The Association between Environmental Justice and Tweet Sentiment: A Case Study in New York City

A thesis submitted by

Yi Zhong

In partial fulfillment of the requirements for the degree of

Master of Science

in

Environmental Policy and Planning

Tufts University

August 2021

Copyright 2021, Yi Zhong

Advisor: Justin Hollander

Reader: Sumeeta Srinivasan

Abstract

This thesis investigates the spatial patterns of Environmental Justice (EJ) and tweet sentiment and the association between them. First, EJ is quantified by using indicators from the EJSCREEN tool and scoring system from the Environmental Justice Screening Method (EJSM). The score of three EJ variables including population vulnerability, proximity to hazard and health risk exposure and overall EJ are calculated at the census block group level. Then, the AFINN dictionary and lexicon-based method are used to analyze the polarity and intensity of sentiment in the content of over 890,000 geotagged tweets in New York City from April to July 2017. Finally, the spatial patterns of EJ and tweet sentiment are explored by descriptive and spatial statistics (Global, Local and Bivariate Moran's I). The association between EJ and tweet sentiment is assessed by Spearman correlation and three regression models (OLS, SLM and SEM). This thesis makes the following observations: 1) Proximity to hazard has the strongest influence on tweet sentiment among three EJ variables. 2) Population vulnerability has a negative association with positive word frequency, positive sentiment score and overall sentiment score, whereas proximity to hazard and health risk exposure have positive associations with these sentiment variables. 3) EJ and tweet sentiment have significant clustering in New York City. Although the spatial regression models have better performance than OLS model, some key explanatory variables are missing to predict tweet sentiment This thesis provides some insights for urban planners and policy makers to understand environmental issues with social media data.

Acknowledgements

First, I would like to express my sincere gratitude to my advisor Justin and my reader Sumeeta. Without their involvement and assistance in every step throughout the process, this thesis will never be accomplished.

Then, I would like to offer my special thanks to my friends and boyfriend. I will never forget their warm support and gentle encouragement when I started to learn Python by myself. In addition, I’m deeply grateful to my parents. They are always there for me no matter what happens.

, thanks to everyone I met in UEP department. The precious memory in Boston lives in me and will always inspire me to face future challenges and overcome difficulties.

Table of Contents

Abstract ……i

Acknowledgments ……ii

List of figures ……iv

List of tables……v

Chapter 1: Introduction ……1

Chapter 2: Literature Review ……4

Chapter 3: Methodology ……18

Chapter 4: Results and Analysis……26

Chapter 5: Conclusion, Limitations and Recommendations……49

Appendix A……55

Appendix B……57

References ……60

List of Figures

Figure 3.1. Flow chart of the research methods for tweet sentiment and EJ……18
Figure 3.2. Sample tweets from New York city……22
Figure 3.3. Sample result of tweet sentiment analysis……23
Figure 3.4. Sample result of tweet sentiment at block group level……24
Figure 4.1. Spatial distribution map and cluster map of Environmental Justice by block group…… 32
Figure 4.2. Spatial distribution map and cluster map of population vulnerability by block group…… 33
Figure 4.3. Spatial distribution map and cluster map of proximity to hazard by block group…… 34
Figure 4.4. Spatial distribution map and cluster map of health risk exposure by block group…… 35
Figure 4.5. Spatial distribution map and cluster map of positive word frequency by block group…… 37
Figure 4.6. Spatial distribution map and cluster map of negative word frequency by block group…… 38
Figure 4.7. Spatial distribution map and cluster map of positive sentiment score by block group…… 39
Figure 4.8. Spatial distribution map and cluster map of negative tweet sentiment by block group…… 40
Figure 4.9. Spatial distribution map and cluster map of overall sentiment score by block group…… 41
Figure 4.10. Map of bivariate association between EJ and positive word frequency by block group……42
Figure 4.11. Map of bivariate association between EJ and positive sentiment score by block group……43
Figure 4.12. Map of bivariate association between EJ and overall sentiment score by block group……43

List of Tables

Table 2.1. Major Groups of Indicators used in EJSCREEN……10
Table 2.2. Major Groups of Indicators used in CalEnvironScreen……12
Table 2.3. Major Groups of Indicators used in EJSM……13
Table 3.1. Summary table of EJ indicators description and sources……21
Table 4.1. Descriptive statistics of EJ and sentiment variables at the census block group level……27
Table 4.2. Results of Spearman correlation coefficient……28
Table 4.3. Multiple linear regression analysis of overall sentiment score and EJ variables……29
Table 4.4. Global Moran's I indices of EJ and sentiment variables……30
Table 4.5. Summary of diagnostics for OLS models at the census block group level…45
Table 4.6. Summary of spatial regression models for sentiment variables at the census block group level……47

