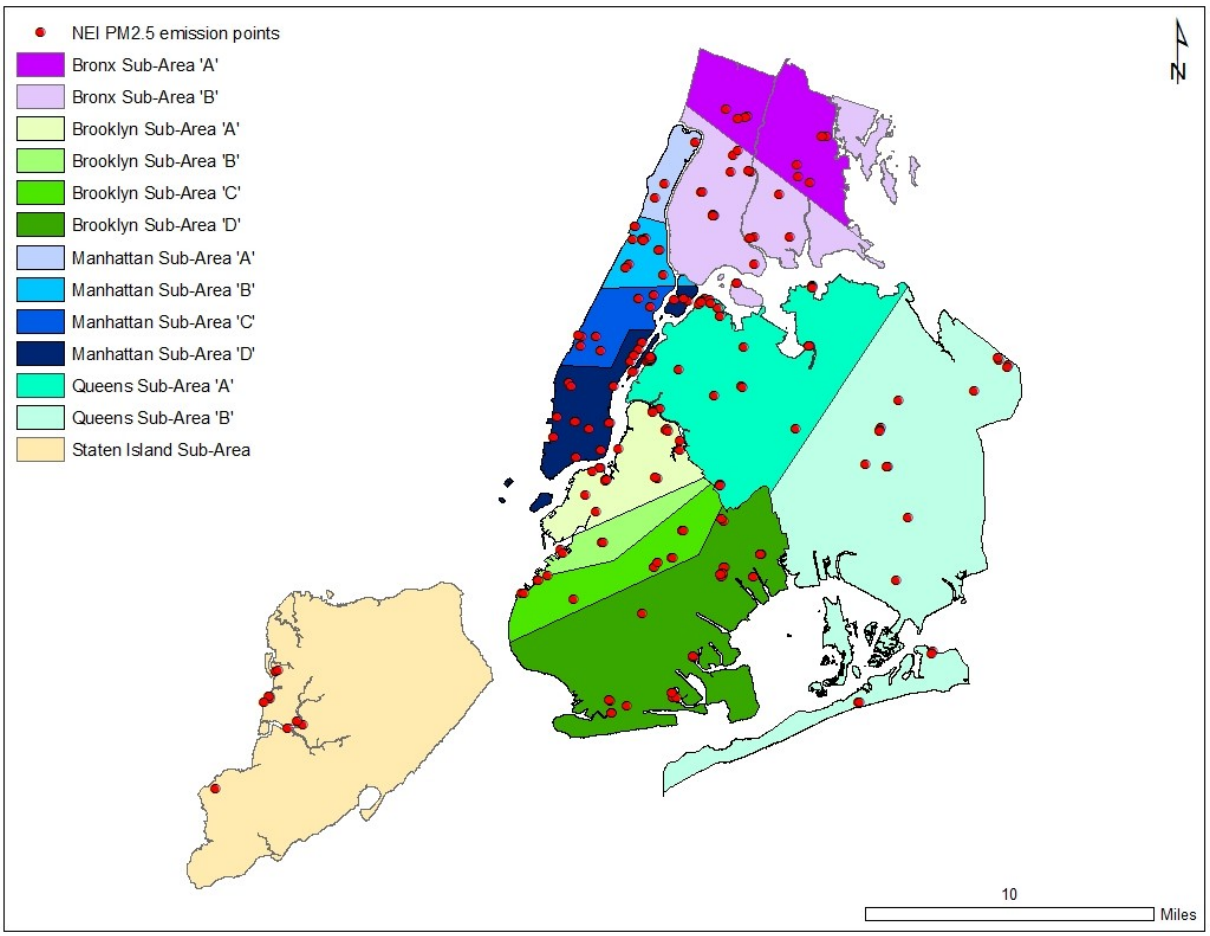


Figure 2-28: Sub-areas delineated for AERMOD processing and NEI $PM_{2.5}$ emission points.

The AERMOD results for $PM_{2.5}$ concentrations originating from NEI point sources are reported as a single estimation for each receptor grid point in each individual sub-area. These values can then be combined so that receptor values across the entire city are in one file (**Figure 2-29**). The values between the points can then be interpolated using inverse distance weighting (IDW), resulting in a continuous surface of modeled $PM_{2.5}$ estimates from NEI sources (**Figure 2-30**). The straight-forward IDW was used as there were enough sample points to belay the requirement

of more robust and complex interpolation techniques. Note that the data in both **Figures 2-29** and **2-30** are classified into groups, but the underlying data is continuous.

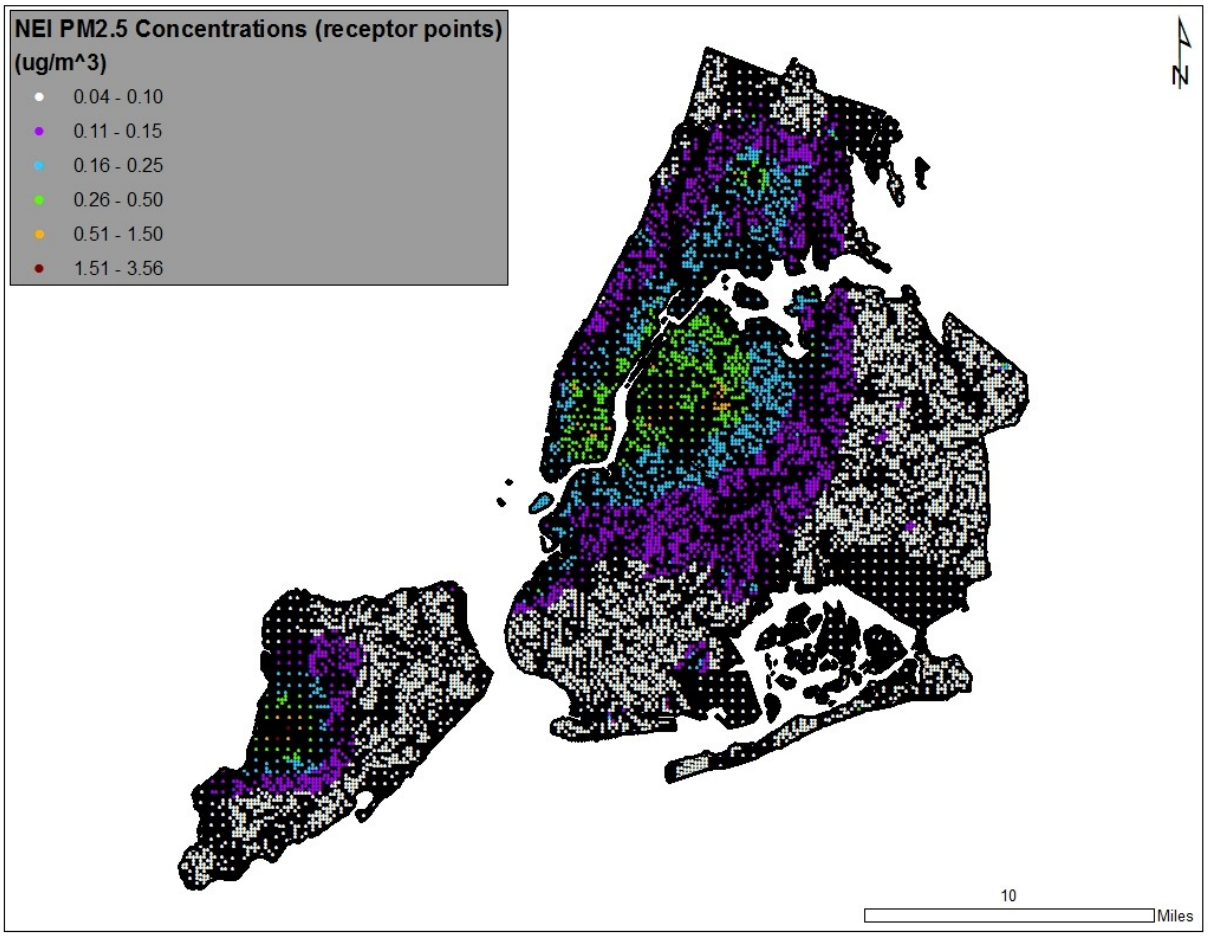


Figure 2-29: AERMOD-estimated $PM_{2.5}$ concentrations from NEI sources at receptor grid points.

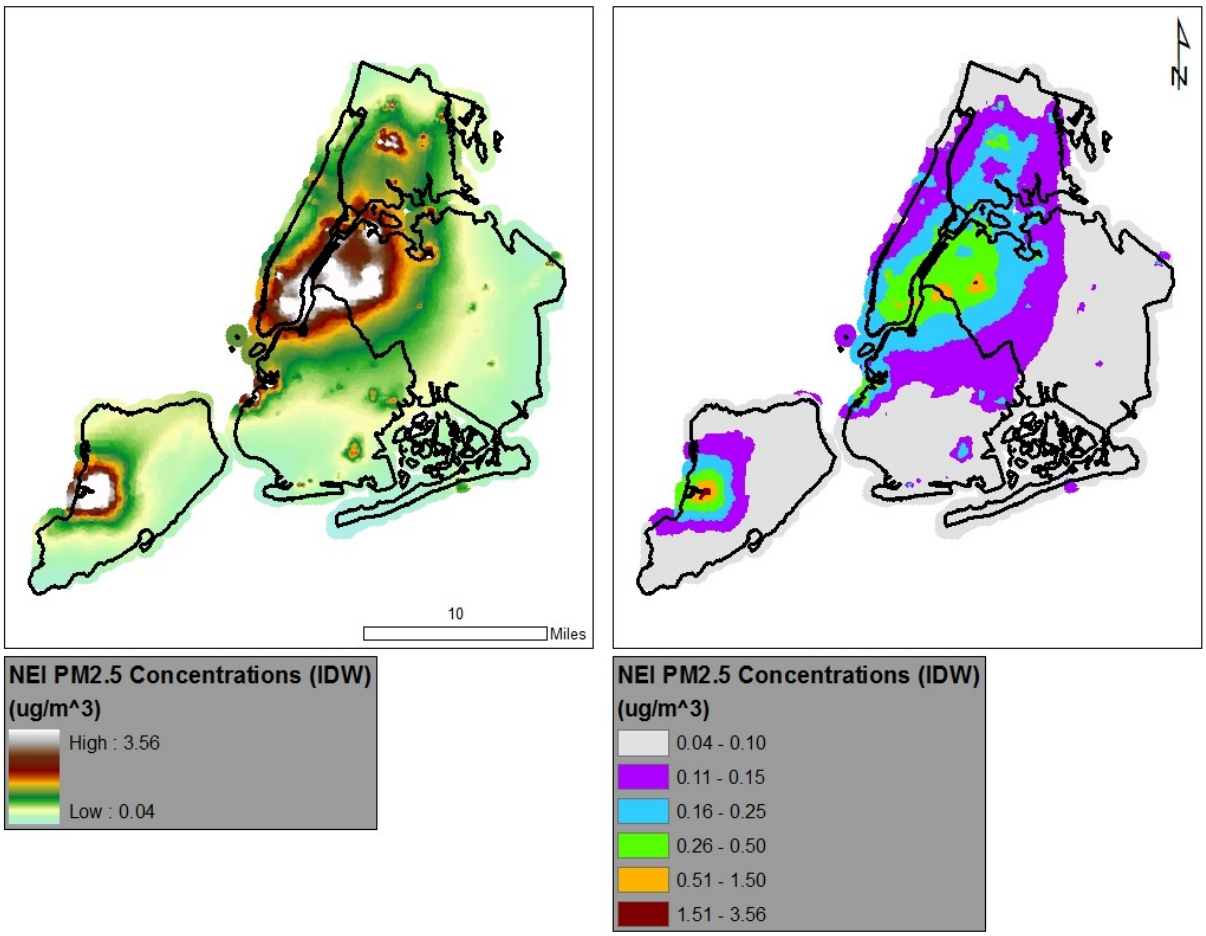


Figure 2-30: AERMOD-estimated $PM_{2.5}$ concentrations from NEI sources interpolated using inverse distance weighting to a continuous surface. Left map shows "stretched" or "unclassified" data. Right map uses a classification scheme identical to **figure 2-29** for comparative purposes.

In order to convert $PM_{2.5}$ concentrations to exposure estimates, it is necessary to join the pollution data with the underlying population data. For environmental justice analysis, the interpolated surface can be spatially linked to the tax-lot level CEDS demographic data within ArcGIS by identifying the concentration value at the centroid of each populated tax lot (**Figure 2-31**). For the health outcome analyses, the $PM_{2.5}$ concentration estimates must be coupled to the

much coarser census tract to match the unit of aggregation of the health data (**Figure 2-32**). This is done by calculating the mean of the interpolated values within the tract boundary.

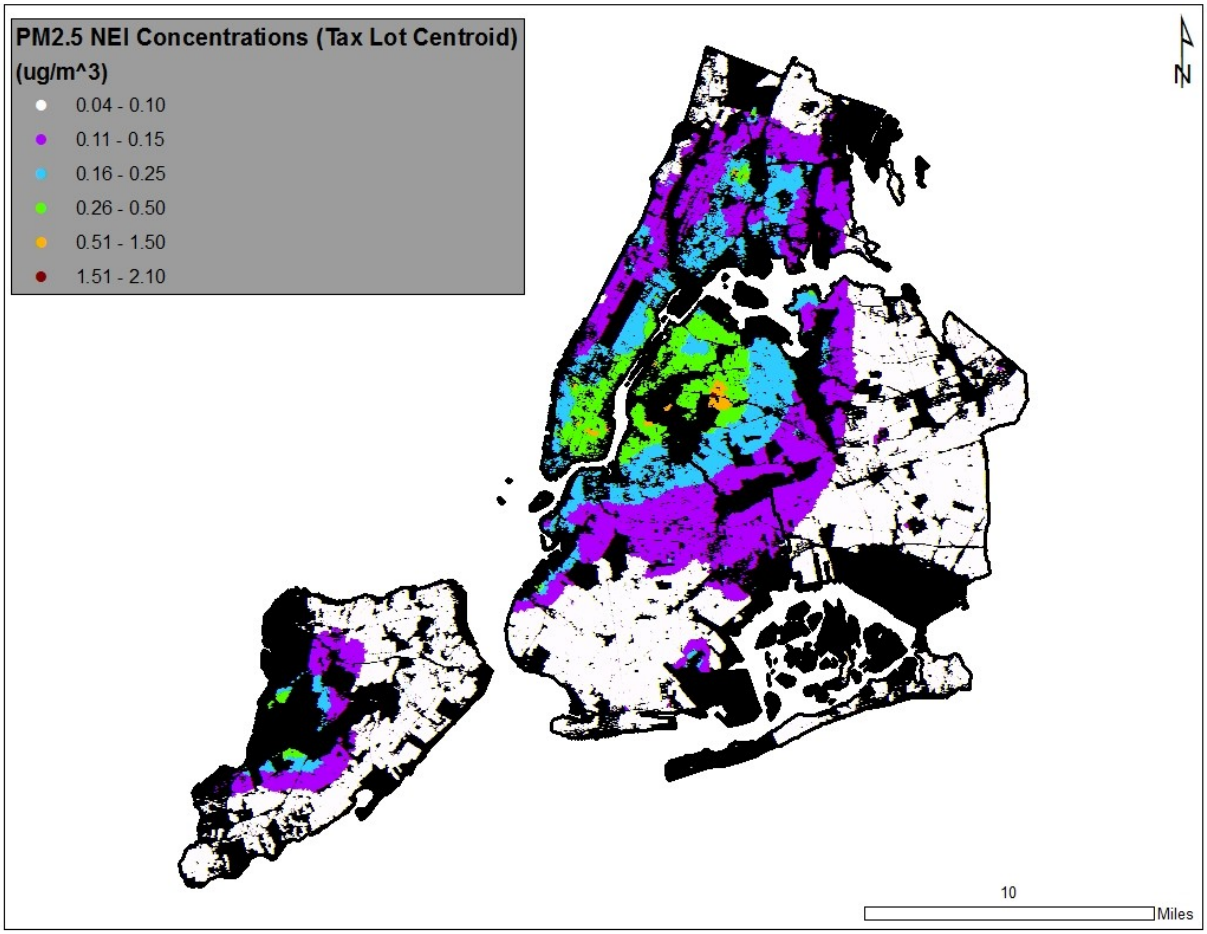


Figure 2-31: AERMOD $PM_{2.5}$ concentration estimates from NEI sources spatially joined to populated tax-lot centroids in order to match the CEDS socio-demographic data (tax lots with no estimated population were not included). Note that the classification scheme for this map is identical to **figures 2-29** and **2-30**.

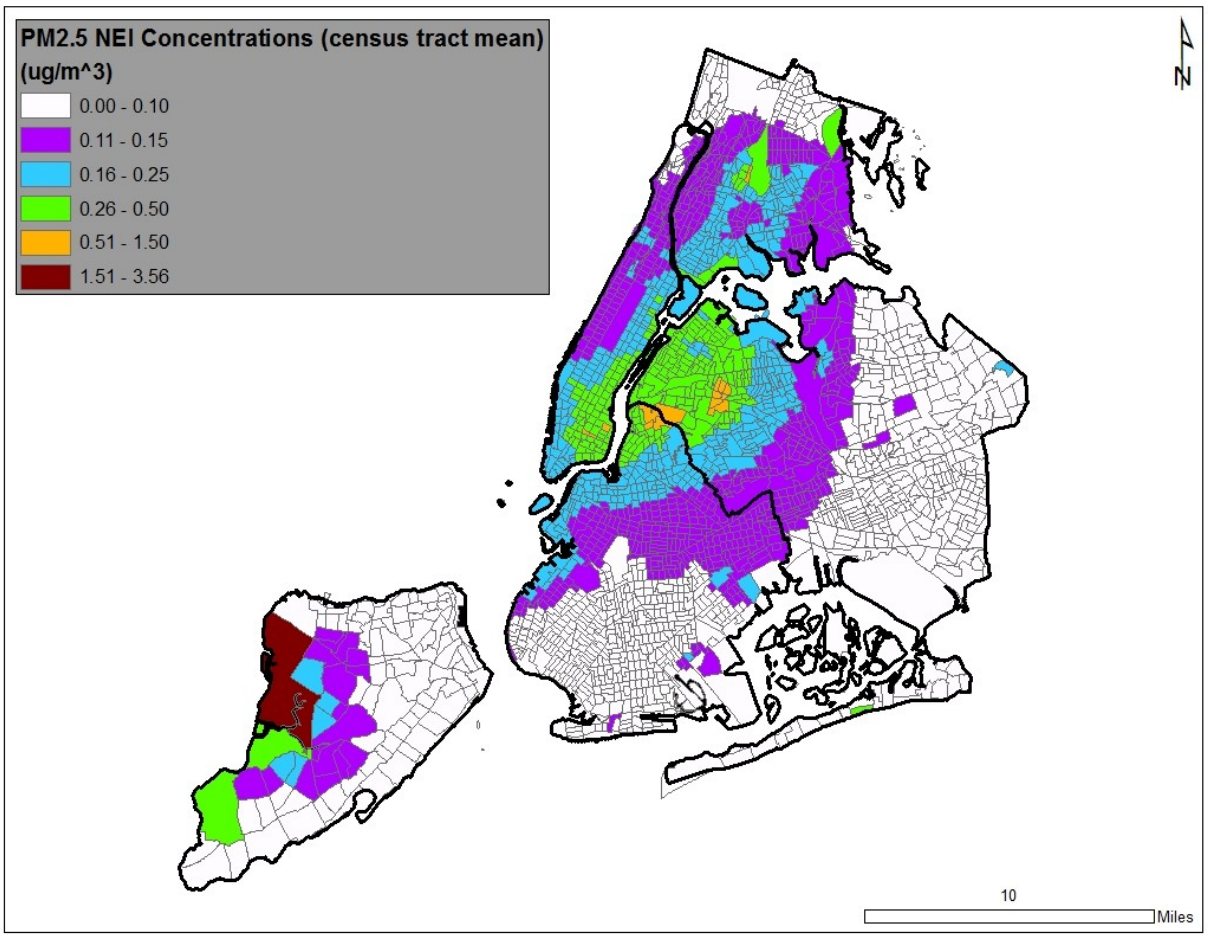


Figure 2-32: AERMOD $PM_{2.5}$ concentration estimates from NEI sources spatially joined to census tracts in order to match the heart failure hospitalization data. Note that the classification scheme for this map is identical to **figures 2-29** through **2-31**.

2.2.2.2 AVERAGE ANNUAL DAILY TRAFFIC (AADT) MOBILE SOURCES

As was stated in **Section 2.1.3.2**, the mobile source dispersion modeling proved not to be as useful as the NEI stationary source modeling for this dissertation due to a lack of data pertaining to major truck routes and other highly trafficked roadways. As such, I will not go into as much detail regarding the modeling of AADT, but rather summarize the major properties of the method.

The main difference between mobile modeling and stationary point source modeling is that line segments (roads) rather than points (stacks) act as emission locations. The estimated $PM_{2.5}$ emissions per street segment, as derived from the AADT data and MOBILE6 outputs (**2.1.3.2**), serve as the emission sources. Unfortunately, AERMOD was not able to model line sources directly, so they had to be converted into a series of individual volume sources. Additionally, AERMOD was not able to process building downwash for volume sources, and as such this was not included in the modeling (EPA, 2004). The input data needed is a bit different from point source dispersion as well (**Table 2-11**).

Srcid	Source ID
Srctyp	Source type (*volume* for mobile sources / line data)
Xs	X coordinate of the source location in meters
Ys	Y coordinate of the source location in meters
Zs	Source elevation location in meters or feet.
Vlemis	Volume emission rate in g/s.
Relhgt	Release height (center of volume) above ground in meters.
Syinit	Initial lateral dimension in meters.
Szinit	Initial vertical dimension in meters.

Table 2-11: Emission inputs for mobile sources in AERMOD

The process for preparing the street segments for modeling was even more labor intensive than the point sources in the previous section. Road widths were measured manually using an orthophoto and appended to the line segment spatial file. Line segments were split at each vertex, and then subdivided into sample points, that would eventually become the volume sources, at intervals of 50m. This interval was chosen since it is approximately twice the average width of the roads as suggested by the AERMOD documentation (EPA, 2004). This value was then used to define the "initial lateral dimension" (Syinit) by using the following formula 2W/2.15, where

"W" equals the road's width. Elevations of the emission volume sources were identified using the nearest point elevation data value from the Department of Information Technology and Telecommunication (DOITT). DOITT point elevations were used rather than a standard digital elevation model in order to accurately represent the elevations of streets not at 'ground level' such as bridges. The number of sample volume sources per line segment were then determined, and the $PM_{2.5}$ emissions were divided equally among each source within the segment. The result of these manipulations was each road segment divided equally into sample points that had attributes which included $PM_{2.5}$ emission, elevation, and width (**Figure 2-33**). There were a total of 9,956 volume sources across NYC.

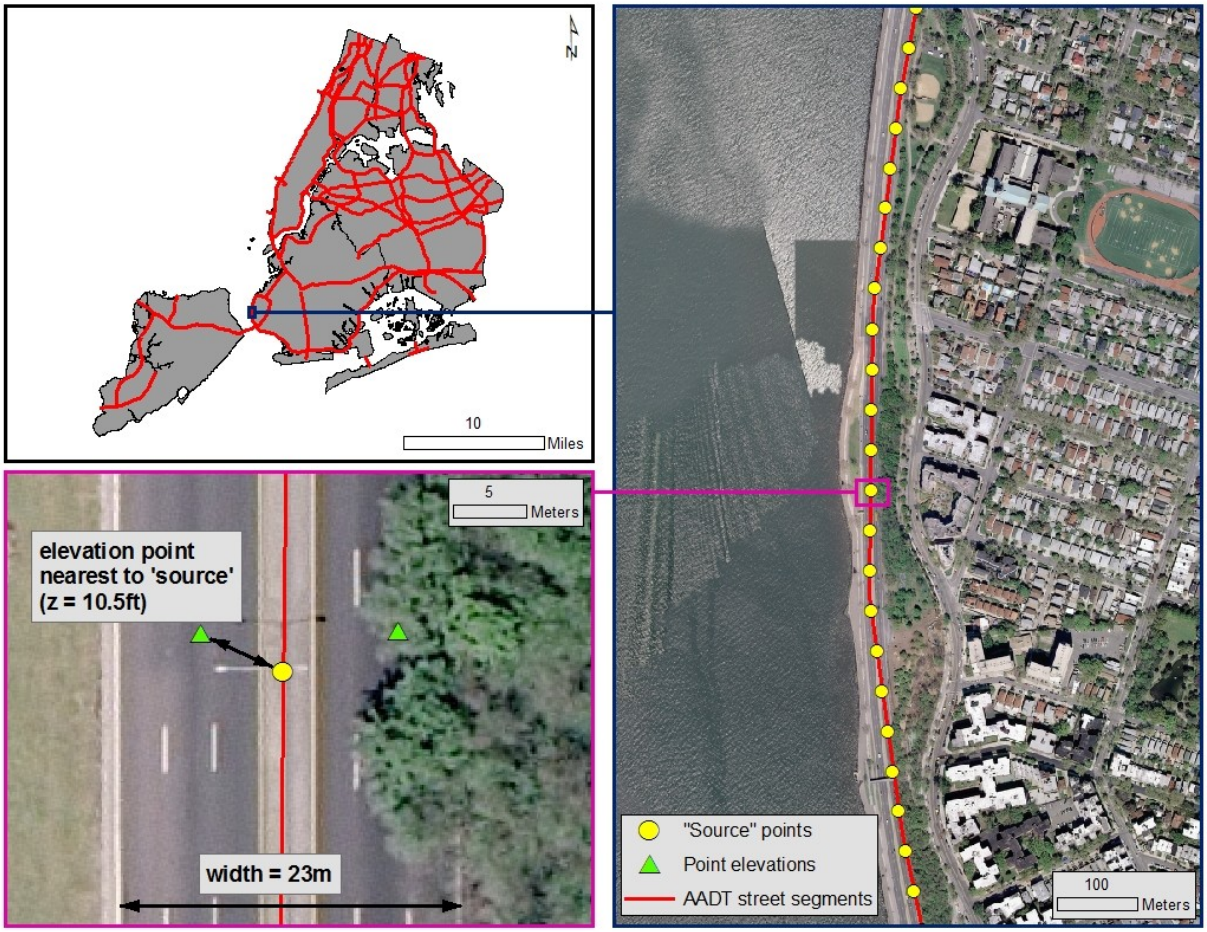


Figure 2-33: Street segments with AADT measurements city-wide (top left); "source" points representing the centers of volume sources for AERMOD input spaced 50 meters from one another (right); and identification of width and elevation of "source" points using orthophotos (NYCMap, 2003) and DOITT elevation points (bottom left).

Even though building downwash, the main cause of hardware failures in the point source modeling, was not calculated for the mobile AERMOD models, the computers were still prone to crashing due to the extremely high number of sources. As such, the data were broken down into boroughs. To mitigate error due to not including proximal sources in neighboring boroughs, a 1000m buffer was created. This resulted in 5 models being run (one per borough), each including the receptors of the borough under study as well as the sources within 1000m from the study

borough's boundary. Processing took quite some time (over two weeks for Queens alone) and resulted an estimated $PM_{2.5}$ concentration emanating from the roads with AADT values for each receptor point in NYC (**Figure 2-34**). Overlapping receptors, those within 1000m of borough boundaries, were added to one another.

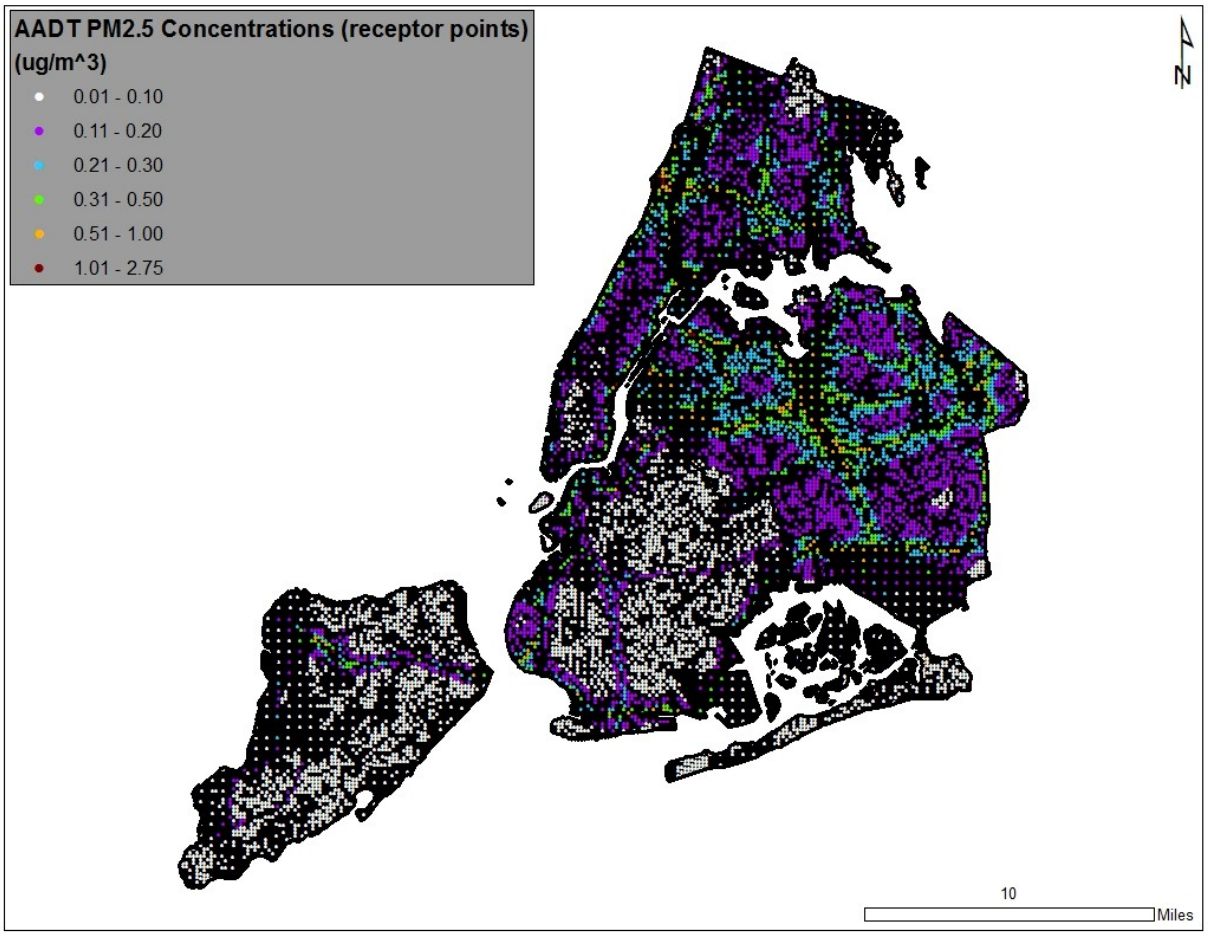


Figure 2-34: AERMOD-estimated $PM_{2.5}$ concentrations from annual average daily traffic (AADT) mobile sources at receptor grid points.

These data were interpolated into a continuous surface, joined with tax lots and census tracts identically to the NEI data above (**Figures 2-35, 2-36,** and **2-37**).