Chapter 1: Introduction

Environmental Justice (EJ) is the effort to ensure that all people, regardless of race, color, or income, are protected from disproportionate impacts of environmental and health hazards. EJ also seeks to ensure that everyone enjoys equal access to democratic participation in the governmental decision-making processes that impact the environment. However, the problem of quantifying the environmental inequities and the health risks of exposure to environmental hazards has become a main challenge in the EJ movement. Research has demonstrated correlations between environmental exposure to specific toxins and specific diseases, but levels of exposure—including toxic concentration, exposure time and extent of impact—are difficult to measure. Regulatory agencies, such as the U.S. Environmental Protection Agency (US EPA), and state authorities, such as the California Air Board (CARB) and the California Office of Environmental Health Hazard Assessment (OEHHA), explore various methodologies to create an EJ index that accounts for the combination and potential interaction of social vulnerability and exposures to environmental hazards. The EJ index better reflects the cumulative impacts of exposures to harmful environmental hazards and factors that make communities more vulnerable to those exposures. To fully understand the applications of the EJ index, I explored and compared the metric categories and scoring system of three widely-used EJ screening tools, including EJSCREEN, CalEnviroScreen and EJSM to figure out their strengths and weaknesses.

Through a case study in New York City, this thesis develops an Environmental Justice mapping method that combines environmental and demographic information to quantify the level of proximity to certain types of environmental hazards and the potential vulnerability of

communities to these hazards. In order to quantify the cumulative impact of EJ, the initial step is to select variables identified in the literature and government reports related to EJ. In Chapter, I discuss the findings of these variables, focusing on how they contribute to EJ and how to combine them into an EJ screening tool.

This study also includes a Twitter data sentiment analysis in New York City and an analysis of the relationship and the spatial correlation between EJ and tweet sentiment. As an increasing number of people share and post their voices and opinions about everyday life and the built environment on social media, sentiment data is now widely used in environment-related fields because of its unique ability to monitor people's attitudes toward different environmental factors.

At present, EJ assessment continues to rely on statistical data from the EPA. With the development of social media, a large amount of user-generated content was produced. The data volume is large and easy to access. Also, it is instantaneously available, which reduces the cost of traditional data collection and solves the problem of data lag. Numerous studies have been done in analyzing the sentiment of social media data to learn about political and environmental issues. But in the field of EJ, the application of social media data is largely unexplored. Although previous sentiment-related environmental studies have investigated the association between environmental pollutants and the sentiment of social media data, few studies addressed spatial autocorrelation. This thesis will examine bivariate correlation between EJ and Twitter sentiment in pairs and build a spatial correlation model to compensate for the limitation of a non-spatial correlation method.

With all the above in mind, I came up with the following research questions:

1. How is Environmental Justice spatially distributed in New York City?

2. How is tweet sentiment spatially distributed in New York City?

3. Is there a relationship between tweet sentiment and Environmental Justice, and how does Environmental Justice affect people's sentiment?

To answer these questions, two sets of data will be used in this thesis. The first is the processed data for EJ indicators at the block group level that was downloaded from the EPA website. The second is a series of tweets collected from April 17 to July 24, 2017. The tweet sentiment will be evaluated based on the AFINN dictionary through Python and then processed into the block group level.

This thesis will be presented as follows. Chapter 2 is a literature review that provides a description of the definition, indicators and screening tools of EJ. This chapter also includes literature reviews on the classification methods of sentiment analysis and applications of social media data at present. Chapter 3 presents the methodologies used in this thesis for data collection, data processing, the scoring system, and the correlations between sentiment variables and EJ variables. Chapter 4 offers an overview of the spatial pattern of EJ and tweet sentiment. The cluster maps of Local Moran's I and Bivariate Moran's I help to make the spatial correlation easier to understand. Chapter 5 describes the conclusions and limitations of this study and offers suggestions for future relevant studies.

Chapter 2: Literature Review

2.1 Background of Environmental Justice in the U.S.

The concept of environmental justice was originated through anti-toxic and civil right movement in the 1980s in the United States (Mah, 2017). Environmental justice activists argue that minorities and low-income communities are always ignored by the mainstream environmental movement and are more likely to face heavier environmental hazard and pollutants (Taylor, 2011; Taylor, 2014). Environmental justice has become a legal issue after the Clinton administration signed Executive Order (E.O.) 12898, "Federal Actions to Address Environmental Justice in Minority Populations and Low-Income Populations", on February 1994 with the purpose to provide environmental protection to all communities (Greene, & Leip, 2002). It requires "each Federal agency to make achieving environmental justice part of its mission by identifying and addressing the disproportionately high and adverse human health or environmental effects of their actions on minority population and low-income population, to the greatest extent practicable and permitted by law (E.O. No. 12989, 1994)." The United States Environmental Protection Agency (EPA, 2011) came up with a clear definition of environmental justice which corresponds to the federal action. The EPA defines environmental justice as:

> It is the fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income with respect to the development, implementation and enforcement of environmental laws, regulations and policies. Fair treatment means no group of people should bear a disproportionate share of the negative environmental consequences resulting from industrial, governmental and commercial operations or policies.

Although the definition has been argued by some environmental justice activists and scholars because there is a lack of consideration in the historical burden of environmental

hazards in neglected communities (Wilson, 2010) and the affected groups under this definition should also include other potential vulnerable groups such as children, the elderly, and the feme (Maantay, & Maroko, 2008), it integrates two important concepts of distributional justice and procedural justice (Schroeder, 2008). The distributional justice refers to equal distribution of environmental hazards and benefits across all people, and the procedural justice refers to equal involvement of all people in environmental decision-making process (Cutter 1995, & Schroeder, 2008). Previous environmental justice activists always focused on fighting against unequal distribution of hazardous facilities, but environmental inequities will still happen when neglected communities share a disproportionately fewer benefits from policies and regulations which aims to reduce environmental hazardous (Bonorris, 2010).

2.2 Indicators related to Environmental Justice

Traditional approaches to examine environmental justice have relied heavily upon source data from industries in connection with socioeconomic data to look for disproportionate sites or chemical exposures in communities that are either economically disadvantaged or heavily populated by minorities. The academic literature has used different kinds of indicators to quantify environmental justice in the context of vulnerable populations and environmental quality.

Some studies develops Social Vulnerability Indices (SVI) based on demographic information to evaluate the vulnerability of communities to a given level of exposure to environmental hazard due to the accessibility to demographic census data (Nelson, Abkowitz,

& Camp, 2015; Chakraborty, Tobin, & Montz, 2005; Cutter, Boruff, & Shirley, 2003). Cutter argued that a higher level of social vulnerability will lead to fewer access to environmental benefits and political power (Cutter, 2000). There are wide variations in selecting demographic indicators to create SVI and the number of indicators "ranging from less than ten to more than fifty depending on the type of analysis and the index construction method (Nelson et al, 2015)". The Social Vulnerability Index (SoVI) is a widely acknowledged analysis method to measure SVI by reducing a large number of demographic indicators to a smaller set of vulnerable variables through principle component analysis, and each of the newly created variable is a integration of some highly-correlated demographic indicators (Cutter et al., 2003; Nelson et al., 2015). In a recent study of mapping the natural risk zones in Greater Lisbon, Cutter, Emrich and Zezere (2014) applied SoVI assessment to extract 7 variables from 46 demographic indicators to measure social vulnerability.

As the concept of environmental justice involves all people, regardless of socioeconomic or demographic character, share an equal proportion of environmental hazards and have equal access to environmental benefits and decision making process, race, income, gender, age and housing tenure measured in census data are the most widely-used demographic indicators to identify vulnerable populations (Azar, & Rain, 2007; Cutter, &Finch, 2008). Many studies conclude that people with a lower socio-economic status are more likely to expose to wider range of environmental hazards and have higher health risk (Gee, 2004; Burger & Gochfeld, 2011; Mitchell & Popham, 2008; Brainard, 2002). Racial and ethnic population face more difficulties to move away from environmental hazard and live in more desirable

neighborhood because of white housing preference and discrimination in real estate market (Galster& Godfrey, 2005; Oliver& Shapiro, 2006; Charles, 2003; Cutter et al., 2003; Downey & Hawkins, 2009). Previous studies in low-income communities found the indoor environment contains more pollutant which lead to disasters like asthma, cancer and respiratory irritation (Downey et al., 2009; Wilson, Kahn, Khoury, & Lanphear, 2007). Some studies found people with lower level of educational attainment or less fluency in English speaking are more vulnerable to environmental hazard because they don't have enough knowledge background to participate in the political decision-making process (Striessnig, Lutz, & Patt, 2013; Silverman, 2014). The Michiana Area Council of Governments (MACOG) uses non-Hispanic minority, household in poverty, linguistic isolation, individual over age 65, careless household and individual with physical disabilities as indicators to identify potentially disadvantage population in the environmental justice analysis (MACOG, 2014). Silverman (2014) used four socioeconomic attributes including median household income, percentage people of color, limited English proficiency and educational attainment less than high school degree to identify vulnerable populations.