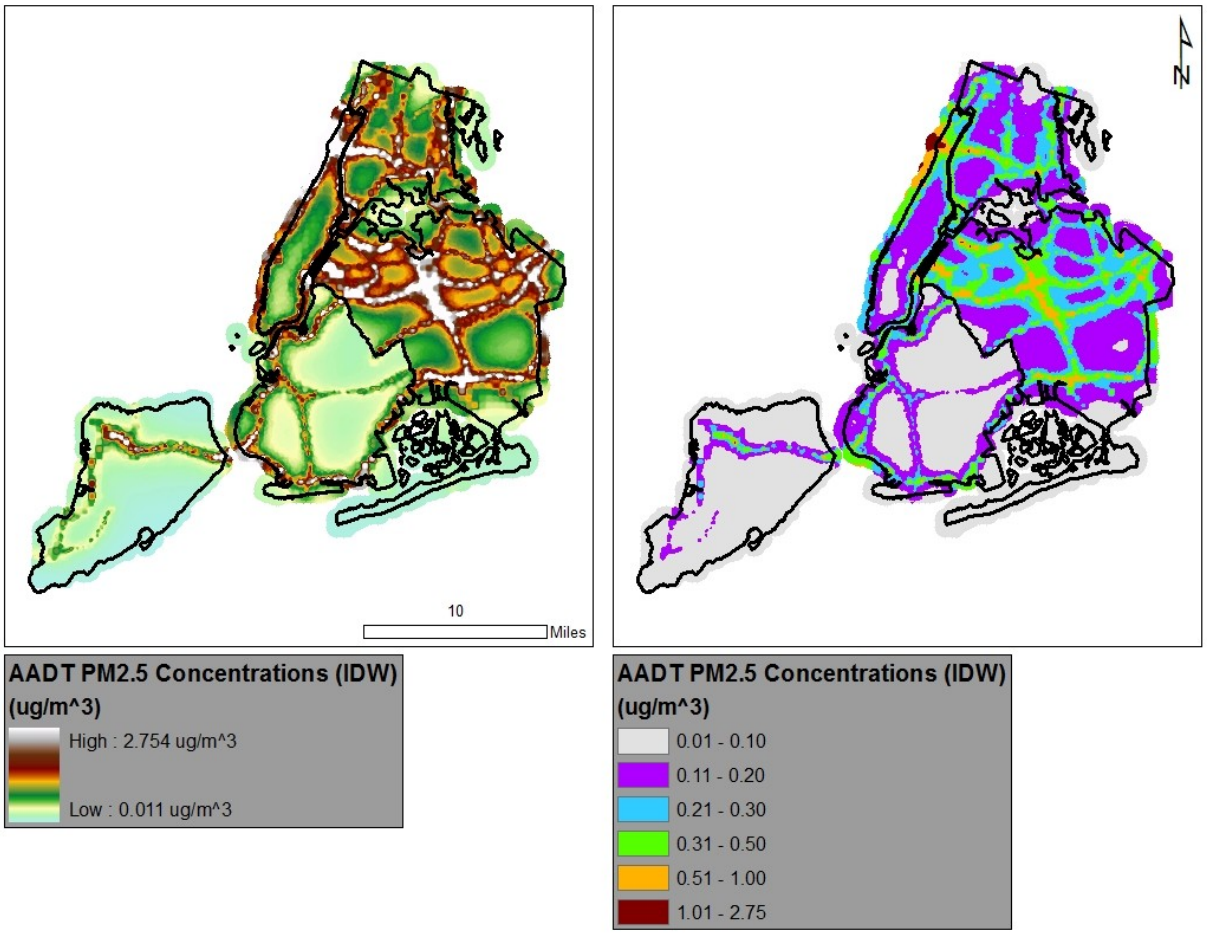


Figure 2-35: AERMOD-estimated $PM_{2.5}$ concentrations from AADT mobile sources interpolated using inverse distance weighting to a continuous surface. Left map shows "stretched" or "unclassified" data. Right map uses a classification scheme identical to **Figures 2-29** to **2-31** for comparative purposes.

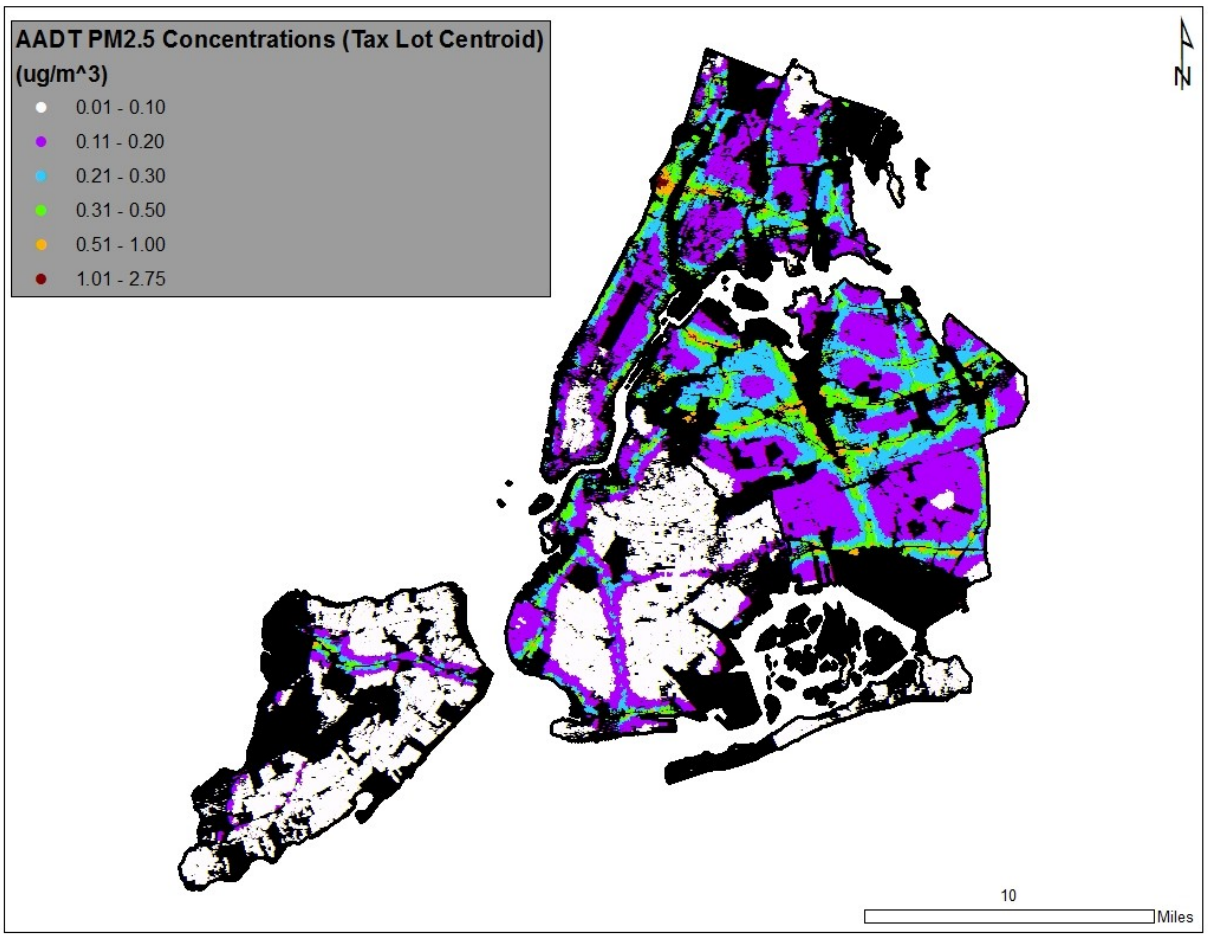


Figure 2-36: AERMOD $PM_{2.5}$ concentration estimates from AADT mobile sources spatially joined to populated tax-lot centroids in order to match the CEDS socio-demographic data (tax lots with no estimated population were not included). Note that the classification scheme for this map is identical to **Figure 2-35**.

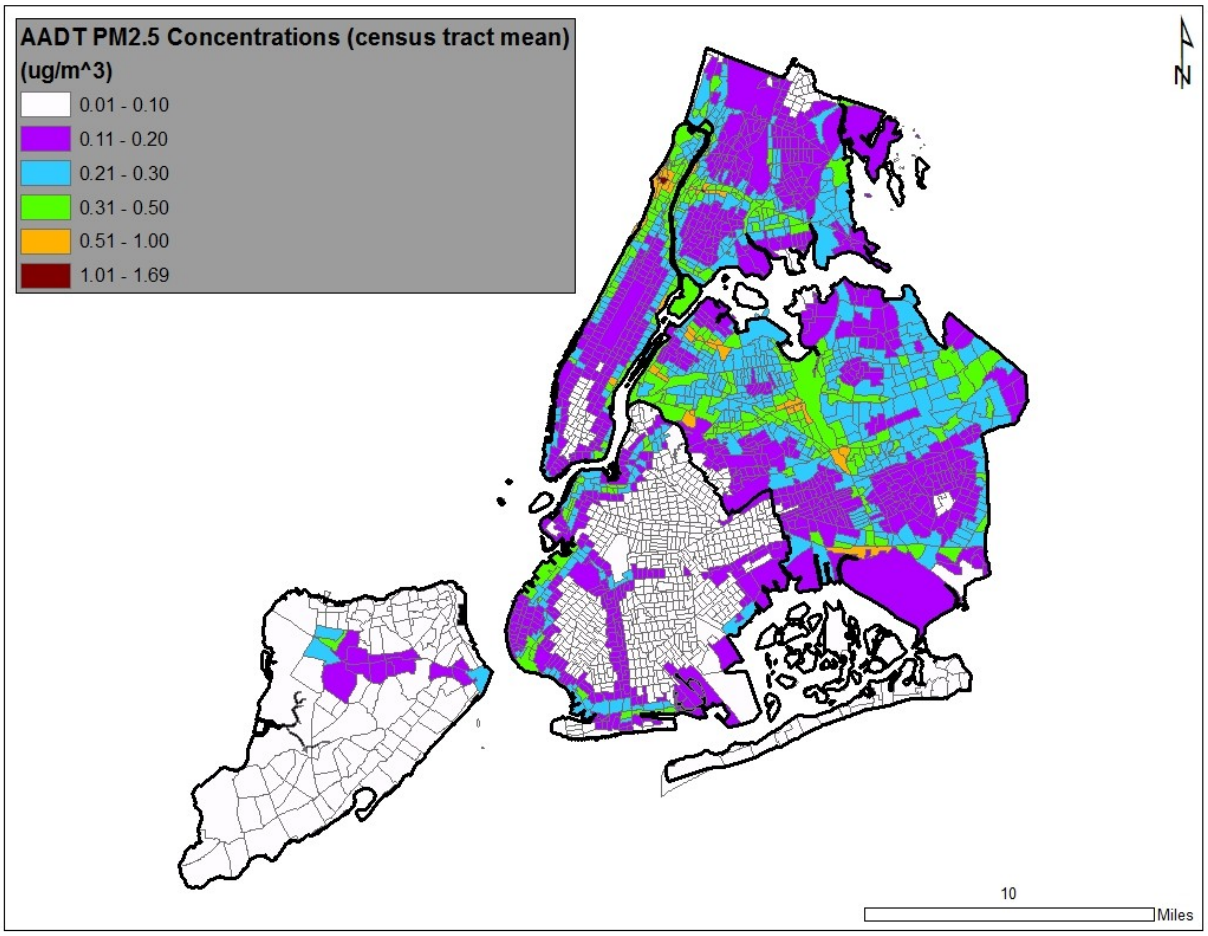


Figure 2-37: AERMOD $PM_{2.5}$ concentration estimates from AADT mobile sources spatially joined to census tracts in order to match the heart failure hospitalization data. Note that the classification scheme for this map is identical to **Figures 2-35** and **2-36**.

The modeled annual average daily traffic and the modeled national emission inventory $PM_{2.5}$ concentration can be combined additively in order to calculate the spatial distribution of $PM_{2.5}$ concentrations from both sets of sources (**Figure 2-38**). If the traffic data were more complete, this would result in a believable and quite useful estimate of the burden of fine particulate matter emanating from major sources in NYC. Unfortunately, as has been mentioned, the AADT data do not appear to be complete enough to justify extensive analyses either on their own or combined with the major point sources.

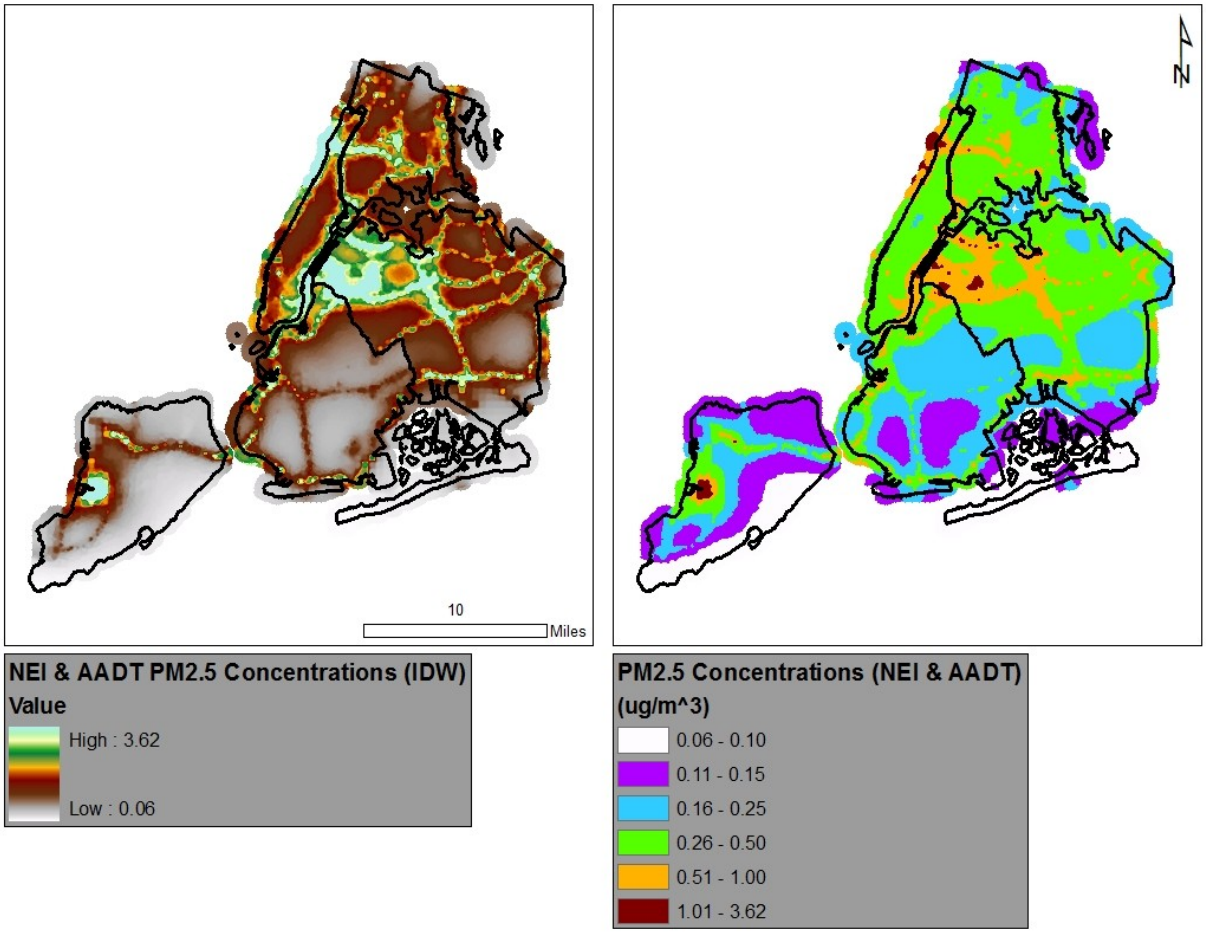


Figure 2-38: Combined AERMOD NEI and AADT $PM_{2.5}$ concentration estimates as a continuous surface (left) and classified identically to **Figure 2-35** to **2-37** for comparative purposes (right).

2.2.3 LAND USE REGRESSION

Land use regression (LUR) can be an extremely useful tool in estimating ambient $PM_{2.5}$ concentrations. It differs from both air dispersion modeling and proximity analysis in that it does not explicitly estimate pollution from any distinct source, but rather attempts to quantify the values by regressing land uses against monitored pollution concentrations. As was discussed in

Chapter 1, this is accomplished by running a regression using the monitored pollution data as the dependent variable and selected land uses or other aspects of the environment as the independent variables. Although there are many potential variables to use, the severe limitation of only having 15 monitored sample points drastically restricts the robustness of the LUR.

A fair number of variables, and versions of variables, were explored. These included roadway variables (lengths of major truck routes, limited access highways, and surface streets), land use variables (population density and measures of amounts of industrial land use and open space), as well as physical environment variables such as elevation. Additional 'modifiers' were also explored in order to attempt to improve the LUR. These included 'correcting' for known amounts of $PM_{2.5}$ by using AERMOD outputs as well as experimentation with using remotely sensed aerosol optical depth (AOD) data from the MODIS instrument. Different bandwidths, or buffers, were also explored using sensitivity analyses to find the optimal distance (geographically) from the monitored $PM_{2.5}$ data from which to collect and process the independent variables.

To prepare the data, first various circular buffers were created around the monitors (100m, 200m, 500m, 1000m, and 2000m). The data (tax lots for CEDS population and land uses; street segments for truck routes, limited access highways and surface streets) were then given identifiers as to which monitor they are proximal, if any. These data were then aggregated by the monitors' identifiers – resulting in a table with 15 records (one for each monitor) that has site specific information such as measured $PM_{2.5}$ and elevation above mean sea level, as well as

summaries of land use, roadways, and population for each buffer distance (**Figure 2-39**). AOD from MODIS was also joined with the monitor sites in order to include this information in the LUR.

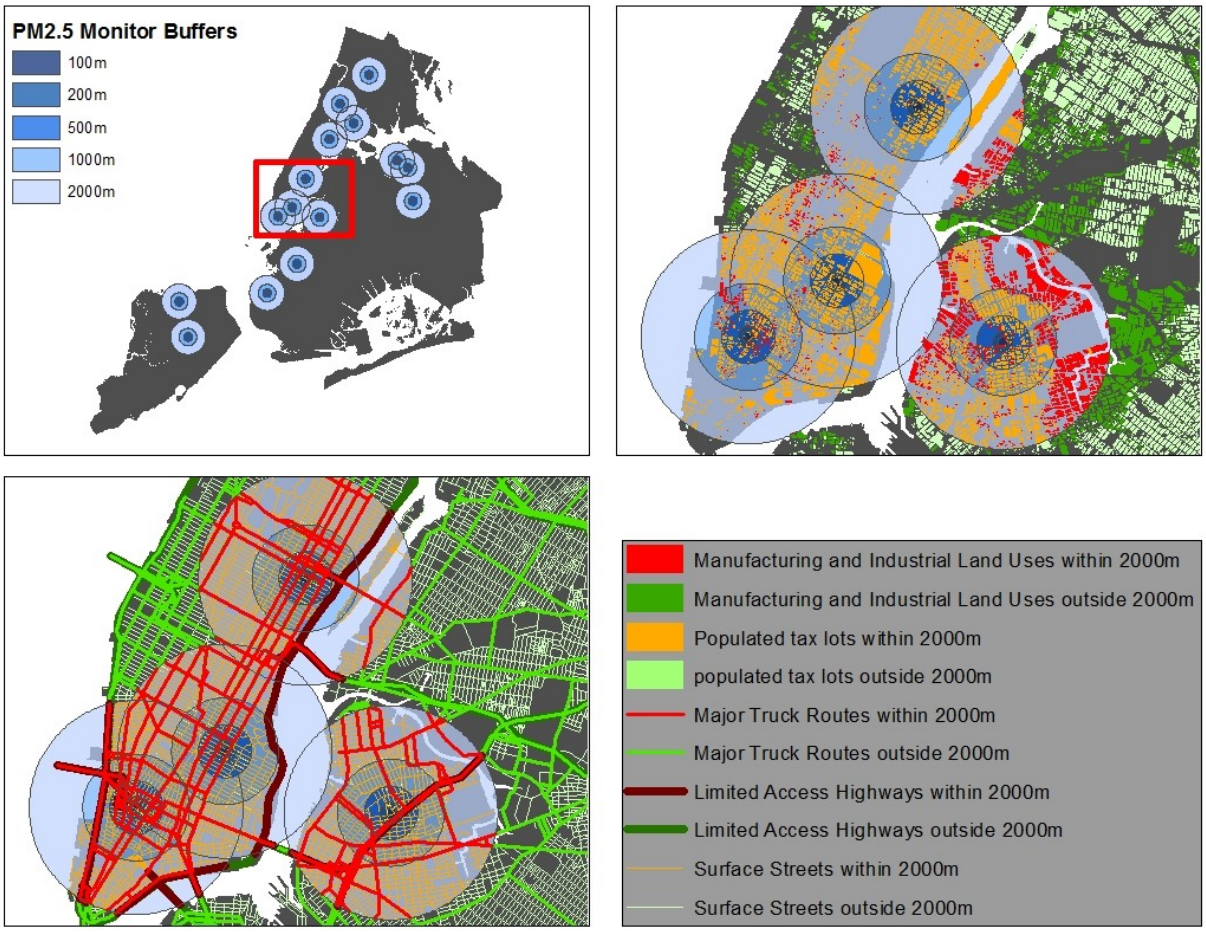


Figure 2-39: A simplified cartographic example of roadways, land use (aggregated to manufacturing/industrial), and population in relation to proximity to $PM_{2.5}$ monitoring sites around lower Manhattan.

Although a fair amount of time was then spent massaging and exploring salient variables within these buffers using scatter plots, correlation analyses, and other statistical and graphical tools, ultimately the ones that appeared most promising were length of major truck routes within 1000m and population density within 1000m. The length of the truck routes is likely related to

the amount of truck traffic and the population density likely acts as a proxy for a number of factors including vehicular traffic, large housing structures (and the associated production of pollution), and any number of additional human activities which are linked to increased $PM_{2.5}$ emissions. Although these variables are not the most novel or groundbreaking, they are strongly supported by the LUR literature, are not collinear, and present consistent relationships across the buffer sizes (e.g. population density is positively correlated with $PM_{2.5}$ concentrations at all buffer sizes). It is the stability of the relationships that mitigates concern for the lack of sample points ($PM_{2.5}$ monitors).

The LUR model itself is a simple OLS regression with monitored $PM_{2.5}$ concentrations acting as the dependent variable and length of major truck routes within 1000m and population density within 1000m acting as the independent variables. The regression outputs suggest a good performing model with an R^2 of .875, standard error of .455, normally distributed errors, and significances for both variables < .001. This indicates that approximately 88% of the variance in $PM_{2.5}$ concentrations is predicted by the length of truck routes and the population density within 1000m. The model coefficients are 11.74 for the constant, $2.47*10^{-5}$ for population density, and $1.10*10^{-4}$ for major truck route length. The equation that describes this relationship (**Equation 2-10**) can now be used to predict $PM_{2.5}$ concentrations across NYC by selecting census tract centroids, calculating the length of truck routes and population density within 1000m, and simply plugging in the appropriate values (**Figure 2-40**).

LUR = 11.738 + .0000246605 * POPDENS + .0001098755 * MTR **Eq. 2-10**

where:
LUR = the $PM_{2.5}$ concentration estimated by the LUR
11.728 = the regression constant
POPDENS = population density within 1000m
MTR = length of major truck routes within 1000m

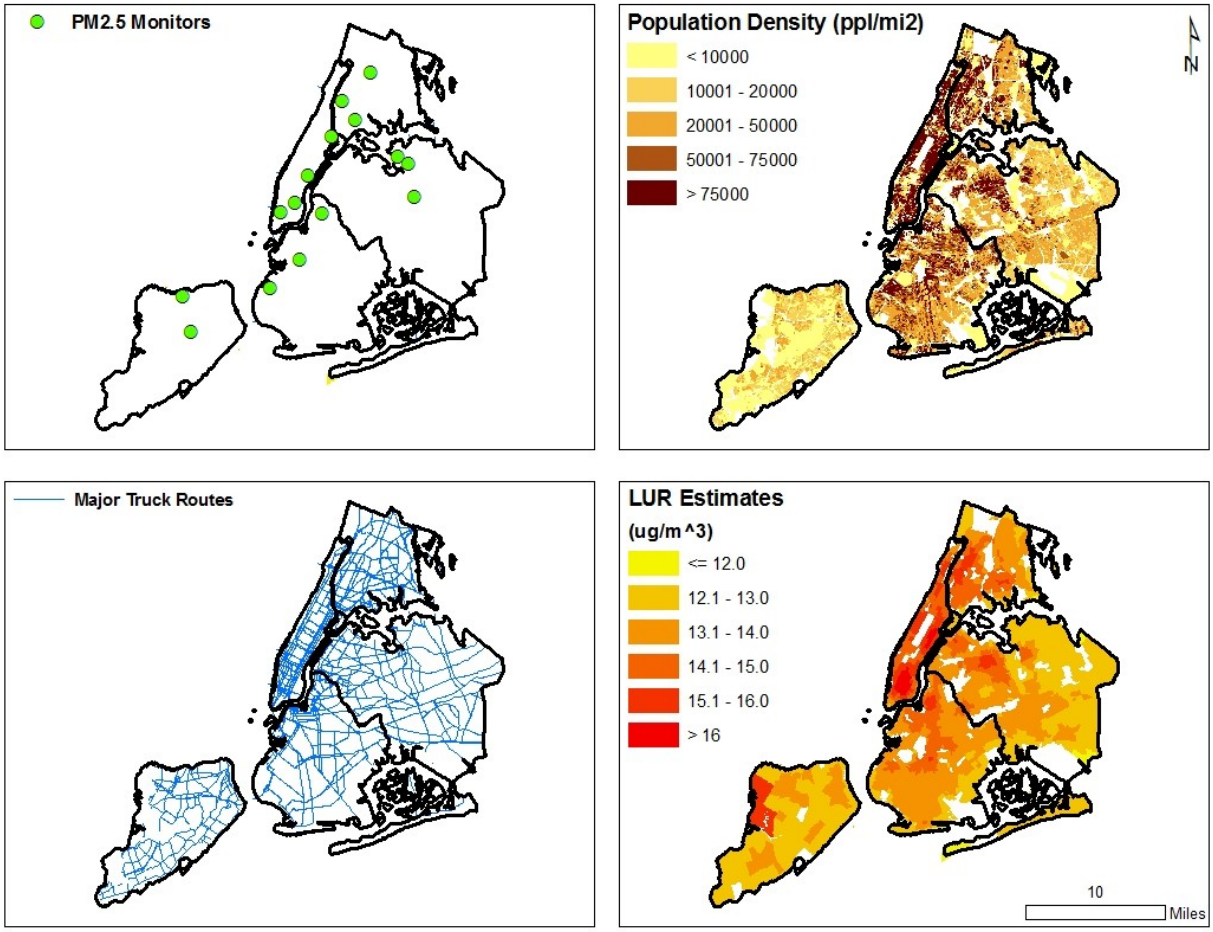


Figure 2-40: LUR $PM_{2.5}$ concentration estimates and input variables.

To explore the potential of utilizing AERMOD $PM_{2.5}$ estimates to improve LUR modeling, two new LUR models were created that "correct" for fine particulate matter emanating from NEI stationary point sources and/or AADT routes. This was done by subtracting the AERMOD concentrations from the monitored $PM_{2.5}$ values, in other words removing the air dispersion

model $PM_{2.5}$ concentration estimates from modeled sources. Regressions were then run, as described above.

One complicating factor in preparing corrected LUR estimates is that some of the major truck routes are also part of the AADT database. As such, these lengths of roads were removed before running the regressions. The corrected LUR models are described by the equations below (**Equations 2-11 and 2-12**)

$LUR_2 = 11.680 + .0000235141 * POPDENS + .0001036302 * MTR$ **Eq. 2-11**

where:
LUR_2 = the $PM_{2.5}$ concentration estimated by the LUR minus the AERMOD estimated $PM_{2.5}$ concentrations from NEI sources
11.680 = the regression constant
POPDENS = population density within 1000m
MTR = length of major truck routes within 1000m

$LUR_3 = 11.667 + .0000235941 * POPDENS + .0001065652 * MTR_{AADT}$ **Eq. 2-12**

where:
LUR_3 = the $PM_{2.5}$ concentration estimated by the LUR minus the AERMOD estimated $PM_{2.5}$ concentrations from all modeled sources
MTR_{AADT} = length of major truck routes within 1000m excluding those segments modeled in AERMOD

The three LUR models can be compared via regression diagnostics (**Table 2-12**). However, this is in some ways "comparing apples to oranges", in that the dependent variables are not identical in any of the models (the former are "raw" monitored data, whereas the latter two are "corrected").

	Land-use only	Land-use, corrected for NEI sources	Land-use, corrected for NEI and mobile sources
R^2	.875	.869	.868
Std. Error	.45490	.44253	.44560
Constant	11.738	11.680	11.667
Population Density Coefficient	.0000246605	.0000235141	.0000235941
MTR Length Coefficient	.0001098755	.0001036302	.0001065600
Population Density, t val.	6.178	6.056	6.069
MTR Length, t val.	4.820	4.673	4.730

Table 2-12: Regression outputs for LUR models. Note that these data are not directly comparable to one another since the dependent variables are not the same

One additional step is needed to calculate the final predicted $PM_{2.5}$ concentrations. The AERMOD estimates from NEI sources, or both NEI and AADT sources, are simply added back to the LUR outputs (**Equations 2-13** and **2-14**).

$$\mathbf{LUR_{NEI} = LUR_2 + AERMOD_{NEI}} \qquad \textbf{Eq. 2-13}$$

where:
LUR_{NEI} = the $PM_{2.5}$ estimate of the AERMOD NEI corrected model
$AERMOD_{NEI}$ = the AERMOD estimated $PM_{2.5}$ concentration from NEI sources

$$\mathbf{LUR_{NEI\_AADT} = LUR_3 + AERMOD_{NEI\_AADT}} \qquad \textbf{Eq. 2-14}$$

where:
LUR_{NEI_AADT} = the $PM_{2.5}$ estimate of the AERMOD corrected model from both sources (NEI and AADT)
$AERMOD_{NEI\_AADT}$ = the AERMOD estimated $PM_{2.5}$ concentration from both sources (NEI and AADT)

Now that the LUR-derived estimates for $PM_{2.5}$ concentration are "apples-to-apples", statistical tests can be utilized to determine which model performed the best with regards to fine particulate matter estimation (i.e., validation). Pearson and Spearman correlations as well as root-mean-square error (RMSE) analyses comparing estimated $PM_{2.5}$ concentration to the monitored data were conducted similar to diagnostics performed on the CEDS-derived population in **Section 2.1.1.2** (**Table 2-13**). It also can be edifying to look at the residuals cartographically. In this case,

the residuals are not the regression residuals, but rather the difference between the LUR-estimated $PM_{2.5}$ (after correction when appropriate) and the observed $PM_{2.5}$ from the monitors. Positive values indicate overestimation of the LUR models, and negative values suggest underestimation (**Figure 2-41**).

Model	Pearson	Spearman	RMSE
LUR	.935**	.936**	0.00010770
LUR_{NEI}	.939**	.929**	0.00006725
LUR_{NEI_AADT}	.938**	.939**	0.00253138

Table 2-13: Correlations and RMSE of LUR estimates vs. monitored $PM_{2.5}$ data. **Correlation is significant at the 0.01 level (2-tailed).

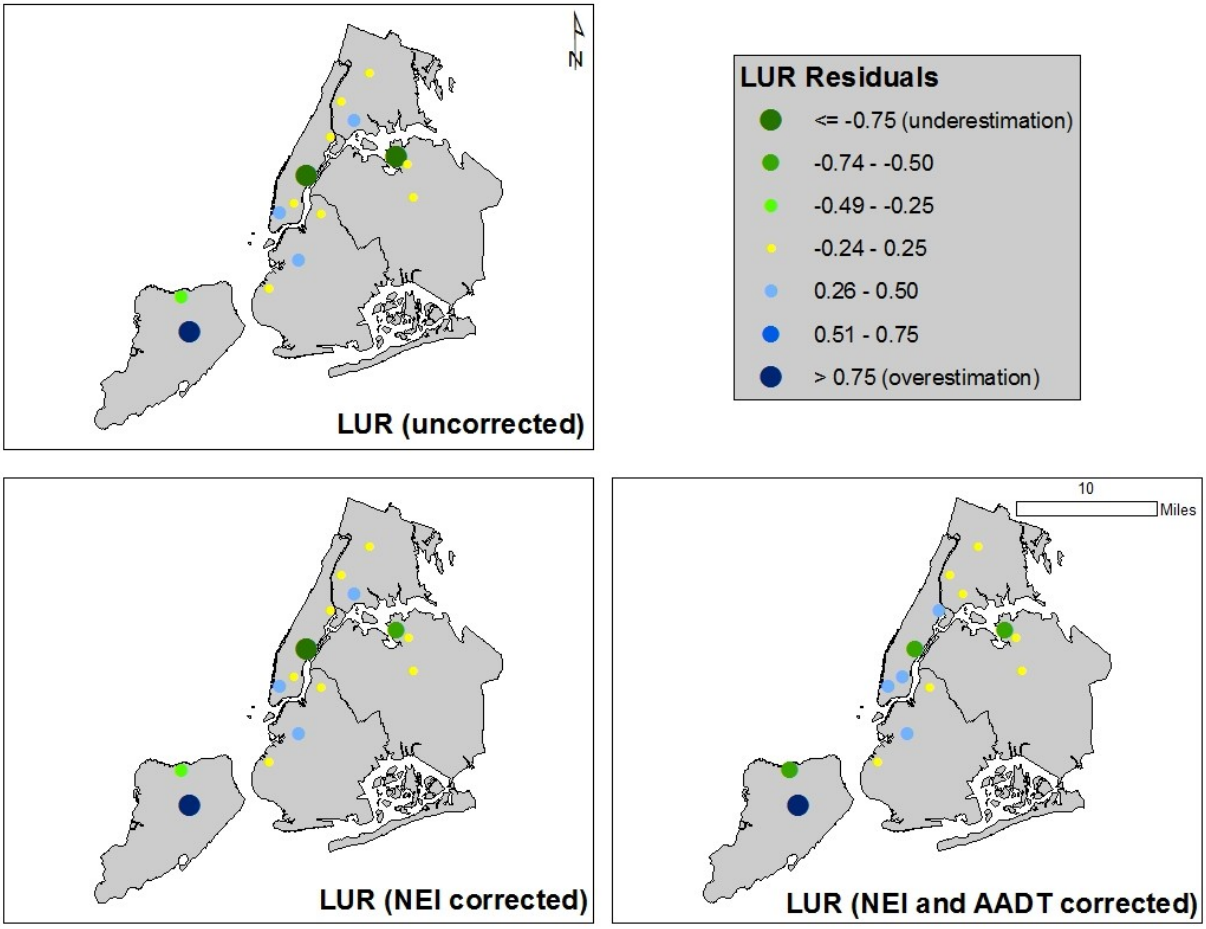


Figure 2-41: LUR residuals. These residuals represent the LUR-derived $PM_{2.5}$ estimates (after correction when applicable) subtracted by the observed $PM_{2.5}$ values from monitors.