Ramsey (2009) identified environmental hazard as "a possible source of danger due to the innate properties of an agent to cause harm". Many environmental justice studies analyzed the proximity to environmental hazardous sources. Some studies used the distance from central point of census tract or block group to the nearest environmental hazard source to measure the proximity, but the most widely used method is buffer analysis, which is "an analytical techniques provides by GIS software for creating new polygons around point, line

or area features on a map (Chakraborty, Maantay, & Brender, 2011).” The buffer radium of 0.5 and 1.0 mile are used very frequently in many studies (Glickman, 1994; Chakraborty & Armstrong, 2013; Boone, 2013; Zimmerman, 1994). Buffer analysis assumes that the exposure to environmental hazards are equal in all directions in despite of the wind and other factors would influence the movement of pollutants. Although Buffer analysis gives a more accurate spatial distribution of potential exposure to environmental hazards, some studies argues that the radius of the buffer is always chosen randomly, and thus making the result highly sensitive to the selection of the buffer radius (Chakraborty et al, 2011). Another limitation is the assumption of buffer analysis method that the affected area of a pollutant is restricted to the buffer boundary, and the area outside the buffer are not affected at all (Chakraborty et al, 2011). Cutter, Holm and Clark (1995) used three indicators to measure inequities in environmental hazardous, which include proximity to hazardous waste treatment, storage and disposal (TSD) facilities; proximity to Toxic Release Inventory (TRI) facilities, and proximity to inactive hazardous waste sites. Cushing (2015) minimize the number of CalEnvironScreen indicators to 6 indicators of exposure group and 4 indicators of environmental effects groups according to the potential overlap among them. The indicators include ozone concentration, PM 2.5 concentration, diesel PM concentration, pesticide use, toxic release, traffic density, and proximity to places such as clean-up sites, groundwater threats, hazardous waste sites, impaired water bodies, and solid waste sites. Silverman (2014) uses three categories such as Hazardous Waste Sites, Air Permit Sites, and Water Discharge Sites to detect environmental hazardous sites.

2.3 Visualization Tools of Environmental Justice

A number of visualization tools have been developed to identify and map the areas with potential environmental justice concerns. This study presents the methodologies of three major screening tools, which are EJSCREEN, CalEnviroScreen and EJSM. From the tools we understand what are the indicators that best reflect the scientific evidence and community concerns about environmental justice and how are the indicators be combined to rank cumulative burdens and show progress on environmental justice (Kuruppuarachchi, Kuma, & Franchetti, 2017). Some studies have applied the indicators or methodologies from these screening tools to identify the disadvantaged communities in the area of interest. For example, the natural resources defense council (NRDC) developed a methodology to map environmental justice for Chicago by combining the EJSCREEN indicators and EJSM quintile ranking method (Meleah, 2018). A research team from the National Center for Smart Growth and School of Public Health incorporated additional indicators based on EJSCREEN and used scoring method in CalEnviroScreen to assess environmental justice in Maryland (Aubree et al., 2019).

2.3.1 EJSCREEN

The EPA develops an environmental justice mapping and screening tool called EJSCREEN to calculate and display the environmental justice index (EJ index) for an area with 2 demographic indicators and 11 environmental indicators as shown in Table 2.1 (EPA 2010).

Demographic indicators	Low-income Minority
Environmental indicators	Air Toxics Cancer Risk Air Toxics Respiratory Hazard Index Air Toxics Diesel PM (DPM) Lead Paint Particulate Matter (PM2.5) Ozone Traffic Proximity and Volume Proximity to Risk Management Plan Sites Proximity to Treatment Storage and Disposal Facilities Proximity to National Priorities List Sites Proximity to Major Direct Water Dischargers

Table 2.1. Major Groups of Indicators used in EJSCREEN

The Demographic Index in EJSCREEN is the average of percent low-income and percent minority for each census block group. EPA combines a single environmental indicator with demographic index to calculate a single EJ index. It looks at the difference of demographic information between local block group and national average, and it considers the population of each block group as well. The EJ index is reported in percentile terms, which shows whether this value of local block group is higher or lower than that of the nation. The EJ indexes are not combined into a single number because there is no object method to decide how much weight should be given to each indicator.