Although the validation values are somewhat similar among the three LUR methods, it can be seen that the land use regression which is corrected for AERMOD-estimated NEI emissions (LUR_{NEI}) seems to perform better according to the Pearson coefficient and the RMSE, whereas the LUR estimate corrected for both NEI and AADT sources (LUR_{NEI_AADT}) has the highest non-parametric Spearman coefficient. Using these diagnostics, and the understanding that the AADT may be biased due to an insufficient representation of roadways as well as the highest RMSE by far, LUR_{NEI} seems to be the most logical choice. These different estimates can also be viewed

cartographically to understand the differences and similarities among the LUR values (**Figure 2-42**). Note that the inset maps show an area of Staten Island that has high $PM_{2.5}$ concentrations from an NEI source but low population density and comparatively low lengths of major truck routes (variables in the LUR). This increased pollution is not accounted for in the standard, non-corrected, LUR model.

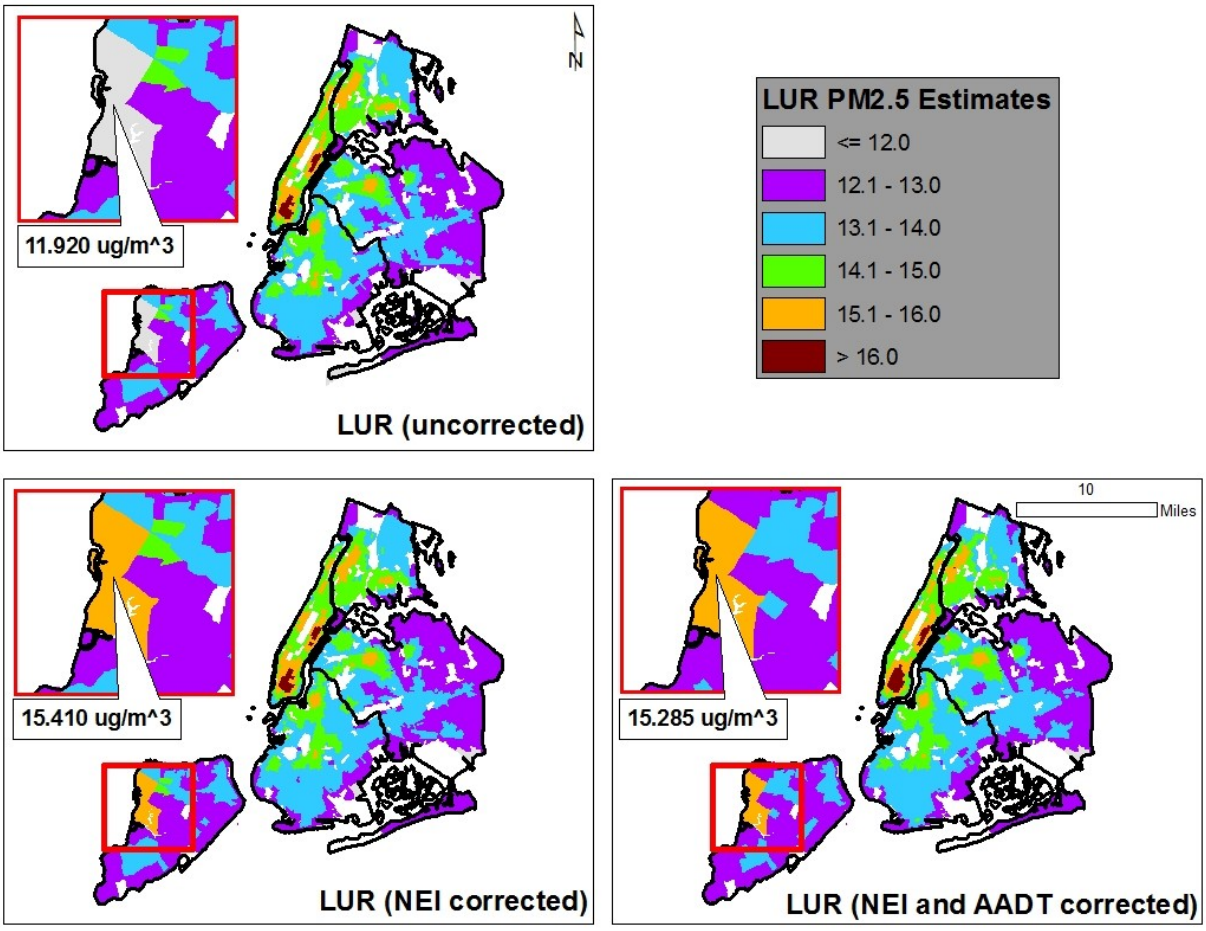


Figure 2-42: Comparison of LUR $PM_{2.5}$ estimates. The "area of interest" shows a section of Staten Island with a high $PM_{2.5}$ concentration emanating from an NEI source.

2.3 METHODS CHAPTER CONCLUSORY STATEMENT

This methods chapter has introduced the data (population, health, and pollution) and the exposure estimation techniques (proximity, air dispersion modeling, and land use regression) used in this dissertation. These data and estimation techniques are employed in the analysis section which follows.

3 ANALYSIS

The analysis chapter of this dissertation takes the output from the Methods sections (**2.1** and **2.2**) and utilizes them in an attempt to examine two main issues: (1) inequitable exposure to $PM_{2.5}$ from major stationary point sources based on socio-demographic characteristics and (2) the effect that chronic $PM_{2.5}$ exposure along with socio-demographics have on cardiovascular disease. The first section of this chapter explores the potential environmental justice issues that surround exposures to fine particulate matter by comparing $PM_{2.5}$ exposure estimates from NEI point sources (proximity analysis and air dispersion modeling) to race/ethnicity, educational attainment, and income/poverty at the tax lot level. The second section examines different statistical techniques (ordinary least squares regression, spatial autoregressive models, and geographically weighted regression) to quantify the associations between heart failure hospitalization rates and $PM_{2.5}$ exposure estimates (air dispersion modeling and land use regression) while adjusting for the socio-demographics listed above at the census tract level.

3.1 ENVIRONMENTAL JUSTICE ANALYSIS

Traditional environmental justice (EJ) studies have used proximity to environmentally burdensome facilities or land uses as a proxy for exposure to pollutants; however this is not the most precise measure of exposure as it does not usually take into account the amount or toxicity of pollution emanating from a particular source, nor does it include other physical or meteorological factors that would pin-point pollution dispersion and therefore potential exposure (more information about EJ and proximity analysis can be found in **Section 2.2**). This section

will compare the estimated exposed socio-demographic groups using proximity analysis versus air dispersion modeling.

Although environmental justice and inequitable exposures are an extremely important aspect of environmental health in NYC, it is not common that EJ analyses are performed on a single pollutant using complex exposure data. In the following sections the socio-demographics discussed in **Section 2.1** and the $PM_{2.5}$ estimates derived in **Section 2.2** are used concomitantly to quantify potential EJ issues. In order to take advantage of the CEDS method (**2.1.1.2**), the analyses use simple descriptive statistics, visualizations, and odds ratios to compare exposure estimates from NEI stationary point sources based on proximity to NEI emission points and air dispersion modeling from those same points.

3.1.1 PROXIMITY ANALYSIS FOR ENVIRONMENTAL JUSTICE

To prepare the data, each populated tax lot, as estimated by CEDS, was given an "exposure value" based on its proximity to the NEI source. If the tax lot centroid was within ¼ mile from the NEI stack, then all the inhabitants of the tax lot were considered exposed, otherwise they were identified as unexposed. The proximity buffer radius of ¼ mile was chosen based on distances established as standards by environmental agencies and used most often by other researchers as the area of greatest potential impact from the sources (MOEC, 2001). These data can then be easily aggregated based on exposure value to calculate the total population as well as

the total number of persons belonging to any particular socio-demographic group that are either exposed or unexposed (**Section 2.2.1**). One caveat is that median household income cannot be disaggregated easily as it is already a summary measure (median); as such, "poverty", a slightly different and less nuanced measure, was used instead.

The total populations and subpopulations can then be converted into rates (e.g., percent of persons who are exposed that do not have a high school degree vs. the percent of people of that group in the entire area of study), allowing the comparison of "observed" populations in close proximity to pollution sources to the "expected" proportions for the entire city and broken down by borough (**Table 3-1**). The interpretation of the chart involves comparing percentages in "exposed" verses "all" populations. For instance, if the proportion of Latino residents in the exposed population is greater than the proportion of Latino residents in the general, city-wide, population, then they could be considered overrepresented in the exposed population. It is important to note that although the race/ethnicity categories are mutually exclusive, they do not capture the entire population. As such, it is possible for all of the groups appear to be over- or underrepresented in the exposed category. This scenario, which occurs in Brooklyn, suggests that the racial/ethnic categories that were not explicitly included in the analysis are the ones who are underrepresented. Staten Island presents the opposite scenario where all the groups explicitly included in the analysis appear underrepresented, suggesting that it is the other groups that are overrepresented. Staten Island, however, must be viewed skeptically since only .1% of the population resides within ¼ mile of an NEI pollution source. The economic and education variables are dichotomous, meaning that if the sub-population (e.g., those living below poverty)

is overrepresented, then the inverse of that sub-population (e.g. those not living below poverty) is underrepresented.

	Exposure	% Total Population	% non-Hispanic White	% non-Hispanic Black	% Hispanic / Latino	% below poverty	% without high school degree
NYC	NYC	100.0	35.0	24.4	27.0	21.2	27.6
	exposed	12.6	32.1	24.4	31.7	24.2	27.7
Brooklyn	Brooklyn	100.0	34.6	34.4	19.8	25.1	31.2
	exposed	8.5	36.1	35.1	20.0	29.6	31.7
Bronx	Bronx	100.0	14.6	30.9	48.6	30.7	37.5
	exposed	13.3	13.6	36.4	43.3	29.5	33.1
Manhattan	Manhattan	100.0	45.7	15.2	27.2	19.9	21.3
	exposed	31.5	38.7	16.3	32.9	21.9	25.1
Queens	Queens	100.0	32.9	18.9	25.0	14.6	25.6
	exposed	6.3	26.5	21.2	30.2	17.7	25.7
Staten Island	Staten Island	100.0	71.2	8.9	12.0	10.1	17.4
	exposed	0.1	70.4	2.8	10.2	6.8	14.4

Table 3-1: Proximity exposure estimates to NEI sources in NYC and its boroughs using CEDS. "Exposed" represents the proportion of the sub-populations within ¼ mile of a source.

The proximity analysis suggests that over 12% of the residential population in NYC is proximal to an NEI $PM_{2.5}$ source. Of the exposed population, Latino residents and those living below poverty are overrepresented, meaning that the exposed residents have higher proportions of these two groups than the population city-wide (e.g., the exposed population is 31% Latino whereas NYC is only 27% Latino; and the exposed population is 24% below poverty whereas the NYC-wide population is only 21% below poverty). When it is examined by borough, the relationships get a little more convoluted. For instance, in the Bronx only the non-Hispanic Black population

appears to suffer from overrepresentation, whereas in Manhattan, the borough with the highest proportion of people living proximal to $PM_{2.5}$ sources (31%), every group other than non-Hispanic White appears to be overrepresented.

These data can be viewed more intuitively by graphs, with one bar representing the total population of the study area (NYC or one of its boroughs) and one bar representing the exposed population (**Figures 3-1 to 3-6**). It can be seen that Manhattan and Queens seem to have the biggest EJ issues across the most, if not all, socio-demographic categories. Brooklyn shows a large overrepresentation of those below poverty in the exposed group and the Bronx demonstrates an overrepresentation of non-Hispanic Blacks. Staten Island and the Bronx are the only boroughs that do not show an increased representation of those below poverty or Latinos in close proximity to NEI $PM_{2.5}$ sources.

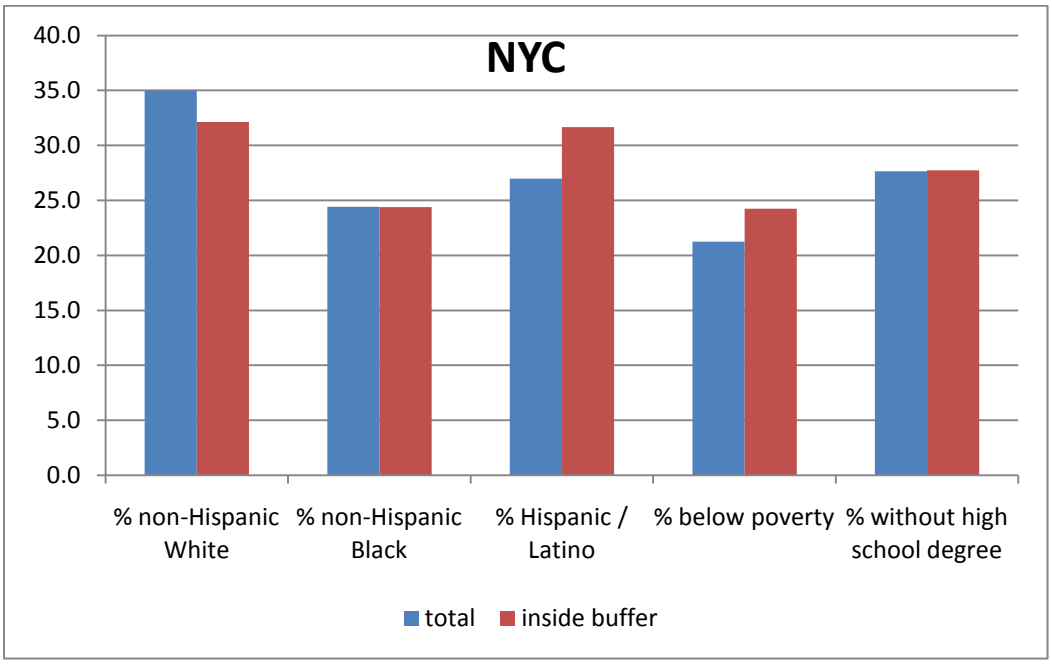


Figure 3-1: Proximity to NEI $PM_{2.5}$ sources and socio-demographics in NYC.

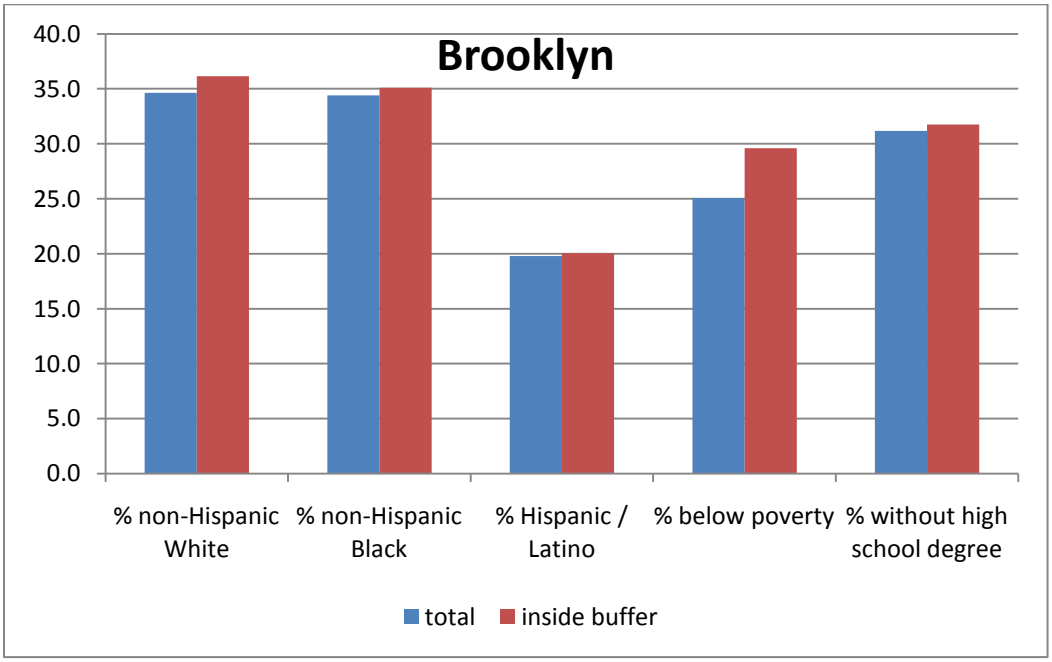


Figure 3-2: Proximity to NEI $PM_{2.5}$ sources and socio-demographics in Brooklyn.

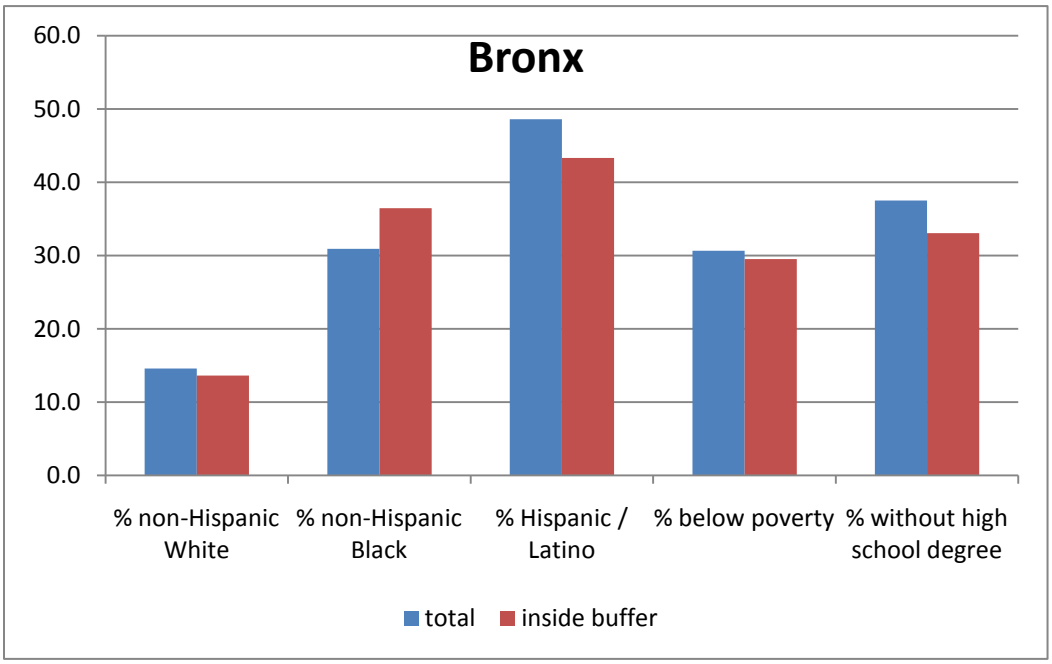


Figure 3-3: Proximity to NEI $PM_{2.5}$ sources and socio-demographics in the Bronx.

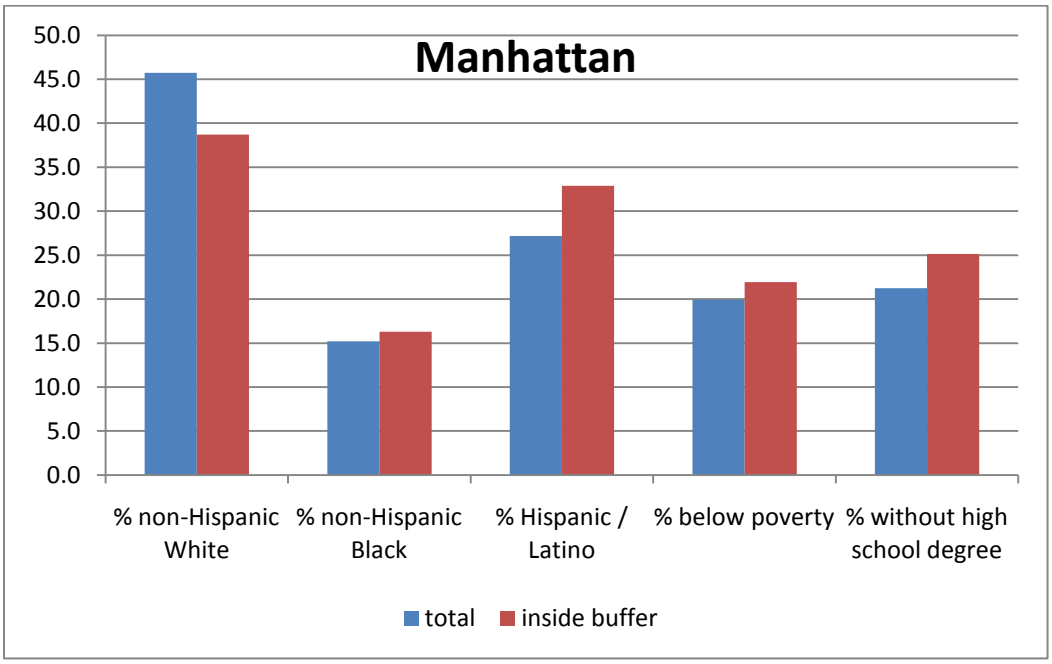


Figure 3-4: Proximity to NEI $PM_{2.5}$ sources and socio-demographics in Manhattan.

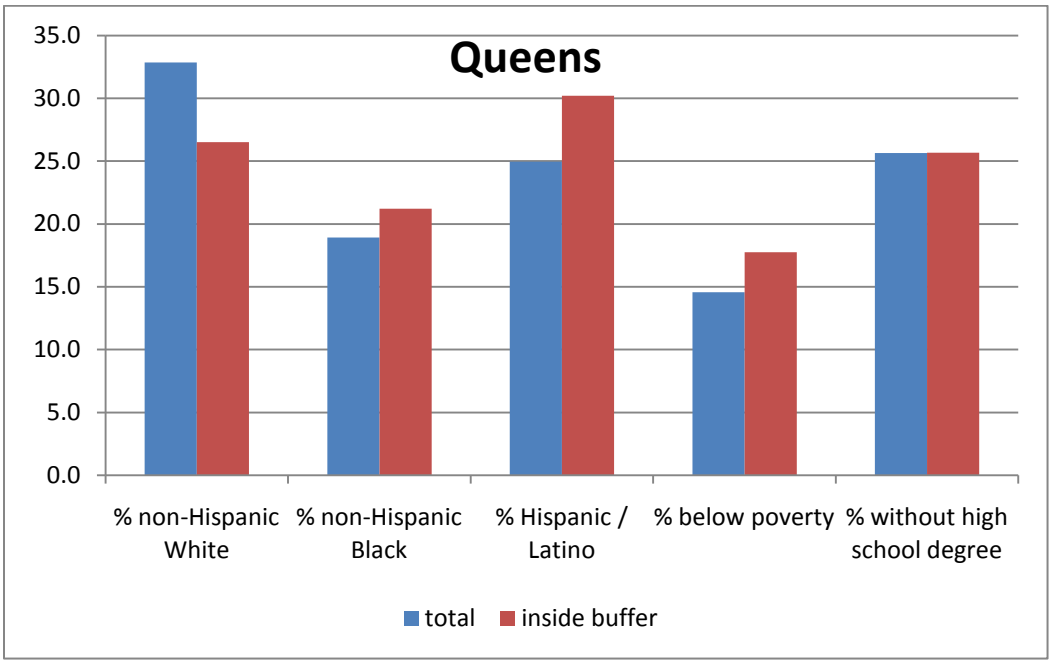


Figure 3-5: Proximity to NEI $PM_{2.5}$ sources and socio-demographics in Queens.

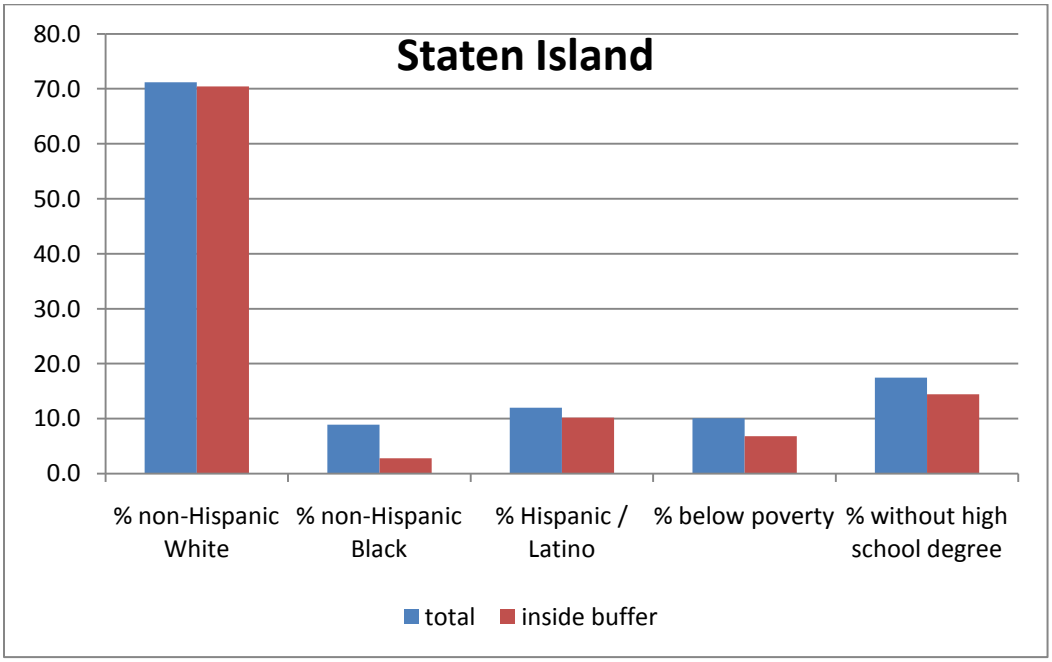


Figure 3-6: Proximity to NEI $PM_{2.5}$ sources and socio-demographics in Staten Island.

Finally, these data can be looked at using odds ratios. An odds ratio (OR) is a traditional statistical technique often used in epidemiological studies which compares the odds of an "exposed" group experiencing an outcome versus a "control" group experiencing the same outcome (Westergren et al., 2001). It is essentially a ratio of the probability that an event will occur versus the probability that the event will not occur. To do this, a 2x2 contingency table was created in Microsoft Excel with cell "A" representing people belonging to a certain socio-demographic group in the "exposed" population and cell "B" representing people of that same group in the "unexposed" population. Cells "C" and "D" represent the people not in the selected socio-demographic groups that are "exposed" (cell C) and "unexposed" (cell D). It is then a simple matter of dividing the odds of being exposed and in the socio-demographic group (A/B) by the odds of being exposed and not in the socio-demographic group (C/D). When the OR is greater than 1, then there is an increased likelihood of the selected demographic being represented in the "exposed" group than the "unexposed" group. When the OR is less than one, the opposite is true. It can be written, and simplified, as **Equation 3-1** (Bland and Altman, 2000).

$$\mathbf{OR = (A/B)/(C/D) = (A*D)/(B*C)} \qquad \mathbf{Eq.\ 3\text{-}1}$$

where:
OR = the odds ratio
A = number of people in a selected socio-demographic group within .25 miles of a NEI source
B = number of people in a selected socio-demographic group not within .25 miles of a NEI source
C = number of people not in a selected socio-demographic group within .25 miles of a NEI source
D = number of people not in a selected socio-demographic group not within .25 miles of a NEI source

Confidence intervals (CI) can then be calculated for each of the odds ratios. This was done in Microsoft Excel by first calculating the 95% CI of the natural log of the OR using **Equation 3-2,** and then converting the output by taking the exponential function of the results (Bland and Altman, 2000). When the 95% CI "straddles" a value of 1, then it is not statistically significant at $p < .05$. Some care, however, must be used when looking at the 95% CI since the sample size for this data is quite large.

$$\textbf{95\% CI of ln(OR)} = \ln(OR) \pm 1.96 * (1/A + 1/B + 1/C + 1/D)^{0.5} \qquad \textbf{Eq. 3-2}$$

where:
95% CI = the 95% confidence interval
ln(OR) = the natural log of the odds ratio

The results of the odds ratios support the implications of the table and graphs above (**Table 3-2**). For instance, in NYC there is a significant (95% CI) decreased likelihood of living in close proximity to an NEI $PM_{2.5}$ source if you are non-Hispanic white, and a significant increased likelihood of being Latino or below poverty and living near such a source. These findings, along with many of the borough-level results, suggest the presence of environmental injustice with regards to proximity to fine particulate matter-producing facilities and socio-demographics.

Geography	Socio-Demographic Group	OR	95% Confident Interval	
			Low	high
NYC	Non-Hispanic White	0.864	0.860	0.868
	Non-Hispanic Black	*0.998*	*0.994*	*1.003*
	Hispanic / Latino	1.299	1.293	1.305
	Below Poverty	1.218	1.212	1.224
	No High School Degree	*1.004*	*0.999*	*1.010*
Brooklyn	Non-Hispanic White	1.075	1.065	1.085
	Non-Hispanic Black	1.033	1.023	1.043
	Hispanic / Latino	1.018	1.006	1.029
	Below Poverty	1.284	1.272	1.297
	No High School Degree	1.030	1.017	1.042
Bronx	Non-Hispanic White	0.915	0.902	0.928
	Non-Hispanic Black	1.334	1.321	1.349
	Hispanic / Latino	0.783	0.775	0.791
	Below Poverty	0.938	0.927	0.948
	No High School Degree	0.799	0.788	0.810
Manhattan	Non-Hispanic White	0.659	0.655	0.664
	Non-Hispanic Black	1.129	1.119	1.140
	Hispanic / Latino	1.507	1.496	1.519
	Below Poverty	1.194	1.183	1.204
	No High School Degree	1.381	1.367	1.394
Queens	Non-Hispanic White	0.723	0.714	0.732
	Non-Hispanic Black	1.167	1.151	1.182
	Hispanic / Latino	1.326	1.310	1.342
	Below Poverty	1.288	1.270	1.307
	No High School Degree	*1.001*	*0.986*	*1.016*
Staten Island	*Non-Hispanic White*	*0.965*	*0.800*	*1.163*
	Non-Hispanic Black	0.294	0.175	0.493
	Hispanic / Latino	*0.829*	*0.625*	*1.100*
	Below Poverty	0.653	0.464	0.917
	No High School Degree	*0.797*	*0.603*	*1.054*

Table 3-2: Odds ratios and 95% confidence intervals of socio-demographics and proximity to NEI $PM_{2.5}$ sources in NYC and its boroughs. Italicized entries are not significant.

3.1.2 AIR DISPERSION MODELING FOR ENVIRONMENTAL JUSTICE

Air dispersion modeling is able to produce more nuanced estimates of exposure to $PM_{2.5}$ since it is based upon estimated $PM_{2.5}$ concentrations rather than using a simple proximity measure as a proxy for exposure. Facilities emitting higher amounts of fine particulate matter will have a greater effect than those emitting little. Additionally, meteorological and physical variables are included in the model (**Section 2.2.2**).

For the first attempt at examination of the relationship between socio-demographic characteristics and $PM_{2.5}$ exposure produced by NEI sources, concentration estimates were attached to each populated tax lot in NYC. As this results in far too many samples to handle (n=733,517), the pollution concentrations were aggregated into percentiles (n=100) based on tax lots. In other words, approximately the same number of tax lots are in each percentile. These data then become essentially aspatial aggregations of samples with each record representing the CEDS-derived socio-demographics associated with the percentile of $PM_{2.5}$ concentration. When the number of people per lot-level pollution percentile is graphed, a positive logarithmic correlation can be seen. This suggests that the tax lots with the densest populations (excluding tax lots with no estimated population) tend to be in the areas with higher concentrations of $PM_{2.5}$ originating from NEI facilities (**Figure 3-7**).

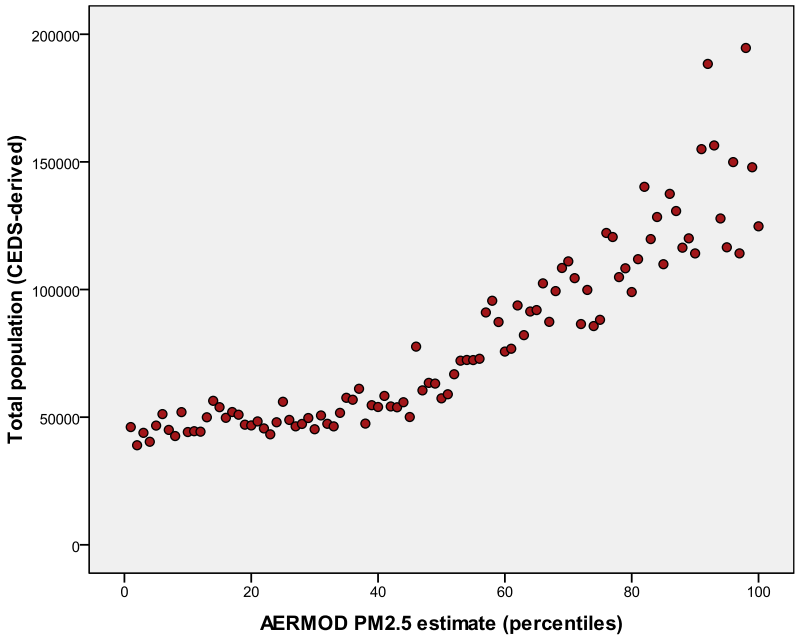


Figure 3-7: Total population versus modeled $PM_{2.5}$ concentration from NEI sources by percentile.

.

The population data can then be converted into rates (e.g. percent Latino) and graphed (**Figures 3-8 through 3-12**). As can be seen by the graphs, the relationships only appear to be linear between the 20th and 80th percentiles (approximately). The relationships between these pollution concentrations suggest environmental injustice with the percentages of non-Hispanic Whites decreasing and the percentages of Latinos, those below poverty, and those without a high school degree increasing. The relationship for the percent non-Hispanic Blacks between the 20th and 80th percentiles does not readily suggest a strong relationship. The relationships below 20th and above the 80th percentile often appear to be the opposite of what would be described as environmental injustice, with the percentage of non-Hispanic Whites increasing while the other socio-demographics are decreasing.

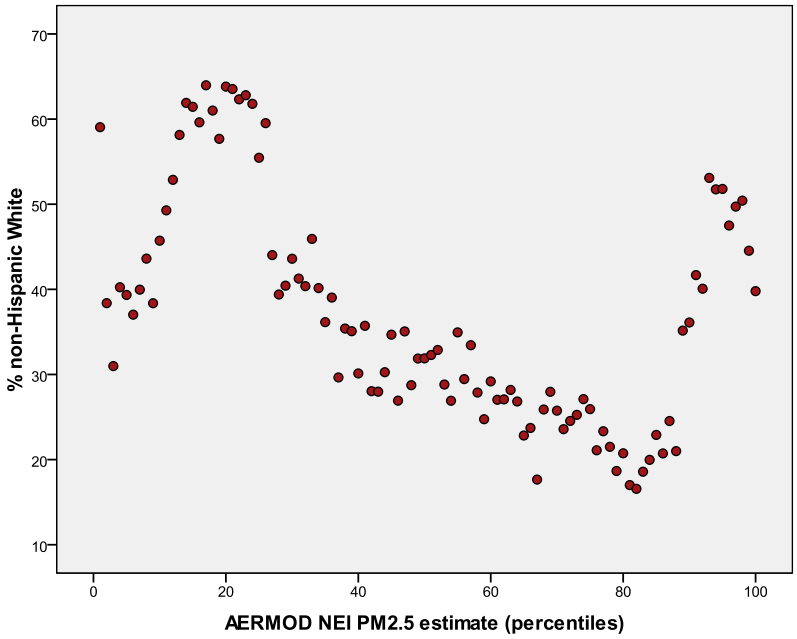


Figure 3-8: Percent non-Hispanic White versus modeled $PM_{2.5}$ concentration from NEI sources by percentile.

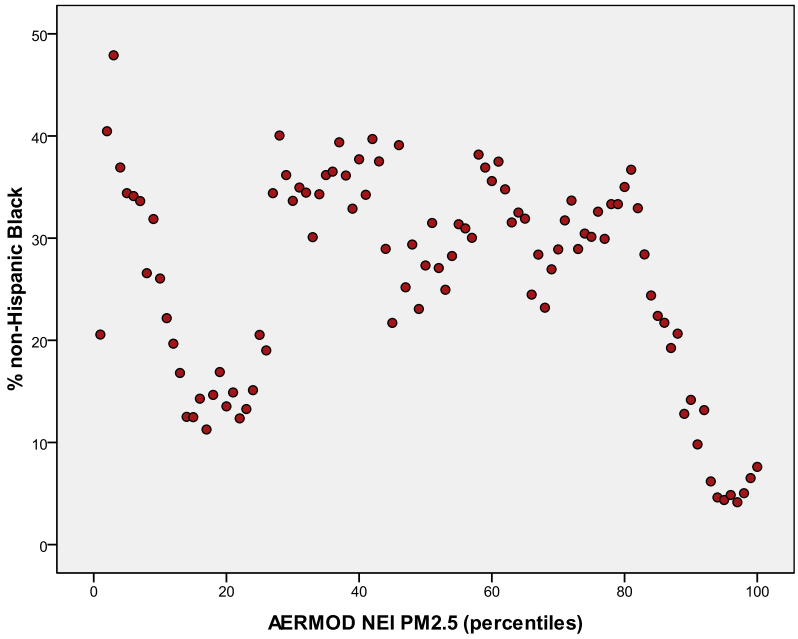


Figure 3-9: Percent non-Hispanic Black versus modeled $PM_{2.5}$ concentration from NEI sources by percentile.

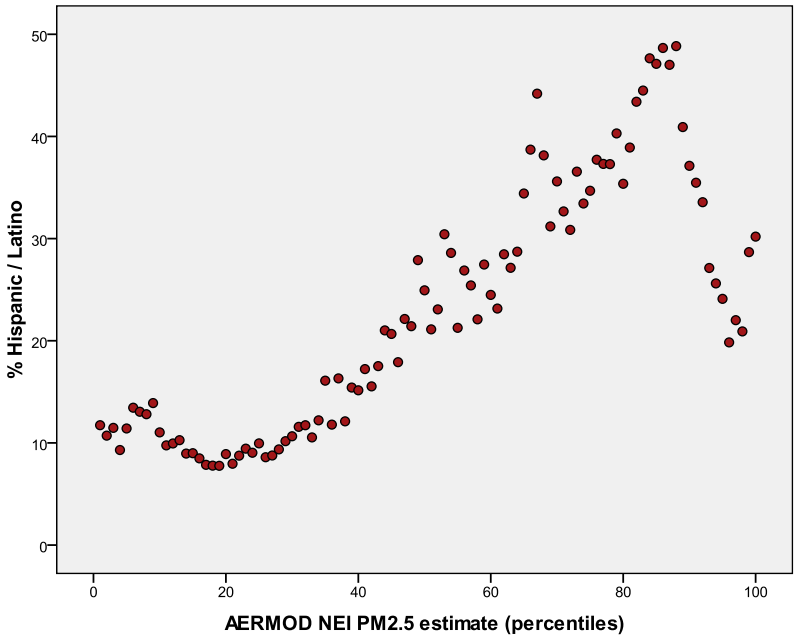


Figure 3-10: Percent Hispanic /Latino versus modeled $PM_{2.5}$ concentration from NEI sources by percentile.

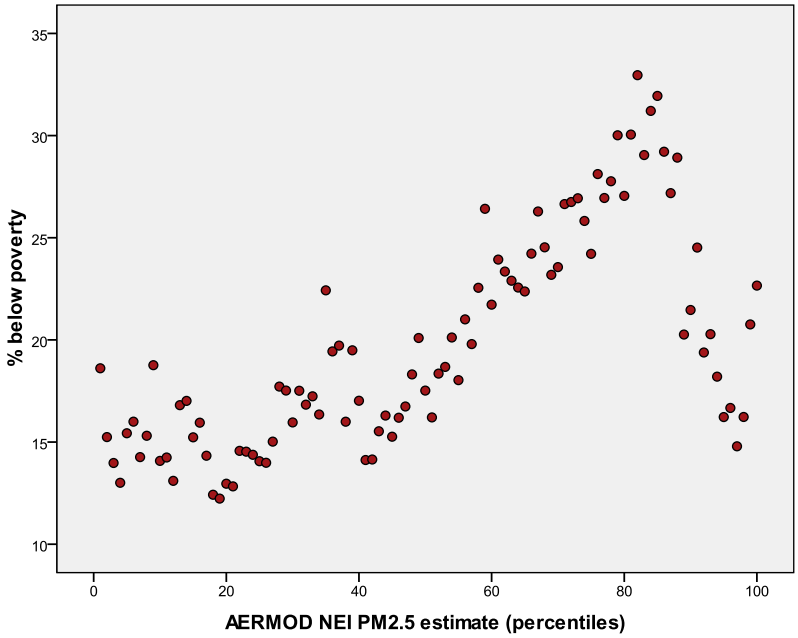


Figure 3-11: Percent below poverty versus modeled $PM_{2.5}$ concentration from NEI sources by percentile.

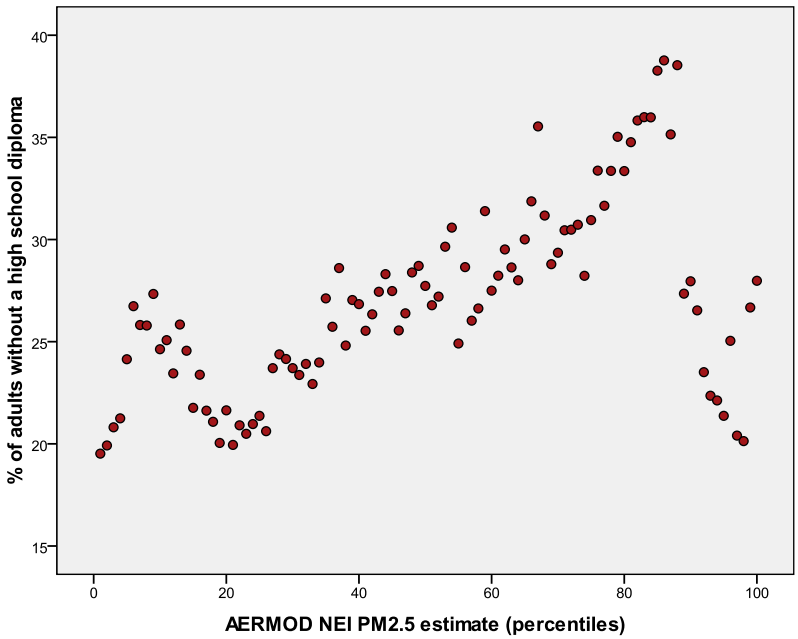


Figure 3-12: Percent of adults without a high school degree versus modeled $PM_{2.5}$ concentration from NEI sources by percentile.

Similar to the proximity analysis, the relationships between modeled $PM_{2.5}$ exposure and socio-demographics can be examined borough by borough in an attempt to detect unusual behavior at a smaller level of aggregation (**Figures 3-13 to 3-17**). It can be seen that some of the boroughs appear to have strong EJ issues, whereas others do not. For instance, the Bronx is the only borough which seems to have a clear negative association between modeled $PM_{2.5}$ exposure from NEI sites and percent non-Hispanic White (**Figure 3-13**). None of the boroughs exhibit consistent relationships between non-Hispanic Blacks and exposure (**Figure 3-14**). The positive correlation between Latinos, percent below poverty, and lack of a high school degree with $PM_{2.5}$

is evident in varying degrees in Brooklyn, the Bronx, and Queens (**Figures 3-15 to 3-17**). Again, these relationships do not always appear linear, particularly for the highest and lowest pollution concentration percentiles.

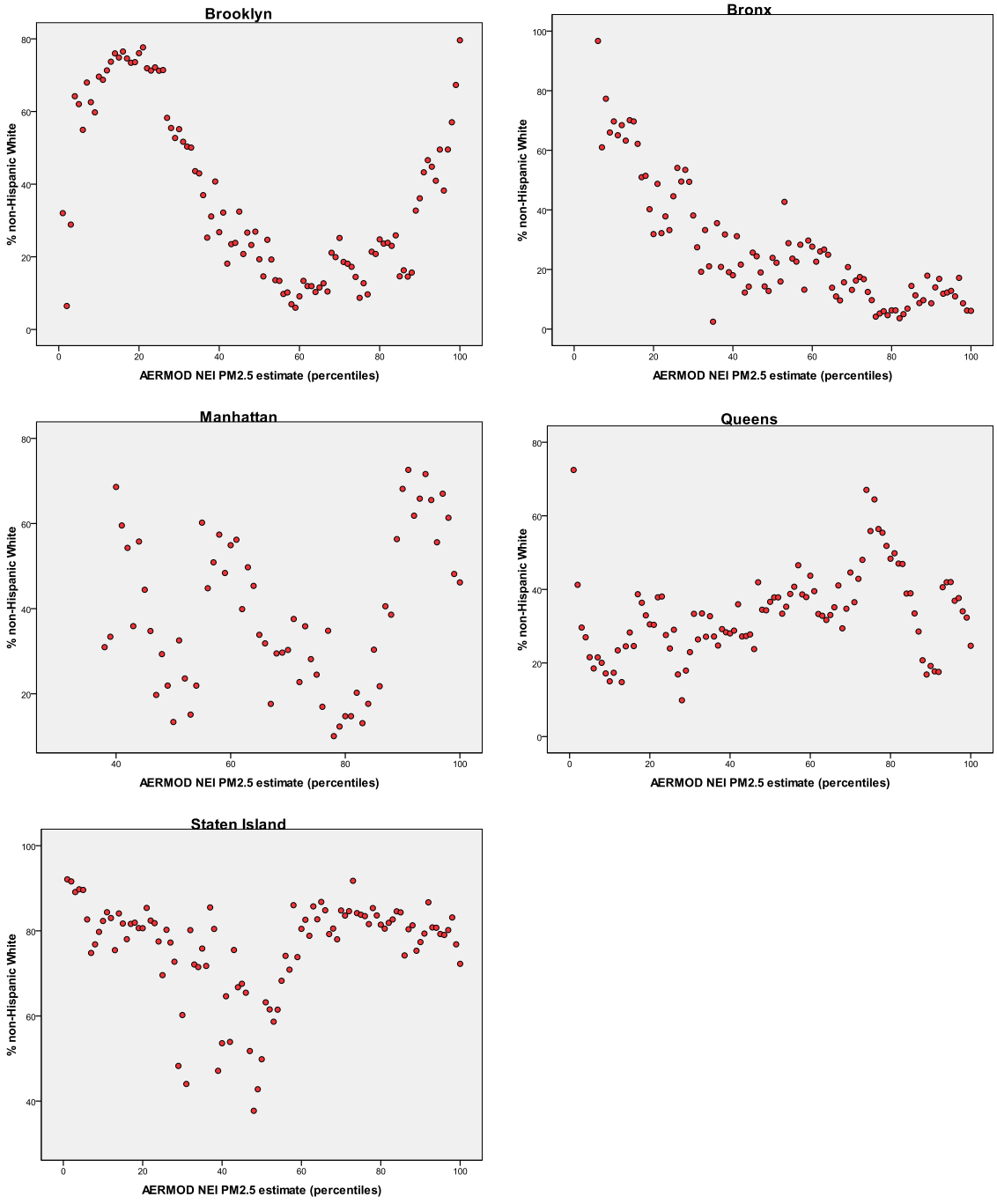


Figure 3-13: Percent non-Hispanic White versus modeled $PM_{2.5}$ concentration from NEI sources by percentile by borough.

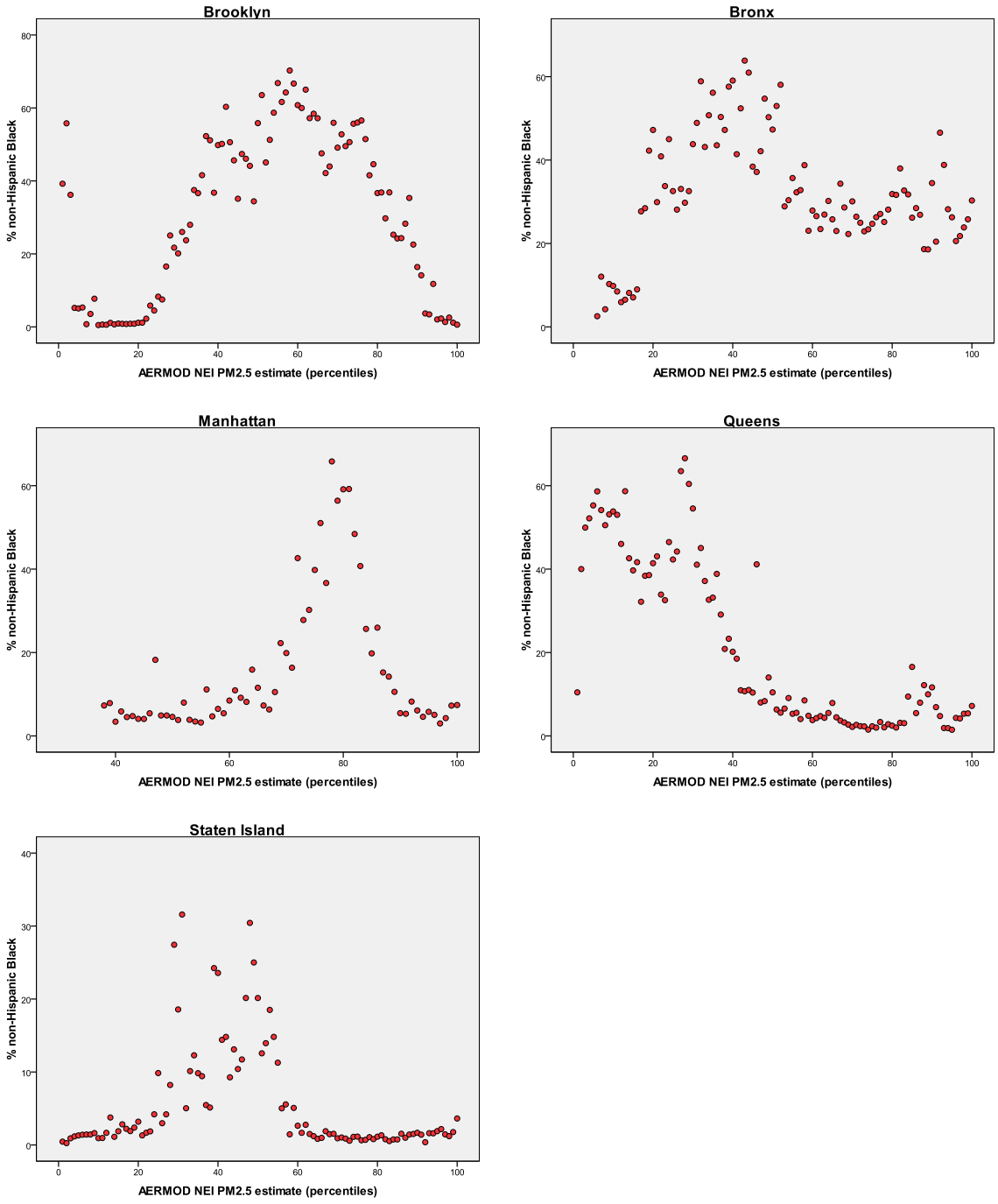


Figure 3-14: Percent non-Hispanic Black versus modeled $PM_{2.5}$ concentration from NEI sources by percentile by borough.

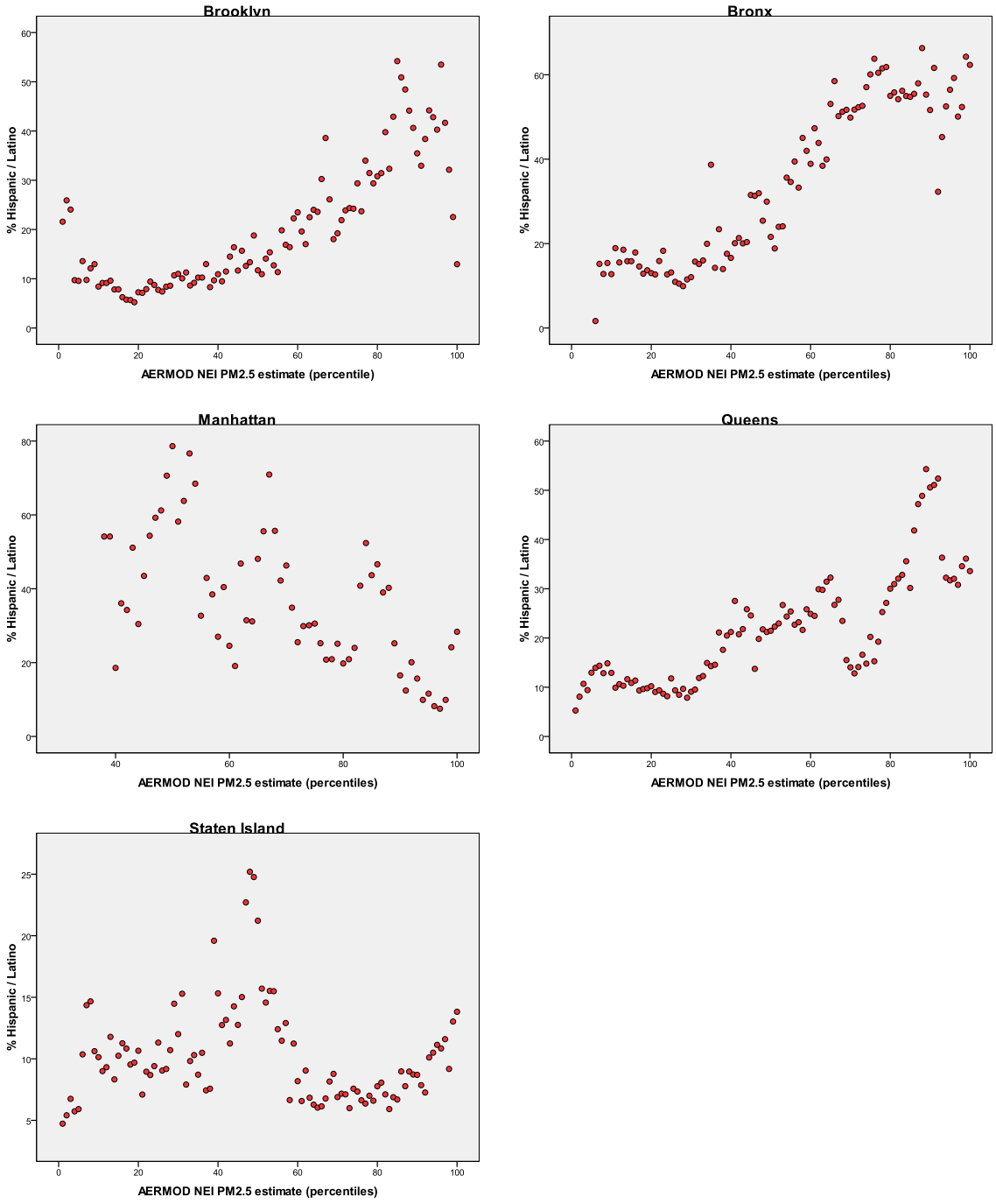


Figure 3-15: Percent Hispanic / Latino versus modeled $PM_{2.5}$ concentration from NEI sources by percentile by borough.

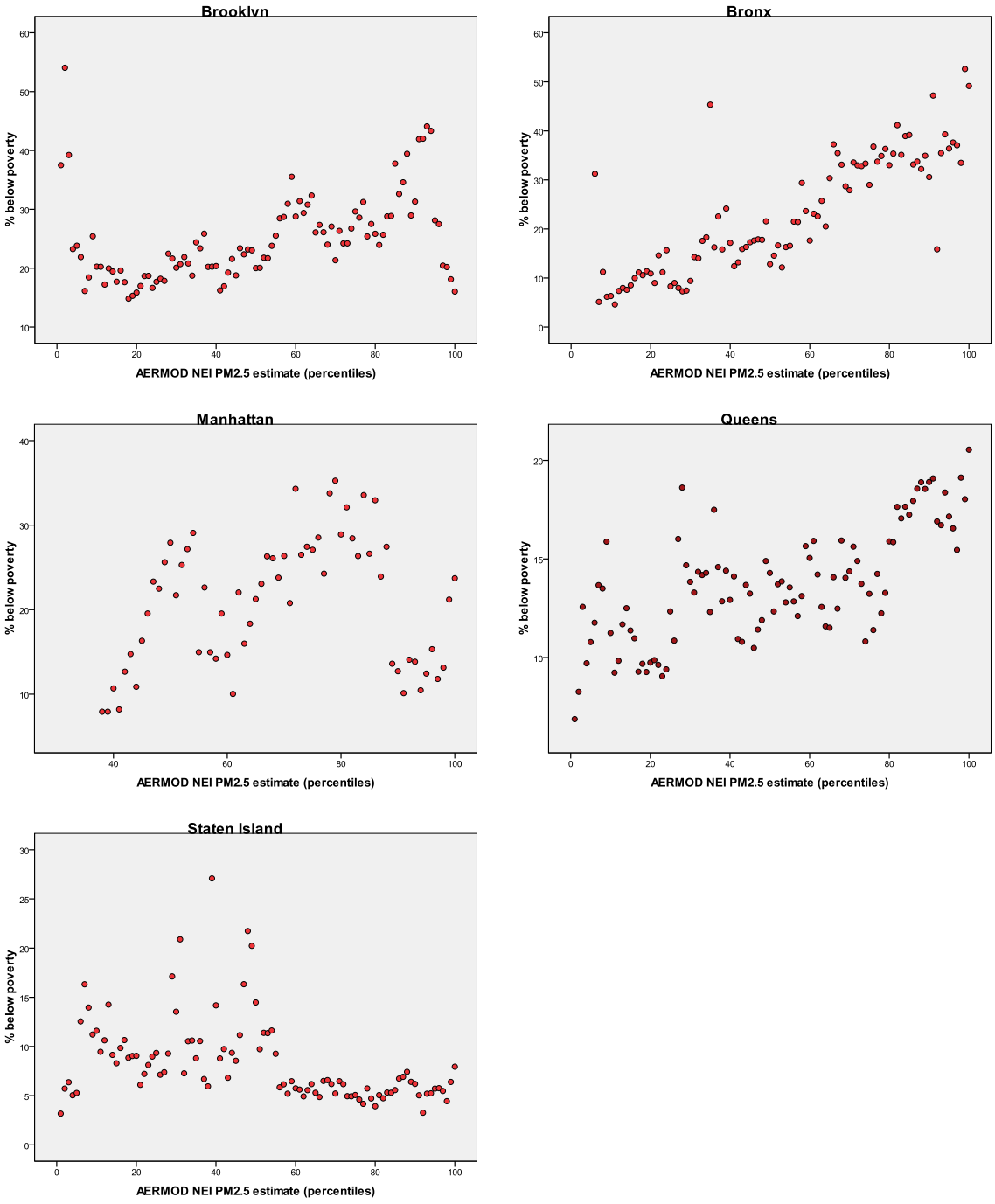


Figure 3-16: Percent below poverty versus modeled $PM_{2.5}$ concentration from NEI sources by percentile by borough.

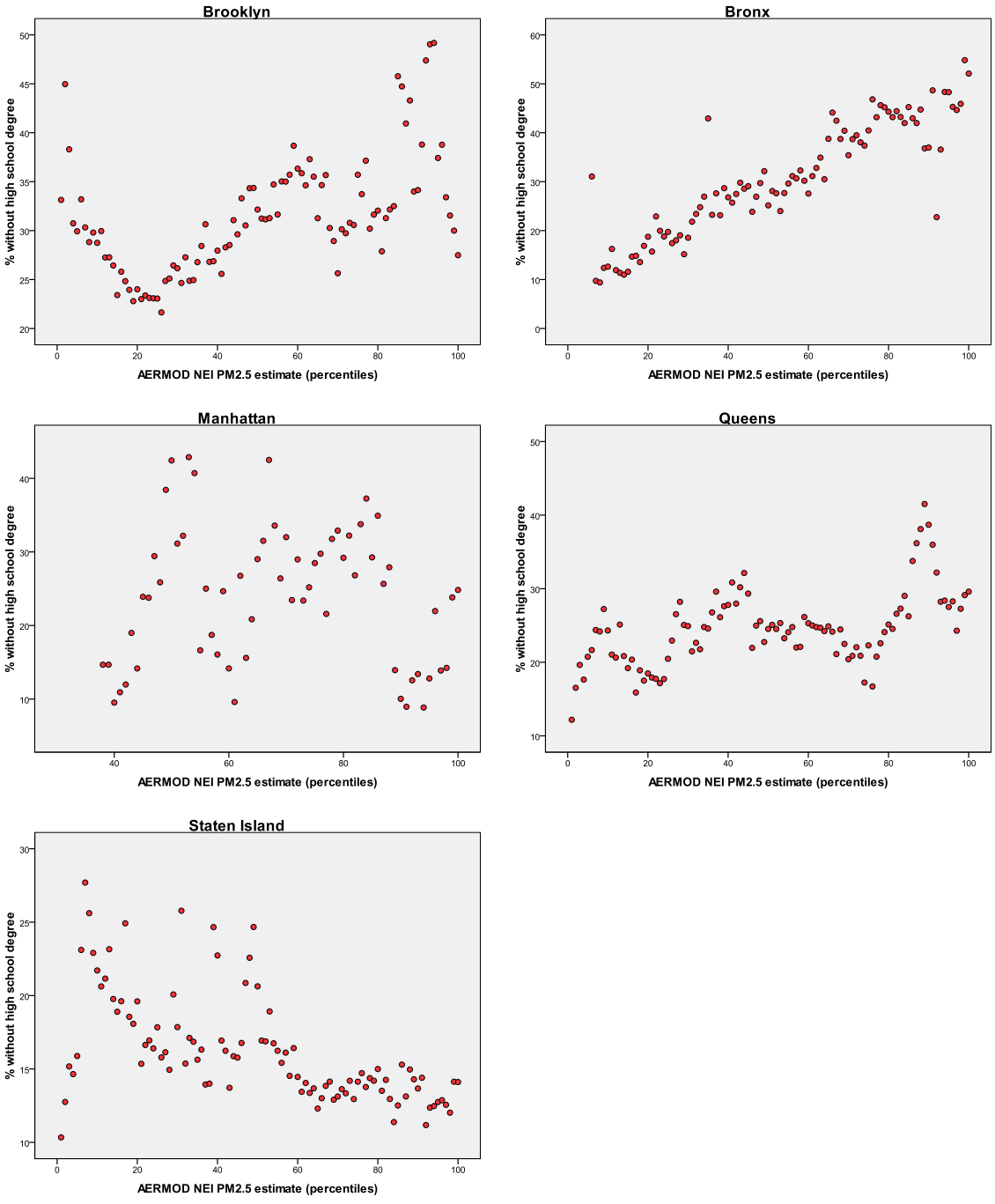


Figure 3-17: Percent of adults without a high school degree versus modeled $PM_{2.5}$ concentration from NEI sources by percentile by borough.

Overall, the AERMOD-estimated $PM_{2.5}$ estimates tell a somewhat different story than the proximity buffers, revealing more subtlety and uncertainty in the pollution/socio-demographic relationships. By examining maps of the proximity buffers and pollution estimates simultaneously, it can be seen that there are resemblances between the buffers and the dispersion estimates (e.g the high concentration of proximity buffers along the northwest portion Queens/Brooklyn border corresponds to high $PM_{2.5}$ concentrations), but it is clear that somewhat different phenomena are being described (**Figure 3-18**).