Demographic Index = (% minority + % low-income) / 2

EJ Index = (Environmental Indicator) * (Demographic Index for Block Group –Demographic Index for US) * (Population count for Block Group)

2.3.2 CalEnvironScreen

The California Environmental Protection Agency (CalEPA) asked Office of Environmental Health Hazard Assessment (OEHHA) to develop CalEnvironScreen, a

screening tool, to identify California communities that are disproportionately burdened by multiple sources of environmental hazard and people who are vulnerable to pollution's effects. CalEnvironScreen indicators are a measure of either environmental conditions, or the vulnerability factors for population characteristic. The indicators fall into four board groups, including exposures, environmental effects, sensitive populations and socioeconomic factors as shown in Table 2.2. The indicators of exposure and environmental effects groups in CalEnvironScreen are based on measurement of different types of pollution that people may come into contact with and locations of toxic chemicals in or neat communities (CalEPA, 2016). The score of each component (exposure, environmental effects, sensitive populations, socioeconomic factors) is calculated by the average percentile of indicators that make up this component. The score of population characteristic group is the average of sensitive population and socioeconomic factors component. The score of population burden group is the average of exposure and environmental effects component. CalEPA thinks the environmental effects has a smaller influence than exposure on pollution burden, so the environmental effects is given a weight of 0.5. The score of each group is divided by the maximum value observed in the state and then multiplied by 10 in order to ensure the population characteristic group and pollution burden group contribute equally to the general EJ score. The EJ score is calculated by multiplying the scores of population characteristic group and pollution burden group.

Pollution burden indicators	Exposure	Ozone Pm2.5 Diesel PM Emissions Drinking water quality Pesticide use Toxic release from facilities Traffic density
	Environmental effects	Cleanup sites Groundwater threats Hazardous waste generators Impaired water bodies Solid waste sites and facilities
Population characteristic indicators	Sensitive populations	Cardiovascular disease Low birth-weight infants Asthma
	Socioeconomic factors	Educational attainment Housing burden Linguistic isolation Poverty Unemployment

Table 2.2. Major Groups of Indicators used in CalEnvironScreen

2.3.3 EJSM

Environmental justice screening method was developed by a group of California-based academic researchers in conjunction with community groups and state environmental agencies (Sadd et.al, 2011). EJSM examines the cumulative impacts and social vulnerability to detect environmental justice based on 23 indicators. These indicators are organized into three categories: hazard proximate and land use, air pollution exposure and estimated health risk, and social and health vulnerability (Table 2.3). The intermedia cumulative impacts (CI) score is calculated for indicators in each category based on quintile distribution rankings (with scores ranging from 1-5). The CI score for each category equals to the sum of intermedia CI scores. The category score is then re-ranked into quintile again, and each

census is assigned a new CI score which ranges from 1 to 5. The final CI score was calculated by adding up the score of each category.

Hazard proximate and land use	Air quality hazards	Facilities in California community health air pollution informatic system Chrome-platers Hazardous waste site
	Hazardous land use	Railroad facilities Ports Airports Refineries Intermodal distribution
Health risk and exposure	Risk Screening Environmental Indicators (RSEI) toxic concentration hazard score National Air Toxics Assessment respiratory hazard for air toxics from mobile and stationary emissions Estimated cancer risks from modeled ambient air toxics concentrations from mobile and stationary emissions PM2.5 estimated concentration interpolated from CARB's monitoring data Ozone estimated concentration interpolated from CARB's monitoring data	
Social and health vulnerability	% people of color (total pop–non-Hispanic white) % below twice the national poverty level Home Ownership–% living in rented households Housing Value–median house value Educational attainment–% >age 24 with age 60 Linguistic isolation–% residents under age 4 in households where no one over age 15 speaks English well Voter turnout–% votes cast in general election Birth outcomes–% preterm and small for gestational age	

Table 2.3. Major Groups of Indicators used in EJSM

2.4 Sentiment Analysis with Social Media data

Sentiment analysis has been defined as the computational study of opinions, feelings, attitudes, and emotions expressed in written language. Sentiment analysis mainly focuses on the opinions which express or imply positive or negative sentiments (Liu, 2012). Social Media is "a group of Internet-based applications that built on the ideological and

technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content (Kaplan & Haenlein, 2010)". The social media has changed the way we interact with people and provided a platform for idea and information exchange. With the rapid development of social media, a growing number of individuals and organizations open accounts in social network sites. According to Statista data, there were 233 million social network users in the United States, which is about three quarters of the country's population. The number of Facebook active user was more than one billion across all platforms in the third quarter of 2012, and the number of monthly active user in the first quarter of 2021 was 2.85 billion (Facebook, 2021). There were about 192 million Twitter daily active users in the fourth quarter of 2020 all over the world, and 55 million were U.S. twitter users (Twitter, 2020). Social media has a large number of opinions which indicate people's reflection to a specific event. These reactions could be quantified by using automatic sentiment classification techniques.

2.5 Key approaches of sentiment classification

There are two common ways to perform sentiment classification, which are machine learning method and lexicon-based method. Both methods classify the text into positive, negative or neutral sentiment based on its polarity (Dhaoui, Webster, & Tan, 2017).