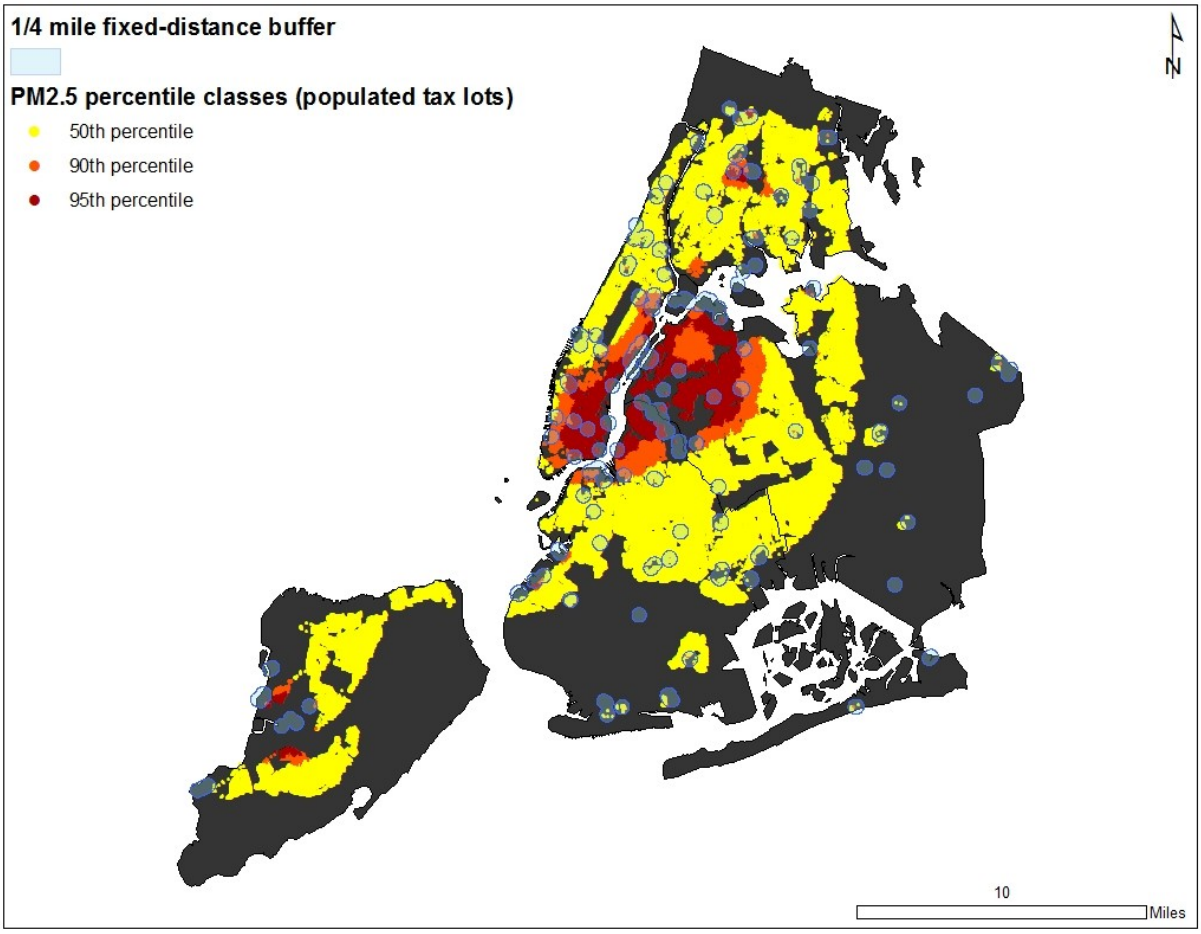


Figure 3-18: Fixed distance proximity buffers and AERMOD-derived $PM_{2.5}$ concentration estimates from NEI sources.

The air dispersion results, however, do not completely refute, and actually bolster the proximity analysis results in certain scenarios. The easiest way to illustrate this is to dichotomize the air dispersion EJ results at different break points and run odds ratios to match those of the proximity analysis. Three break points were chosen based on lot-level pollution percentiles, the 50th (median, where 68% of the population is in the "exposed" class NYC-wide), 90th percentile (where 19% of the population is in the "exposed" class NYC-wide), and 95th percentile (where 9% of the population is in the "exposed" class NYC-wide). Each aggregation was summarized and odds ratios with confidence intervals calculated (**Table 3-3**).

Break Point	Socio-Demographic Group	OR	95% CI	
			lower	upper
50 PERCENTILE	Non-Hispanic White	0.574	0.573	0.576
	Non-Hispanic Black	0.765	0.762	0.768
	Hispanic / Latino	3.305	3.291	3.318
	Below Poverty	1.622	1.615	1.628
	No High School Degree	1.280	1.275	1.286
90 PERCENTILE	Non-Hispanic White	1.849	1.842	1.856
	Non-Hispanic Black	0.187	0.185	0.188
	Hispanic / Latino	*0.997*	*0.993*	*1.001*
	Below Poverty	0.847	0.843	0.851
	No High School Degree	0.765	0.761	0.769
95 PERCENTILE	Non-Hispanic White	1.717	1.709	1.726
	Non-Hispanic Black	0.166	0.164	0.168
	Hispanic / Latino	0.843	0.839	0.848
	Below Poverty	0.805	0.800	0.810
	No High School Degree	0.800	0.794	0.805

Table 3-3: Odds ratios and 95% confidence intervals of socio-demographics and AERMOD-derived $PM_{2.5}$ concentration estimates from NEI sources in NYC. Italicized entries are not significant.

It is very interesting to note that it is only the odds ratios that use the median break point (50th percentile) that detect any sort of environmental injustice with Latinos, those below poverty, and those lacking a high school degree showing increased likelihoods of being present in the "exposed" group. Note that this measure includes over two thirds of the total population of NYC in the "exposed" group. When higher lot-level pollution percentile values are chosen as break points, it is clearly only the non-Hispanic White population that appears to have a disproportionate exposure to $PM_{2.5}$ from NEI sources – a finding that is supported by the percentile graphs above (**Figures 3-8 to 3-12**). These higher break values may be more directly comparable to the proximity OR results since the proportions of the total population that are "exposed" are more similar (19% and 9% of the total population are exposed in the 90th and 95th percentiles, respectively, as compared to 12% for the proximity method).

Borough-specific odds ratios can be calculated as well (similar to what was done with the proximity analysis results). When the estimated $PM_{2.5}$ concentrations from the air dispersion model are compared with socio-demographics in this way, it becomes clear that the relationships are not only complex in terms of non-linearity, but also in terms of spatial heterogeneity (**Table 3-4**). For instance, Brooklyn shows the non-Hispanic White population underrepresented in the 50th percentile of $PM_{2.5}$ concentrations while all the other socio-demographics appear overrepresented. In the 95th percentile of exposure, only non-Hispanic Black is underrepresented, with the rest of the groups showing overrepresentation. Manhattan is reversed in the 50th

percentile of pollution, with only the non-Hispanic White population being overrepresented. These odds ratios, which shift as the break points are changed and the different boroughs are analyzed, clearly demonstrate a very complex series of associations.

Break Point	Socio-demographic Group	NYC	Brooklyn	Bronx	Manhattan	Queens	Staten Island
50 PERCENTILE	Non-Hispanic White	0.574	0.212	0.302	3.330	1.480	0.747
	Non-Hispanic Black	0.765	3.030	0.754	0.196	0.098	1.286
	Hispanic / Latino	3.305	3.371	2.546	0.284	3.068	1.291
	Below Poverty	1.622	1.595	1.927	0.530	1.362	0.930
	No High School Degree	1.280	1.440	1.721	0.499	1.275	0.878
90 PERCENTILE	Non-Hispanic White	1.849	0.936	0.833	1.131	1.125	1.786
	Non-Hispanic Black	0.187	0.380	0.838	0.385	0.175	0.111
	Hispanic / Latino	0.997	3.506	1.287	0.812	1.612	0.616
	Below Poverty	0.847	1.622	1.153	1.097	1.330	0.499
	No High School Degree	0.765	1.489	1.092	1.151	1.122	0.740
95 PERCENTILE	Non-Hispanic White	1.717	1.873	0.888	*0.987*	0.938	1.500
	Non-Hispanic Black	0.166	0.143	1.069	0.428	0.257	0.146
	Hispanic / Latino	0.843	2.365	1.027	*0.987*	1.636	0.726
	Below Poverty	0.805	1.538	1.121	1.239	1.424	0.582
	No High School Degree	0.800	1.396	0.969	1.264	1.148	0.756

Table 3-4: Odds ratios of socio-demographics and AERMOD-derived $PM_{2.5}$ concentration estimates from NEI sources in NYC and its boroughs. Italicized entries are not significant (non-Hispanic White and Hispanic/Latino in Manhattan using the 95th percentile break point). All other entries are significant ($p<.05$).

Another way to visualize the odds ratios for NYC and the boroughs is by graphing them as a function of the break points (**Figure 3-19**). The trends of the changes in OR become a bit more evident in this fashion. When the ORs are above "1" (signified by the dotted black line) the

socio-demographic group is overrepresented in the "exposed" category (50^{th}, 90^{th}, or 95^{th} percentile).

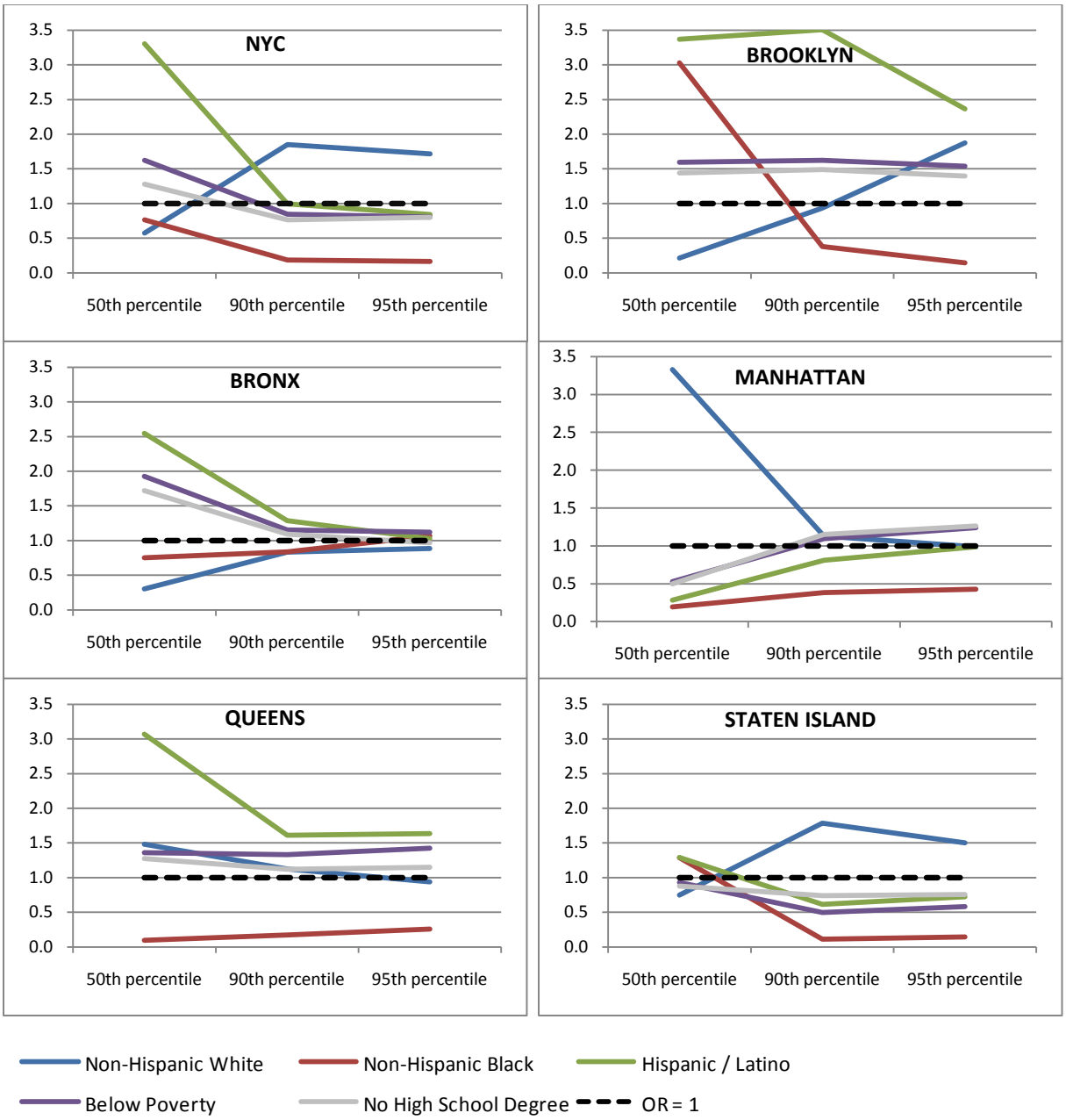


Figure 3-19: Odds ratios of socio-demographics and AERMOD-derived $PM_{2.5}$ concentration estimates from NEI sources in NYC and its boroughs.

3.2 ENVIRONMENTAL HEALTH ANALYSIS

This section quantifies the association between heart failure hospitalization rates with $PM_{2.5}$ concentration estimates and socio-demographics. These analyses were performed using census tracts as the units of analysis in order to match the level of aggregation of the health data (**Section 2.1.2**). Three regression models are used: ordinary least squares (**Section 3.2.1**), spatial autoregressive models (**Section 3.2.2**), and geographically weighted regression (**Section 3.2.3**). Each regression type was run multiple times, using a series of different pollution estimations which include air dispersion estimates from NEI sources as well as land use regression estimates ("raw" and "corrected" for NEI and AADT sources**)**.

3.2.1 ORDINARY LEAST SQUARES REGRESSION

Although ordinary least squares regressions (OLS) are comparatively straightforward, it proved difficult to avoid breaking the requirement of normally distributed residuals. The highly skewed nature of the heart failure hospitalization data (**Figure 3-20**, also see **Section 2.1.2.2**) contributed to highly skewed residuals in the models. As such, the first goal of this section is to find a version of the heart failure data that will result in a normal distribution of regression residuals.

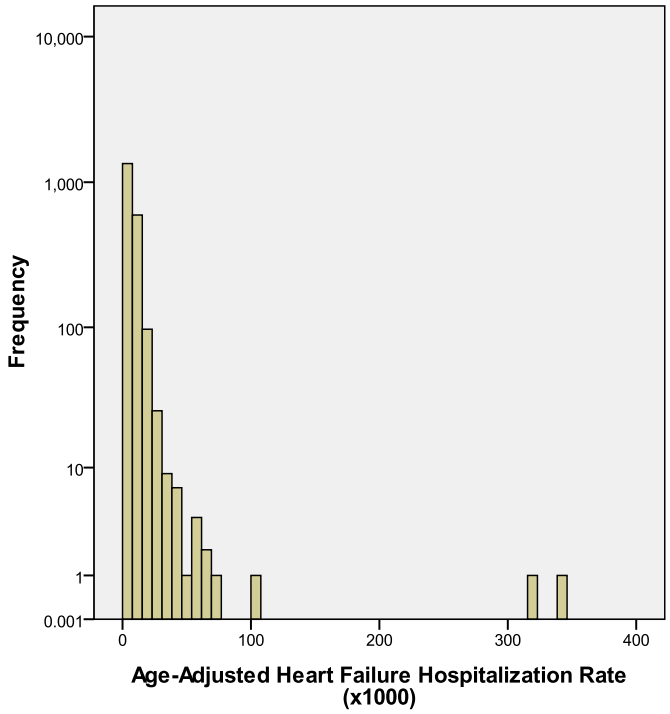


Figure 3-20: Histogram for age-adjusted heart failure rates (2001-2003, inclusive). Vertical axis is logarithmic to enable identification of the positive tail.

3.2.1.1 OLS: AERMOD $PM_{2.5}$ CONCENTRATIONS FROM NEI FACILITIES

The first model explored was an OLS with the complete heart failure hospitalization data per census tract (2001-2003) as the dependent variable and percent non-Hispanic Black, percent Hispanic/Latino, median household income, percent without a high school degree, percent foreign-born, and AERMOD $PM_{2.5}$ concentrations from NEI sources as the independent variables. Although the mobile sources (AADT) and combined NEI and AADT sources were examined, they did not perform as well as the NEI sources alone. This was expected, (**Sections 2.1.3.2** and **2.2.2.2**) and as such are not described in detail in this section. Note that the percent non-Hispanic White variable was not included in any of the models as it demonstrated too much

collinearity with percent non-Hispanic Black and percent-Hispanic/Latino. As was the case in the earlier analyses, census tracts with less than 415 people (lowest 5%) or no hospitalizations over the three years were removed in order to stabilize the data (n=2,048). The R^2 of the regression was very low (.06). There were significant positive associations between hospitalization rates and percent Hispanic/Latino and AERMOD $PM_{2.5}$ concentrations from NEI sources ($p<.05$) and a negative association with percent foreign-born ($p<.05$). These relationships are in the expected directions suggesting increased hospitalization rates for census tracts with higher proportions of Latino populations and increased exposure to fine particulate matter from NEI sources and lower hospitalization rates for tracts with higher proportions of foreign-born residents (while adjusting for other socio-demographics). The issue, aside from the very low R^2, is that the residual distribution was strongly positively skewed (**Figure 3-21**). This violates the assumption of normality and suggests that the data need to be altered in some way in order to function properly in an OLS.

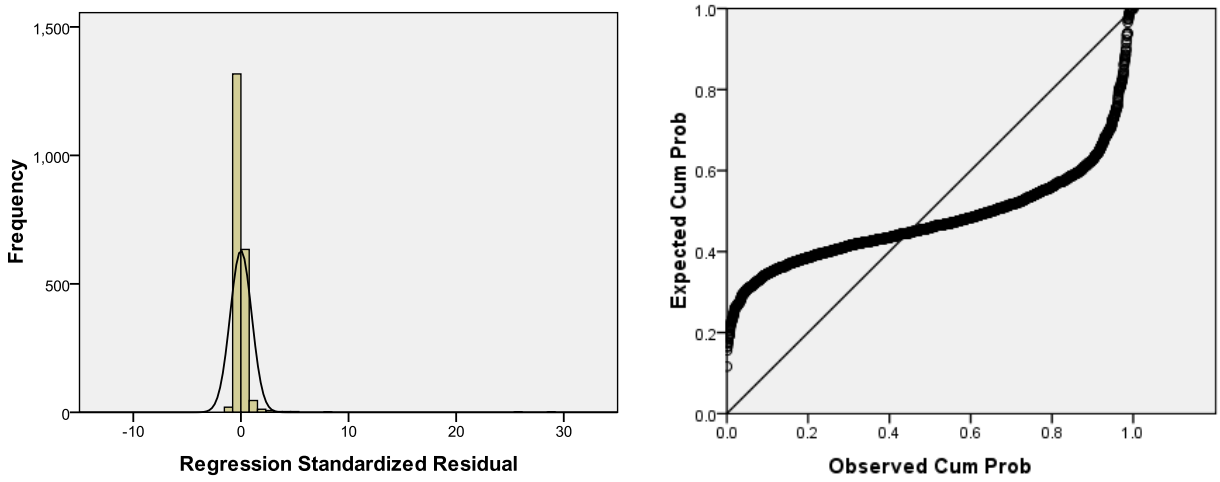


Figure 3-21: Histogram (left) and P-P plot (right) of standardized regression residuals of OLS with untrimmed heart failure hospitalization data. AERMOD NEI $PM_{2.5}$ estimates served as the pollution variable.

One way to resolve the non-normal residual distribution is to transform the dependent variable using a LOG_{10} transformation. When an OLS is run using the log-transformed heart failure hospitalization rate as the dependent variable (n=2,048) and the same independent variables as above, the R^2 improves to .27 (from .06 in the un-transformed model). All of the independent variables were significant ($p<.01$) with percent non-Hispanic Black, percent Hispanic/Latino, percent without a high school degree, and AERMOD $PM_{2.5}$ concentrations from NEI sources showing positive associations. Median household income and percent foreign-born showed negative associations. These relationships behave in accordance with the literature with racial and ethnic minorities, low educational status, and increased exposure being associated with higher hospitalization rates as well as inverse associations between poor health with income and higher proportions of foreign-born residents. The residuals for the log-transformed OLS are more normally distributed as well with only a very slight positive tail (**Figure 3-22**).

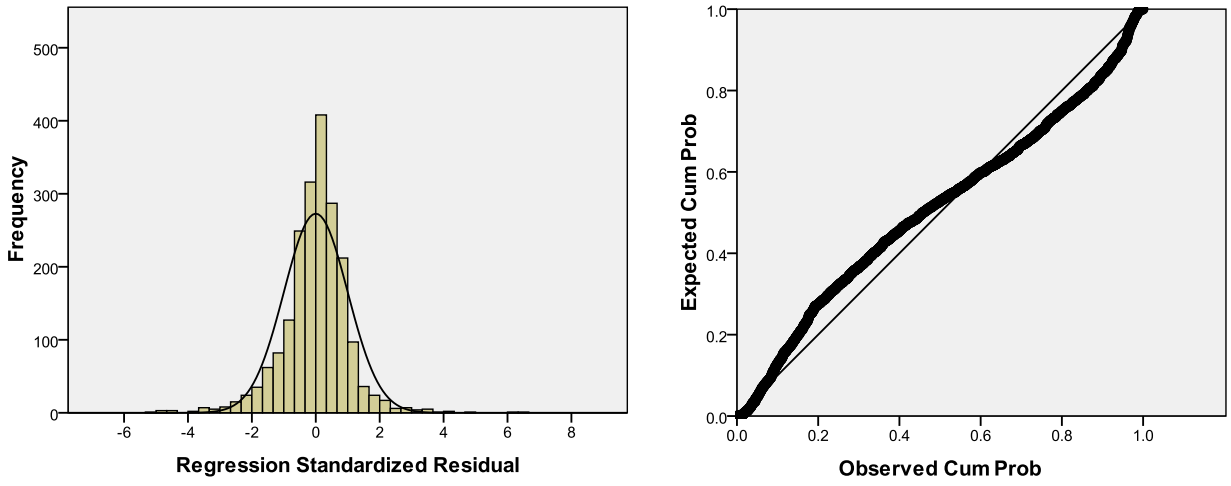


Figure 3-22: Histogram (left) and P-P plot (right) of standardized regression residuals of OLS with untrimmed log-transformed heart failure hospitalization data. AERMOD NEI $PM_{2.5}$ estimates served as the pollution variable.

Although the log-transformed OLS is an acceptable regression, it is not truly modeling the heart failure hospitalization rate, but rather the LOG_{10} of the hospitalization rate. As such, it is inclined towards bias in the predicted values. Another method for normalizing residuals is by trimming some of the high values of the dependent variable. Three trims were explored: (1) removing only the top two heart failure hospitalization rate outliers, (2) removing the census tracts with the highest 3% of hospitalization rates, and (3) removing tracts with the highest 5% of rates (**Figure 3-23**).

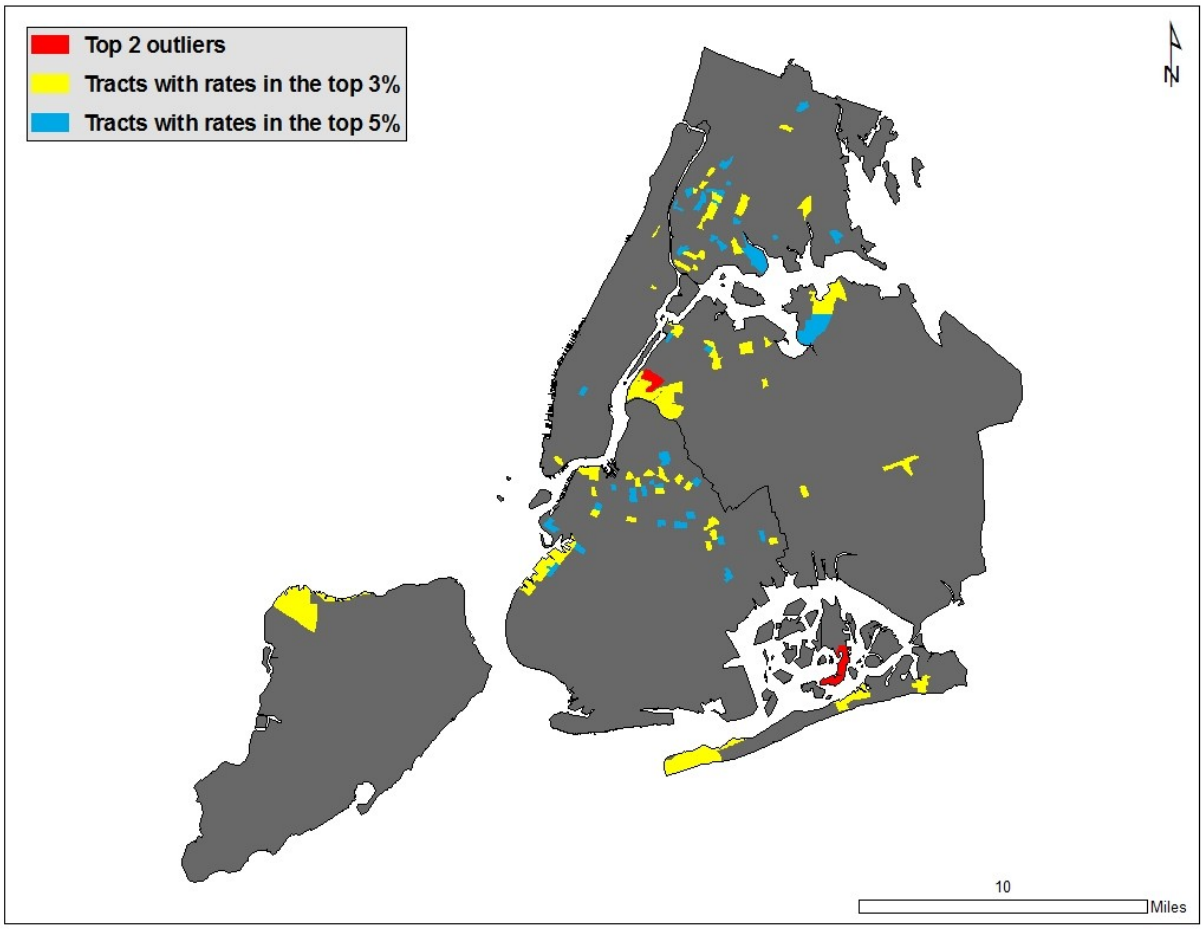


Figure 3-23: Map of heart failure hospitalization rate "trims". Note that the more inclusive trims add to the tracts from less inclusive sets (e.g. tracts with rates in the top 5% include all the tracts from the top 3% and top 2 outliers).

When an OLS is run omitting the two tracts with the highest heart failure hospitalization rates (n=2,046), the R^2 improves to .19 (from .06 in the untrimmed data OLS). Significant positive relationships are detected with the percent Hispanic/Latino, percent who lack a high school degree, and $PM_{2.5}$ concentration from NEI sources ($p<.01$). Percent foreign-born was negatively associated ($p<.01$). Although this is certainly an improvement over the untrimmed data model, the residuals were still positively skewed (**Figure 3-24**).

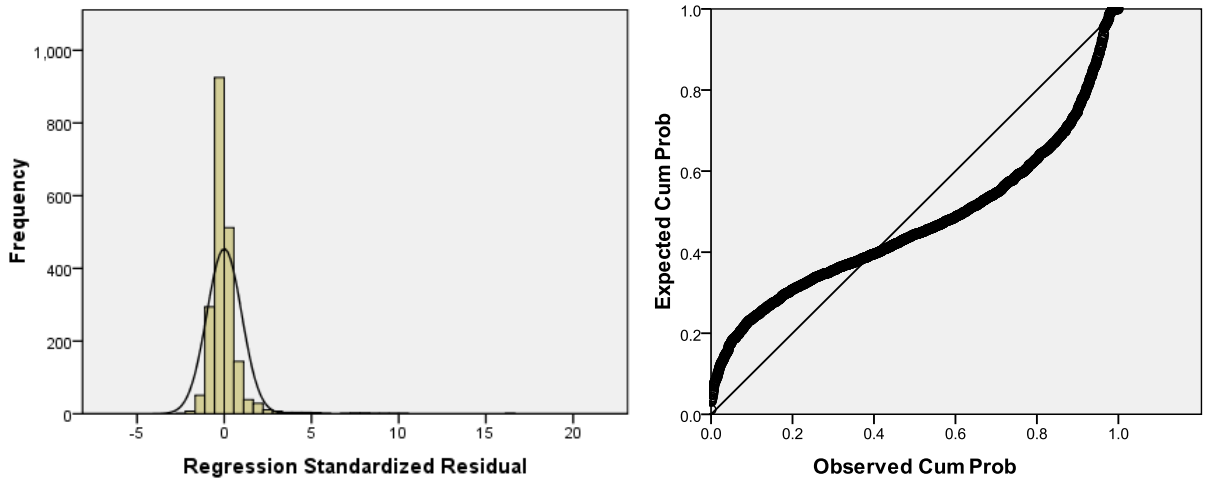


Figure 3-24: Histogram (left) and P-P plot (right) of standardized regression residuals of OLS with the top two outliers removed. AERMOD NEI $PM_{2.5}$ estimates served as the pollution variable.

When the tracts with heart failure hospitalization rates in the highest 3% were trimmed (n=1,987), the OLS R^2 improved dramatically to .35 (from .06 in the untrimmed model). All the variables were significant ($p<.01$) and displayed the same directionality as the log-transformed model. The proportion of racial and ethnic minorities, proportion of residents with low educational status, and higher concentrations of $PM_{2.5}$ from NEI sources were associated with higher hospitalization rates. Higher incomes and higher proportions of foreign-born residents were associated with lower rates. The residuals for the 3% trim OLS approach a normal distribution by losing most of the positive tail present in the previous models (**Figure 3-25**).

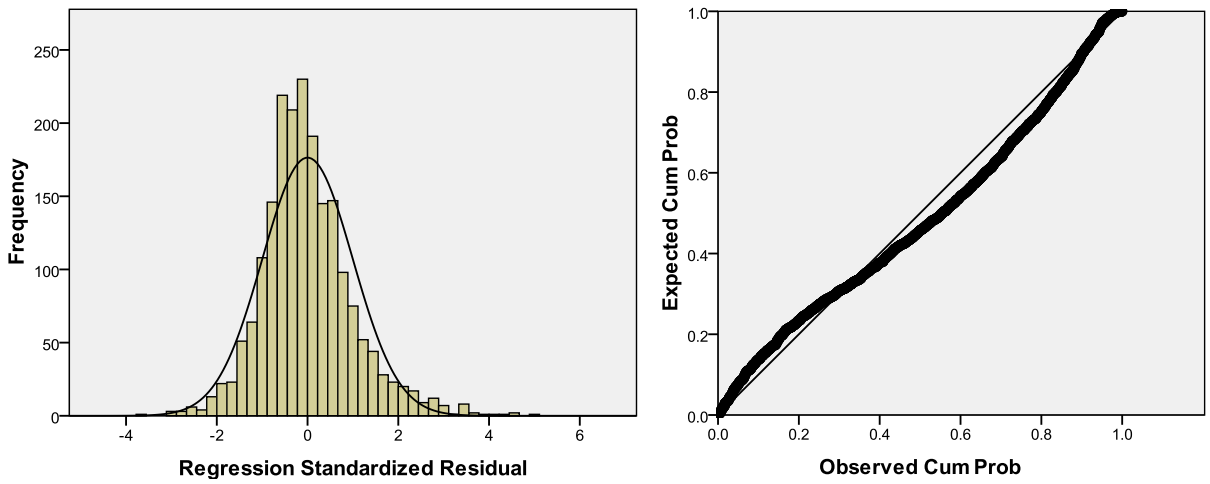


Figure 3-25: Histogram (left) and P-P plot (right) of standardized regression residuals of OLS with the tracts with heart failure hospitalization rates in the top 3% trimmed. AERMOD NEI $PM_{2.5}$ estimates served as the pollution variable.

The final modification of the dependent variable is the trimming of tracts having heart failure hospitalization rate in the highest 5% (n=1,946). This OLS had an R^2 of .33, slightly lower than the 3% trim model (R^2 = .35). The coefficients of the independent variables were all significant and in the same directions as the log-transformed and 3% trim models. The residual distribution was nearly identical to that of the 3% trimmed distribution (**Figure 3-26**).

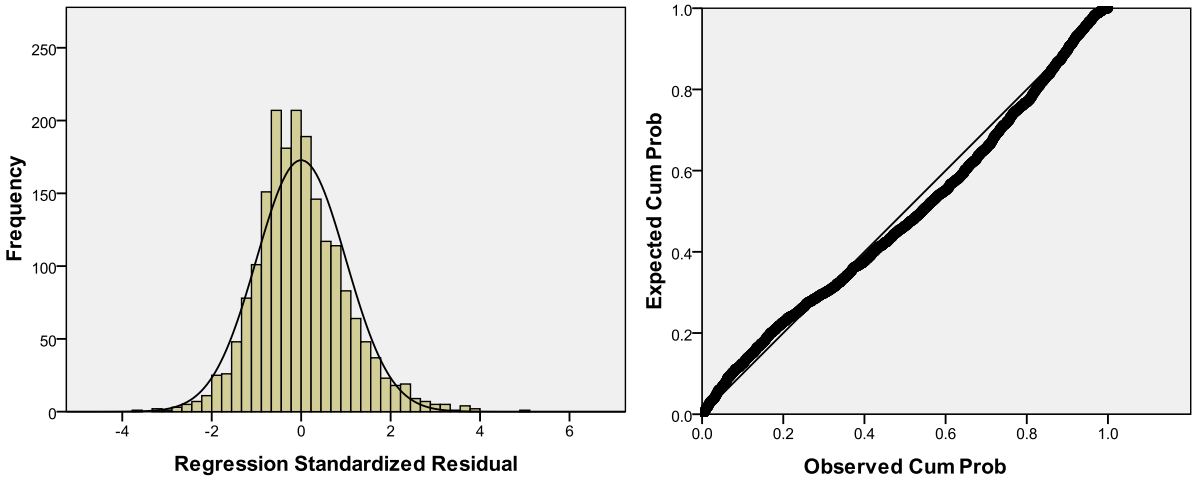


Figure 3-26: Histogram (left) and P-P plot (right) of standardized regression residuals of OLS with the tracts with heart failure hospitalization rates in the top 5% trimmed. AERMOD NEI $PM_{2.5}$ estimates served as the pollution variable.

When the 5 OLS models (untrimmed data, logged hospitalization data, 2 highest outliers trimmed, 3% trimmed, and 5% trimmed) are compared side-by-side (**Table 3-5**), the diagnostics suggest that the 3% trimmed OLS is the best performing since it has the highest R^2, lowest standard error, a normal residual distribution, and all the dependent variables maintain significance ($p<.01$).

MODEL	DIAGNOSTICS				BETAS						
	n	R2	std. err. of estimate	residual distribution	% non-Hispanic Black	% Hispanic/ Latino	% with no high school degree	median household income	% foreign-born	NEI $PM_{2.5}$ concentration	NEI PM2.5 t-value
untrimmed data	2048	.060	.012	positive skew	.038*	.103*	.076	-.054	-.139*	.046*	2.098
log10 of rate	2048	.268	.273	normal	.143*	.170*	.197*	-.145*	-.225*	.052*	2.656
top 2 outliers trimmed	2046	.190	.006	positive skew	.107*	.165*	.214*	-.043	-.199*	.069*	3.372
top 3% trimmed	1987	.351	.003	normal	.162*	.204*	.213*	-.176*	-.265*	.054*	2.870
top 5% trimmed	1946	.331	.003	normal	.170*	.194*	.208*	-.181*	-.249*	.053*	2.768

* Indicates significant ($p<.01$)

Table 3-5: Model comparisons of OLS regressions using untrimmed heart failure data, log-transformed data, and trimmed data.

Even though the 3% trimmed model appears to be the most trustworthy, it should be noted that the hospitalization rates in the excluded tracts do not all appear to be contrary to the OLS results, but rather that the extremely high rates are not fully described by the independent variables in the context of linear regression. It can be seen cartographically that many of the excluded tracts fall in, or near, areas with high $PM_{2.5}$ concentrations (**Figure 3-27**), and some, particularly those in the South Bronx, have classic EJ characteristics in terms of race/ethnicity, education, and income. However, when just the trimmed data (n=61 for the 3% model) are analyzed with either regressions or simple exploratory analysis (e.g., box plots, scatter plots, etc.), there are no readily discernable associations. It is possible that the highest part of the range of hospitalization rates operates in a non-linear fashion to the rest of the data, that those areas have other forces acting upon them that are not captured in the models, or even that there was error in the recording of the raw information from SPARCS.

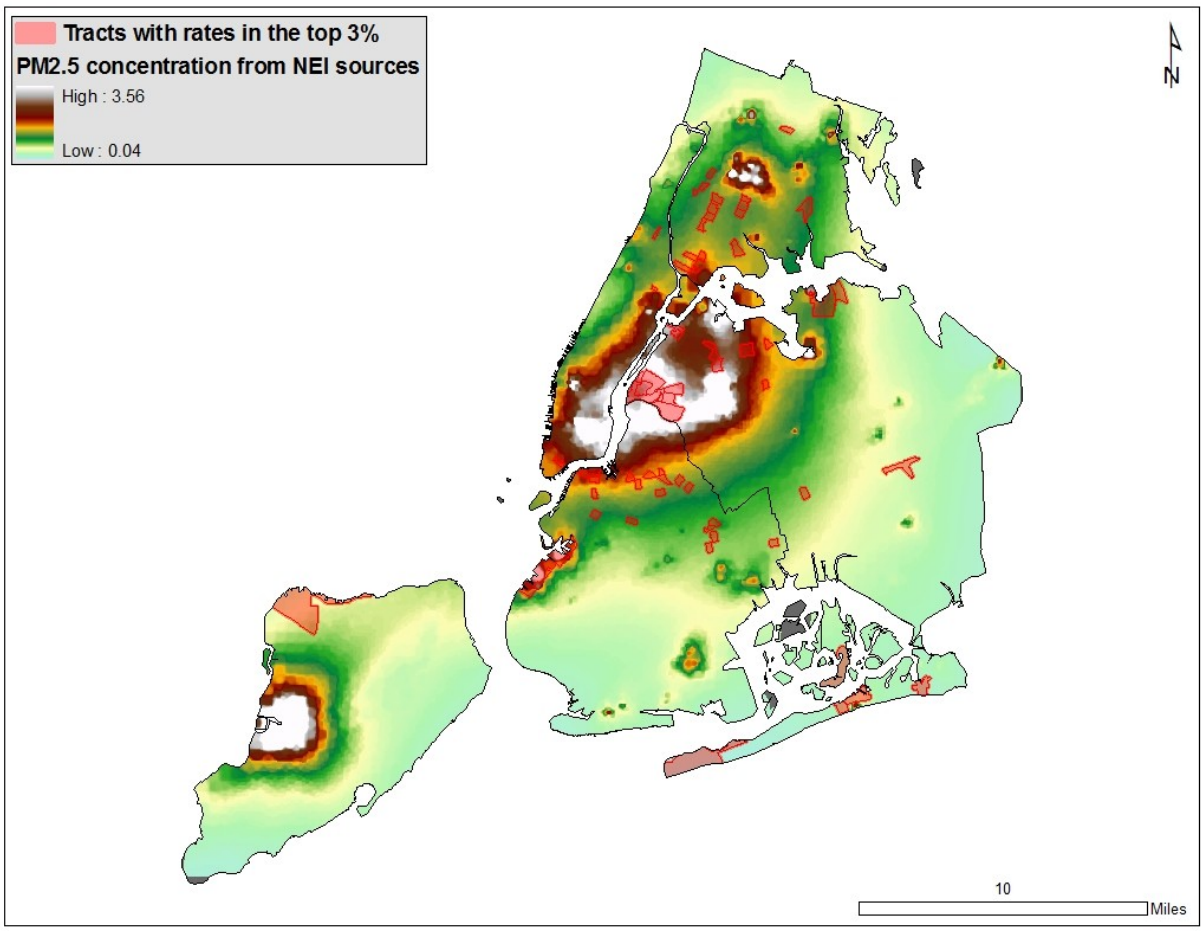


Figure 3-27: Modeled $PM_{2.5}$ concentrations from NEI sources overlaid with tracts with the highest 3% of heart failure hospitalization rates (omitted from the 3% trim models).

3.2.1.2 OLS: LAND USE REGRESSION $PM_{2.5}$ CONCENTRATIONS

The next set of OLS regression run used LUR-derived $PM_{2.5}$ estimates, rather than AERMOD estimates of NEI sources. This will ostensibly result in the quantification of the contribution of ambient $PM_{2.5}$ to heart failure hospitalization rates rather than focusing on the effect just from one specific set of sources. Only the 3% trimmed models will be presented here as the behaviors

of the other versions of the dependent variables (untrimmed data, log-transformed, 2 outliers removed, and top 5% removed) were very similar to the regressions above. Although only the 3% model will be reported with regard to the dependent variable, there were three different versions of the LUR estimate used (uncorrected (LUR), corrected for NEI (LUR_{NEI}), and corrected for AADT and NEI ($LUR_{NEIAADT}$), see **Section 2.2.3** for more information).

The first LUR OLS uses the 3% trimmed heart failure hospitalization rate as the dependent variable, and percent non-Hispanic Black, percent Hispanic/Latino, median household income, percent of adults without a high school degree, percent foreign-born, and ambient $PM_{2.5}$ concentrations as estimated by an 'uncorrected' land use regression model. This resulted in an R^2 of .352 with all independent variables showing significance ($p<.01$) in the same directions as were described in the OLS regressions above. The LUR $PM_{2.5}$ estimates were also significant and positively correlated with heart failure hospitalization rates. The residual were normally distributed (**Figure 3-28**).

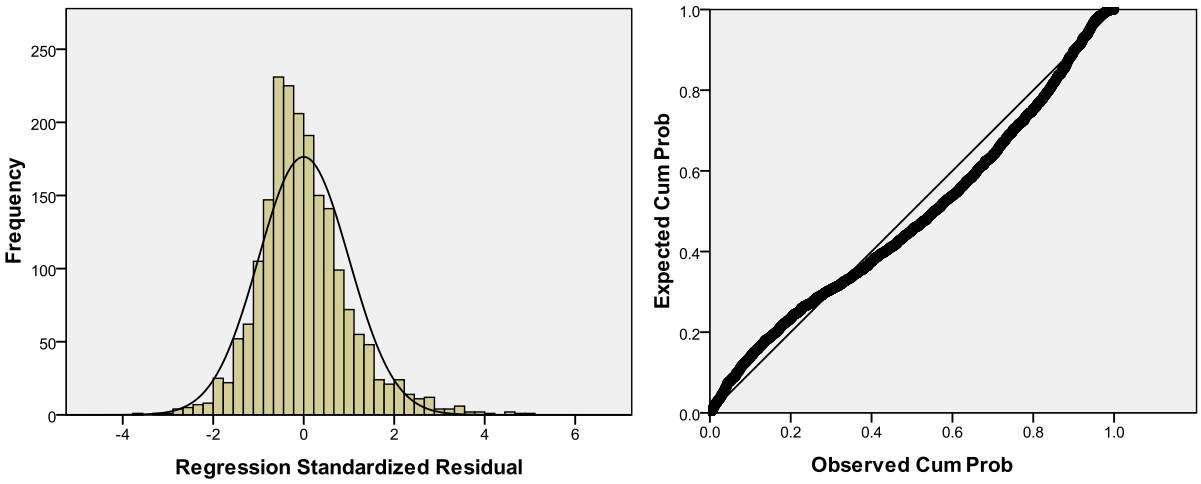


Figure 3-28: Histogram (left) and P-P plot (right) of standardized regression residuals of OLS with the tracts with heart failure hospitalization rates in the top 3% trimmed. LUR $PM_{2.5}$ estimates served as the pollution variable.

When the LUR_{NEI} variable was used (LUR corrected for NEI $PM_{2.5}$ concentrations as estimated by AERMOD), the R^2 improved very slightly (from .352 to .354) with little else in terms of model diagnostics changing. The use of the $LUR_{NEIAADT}$ variable (LUR corrected for both NEI and selected mobile sources) resulted in a model which was nearly identical to the model which utilized LUR_{NEI} variable. The differences between these three LUR variables are most visible in terms of the variable coefficients, t-values, and partial correlations of the pollution variable (**Table 3-6**).

	DIAGNOSTICS				BETAS							
LUR VARIABLE	n	R2	std. err. of estimate	residual distribution	% non-Hispanic Black	% Hispanic/ Latino	% with no high school degree	median household income	% foreign-born	LUR_{NEI} $PM_{2.5}$ concentration	LUR_{NEI} PM2.5 t-value	LUR_{NEI} PM2.5 partial correlation
LUR	1987	.352	.003	normal	.157*	.192*	.212*	-.177*	-.267*	.066*	3.479	.078
LUR_{NEI}	1987	.354	.003	normal	.157*	.184*	.213*	-.177*	-.265*	.083*	4.304	.096
$LUR_{NEIAADT}$	1987	.354	.003	normal	.156*	.186*	.212*	-.177*	-.266*	.079*	4.102	.092

* Indicates significant (p<.01)

Table 3-6: Model comparisons of OLS regressions using 3% trimmed hospitalization data and LUR-derived $PM_{2.5}$ estimates.

Notice that the main difference between the OLS regressions using LUR estimates is the information regarding the LUR variables themselves. For instance, it is the LUR_{NEI}-based model which has the highest standardized coefficient for the $PM_{2.5}$ concentration estimate (.083) suggesting that it is that model in which pollution has the biggest influence. The LUR_{NEI} model also has the highest t-value (4.3) for the $PM_{2.5}$ estimate, indicating that there can be more confidence in the LUR_{NEI} variable than either LUR or $LUR_{NEIAADT}$, and the highest partial correlation (.096) signifying that LUR_{NEI} is most correlated with the hospitalization rate when adjusting for the other independent variables. While this correlation is still quite low, it is higher than the other LUR variables, particularly the 'uncorrected' version (.078). These results, although subtle, support the ideas put forth in the LUR section of this dissertation (**Section 2.2.3**), namely that the annual average daily traffic data is not complete enough to rely upon and that correction for estimated $PM_{2.5}$ from NEI sources improves the LUR estimate.

The remainder of the models (SAR and GWR) will focus solely on the better performing OLS regressions, namely the 3% data with AERMOD NEI $PM_{2.5}$ estimates and the 3% data with the LUR_{NEI} $PM_{2.5}$ estimates.

3.2.2 SPATIAL AUTOREGRESSIVE MODELS

Spatial autoregressive modeling (SAR) was the next type of regression explored. As was discussed earlier in this dissertation (**Section 6.1.2**), an SAR is quite similar to an OLS with the addition of a spatial term which either represents a spatially lagged version of the dependent variable (rho) or a spatial error term (lambda). The choice of model (spatial lag or spatial error) is most often determined by the LM statistics. The spatial weights matrix was defined by exploring different options based on the dependent variable and selecting the one with the highest Moran's I. This was determined to be 1st order queens contiguity with a Moran's I value of .45 (**Figure 3-29**)

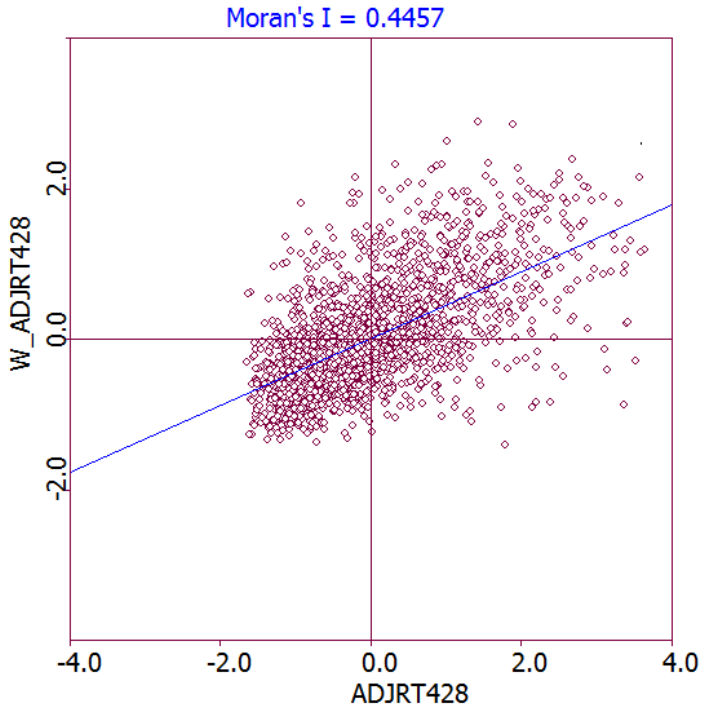


Figure 3-29: Moran's I using 1st order queens contiguity. ADJRT428 = standardized heart failure hospitalization rate; W_ADJRT428 = standardized spatially lagged version of the hospitalization rate.

The first SAR run used the trimmed (3%) heart failure hospitalization data per census tract (2001-2003) as the dependent variable and percent non-Hispanic Black, percent Hispanic/Latino, median household income, percent without a high school degree, percent foreign-born, and AERMOD $PM_{2.5}$ concentrations from NEI sources as the independent variables (n=1987). Census tracts with low populations (<415) were omitted. By examining the Lagrange Multiplier statistics (LM), the spatial lag model was determined to be the most appropriate since both LM(lag) and LM(error) are significant, but robust LM(error) does not retain significance whereas robust LM(lag) does (**Table 3-7**).

Test	Value	Probability
Lagrange Multiplier (lag)	181.67	0.000
Robust LM (lag)	45.07	0.000
Lagrange Multiplier (error)	137.88	0.000
Robust LM (error)	1.28	0.259

Table 3-7: Lagrange Multipliers using 1st order queen's contiguity. 3% heart failure trimmed dependent variable, AERMOD $PM_{2.5}$ estimate from NEI sources is the pollution variable.

The spatial lag model resulted in an R^2 of .412 with all the independent variables retaining significance ($p<.01$) and directionality (percent Hispanic/Latino, percent non-Hispanic Black, percent without a high school degree, and AERMOD NEI $PM_{2.5}$ concentration were positively associated with heart failure hospitalization rate, whereas percent foreign-born and median household income were negatively associated). The newly introduced lagged version of the heart

failure hospitalization rate (rho) also showed a significant positive correlation. It is the addition of this variable that is the main cause of the increase in the model's R^2 from that of the OLS (OLS R^2 = .351; SAR R^2 = .412).

When an SAR is run replacing the AERMOD $PM_{2.5}$ concentration estimate from NEI facilities with the land use regression-derived ambient $PM_{2.5}$ concentration estimate corrected for NEI $PM_{2.5}$ (LUR_{NEI}), the LM statistics once again point to the spatial lag model (**Table 3-8**). The R^2 was .411 with all the independent variables showing significance and behaving the same way as the SAR above.

Test	Value	Probability
Lagrange Multiplier (lag)	167.42	0.000
Robust LM (lag)	45.91	0.000
Lagrange Multiplier (error)	123.96	0.000
Robust LM (error)	2.44	0.118

Table 3-8: Lagrange Multipliers using 1st order queen's contiguity. 3% heart failure trimmed dependent variable, LUR_{NEI} $PM_{2.5}$ estimate (corrected for AERMOD NEI) is the pollution variable.

It can be interesting and informative to chart the OLS and SAR results side-by-side. One issue was that OpenGeoDa, the software running the SAR models, does not report standardized regression coefficients (betas) in the outputs. In order to create them, the variables were

standardized by subtracting the mean and dividing by the standard deviation, and input into the SAR again and re-run. This way, it is easier to make direct comparisons between both the regression techniques and the pollution variables (**Table 3-9**).

MODEL	DIAGNOSTICS				BETAS							$PM_{2.5}$ t-values
	n	R^2	std. err. of estimate	AIC	spatial lag (rho)	% non-Hispanic Black	% Hispanic/ Latino	% with no high school	median household income	% foreign-born	$PM_{2.5}$ Estimate	
OLS: AERMOD NEI	1987	.351	.003	-17088	--	.162*	.204*	.213*	-.176*	-.265*	.054*	2.870
SAR: AERMOD NEI	1987	.412	.003	-17238	.350*	.104*	.128*	.175*	-.108*	-.178*	.049*	2.767
OLS: LUR_{NEI}	1987	.354	.003	-17098	--	.157*	.184*	.213*	-.177*	-.265*	.083*	4.304
SAR: LUR_{NEI}	1987	.411	.003	-17237	.341*	.100*	.121*	.173*	-.111*	-.180*	.050*	2.676

* Indicates significant (p<.01)

Table 3-9: Model comparisons of OLS and SAR regressions using 3% trimmed hospitalization data, AERMOD-derived NEI $PM_{2.5}$ estimates, and LUR-derived $PM_{2.5}$ estimates (corrected for NEI sources).

Notice that the R^2 values for both SAR models are higher than their OLS counterparts. This is due to the significance, and high BETA, for the spatial lag variables (rho). More importantly, the Akaike Information Criterion (AIC) is lower in the SAR models as compared to the OLS models. This is thought to be a better diagnostic for a model's fit in these scenarios. The addition of the spatial lag variables reduces the influence of all the other independent variables; however they all retain their significance and direction. For instance, the BETAS and t-values for the $PM_{2.5}$ estimates are lowered in the SAR models, but the generic model is robust enough regarding the impact of $PM_{2.5}$, either from NEI sources alone or ambient concentrations, to still be detectible and significant.

3.2.3 GEOGRAPHICALLY WEIGHTED REGRESSION MODELS

The final type of regression that was employed is geographically weighted regression (GWR). As was discussed earlier in this dissertation (**Section 1.6.3**), GWR allows the regression coefficients to vary over space. This can be quite interesting and informative if your model is truly spatially non-stationary, misspecified, or otherwise spatially biased in some way. As with the SAR models, only the 3% trimmed AERMOD NEI and 3% trimmed LUR_{NEI} models will be reported on.

The centroid for each census tract (n=1,987) served as the local regression points in the GWR models. Both adaptive bandwidths (nearest neighbors) and fixed bandwidth (distance) were explored. Ultimately, the adaptive kernel produced better fitting models and as such was used for these analyses. When the AERMOD NEI $PM_{2.5}$ concentration is used as the pollution variable, using the same socio-demographics and health outcome variables as the previous models, the GWR used 873 nearest neighbors (out of 1987). This suggests a somewhat 'global' and stable model since nearly half of the samples are used in each local regression. Both the R^2 and AIC were improvements over the OLS (GWR R^2 = .41, OLS R^2 = .35; GWR AIC = -17289, OLS AIC = -17088). The 5-number summaries of the parameter estimates (minimum, 1st quartile, median, 3rd quartile, and maximum) are relatively stable, with no sign changes between the first and third quartiles (**Table 3-10**). The parameter estimates also demonstrate statistically significant spatial variability ($p<.001$) according to the Monte Carlo test.

Variable	Minimum	1st quartile	Median	3rd quartile	Maximum
% non-Hispanic Black	-1.70E-05	1.96E-05	2.33E-05	3.35E-05	5.30E-05
% Hispanic/ Latino	3.61E-08	2.70E-05	3.37E-05	5.53E-05	7.78E-05
% with no high school	-3.52E-06	4.02E-05	6.68E-05	9.30E-05	1.22E-04
median household income	-7.37E-08	-4.76E-08	-3.18E-08	-1.77E-08	1.41E-08
% foreign-born	-9.56E-05	-7.36E-05	-5.69E-05	-4.36E-05	-1.18E-05
AERMOD $PM_{2.5}$	-1.76E-03	4.92E-04	1.84E-03	4.19E-03	2.14E-02

Table 3-10: 5-number summaries of GWR parameter estimates from the 3% trim model using AERMOD $PM_{2.5}$ estimates.

The GWR that included land use regression estimates corrected for NEI $PM_{2.5}$ for the pollution variable (LUR_{NEI}) utilized only 407 out of 1987 nearest neighbors – a more 'local' model than the AERMOD version which used 873 samples per local regression. It had an R^2 of .44 and an AIC of -17334, both of which are improvements over the corresponding OLS model. The 5-number summaries (**Table 3-11**) display more sign-switching than the AERMOD NEI model, particularly in the LUR_{NEI} variable which switched from negative to positive between the 1st quartile and median (however, the local regressions that resulted in negative $PM_{2.5}$ parameter estimates were only significant in a very small area of south-western Brooklyn). All variables showed spatial variability ($p<.01$) except for percent without a high school degree which was significant only at the 5% level.

Variable	Minimum	1st quartile	Median	3rd quartile	Maximum
% non-Hispanic Black	-5.72E-05	1.45E-05	2.67E-05	3.84E-05	7.06E-05
% Hispanic/ Latino	-1.14E-05	2.84E-05	4.09E-05	5.73E-05	9.88E-05
% with no high school	-7.18E-05	2.42E-05	6.43E-05	8.63E-05	1.55E-04
median household income	-1.10E-07	-5.95E-08	-2.27E-08	-4.74E-09	4.38E-08
% foreign-born	-1.20E-04	-7.28E-05	-3.98E-05	-1.90E-05	1.16E-05
LUR_{NEI} $PM_{2.5}$	-9.36E-04	-6.25E-05	1.34E-04	3.77E-04	2.11E-03

Table 3-11: 5-number summaries of GWR parameter estimates from the 3% trim model using LUR_{NEI} $PM_{2.5}$ estimates.

The model diagnostics can be viewed in tabular form similar to the OLS and SAR results (**Table 3-12**). Note that the parameter estimates are not standardized (BETAS) in this table and ranges are provided for parameter estimates and t-values since a statistic was produced at each local regression point (census tract centroid).

MODEL	DIAGNOSTICS				PARAMETER ESTIMATES						$PM_{2.5}$ t-values
	n	R2	std. err. of estimate	AIC	% non-Hispanic Black	% Hispanic/ Latino	% with no high school	median household income	% foreign-born	PM2.5 Estimate	
AERMOD NEI	1987	.429	.00309	-17289	-1.70E-05 to 5.30E-05	3.61E-08 to 7.78E-05	-3.5E-06 to 1.22E-04	-7.37E-08 to 1.41E-08	-9.56E-05 to -1.18E-05	-1.76E-03 to 2.14E-02	-1.19 to 5.61
LUR_{NEI}	1987	.464	.00303	-17334	-5.72E-05 to 7.06E-05	-1.14E-05 to 9.88E-05	-7.2E-05 to 1.55E-04	-1.10E-07 to 4.38E-08	-1.20E-04 to 1.16E-05	-9.36E-04 to 2.11E-03	-2.06 to 2.98

Table 3-12: Model comparisons of GWR models using 3% trimmed hospitalization data, AERMOD-derived NEI PM2.5 estimates, and LUR-derived $PM_{2.5}$ estimates (corrected for NEI sources).

As was written earlier, it is often most informative to map the GWR output including t-values (**Figures 3-30** and **3-31**), residuals, and local R^2 (**Figure 3-32**). It is readily apparent that the GWR model that uses LUR_{NEI} as the $PM_{2.5}$ variable behaves much more locally than the AERMOD NEI model, with more variation in the outputs.

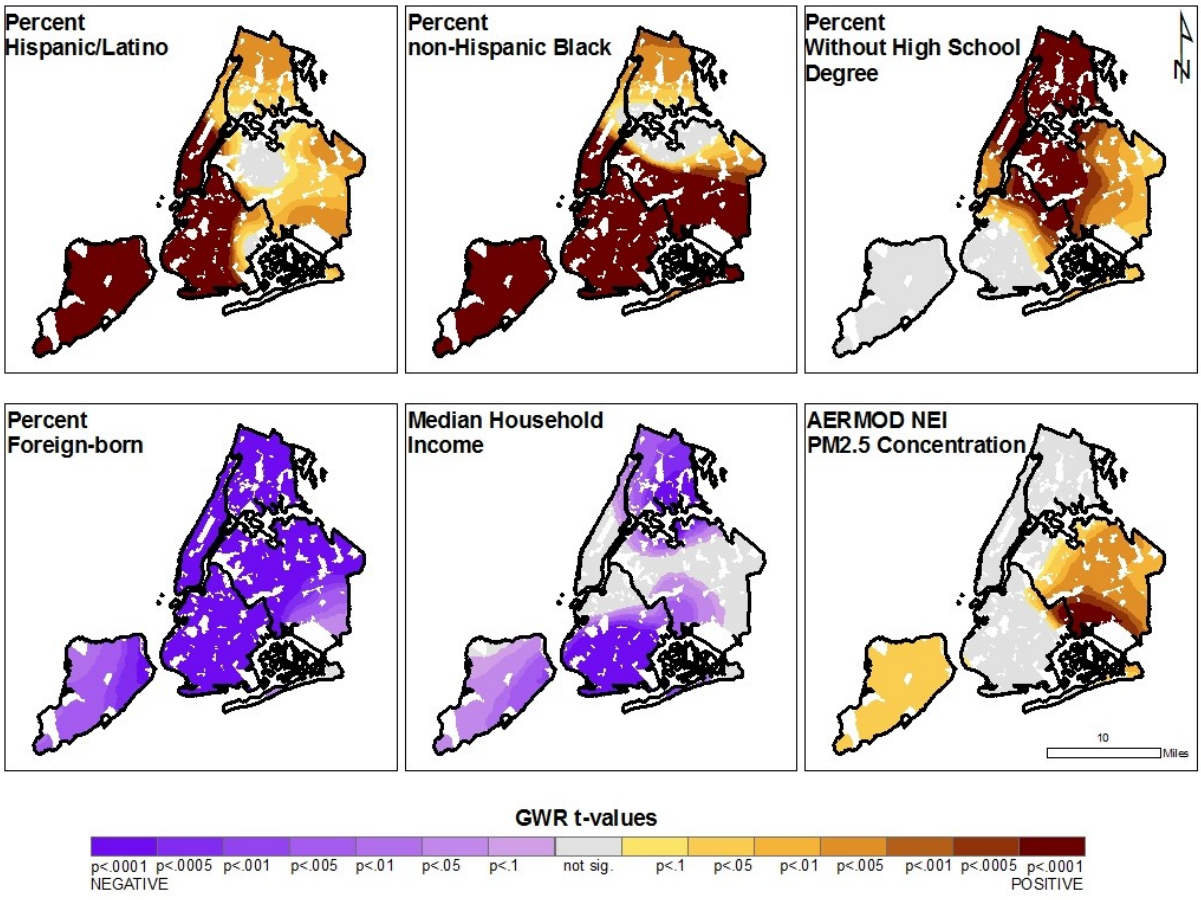


Figure 3-30: t-values from GWR analysis using AERMOD NEI $PM_{2.5}$ concentration estimates. The t-values are mapped by level of significance with bluish areas representing negative associations with heart failure hospitalization rates (3% trimmed) and brown areas representing positive associations. "White" areas were not included in the model due to high hospitalization rates (top 3%) or low populations (<415 people).