In machine learning method, two datasets are involved. The training dataset, collected from various type of sources and whose output values are known to us, is used by classifier to learn recognizing the distinctive features of text. The test dataset whose output values are unknown is then used to examine whether the classifier works well after training (Kolchyna, Souza,

Treleaven, & Aste, 2015). Machine learning method uses features such as unigrams or bigrams to train classifier by applying different learning techniques such as Naive Bayes, maximum entropy or support vector machines (Taboada et al., 2011). Machine learning method "tend to be highly domain/genre specific, performing well on the domain/genre of training but poorly on a different domain/genre (Muhammad, Wiratunga, & Lothian, 2016)." The size and quality of the training dataset have a great effect on the performance of training result. Social media text has various sorts of domain and genre in different topics with short or long posts, thus the performance of training result is determined by the number of domain or genre that is sampled in the training dataset. Although using high-quality and large training dataset is time-consuming, it leads to higher accuracy in classification (Muhammad et al., 2016).

In lexicon-based method the unigrams are extracted and given polarity scores according to a sentiment dictionary, and the overall sentiment of the text is predicted by aggregating all the polarity scores (Muhammad et al., 2016). The sentiment dictionary such as SentiWorldNet, Linguistic Inquiry and Word Count (LIWC), Opinion Lexicon world list, and AFINN polarity world list contains different types of sentiment words that express people's opinions or feelings with quantified polarity scores (Kolchyna et al., 2015). Some limitations exist in lexicon-based method: a same sentiment word might have opposite sentiment orientations in different domains, but sentiment words in the dictionaries are not specific to different topics; sometimes a sentiment words might not indicate any sentiment, especially in

interrogative and conditional sentences; some texts have a sentiment orientation although none sentiment words exist in the text (Liu, 2010, 2012; Tanvi & Dilipkumar, 2016).

2.6 Sentiment Analysis Application in Urban and Environmental Issues

Sentiment analysis over social media data has been applied to various urban and environmental topics. López-Ornelas, Abascal-Mena and Zepeda-Hernández (2017) used sentiment analysis in proposing a new indicator to evaluate subject well-being, and the new measure shows similarities with the traditional method to detect well-being. Hollander and Renski (2017) compared attitudes of people between shrinking cities and growing cities through sentiment analysis and found population decline might not contribute to lower sentiment levels, suggesting that more investigation should be done to explore how does decline impact the community. Hollander and his college (2020) identified a number of risks associated with AI powered automated social media accounts and argued that social bots could affect community well-being. Plunz et.al (2019) explored visitor sentiment in areas inside and outside the park to see how urban green space contribute to general well-being. Hollander and Shen (2017) used twitter data to analyze if relationship exist between biking ability and people's attitudes about biking in Washington. Bertrand et.al (2013) generated a sentiment map of New York City to locate the area of the highest and the lowest public mood. In transportation research, Das et.al (2019) investigated how to use biker sentiment to help transportation planners or policy makers to promote biking and bike commuting. The findings from Bao et.al (2017) study showed the advantages of incorporating human activity information in spatial analysis of crashes through twitter sentiment data. Sentiment analysis

of microblog data has also been used to investigate air quality in many studies. For instance, a recent study analyzed the sentiment strength of dataset from a Chinese social media platform Microblog and found the sharp bursts of sentiment strength happened simultaneously with the occurrence of PM2.5 air pollution events (Sha, Yan, & Cai, 2015). Zheng and her colleagues (2018) found a strong negative correlation between Microblog sentiment and air quality after collecting 210 million microblog data from 144 Chinese cities. Another similar study used Chinese Microblog data to explore air pollution issue also found the air pollution-related Microblog sentiment correspond to Air Quality Index (Jiang, Wang, Tsou, & Fu, 2015). Poor air quality can stimulate depression, which affect people's mood and social lives negatively (Shan, Ju, Wei, & Wang, 2021).

Chapter 3: Data and Methodology

Figure 3.1 shows the theoretical and methodological process of the study: 1) Environmental Justice data collection and Environmental Justice score calculation. 2) Twitter data collection and pre-processing in Python: filter tweets without sentimental words. 3) Tweet sentiment analysis: use AFINN dictionary to judge the sentiment polarity and intensity. 4) Descriptive statistic: summarize data using index such as mean, standard deviation, minimum and maximum values of sentiment variables and Environmental justice score. 5) Spearman correlation and multiple linear regression: verify the association between sentiment variables and Environmental Justice score 6) Spatial correlation analysis: use global Moran's I, Local Moran's I, and Bivariate Moran's I to investigate the spatial correlation between sentiment and Environmental Justice.