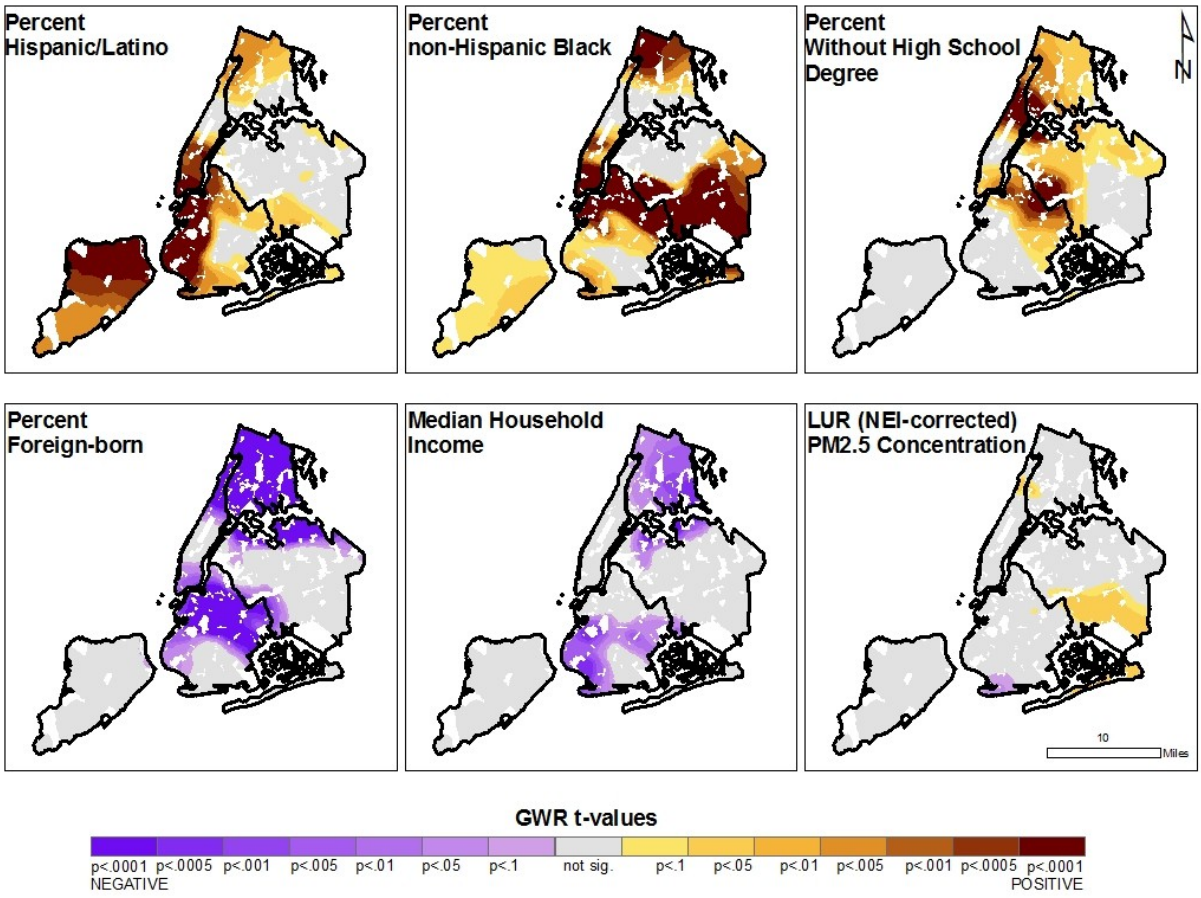


Figure 3-31: t-values from GWR analysis using LUR_{NEI} $PM_{2.5}$ concentration estimates. The t-values are mapped by level of significance with bluish areas representing negative associations with heart failure hospitalization rates (3% trimmed) and brown areas representing positive associations. "White" areas were not included in the model due to high hospitalization rates (top 3%) or low populations (<415 people).

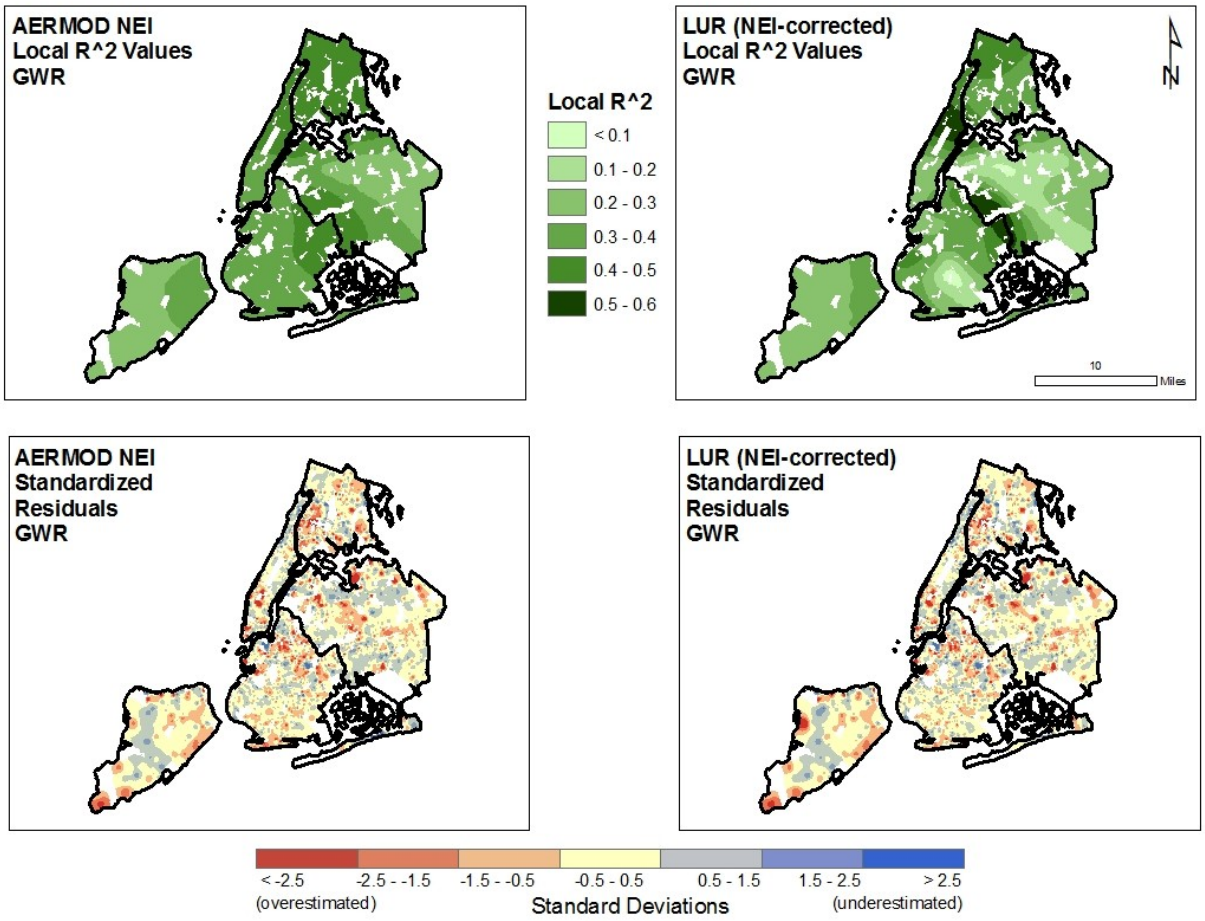


Figure 3-32: GWR model diagnostics (local R^2 and residuals) for model using AERMOD NEI $PM_{2.5}$ estimates and LUR_{NEI} $PM_{2.5}$ estimates. "White" areas were not included in the models due to high hospitalization rates (top 3%) or low populations (<415 people).

3.3 ANALYSIS CHAPTER CONCLUSORY STATEMENT

This analysis chapter has examined potential environmental injustice with regards to $PM_{2.5}$ exposure using both proximity and air dispersion modeling using exploratory analyses and odds ratios. The environmental health outcomes vis-à-vis fine particulate exposure and socio-demographics were also examined using various regression techniques (OLS, SAR, and GWR). The next two chapters (Results and Conclusions) summarize and interpret the research findings.

4 RESULTS

Although there are many different findings that could be considered "results" in this dissertation, what will be focused on in this section is simply the outputs of the environmental justice analysis (proximity buffers and AERMOD air dispersion modeling of NEI sources from **Section 3.1**) and the environmental health analysis (OLS, SAR, and GWR regression models from **Section 3.2**). Bear in mind that many of the "results" were already presented in the various analysis sections, but are discussed here in a more comparative and comprehensive fashion.

4.1 ENVIRONMENTAL JUSTICE RESULTS

Comparing proximity analysis and air dispersion modeling for the assessment of environmental injustice leads one to believe that there really are two different phenomena being examined: proximity measuring distance vs. air dispersion modeling measuring $PM_{2.5}$ concentrations. There certainly appears to be environmental injustices occurring both city-wide and in specific boroughs, particularly with regards to Latinos, poverty status, and educational attainment, however the relationships are quite complex both in terms of non-linearity and spatiality. A table showing the results of odds ratio analyses together for proximity (1/4 mile from source) and air dispersion modeling (50%, 90%, and 95% break values) show that proximity analysis odds ratios are most similar city-wide and by borough to the air dispersion model which uses the median (50^{th} percentile) as the break point (**Table 4-1**). Although these are measuring distinctly different things with proximity examining populations residing within ¼ mile of an NEI $PM_{2.5}$ source and

50th percentile dispersion modeling estimating the populations residing in areas that are above or below the median concentrations of $PM_{2.5}$ from NEI sources, the odds ratios both suggest environmental injustice may be present. The proximity analysis suggests that there is an increased likelihood of predominantly Latino residents and those below poverty living within ¼ mile from NEI $PM_{2.5}$ sources city-wide. Most of the boroughs behave similarly with the exception of the Bronx (only non-Hispanic Blacks appear over-represented) and Staten Island (a low number of samples and wide confidence intervals make these results suspect).

Break Point	Socio-demographic Group	NYC	Brooklyn	Bronx	Manhattan	Queens	Staten Island
PROXIMITY .25 miles	Non-Hispanic White	0.864	1.075	0.915	0.659	0.723	*0.965*
	Non-Hispanic Black	*0.998*	1.033	1.334	1.129	1.167	0.294
	Hispanic / Latino	1.299	1.018	0.783	1.507	1.326	*0.829*
	Below Poverty	1.218	1.284	0.938	1.194	1.288	0.653
	No High School Degree	*1.004*	1.030	0.799	1.381	*1.001*	*0.797*
50 PERCENTILE	Non-Hispanic White	0.574	0.212	0.302	3.330	1.480	0.747
	Non-Hispanic Black	0.765	3.030	0.754	0.196	0.098	1.286
	Hispanic / Latino	3.305	3.371	2.546	0.284	3.068	1.291
	Below Poverty	1.622	1.595	1.927	0.530	1.362	0.930
	No High School Degree	1.280	1.440	1.721	0.499	1.275	0.878
90 PERCENTILE	Non-Hispanic White	1.849	0.936	0.833	1.131	1.125	1.786
	Non-Hispanic Black	0.187	0.380	0.838	0.385	0.175	0.111
	Hispanic / Latino	0.997	3.506	1.287	0.812	1.612	0.616
	Below Poverty	0.847	1.622	1.153	1.097	1.330	0.499
	No High School Degree	0.765	1.489	1.092	1.151	1.122	0.740
95 PERCENTILE	Non-Hispanic White	1.717	1.873	0.888	*0.987*	0.938	1.500
	Non-Hispanic Black	0.166	0.143	1.069	0.428	0.257	0.146
	Hispanic / Latino	0.843	2.365	1.027	*0.987*	1.636	0.726
	Below Poverty	0.805	1.538	1.121	1.239	1.424	0.582
	No High School Degree	0.800	1.396	0.969	1.264	1.148	0.756

Table 4-1: Odds ratios of socio-demographics for proximity analysis and AERMOD-derived $PM_{2.5}$ concentration estimates (break values at 50%, 90%, and 95%) from NEI sources in NYC and its boroughs. Italicized entries are not significant, all other entries are significant ($p<.05$).

As the break value for the dispersion modeling results is increased to 90% and 95%, the apparent environmental injustice disappears city-wide. At those break values, it appears to be only the non-Hispanic White population, which is over-represented in the "exposed" groups. When examined by borough, however, some of the EJ issues reappear with Latinos being over-represented in Brooklyn, the Bronx, and Queens; those below poverty being over-represented in all boroughs except Staten Island; and those without a high school degree being over-represented in all boroughs but Staten Island in the 90% model and Staten Island and the Bronx in the 95% model. These results, although present in the table above (**Table 4-1**), are more easily seen graphically (**Figure 4-1**). Note that these tables and graphs differ from those which appeared in **Section 3.1** since they include the proximity analysis results as well as the modeled concentrations.

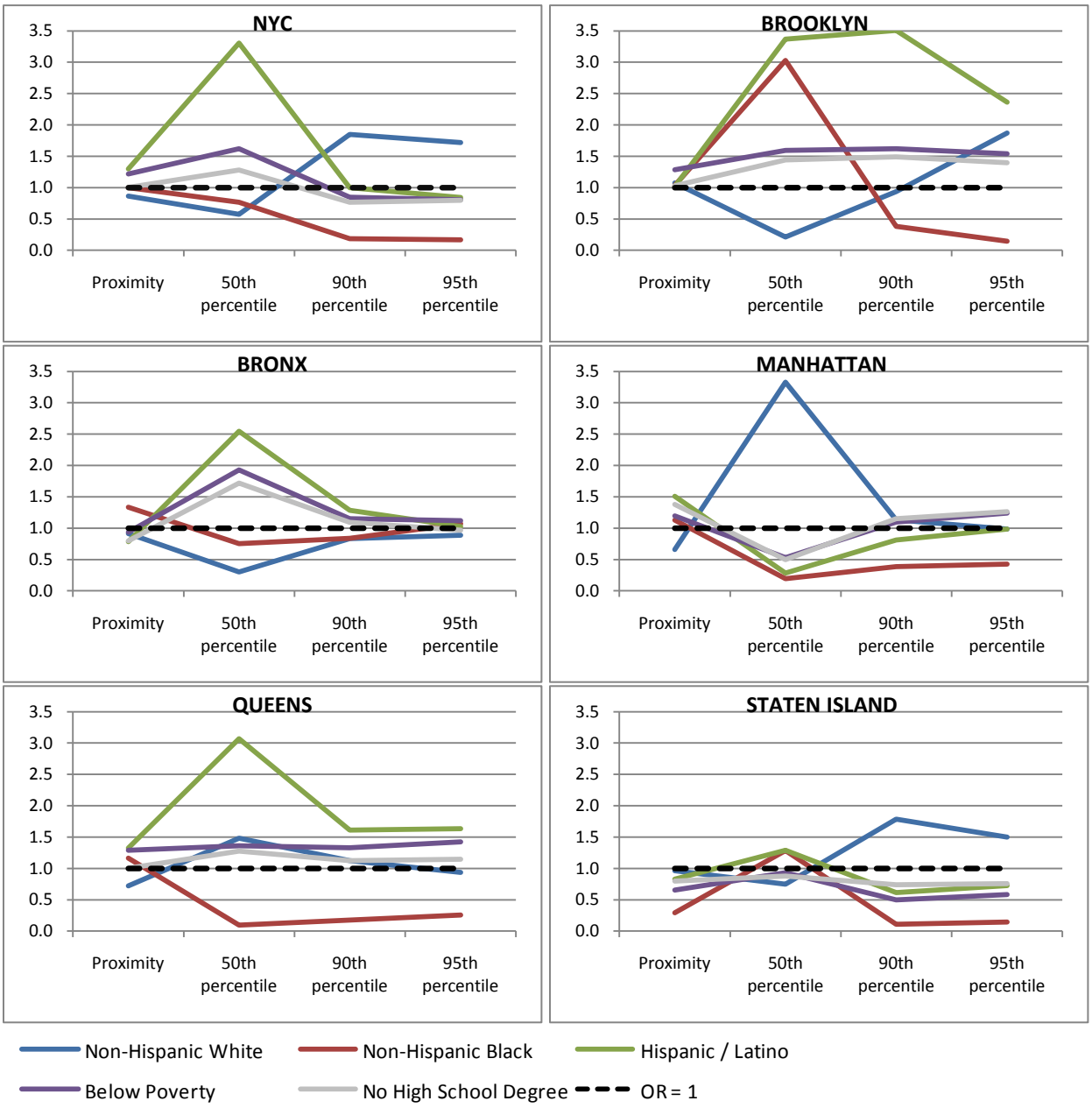


Figure 4-1: Odds ratios of socio-demographics for proximity analysis and AERMOD-derived $PM_{2.5}$ concentration estimates (break values at 50%, 90%, and 95%) from NEI sources in NYC and its boroughs.

These findings suggest that although there appear to be environmental justice issues throughout NYC and its boroughs, the relationships between exposure to fine particulate matter and socio-

demographics are spatially complex with different areas presenting different results. The non-linearity of the relationships raises the question about the definition of "exposure". For instance, if the $PM_{2.5}$ concentration values above the median are considered exposed, Latinos, those without a high school degree, and those below poverty would be considered overrepresented in the exposed category suggesting environmental injustices in NYC. Whereas if only the most extreme concentrations of $PM_{2.5}$ (e.g., 95^{th} percentile) constitute an exposed categorization, then it is only the non-Hispanic white group which is overrepresented suggesting a lack of disproportionate burdens on traditional EJ populations.

4.2 ENVIRONMENTAL HEALTH RESULTS

The environmental health ramifications of exposure to $PM_{2.5}$ were quantified using two estimates of $PM_{2.5}$ exposure (pollution only from NEI sources and ambient pollution) and three regression techniques (OLS, SAR, and GWR). When all three regression models are looked at simultaneously to compare overall model performance and parameter estimations, similarities and differences can be readily seen (**Table 4-2**).

MODEL	DIAGNOSTICS				Parameter Estimates							$PM_{2.5}$ t-values
	n	R^2	std. err. of estimate	AIC	spatial lag (rho)	% non-Hispanic Black	% Hispanic/ Latino	% with no high school	median household income	% foreign-born	$PM_{2.5}$ Estimate	
OLS: AERMOD NEI	1987	.351	.00328	-17088	--	2.07E-05	3.73E-05	5.99E-05	-3.87E-08	6.62E-05	2.00E-03	2.87
SAR: AERMOD NEI	1987	.412	.00311	-17238	3.45E-01	2.36E-05	1.35E-05	4.89E-05	-4.471E-05	-2.4E-08	1.70E-03	2.767
GWR: AERMOD NEI	1987	.429	.00309	-17289	--	-1.7E-05 to 5.3E-05	3.6E-08 to 7.8E-05	-3.5E-06 to 1.2E-04	-7.4E-08 to 1.4E-08	-9.6E-05 to -1.2E-05	-1.8E-03 to 2.1E-02	-1.19 to 5.61
OLS: LUR_{NEI}	1987	.354	.00327	-17098	--	2.01E-05	3.36E-05	5.98E-05	-3.88E-08	-6.62E-05	3.77E-04	4.30
SAR: LUR_{NEI}	1987	.411	.00311	-17237	3.36E-01	2.24E-05	1.30E-05	4.85E-05	-4.54E-05	-2.47E-08	2.25E-04	2.676
GWR: LUR_{NEI}	1987	.464	.00303	-17334	--	-5.7E-05 to 7.1E-05	-1.1E-05 to 9.9E-05	-7.1E-05 to 1.6E-04	-1.1E-07 to 4.4E-08	-1.2E-04 to 1.2E-05	-9.4E-04 to 2.1E-03	-2.06 to 2.98

Table 4-2: Model comparisons of OLS, SAR, and GWR models using 3% trimmed hospitalization data, AERMOD-derived NEI $PM_{2.5}$ estimates, and LUR-derived $PM_{2.5}$ estimates (corrected for NEI sources).

A trend can be seen with increases in R^2 values and corresponding decreases in AIC values (both suggesting improved models) from OLS to SAR and again from SAR to GWR. The spatial autoregressive model (lag) clearly outperforms the OLS as it takes the effect of neighboring heart failure hospitalization rates into account by creating the lagged version of the health outcome variable. This improves the fit of the model with a highly correlated lagged variable; however the other independent variables, including the $PM_{2.5}$ estimates, retain significance and logical directionality even though their magnitudes are reduced. The geographically weighted regression improves upon the SAR by allowing the relationships to vary over space. The range in the parameter estimates (and other statistics) show areas where one or more of the variables is significant, or more interestingly, where the variables identified by the OLS and SAR do not appear to be significant. For instance, **Figure 3-30**, which shows the GWR results for the AERMOD NEI model, suggests that the majority of NYC follows the classic environmental health justice trend that census tracts with higher proportions of racial and ethnic minorities (percent non-Hispanic Black and percent Hispanic/Latino) are more likely to have elevated hospitalization rates for heart failure. This trend only appears to fail around north/central Queens.

Lack of a high school education is also positively associated with increased hospitalization rates in the majority of the city with the exception of the southern section of Brooklyn and all of Staten Island. The percent foreign-born variable is consistently negatively associated with heart failure hospitalizations across the entire city, supporting the notion that immigrants, when adjusting for race and ethnicity, tend to be healthier (or at least less hospitalized) than their domestic-born counterparts. The negative correlation between income and hospitalization rates appears strongest in and around the Bronx, Brooklyn, and Staten Island, suggesting that in these areas economic status is strongly associated with health. Lastly, the AERMOD NEI concentration is significantly positively correlated with heart failure in most of Queens and Staten Island. Although the GWR tends to support the results of the OLS and SAR, meaning that the model is relatively stable and 'believable', it does offer insight into the fluctuation of the relationships and the regions where one variable may gain significance while another loses it.

The land use regression ambient $PM_{2.5}$ estimates (corrected for NEI) tell an almost identical story to the AERMOD NEI models in the OLS and SAR analyses with very similar R^2 values, AICs, parameter estimates, and t-values. It is the GWR that appears to stray farthest from the AERMOD NEI results by detecting a much more 'local' phenomenon. The parameter estimates and significances are similar between the AERMOD NEI and LUR_{NEI} models, but the maps produced by the latter tend to be more dynamic and "spotty" than the former, even including a small area of south-west Brooklyn which has a weak negative relationship between LUR_{NEI} and hospitalization rates. As was demonstrated in the hypothetical example in **Section 1.6.4**, this is

more likely due to a locally misspecified model or another variable demonstrating spatial non-stationarity, rather than a truly negative relationship between $PM_{2.5}$ exposure and hospitalization rates. When looking at the mapped GWR model diagnostics (**Figure 3-32**), when AERMOD NEI is used as the $PM_{2.5}$ concentration estimate the model appears more 'fluid' (i.e. global), with moderate R^2 across the majority of the city, whereas the when LUR_{NEI} acts as the $PM_{2.5}$ estimate, the model has highest local R^2 values (between .5 and .6) around Harlem in Manhattan and the Queens/Brooklyn border and much lower local R^2 values in southern Brooklyn and parts of Queens.

Although the results from the three regression types do support each other, the magnitude of the relationships between the dependent and independent variables are different for each. For instance, the OLS suggests that every 5.0 $\mu g/m^3$ increase in $PM_{2.5}$ concentration from NEI sources is associated with an increase of one hospitalization per 100,000 residents over three years. SAR implies that the increase in $PM_{2.5}$ concentration from NEI sources would be 5.9 $\mu g/m^3$. GWR, on the other hand, suggests that there are areas in the city where an increase of as little as 0.5 $\mu g/m^3$ could result in an extra hospitalization per 100,000 residents over a three year period. Regarding ambient $PM_{2.5}$, the estimated concentrations required to increase the hospitalization rate for heart failure per 100,000 residents over three years by one is higher, with OLS estimating 26.5 $\mu g/m^3$, SAR estimating 44.4 $\mu g/m^3$, and GWR estimating concentrations as low as 4.7 $\mu g/m^3$ in certain parts of NYC. The results presented in this simplified form, however, assume that the relationship between $PM_{2.5}$ concentration and heart failure hospitalization rates is linear throughout an infinite range, which is unlikely to be true. These numbers do seem to be

believable when the range and variance of $PM_{2.5}$ is relatively small, but care must be taken in over-interpretation of these findings.

5 CONCLUSIONS AND FUTURE STEPS

The conclusions for this dissertation can be looked at in three semi-discreet sections: (1) methods of $PM_{2.5}$ estimation including pros and cons of each technique; (2) environmental justice analysis; and (3) environmental health analysis. The latter two items include discussion not just regarding the findings, but also the statistical methods used to quantify the associations and their relative strengths and weaknesses. These are followed by sections containing brief discussions on policy implications, future steps, and a final statement.

5.1 $PM_{2.5}$ ESTIMATION

The estimation of fine particulate matter in NYC was attempted in three ways: (1) proximity, (2) air dispersion modeling, and (3) land use regression. Proximity analysis was very quick and easy to do, however it proved to be somewhat inefficient for estimating actual $PM_{2.5}$ exposure as was made evident in the environmental justice analysis below. It is severely limited in that it makes the assumption that all emission points release equal amounts of fine particulate matter. Simple proximity analysis also assumes that the pollution will disperse homogeneously in all directions from the source, not accounting for physical or meteorological factors. Lastly, this type of analysis treats exposure as a dichotomized phenomenon, with residents living within ¼ mile of a source treated as “exposed”, and those beyond ¼ mile being treated as “unexposed”.

This contrasts sharply with air dispersion modeling. The use of AERMOD required a tremendous volume of input data for the emissions, physical surroundings, and meteorology. It also necessitates a huge amount of computer resources and time to run for a study area as large as New York City. The output, however, is quite a bit more subtle and believable when compared to simple proximity analysis. AERMOD is able to produce a $PM_{2.5}$ concentration estimate at any point in the study area, enabling the production of a continuous surface of estimated pollution concentrations. This process, however, is also not free from assumptions and error. For instance, the emission rate from NEI sources is only provided as a yearly total (tons/year), and as such the assumption had to be made that the $PM_{2.5}$ was being emitted from the stack at a constant rate for the entire year (g/s). AERMOD was also not able to properly estimate pollution from mobile sources in NYC due to a lack of daily traffic data from the NYSDOT (only a small percentage of roads had measured traffic). Proximity analysis, on the other hand, would easily be able to buffer any road that was chosen. Finally, although AERMOD takes building downwash into account when modeling pollution dispersion, it does not explicitly account for the "canyon effect" – pollution being trapped between tall buildings and "travelling" along the roads of the city creating higher pollution concentrations in those areas. Both of these methods also suffer from the "edge effect" problem. This means that sources outside of NYC are not taken into account, and as such there is a possibility for the underestimation of $PM_{2.5}$ exposure estimates for populations living near the city's borders. This issue may be particularly problematic along the western shore of Staten Island, which is likely to be affected by the heavily industrial land uses in New Jersey.

The final $PM_{2.5}$ estimation technique utilized was land use regression. Although this method is quite a bit more demanding than simple proximity analysis, in its basic form it is faster and less data intensive than air dispersion modeling. Regarding outputs, the fundamental difference between LUR and the other two methods is that LUR estimates ambient $PM_{2.5}$ concentration, rather than the amount emanating from any particular source (e.g., NEI facilities). A central limitation to LUR in this dissertation was the lack of monitoring stations with which to calibrate the model (n=15). This very small sample size necessitated a very simple regression with which to estimate the fine particulate matter concentrations. Aside from the lack of monitors in terms of number of samples, those that existed were not evenly spread throughout the city, potentially biasing the LUR outputs. A novel aspect to the LUR calculations in this dissertation was the incorporation of the AERMOD outputs to 'correct' for $PM_{2.5}$ being emitted from major sources in NYC. This modification of the LUR seemed to produce more nuanced and realistic results. These estimates may be improved in the future as there are now over 50 $PM_{2.5}$ monitors in the city, as compared to the 15 that were available in 2002, whose data will be available by the end of the summer of 2010 according to the New York City Department of Health and Mental Hygiene. This increased sample size could drastically improve land use regression modeling-derived estimates of ambient pollution concentrations.

A serious limitation that affected all of the chronic exposure estimates is the fact that only residential locations were used. It was not possible to consider exposure when individuals were

at other locations (e.g., work, school, or vacation). Another shortcoming was that indoor air quality was not taken into account and only the $PM_{2.5}$ concentrations in the outdoor environments were examined.

It seems that the study question will more often than not suggest the proper $PM_{2.5}$ estimation technique. For instance, if specific pollution sources are not a concern, then land use regression (without correction) may be the most practical solution for fine particulate matter concentration estimation since it is far faster and less demanding than dispersion modeling, but more robust and realistic than simple proximity analysis. If specific pollution sources are of interest, then dispersion modeling would be the preferred model to use assuming the necessary data, time, resources, and expertise are available. If a specific pollution type is not a concern, then proximity analysis may be the best option to act as a proxy for a multitude of environmental issues (e.g., general pollution levels and environmentally burdensome land uses).

5.2 ENVIRONMENTAL JUSTICE

Comparing proximity analysis and air dispersion modeling for the assessment of environmental injustice leads one to believe that there really are two different phenomena being examined. Even when the proximity analysis utilizes the stacks which are emitting $PM_{2.5}$ from NEI facilities as the point of origin of the buffers, the output is not specifically about fine particulate matter.

These simple distances are fundamentally serving as proxies. When you have a facility which is a polluter, you not only are likely to have more pollution, but also often have increased truck traffic, reduced land value, "eye sores," and other generally undesirable land uses. It seems that these implied characteristics are what are really being measured. Conversely, air dispersion modeling is truly only measuring the $PM_{2.5}$ being emitted from the specific stacks belonging to the specific NEI facilities. Ultimately, it is simply a measure of the pollutant concentration and does not attempt to imply characteristics such as those listed above. As such, it is hard to say that one method is truly better than another since they are not really able to quantify the same phenomena.

Another limitation to this analysis is the use of odds ratios to assess the EJ status of socio-demographic groups in NYC. Odds ratios are dichotomized as 'exposed groups' and 'unexposed groups'. As such, a threshold for exposure must be defined. In the case of proximity analysis, the choice is based upon the buffer distance (e.g. ¼ mile from an NEI point source). In the case of air dispersion modeling, however, the choice is not as cut-and-dry. It was found that when using different lot-level pollution concentration break points for the dichotomization (median, 90^{th} percentile, and 95^{th} percentile) the results can change dramatically.

The EJ analysis suggests that there are groups that are over-represented in the 'exposed' categories for both the proximity analysis and the dispersion analyses (with different break points). These findings, however, differ wildly from borough to borough with certain areas

showing extreme EJ issues and others showing the exact opposite. To further complicate the issue, as the AERMOD $PM_{2.5}$ estimates from NEI sources break points are altered what may have appeared to be environmental injustice in NYC when comparing populations below the lot-level mean of $PM_{2.5}$ concentrations to those above the mean reverse direction when residents in the lower 90 percentile and contrasted to those in the top 10 percentile.

Ultimately, I believe it is hard to claim that there is a city-wide environmental justice issue with regards to chronic $PM_{2.5}$ exposure from major stationary sources (based on the 90 and 95 percentile AERMOD analysis). What is more defendable is that there appear to be environmental injustices in the Bronx and Brooklyn with more subtle EJ issues in Queens and Manhattan. These injustices are mostly associated with Latinos, poverty status, and educational attainment, however the relationships are quite complex both in terms of non-linearity and spatiality. The ambiguity and complexity of these findings suggest the need for further analyses. These could include different levels of geographic aggregation, more refined socio-demographic groupings (e.g., break down the “non-Hispanic blacks” or “Hispanic / Latinos” into multiple mutually exclusive categories), and additional qualitative exploration.

5.3 ENVIRONMENTAL HEALTH

At the onset of this dissertation, I was uncertain that working with the data I had access to would allow for the detection of the influence of chronic exposure to fine particulate matter concentrations on heart failure hospitalization rates. Thankfully the associations, although subtle, were quantifiable and statistically significant. By using a variety of $PM_{2.5}$ estimates (air dispersion modeling and land use regression) and a variety of regression methods (ordinary least squares, spatial autoregressive models, and geographically weighted regressions), the positive correlation between fine particulate matter and heart failure hospitalization rates were shown in an ecological framework while controlling for race, ethnicity, income, education, and foreign-born status.

Regarding the statistical models themselves, the outputs suggest that data such as this benefits from accounting for elements of space in the regressions. OLS tended to overemphasize the influence of all of the independent variables when compared with SAR, since the latter introduced a spatially-lagged version of the heart failure variable as an additional variable. When the model was adjusted for spatial autocorrelation via the SAR, the slightly more modest parameter estimates became more believable. GWR, on the other hand, was able to allow the relationships to vary over space identifying areas where any given variable was significant or not significant. The GWR results were quite interesting, exposing locations where $PM_{2.5}$ was significantly associated with heart failure hospitalization rates and other areas where the rates were best predicted solely by socio-demographics without taking pollution exposure into

account. These results may be used to inform future research, particularly qualitative field studies, by identifying the behaviors unique to specific areas. To reiterate, although each regression model type produced somewhat different results, each tended to support the other two's results.

As has been mentioned, AERMOD $PM_{2.5}$ estimates from NEI sources quantify only the pollution concentration from specific NEI sources whereas LUR_{NEI} estimates ambient $PM_{2.5}$ concentrations (correcting for NEI sources). The socio-demographic variables' relationships with heart failure hospitalization rates in all of the models, independent of regression technique or pollution variable, imply a classic environmental health justice scenario. Areas with high proportions of racial and ethnic minorities and the less-educated tend to exhibit increased hospitalization rates. This suggests a number of possibilities, including institutionalized racism, classism, and a lack of reliable, high quality medical access. Conversely, tracts with high median incomes or higher proportions of foreign-born residents showed lower hospitalization rates. It is possible that those with higher incomes have increased abilities to control their health issues due to better access to physicians, medicines, information, and health insurance whereas those with low incomes may need to rely upon hospitalizations for access to any sort of substantive health care. The inverse relationship between foreign-born populations and hospitalization rates may be due to selective migration, where it tends to be the healthier individuals who are able to move to, and stay in, the destination country. It is also possible that foreign-born residents of NYC have

not had the same chronic exposures to certain environments (physical, cultural, nutritional) that lead to poor health outcomes typical of native-born residents.