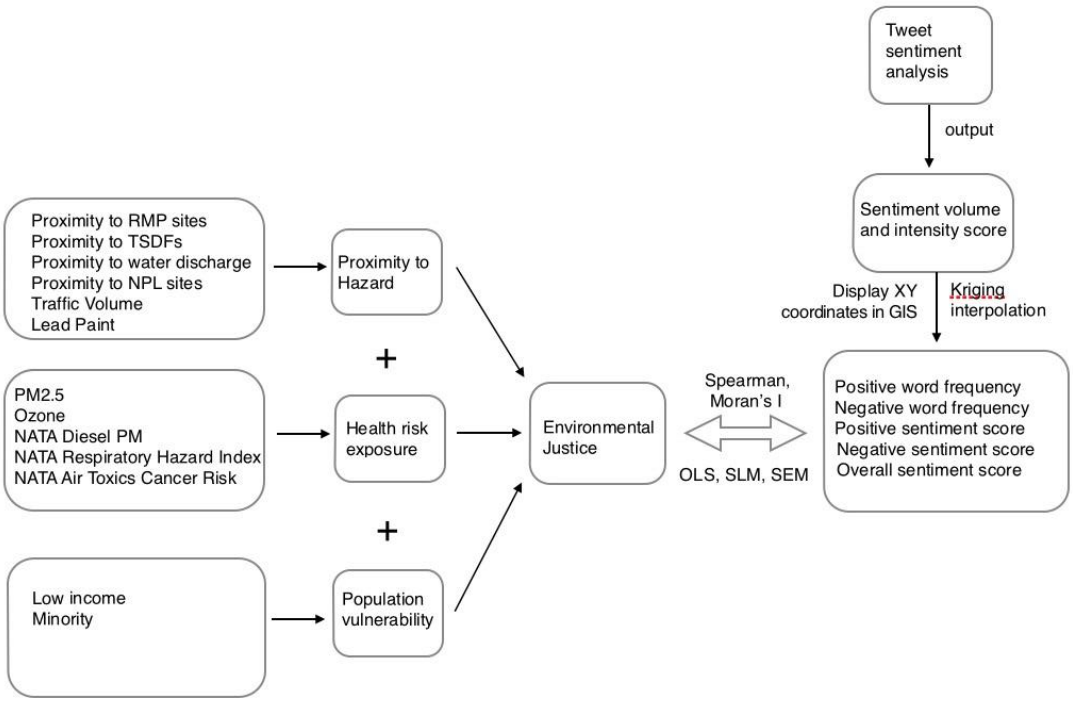


Figure 3.1. Flow chart of the research methods for tweet sentiment and EJ

3.1 Mapping Environmental Justice

In this study, I will follow the methodology based on EJSM to calculate a composite score that accounts for 13 EJSCREEN indicators. According to the literature review, indicators in EJSM and CalEnvironScreens are more regional-specific and suitable for California. EJSCREEN, developed by the US EPA, could be used in every state. However, unlike the other two tools calculating the cumulative EJ score, EJSCREEN evaluated a single EJ index for each indicator, which is not consistent with the purpose of this study. Finally, I adopt the scoring system of EJSM because there is no weight attached to its indicators. Based on the literature review, very few studies provided guidelines for the weighting scheme of EJ indicators. The 2017 EJSCREEN raw data could be downloaded from the EPA website. Table 3.1 summarizes the description and data sources of indicators of demographics, proximity to hazards, and health risks and exposures (EPA, 2017).

An EPA 2017 data file provides the values of each block group in the U.S. First, I will extract the values of block groups in New York City. For each indicator, each block group is assigned a score ranging from 1 to 5 based on the quintile ranking. The first quintile, which represents the lowest 20% of the data, receives a score of 1, while the fifth quintile, which represents the highest 20% of the data, receives a score of 5. The block group with the higher score has greater vulnerability or exposure.

The quintile scores for indicators in each group are added up respectively. The score for the hazard proximity group ranges from 6 to 30, the score for the health risk and exposure group ranges from 5 to 25, and the score for the demographic group ranges from 2 to 10. The

scores of each group are then re-ranked into the quintile, and each block group is assigned a new score ranging from 1 to 5. The final EJ score is calculated by adding up the new score for each group.