There are some important limitations associated with the environmental health analysis. In addition to the limitations described above (indoor air quality, residential locations, etc.) there is a real likelihood of a type of model misspecification. Although $PM_{2.5}$ concentration was estimated and accounted for in all the models, there are many other pollutants associated with heart disease – many of which may also be correlated with $PM_{2.5}$. In other words, locations that have very high concentrations of fine particulate matter may also have high concentrations of other pollutants which influence cardiovascular health. Without explicitly controlling for these other pollutants it is difficult to say with certainty that it is the $PM_{2.5}$ alone which is driving the associations detected in this dissertation.

In the end, although there are caveats, limitations, and stipulations, all three statistical models were able to show the statistically significant positive association between heart failure hospitalization rates and chronic exposure to $PM_{2.5}$ from NEI sources as well as ambient $PM_{2.5}$ concentrations, however they do so in different ways and with different estimated magnitudes of effect.

5.4 POLICY IMPLICATIONS

Since the association between chronic exposure to fine particulate matter and heart disease has been well established, and re-confirmed using various techniques and methods in this dissertation, there are policy implications that can be stated. For instance, the AERMOD analysis suggests that it would prudent to try and reduce the emissions of $PM_{2.5}$ from the major stationary sources in the city. The LUR analysis suggests that reduction of major truck routes would also aid in the lowering of heart failure hospitalizations in NYC. Lastly, both the environmental justice and the environmental health analyses suggest that socio-demographics play a role in exposure (particularly in the Bronx and Brooklyn) as well as health outcomes (city-wide). Clearly it would be optimal to have alternatives to pollution producing facilities and land uses (e.g., freight trains or barges rather than trucks, and non-polluting energy generation rather than countless Con Edison power stations), but if these land uses must exist, then great care should be used when designating locations for facilities producing $PM_{2.5}$ and the routing of major roadways to insure an equitable distribution of environmental hazards among all NYC residents rather than a bias towards to racial and ethnic minorities, the less-educated, or the lower income populations.

5.5 FUTURE STEPS

In order to mitigate some of the limitations described throughout this dissertation, there are some additional analyses and techniques that could be employed. For instance, it would be informative

to analyze the environmental justice issues using continuous rather than dichotomous data. This could be done using any number of regression techniques with exposure estimates acting as the dependent variables and socio-demographics acting as the independent variables. It seems that this may be able to provide a 'truer' understanding of the relationship between chronic fine particulate matter exposure originating major stationary sources and socio-demographics in NYC. It would also be extremely useful to include mobile sources in the EJ analyses as well as the environmental health analyses. If more complete and reliable data regarding traffic in NYC were acquired, created, or collected then I believe that all of the results in this dissertation would be greatly improved.

Regarding the environmental health analyses, it would be useful to model additional pollutants (using AERMOD and/or LUR) in order to further tease out the relationships between chronic exposures and cardiovascular disease. Other regression techniques, such as 'piece-wise regression' may be useful to explore as well since it may be able to account for the non-linearity of the associations in the areas of highest hospitalization rates. Regarding this same issue of the highest 3% of the health outcome data, it may be beneficial to conduct a more qualitative or observational assessment in an attempt to determine if there are other issues of the physical or social environments in or near these locations that may be driving the extremely high rates. It may also prove valuable to re-run this study using more modern data to take advantage of the new monitoring stations placed around New York City. Unfortunately, this would require health data from matching years, which will most likely not be available in the near future.

5.6 FINAL STATEMENT

Ultimately, this dissertation has shown that when looking at spatial processes, it is important to accommodate "space" in the quantification of the associations. Without using geography in the data preparation and the statistical analyses, there is a real possibility of unreliable and misleading results. This dissertation has also confirmed the intra-urban correlations between heart failure hospitalization rates and chronic exposure to fine particulate matter from major stationary sources and ambient $PM_{2.5}$ estimations in New York City. The association between mobile sources of fine particulate matter and heart failure hospitalization rates were not quantifiable in this work due to a lack of comprehensive data; however some of the analyses (e.g., land use regression) suggest that the relationship does indeed exist. This dissertation also found the existence of environmental injustices, although the relationships are more complex in terms of non-linearity and geographic variation than was previously thought.

Although these types of analyses are not able to definitively demonstrate causation it seems likely, given the growing body of research, that the relationship is truly a causative one. If we, as human beings and New Yorkers, are concerned about health equity and the reduction of disease, it would be prudent to adhere to the tenets of the Precautionary Principle and lessen the production of fine particulate matter in our city while trying to ensure an equitable distribution of environmental burdens.

6 REFERENCES

Anselin, L. (1995). Local Indicators of Spatial Association – LISA. *Geographical Analysis, 27*, 93-115.

Anselin, L. (2003). *GeoDa User's Guide*. University of Illinois, Urbana, IL: Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics and CSISS.

Anselin, L. (2005). *Exploring Spatial Data with GeoDa: A Workbook*. Urbana, IL: Spatial Analysis Laboratory Department of Geography University of Illinois / Center for Spatially Integrated Social Science.

Anselin, L., & Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah & D. E. Giles (Eds.), *Handbook of Applied Economic Statistics* (pp. 237-289). New York: Marcel Dekker.

Anselin, L., Syabri, I., & Kho, Y. (2004). *GeoDa: An Introduction to Spatial Data Analysis*. Urbana, IL: Spatial Analysis Laboratory Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

Antecol, H., & Bedard, K. (2006). Unhealthy Assimilation: Why do immigrants converge to American health status levels? *Demography, 43*(2), 337-360.

Been, V., & Gupta, F. (1996). Coming to the nuisance or going to the barrios? A longitudinal analysis of environmental justice claims. *Ecology Law Quarterly, 24*(1), 1–35.

Bellander, T., Berglind, N., Gustavsson, P., Jonson, T., Nyberg, F., Pershagen, G., et al. (2001). Using Geographic Information Systems To Assess Individual Historical Exposure to Air Pollution from Traffic and House Heating in Stockholm. *Environmental Health Perspectives, 109*(6), 633-639.

Bielecka, E. (2005). *A Dasymetric Population Density Map of Poland.* Paper presented at the the 22nd Annual International Cartographic Conference, July 9-15, A Coruna, Spain.

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *British Medical Journal, 320*(7247), 1468.

Bowen, W. M., Salling, M. J., Haynes, K. E., & Cyran, E. J. (1995). Towards environmental justice: spatial equity in Ohio and Cleveland. *Annals of the Association of American Geographers, 85*(4), 641-663.

Brauer, M., Hoek, G., van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., et al. (2003).

Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems. *Epidemiology, 14*, 228-239.

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., et al. (1997). Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographical Information Science, 11*, 699–718.

Briggs, D. J., de Hough, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., et al. (2000). A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. *Science of the Total Environment, 253*, 151–167.

Brook, R., Franklin, B., Cascio, W., & al., e. (2004). Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation, 109*, 2655-2675.

Burke, L. (1993). Race and environmental equity: a geographic analysis in Los Angeles. *Geo Info Systems, 3*, 44-50.

CDC. (2004). Trends in Tuberculosis - United States, 1998-2003. Retrieved 6/9/2010, from http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5310a2.htm

Centner, T., Kriesle, W., & Keeler, A. (1996). Environmental justice and toxic releases: establishing evidence of discriminatory effect based on race and not income. *Wisconsin Environmental Law Journal, 3*(2), 119–158.

Chakraborty, J., & Armstrong, M. (1997). Exploring the use of buffer analysis for the identification of impacted areas in environmental equity assessment. *Cartography and Geographic Information Systems 24*(3), 145-157.

Chen, J., Ng, E., & Wilkins, R. (1996). The Health of Canada's Immigrants in 1994-95. *Health Reports 7*(4), 33-45.

Chen, L., Knutsen, S., Shavlik, D., & al, e. (2005). The association between fatal coronary heart disease and ambient particulate air pollution: are females at greater risk? . *Environmental Health Perspectives, 113*, 1723-1729.

Ciccone, G., Forastiere, F., Agabiti, N., Biggeri, A., Bisanti, L., Chellini, E., et al. (1998). Road traffic and adverse respiratory effects in children. *Occupational and Environmental Medicine, 55*, 771–778.

Cimorelli, A., Perry, S., Venkatram, A., Weil, J., Paine, R., Wilson, R., et al. (2005). AERMOD: A Dispersion Model for Industrial Source Applications. Part I: General Model

Formulation and Boundary Layer Characterization. *Journal of Applied Meteorology, 44*, 682-693.

Clancy, L., Goodman, P., Sinclair, H., & Dockery, D. W. (2002). Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet, 360*(9341), 1210-1214.

Cliff, A. D., & Ord, J. (1973). *Spatial autocorrelation*. London: Pioneer.

Curtin, L. R., & Klein, R. J. (1995). *Direct Standardization (Age-Adjusted Death Rates). Statistical notes.* (Vol. 6). Hyattsville, Maryland: National Center for Health Statistics.

Delfino, R. J., Gong Jr., H., Linn, W. S., Pellizzari, E. D., & Hu, Y. (2003). Asthma symptoms in Hispanic children and daily ambient exposures in toxic and criteria air pollutants. *Environmental Health Perspectives, 111*(4), 647-656.

Dent, A. L., Fowler, D. A., Kaplan, B. M., & Zarus, G. M. (1998). *Using GIS to Study the Health Impact of Air Emissions.* Paper presented at the Geographic Information Systems in Public Health, Third National Conference, San Diego, CA.

Deri, C. (2003). Understanding the 'Healthy Immigrant Effect' in Canada. Working paper 0502E. Department of Economics, University of Ottawa

Dockery, D., Pope, C. I., Xu, X., & al., e. (1993). An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine 329*, 1753-1759.

Donovan, J., d'Espaignet, E., Metron, C., & van Ommeren, M. (1992). *Immigrants in Australia: A Health Profile*. Canberra: Australian Institute of Health and Welfare Ethnic Health Series.

Drucker, E., Alcabes, P., Bosworth, W., & Sckell, B. (1994). Childhood tuberculosis in the Bronx, New York. *Lancet, 343*, 1482–1485.

Edwards, J., Walters, S., & Griffiths, R. C. (1994). Hospital admissions for asthma in pre-school children: relationship to major roads in Birmingham, UK. *Archives of Environmental Health, 49*, 223–227.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and aerial interpolation: implementation and evaluation. *Cartography and Geographic Information Science, 28*(2), 125-138.

Elliott, P., Shaddick, G., Wakefield, J., de Hoogh, C., & Briggs, D. (2007). Longterm associations of outdoor air pollution with mortality in Great Britain. *Thorax, 62*(12), 1088-1094.

English, P., Neutra, R., Scalf, R., Sullivan, M., Waller, L., & Zhu, L. (1997). Examining associations between childhood asthma and traffic flow using a geographic information system. *Environmental Health Perspectives, 107*, 761–767.

EPA. (2004). *User's Guide for the AMS/EPA Regulatory Model - AERMOD*. Research Triangle Park, NC: United States Environmental Protection Agency Office of Air Quality Planning and Standards Emissions Monitoring and Analysis Division.

EPA. (2006a). *Documentation for the final 2002 point source national emissions inventory.* Research Triangle Park, NC: Emission Inventory and Analysis Group, Air Quality and Analysis Division, U.S. Environmental Protection Agency.

EPA. (2006b). *Guidelines for the Reporting of Daily Air Quality – the Air Quality Index (AQI)*. Research Triangle Park, North Carolina: Emission Inventory and Analysis Group, Air Quality and Analysis Division, U.S. Environmental Protection Agency.

EPA. (2008). Fine Particle (PM2.5) Designations: Basic Information. Retrieved 6/11/2010, from http://www.epa.gov/pmdesignations/basicinfo.htm

EPA. (2010a). Particulate Matter (PM) Research. Retrieved 6/11/2010, from http://www.epa.gov/airscience/quick-finder/particulate-matter.htm

EPA. (2010b). Technology Transfer Network Support Center for Regulatory Atmospheric Modeling - Preferred/Recommended Models Retrieved 6/17/2010

EPA. (2010c). Environmental Justice - Frequently Asked Questions. Retrieved 6/17/2010, from www.epa.gov/compliance/resources/faqs/ej/

Forster, B. C. (1985). An examination of some problems and solutions in monitoring urban areas from satellite platforms. *International Journal of Remote Sensing, 6*(1), 139-151.

Fothergill, A., Maestas, E., & Darlington, J. D. (1999). Race, Ethnicity and Disasters in the United States: AReviewof the Literature. *Disasters, 23*(2), 156-173.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted Regression: the analysis of spatially varying relationships*. West Sussex, England: John Wiley & Sons Ltd,.

Friedman, M. S., Powell, K. E., Hutwagner, L., Graham, L. M., & Teague, W. G. (2001). Impact of changes in transportation and commuting behaviors during the 1996 summer Olympic games in Atlanta on air quality and childhood asthma. *Journal of the American Medical Association, 285*(7), 897–905.

GeodaCenter. (2010). Geoda Center Glossary of Terms. Retrieved 6/27/2010, from http://geodacenter.asu.edu/node/390

Gibson, J., & Olivia, S. (2010). *Spatial Autocorrelation and Household Choices in Indonesia*. Paper presented at the Economics of Trade and Development Seminar Series.

Gilbert, N. L., Goldberg, M. S., Beckerman, B., Brook, J. R., & Jerrett, M. (2005). Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land-use regression model. *Journal of the Air and Waste Management Association, 55*, 1059-1063.

Gilliland, F., Avol, E., Kinney, P., Jerrett, M., Dvonch, T., Lurmann, F., et al. (2005). Air pollution exposure assessment for epidemiologic studies of pregnant women and children: Lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research. *Environmental Health Perspectives, 113*, 1447–1454.

Guo, Y., & al., e. (1999). Climate, traffic-related air pollutants, and asthma prevalence in middle-school children in Taiwan. *Environmental Health Perspectives, 107*(12), 1001–1006.

Hamilton, L. C. (1990). *Modern Data Analysis: A First Course in Applied Statistics* Belmont, CA: Brooks/Cole.

Henneberger, A., Zareba, W., Ibald-Mulli, A., Ruckerl, R., Cyrys, J., Couderc, J.-P., et al. (2005). Repolarization Changes Induced by Air Pollution in ischemic Heart Disease Patients. *Environmental Health Perspectives, 113*(4), 440-446.

Herrmann, C., & Maroko, A. R. (2006). Crime Pattern Analysis: Exploring Bronx Auto Thefts using GIS. In J. A. Maantay, and Ziegler, J (Ed.), *GIS for the Urban Environment* (pp. 407-413). Redlands, CA: Environmental Systems Research Institute (ESRI).

Hodgson, S., Nieuwenhuijsen, M. J., Colvile, R., & Jarup, L. (2007). Assessment of exposure to mercury from industrial emissions: comparing "distance as a proxy" and dispersion modelling approaches. *Occupational and Environmental Medicine, 64*, 380-388.

Holloway, S. R., Schumacher, J., and Redmond, R. L. (1997). *People and Place: Dasymetric Mapping Using Arc/Info. Cartographic Design Using ArcView and ARC/INFO*: High Mountain Press, NM. Wildlife Spatial Analysis Lab, University of Montana, Missoula, MT. .

Holt, J. B., Lo, C. P., & Hodler, R. W. (2004). Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science, 31*(2), 103-121.

House, J. S., Kessler, R. C., Herzog, A. R., Mero, R. P., Kinney, A. M., & Breslow, M. J. (1990). Age, Socioeconomic Status and Health. *The Milbank Quarterly 68*, 383-411.

Hruba, F., Fabianova, E., Koppova, K., & Vandenberg, J. (2001). Childhood respiratory symptoms, hospital admissions, and long-term exposure to airborne particulate matter. *Journal of Exposure Analysis and Environmental Epidemiology, 11*, 33-40.

Hu, Z. (2009). Spatial analysis of MODIS aerosol optical depth, PM2.5, and chronic coronary heart disease. *International Journal of Health Geographics, 8*(27).

Hu, Z., Liebens, J., & Rao, K. (2008). Linking stroke mortality with air pollution, income, and greenness in northwest Florida: an ecological geographical study. *International Journal of Health Geographics, 7*(20).

Hu, Z., & Rao, K. (2009). Particulate air pollution and chronic ischemic heart disease in the eastern United States: a county level ecological study using satellite aerosol data. *Environmental Health, 8*(26).

Ihrig, M., Shalat, S., & Baynes, C. (1998). A hospital-based case control study of stillbirths and environmental exposure to arsenic using an atmospheric dispersion model and a geographical information system. *Epidemiology, 9*(3), 290-294.

Jalaludin, B., Morgan, G., Lincoln, D., Sheppeard, V., Simpson, R., & Corbett, S. (2006). Associations between ambient air pollution and daily emergency department attendances for cardiovascular disease in the elderly (65+ years), Sydney, Australia. *Journal of Exposure Science and Environmental Epidemiology, 16*, 225–237.

Janssen, N., Vliet, P. v., Aarts, F., Harssema, H., & Brunekreef, B. (2001). Assessment of exposure to traffic relatedair pollution of children attending schools near motorways. *Atmospheric Environment, 35*, 3875-3884.

Jasso, G., Massey, D. S., Rosenzweig, M. R., & Smith, J. P. (2004). Immigrant Health-Selectivity and Acculturation. RAND.

Jerrett, M., Burnett, R. T., Renjun, M., Pope, C., Arden, I., Krewski, D., et al. (2005). Spatial Analysis of Air Pollution and Mortality in Los Angeles. *Epidemiology, 16*(6), 727-736.

Johnston, B. (Ed.). (1994). *Who Pays the Price? The Sociocultural Context of Environmental Crisis*. Washington, DC: Island Press.

Kasl, S. V., & Berkman, L. (1983). Health Consequences of The Experiences of Migration. *Annual Review of Public Health, 4*, 69-90.

Kesarkar, A. P., Dalvi, M., Kaginalkar, A., & Ojha, A. (2007). Coupling of the weather research and forecasting model with AERMOD for pollutant dispersion modeling. A case study for PM10 dispersion over Pune, India. *Atmospheric Environment, 41*, 1976-1988.

Kim, J. J. (2004). Ambient air pollution: health hazards to children. *Pediatrics, 114*, 1699-1707.

Kunzli, N., Jerrett, M., Mack, W., & al, e. (2005). Ambient air pollution and atherosclerosis in Los Angeles. *Environmental Health Perspectives, 113*, 201-206.

Langford, M., Maguire, D. J., & Unwin, D. (1991). The areal interpolation problem: estimating population using remote sensing in a GIS framework. In I. Masser & M. Blakemore (Eds.), *Handling Geographic Information: Methodology and Potential Applications*. London: Longman.

Lin, S., Munsie, J. P., Hwang, S.-A., Fitzgerald, E., & Cayo, M. R. (2002). Childhood Asthma Hospitalization and Residential Exposure to State Route Traffic. *Environmental Research, 88*(2), 73-81.

Liu, Y., Paciorek, C., & Koutrakis, P. (2009). Estimating Regional Spatial and Temporal Variability of PM2.5 Concentrations Using Satellite Data, Meteorology, and Land Use Information. *Environmental Health Perspectives, 117*(6), 886-892.

Maantay, J. (2002). Mapping Environmental Injustices: Pitfalls and Potential of Geographic Information Systems in Assessing Environmental Health and Equity. *Environmental Health Perspectives, 110*(2), 161-171.

Maantay, J. (2007). Asthma and Air Pollution in the Bronx: Methodological and Data Considerations in Using GIS for Environmental Justice and Health Research. *Health and Place, special issue: Linking Environmental Justice, Population Health, and Geographical Information Science*(13), 32-56.

Maantay, J., Tu, J., & Maroko, A. (2009). Loose-coupling an air dispersion model and a geographic information system (GIS) for studying air pollution and asthma in the Bronx, New York City. *International Journal of Environmental Health Research, 19*(1), 59-79.

Maantay, J. A., & Maroko, A. R. (2008). Mapping urban risk: Flood hazards, race, & environmental justice in New York. *Applied Geography, doi:10.1016/j.apgeog.2008.08.002*.

Maantay, J. A., Chakraborty, J., & Brender, J. (2010). Proximity to Environmental Hazards: Environmental Justice and Adverse Health Outcomes, *monograph prepared for the U.S. Environmental Protection Agency's Symposium "Strengthening Environmental Justice*

Research and Decision Making: A Symposium on the Science of Disproportionate Environmental Health Impacts." (pp. 165). Washington, DC: United States Environmental Protection Agency.

Macleod, C., Duarte-Davidson, R., Fisher, B., Ng, B., Willey, D., Shi, J. P., et al. (2006). Modeling human exposures to air pollution control (APC) residues released from landfills in England and Wales. *Environment International, 32*, 500-509.

Mahaffey, K. R. (1995). Nutrition and lead: strategies for public health. *Environmental Health Perspectives, 103*(5), 191–196.

Maheswaran, R., Haining, R., Brindley, P., Law, J., Pearson, T., Fryers, P., et al. (2005a). Outdoor air pollution, mortality, and hospital admissions from coronary heart disease in Sheffield, UK: a small-area level ecological study. *European Heart Journal, 26*(23), 2543-2549.

Maheswaran, R., Haining, R., Brindley, P., Law, J., Pearson, T., Fryers, P., et al. (2005b). Outdoor air pollution and stroke in Sheffield, United Kingdom: a small-area level geographical study. *Stroke, 36*(2), 239-243.

Marmot, M. G., Adelstein, A. M., & Bulusu, L. (1984). Lessons From the Study of Immigrant Mortality. *Lancet, 30*, 1455-1457.

Marmot, M. G., & Syme, S. L. (1976). Acculturation and Coronary Heart Disease in Japanese-Americans. *American Journal of Epidemiology, 104*, 225-247.

McDonald, J. T. (2003). The Health of Immigrants to Canada. Department of Economics, University of New Brunswick.

McDonald, J. T. (2004). BMI and the Incidence of Being Overweight and Obese Among Canadian Immigrants: Is Acculturation Associated with Unhealthy Weight Gain? Department of Economics, University of New Brunswick.

Mennis, J. (2003). Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer, 55*(1), 31-42.

Miller, K. A., Siscovick, D. S., Sheppard, L., Shepherd, K., Sullivan, J. H., Anderson, G. L., et al. (2007). Long-term exposure to air pollution and incidence of cardiovascular events in women. *New England Journal of Medicine, 356*(5), 447-458.

MOEC. (2001). *City Environmental Quality Review (CEQR) Technical Manual*. New York, NY: New York City Mayor's Office of Environmental Coordination.

Morello-Frosch, R., Pastor, M., & Sadd, J. (2001). Environmental justice and southern California's ''riskscape''—the distribution of air toxics exposures and health risks among diverse communities. *Urban Affairs Review, 36*(4), 551-578.

NASA. (2009). Aerosol Optical Thickness. Retrieved 6/16/2010, from http://disc.sci.gsfc.nasa.gov/data-holdings/PIP/aerosol_optical_thickness_or_depth.shtml

Neumann, C. M., Forman, D. L., & Rothlein, J. E. (1998). Hazard screening of chemical releases and environmental equity analysis of populations proximate to toxic release inventory facilities in Oregon. *Environmental Health Perspectives, 106*(4), 217-226.

Neutra, P. (1999). Examining associations between childhood asthma and traffic flow using a geographic information system. *Environmental Health Perspectives, 107*(9), 761–767.

Nichol, J., & Wong, M. (2009). High Resolution Remote Sensing of Densely Urbanised Regions: a Case Study of Hong Kong. *Sensors, 9*, 4695-4708.

Nitta, H., Sato, T., Nakai, S., Maeda, K., Aoko, S., & Oho, M. (1993). Respiratory health associated with exposure to automobile exhaust. Results of cross-sectional studies in 1979, 1982, and 1983. *Archives of Environmental Health, 48*, 53–58.

Nyberg, F., Gustavsson, P., Jarup, L., Bellander, T., Berglind, N., Jakobsson, R., et al. (2000). Urban Air Pollution and Lung Cancer in Stockholm. *Epidemiology, 11*(5), 487-495.

NYSDEC. (2010). Fine Particulate Matter Monitoring. Retrieved 6/11/2010, from http://www.dec.ny.gov/chemical/8539.html

NYSDOT. (2008). MOBILE6.2 PM Emission Factor Tables Look Up and Calculation Program. Retrieved 6/25/2010, from https://www.nysdot.gov/divisions/engineering/environmental-analysis/repository/mobile6/pm/pmtable.html

Oosterlee, A., Drijver, M., Lebret, E., & Brunekreff, B. (1996). Chronic respiratory symptoms in children and adults living along streets with high traffic density. *Occupational and Environmental Medicine, 53*, 241-247.

Ostro, B., Lipsett, M., Reynolds, P., Goldberg, D., Hertz, A., Garcia, C., et al. (2010). Long-Term Exposure to Constituents of Fine Particulate Air Pollution and Mortality: Results from the California Teachers Study. *Environmental Health Perspectives, 118*(3), 363-369.

Paciorek, C., & Liu, Y. (2009). Limitations of Remotely Sensed Aerosol as a Spatial Proxy for Fine Particulate Matter. *Environmental Health Perspectives, 117*(6), 904-909.

Palloni, A., & Arias, E. (2003). A Re-Examination of the Hispanic Mortality Paradox. CDE Working Paper No. 2003-01. . Center for Demography and Ecology, University of Wisconsin-Madison.

Perez, C. E. (2002). Health Status and Health Behaviour Among Immigrants. *Health Reports 13*, 1-12.

Perlin, S. A., Setzer, R. W., Creason, J., & Sexton, K. (1995). of industrial air emissions by income and race in the United States: an approach using the toxic release inventory. *Environmental Science Technology, 29*(1), 69-80.

Perry, G. P., Cimorelli, A., Paine, R., Brode, R., Weil, J., Venkatram, A., et al. (2005). AERMOD: A Dispersion Model for Industrial Source Applications. Part II: Model Performance against 17 Field Study Databases. *Journal of Applied Meteorology, 44*, 694-708.

Peters, A., von Klot, S., Heier, M., Trentinaglia, I., Hörmann, A., Wichmann, H. E., et al. (2004). Exposure to Traffic and the Onset of Myocardial Infarction. *New England Journal of Medicine, 351*(17), 1721-1730.

Pollock, P. H., & Vittas, M. E. (1995). Who bears the burden of environmental pollution? Race, ethnicity, and environmental equity in Florida. *Social Science Quarterly, 76*(2), 294-309.

Pope, C. I., Burnett, R., Thurston, G., & al., e. (2004). Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general patho-physiological pathways of disease. *Circulation, . 109*, 71-77.

Poulstrup, A., & Hansen, H. L. (2004). Use of GIS and Exposure Modeling as Tools in a Study of Cancer Incidence in a Population Exposed to Airborne Dioxin. *Environmental Health Perspectives, 112*(9), 1032-1036.

Puett, R. C., Hart, J. E., Yanosky, J. D., Paciorek, C., Schwartz, J., Suh, H. H., et al. (2009). Chronic Fine and Coarse Particulate Exposure, Mortality, and Coronary Heart Disease in the Nurses' Health Study. *Environmental Health Perspectives, 117*(11), 1697-1701.

Pulido, L. (1996). A critical review of the methodology of environmental racism research. *Antipode, 28*(2), 142–159.

Rosenthal, F. S., Carney, J. P., & Olinger, M. L. (2008). Out-of-Hospital Cardiac Arrest and Airborne Fine Particulate Matter: A Case-Crossover Analysis of Emergency Medical Services Data in Indianapolis, Indiana. *Environmental Health Perspectives, 116*(5), 631-

636.

Ross, Z., English, P. B., Scalf, R., Gunier, R., Smorodinsky, S., Wall, S., et al. (2006). Nitrogen dioxide prediction in Southern California using land use regression modeling: Potential for environmental health analyses. *Journal of Exposure Analysis and Environmental Epidemiology, 16*, 106-114.

Ruckerl, R., Greven, S., Ljungman, P., Aalto, P., Antoniades, C., Bellander, T., et al. (2007). Air Pollution and Inflammation (Interleukin-6, C-Reactive Protein, Fibrinogen) in Myocardial Infarction Survivors. *Environmental Health Perspectives, 115*(7), 1072-1080.

Ryan, P., & LeMasters, G. (2007). A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure *Inhalation Toxicology, 19*(1), 127-133.

Ryan, P., LeMasters, G., Levin, L., Burkle, J., Biswas, P., Hu, S., et al. (2008). A land-use regression model for estimating microenvironmental diesel exposure given multiple addresses from birth through childhood. *Science of the Total Environment, 404*(1), 139-147.

Ryan, P. H., LeMasters, G., Biagini, J., Bernstein, D., Grinshpun, S. A., Shukla, R., et al. (2006). Is it traffic type, volume, or distance? Wheezing in infants living near truck and bus traffic. *Journal of Allergy and Clinical Immunology, 116*(2), 279-284.

Schwartz, J. (2004). Air pollution and children's health. *Pediatrics, 113*, 1037-1043.

Schwartz, J., Slater, D., & Larson, T. V. (1993). Particulate air pollution and hospital emergency room visits for asthma in Seattle. *American Review of Respiratory Disease, 147*, 826-831.

Singh, R. B., Desloges, C., & Sloan, J. J. (2006). Application of a microscale emission factor model for particulate matter to calculate vehicle-generated contributions to fine particulate emissions. *Journal of the Air and Waste Management Association, 56*, 37-47.

Sleeter, R. (2004). *Dasymetric Mapping Techniques for the San Francisco Bay Region, California.* Paper presented at the Urban and Regional Information Systems Association Annual Conference, Reno, NV.

Stephen, E. H., Foote, K., Hendershot, G. E., & Schoenborn, C. A. (1994). Health of the Foreign-Born Population. *Advance Data From Vital and Health Statistics, 241*, 1-10.

Stephens, C. (1996). Healthy cities or unhealthy island: The health and social implications of urban lnequality. *Environ Urban, 8*(2), 9-30.

Timney, M. (1998). Environmental injustices In D. Camacho (Ed.), *Environmental Injustices,*

Political Struggles: Race Class, and the Environment. Durham, NC: Duke University Press.

Touma, J. S., Isakov, V., Cimorelli, A. J., Brode, R. W., & Anderson, B. (2007). Using Prognostic Model-Generated Meteorological Output in the AERMOD Dispersion Model: An Illustrative Application in Philadelphia, PA. *Journal of the Air and Waste Management Association, 57*, 586-595.

Van Vliet, P., Knape, M., de Hartog, J., Janssen, N., Harssema, H., & Brunekreef, B. (1997). Motorvehicle exhaust and chronic respiratory symptoms in children living near freeways. *Environmental Research, 74*, 122–132.

van Vliet, P., Knape, M., Hartog, J. d., Janssen, N., Harssema, H., & Brunekreef, B. (1997). Motor Vehicle Exhaust and Chronic Respiratory Symptoms in Children Living near Freeways. *Environmental Research, 74*, 122-132.

Venn, A. J., Lewis, S. A., Cooper, M., Hubbard, R., & Britton, J. (2001). Living near a main road and the risk of wheezing illness in children. *American Journal of Respiratory and Critical Care Medicine, 164*, 2177–2180.

Westergren, A., Karlsson, S., Andersson, P., Ohlsson, O., & Hallberg, I. R. (2001). Eating difficulties, need for assisted eating, nutritional status and pressure ulcers in patients admitted for stroke rehabilitation. *Journal of Clinical Nursing, 10*, 257-269.

WHO. (1996). *Groups at Risk: WHO Report on the Tuberculosis Epidemic*. Geneva: World Health Organization.

Wilkinson, P., Elliott, P., Grundy, C., Shaddick, G., Thakrar, B., Walls, P., et al. (1999). Case-control study of hospital admission with asthma in children aged 5–14 years: relation with road traffic in north west London. *Thorax, 54*(12), 1070–1074.

Wu, C., & Murray, A. T. (2007). Population Estimation Using Landsat Enhanced Thematic Mapper Imagery. *Geographical Analysis, 39*, 26-43.

Wu, S., Qiu, X., & Wang, L. (2005). Population Estimation Methods in GIS and Remote Sensing: A Review. *GIScience and Remote Sensing, 42*(1).

Wyst, J. H., Reitmeir, P., Dold, S., Wulff, A., Nicolai, T., Von Loeffelholz-Colberg, E., et al. (1993). Road traffic and adverse effects on respiratory health in children. *British Medical Journal, 307*, 596-600.

Zanobetti, A., Gold, D. R., Stone, P. H., Suh, H. H., Schwartz, J., Coull, B. A., et al. (2010). Reduction in Heart Rate Variability with Traffic and Air Pollution in Patients with Coronary Artery Disease. *Environmental Health Perspectives, 118*(3), 324-330.