EJ indicator		**Description**	**Source**
Demographic indicators	Low income	The percent of a block group's population in households where the household income is less than or equal to twice the federal "poverty level."	ACS five-year estimates 2017
	minority	The percent of individuals in a block group who list their racial status as a race other than white alone and/or list their ethnicity as Hispanic or Latino.	ACS five-year estimates 2017
Hazard proximity indicators	Proximity to RMP sites	Count of RMP (potential chemical accident management plans) facilities within 5 kilometers or close to 5 kilometers, divided by distance in kilometers	EPA RMP database 2017
	Proximity to TSDFs	Count of TSDF (hazardous waste management facilities) within 5 kilometers or closest to 5 kilometers, divided by distance in kilometers.	EPA RCRAInfo database 2017
	Proximity to wastewater discharge	Toxic concentrations in stream segments within 500 meters, divided by distance in kilometers.	RSEI 2017
	Proximity to NPL sites	Count of NPL/Superfund sites (polluted sites that pose a risk to human health and/or the environment) within 5 kilometers or close to 5 kilometers, divided by distance in kilometers.	EPA CERCUS database 2017
	Traffic Volume	Count of vehicles (average annual daily traffic) at major roads within 500 meters or close to 500 meters, divided by distance in meters	U.S. DOT traffic data 2014
	Lead Paint	Percent of houses built before 1960, which likely contain lead paint.	Census data 2017

Health risk and exposure indicators	Particulate Matter 2.5	Levels of particulate matter with a diameter of 2.5 micrometers or smaller in air.	EPA 2013
	Ozone	Summer seasonal average of the maximum daily 8-hour concentration of ozone in air in parts per billion.	EPA 2013
	NATA Diesel PM	Levels of diesel particulate matter in the air.	EPA 2011
	NATA Respiratory Hazard Index	Ratio of exposure concentration in the air to the health-based reference.	EPA 2011
	NATA Air Toxics Cancer Risks	Lifetime risk of developing cancer from inhalation of air toxins.	EPA 2011

Table 3.1. Summary table of EJ indicators description and sources

3.2 Tweet Sentiment Analysis

The Urban Attitude Lab research team of Tufts University developed a software that connects to Twitter application program interface (API) to collect individual tweets from April 17 to July 24, 2017. In order to extract the tweets that truly indicate people's opinions, the dataset was filtered by excluding the commercial tweets like advertisements and job announcements. The tweets were exported to Microsoft Excel in .cvs file format. Figure 3.2 represents a small set sample of the filtered tweets. Column A is the twitter id for users instead of using their actual user's name. column b is the text part of tweets. Column c and column d show the longitude and latitude coordinates of tweets. Column f shows the time that tweets were posted.

A	B	C	D	E	F
362629413	I can't believe my bug is going to be 11 in a few days @ Amazura https://t.co/5V7n6ZD	-73.81037	40.70017	2017-05-24	09:58:20
10254122	Iris. Owen's A Photo A Day. Day 144. Taken on May 22 2017 at 6:47 https://t.co/58V	-73.94472222	40.66305556	2017-05-24	09:58:37
11101232	I'm at Astoria Park - @nycparks in Astoria NY https://t.co/qsblm8AzWF	-73.920682	40.780519	2017-05-24	09:59:06
8.26784E+17	Peter Hunter og Julie Hendel https://t.co/a9wkNTH7Mm	-74.0007613	40.7207559	2017-05-24	09:59:09
15997164	I need a little country! (at @DunkinDonuts in New York NY) https://t.co/LGJ0jYT5kd	-73.9750886	40.7902246	2017-05-24	10:00:08
3214786035	Storytime: Letter F https://t.co/pmM4VBZPAk	-74.0007613	40.7207559	2017-05-24	10:00:09
30209392	Why does God love me? https://t.co/9DOVClhubn	-74.0007613	40.7207559	2017-05-24	10:00:09
188149127	catching the sunset https://t.co/izH10fZXZ9	-74.0007613	40.7207559	2017-05-24	10:00:10
992126359	Just posted a photo @ Mount Carmel Pentecostal Church https://t.co/SVMsQYESfi	-73.86698	40.87097	2017-05-24	10:00:12
7.32428E+17	FLEET WEEK MATCH! Come watch Combined Sea Services Play NYPD/FDNY 6pm at Rar	-73.92483441	40.79473147	2017-05-24	10:00:15

Figure 3.2. Sample tweets from New York city

The approach implemented for sentiment analysis in this study is lexicon-based method using the AFINN dictionary on Python. The AFINN dictionary was developed by Finn Årup Nielsen, including 2477 unique words in the newest version (Nielsen, 2011). Each word is assigned with an integer score ranging from -5 to 5 based on its sentiment polarity and strength. The negative sentiments receive negative scores and positive sentiments receive positive scores. A higher absolute value of the score indicates stronger sentiment. For instance, the word "pollution" is given a score of -2, while the word "happy" is given a score of 3. The AFINN dictionary also contains some obscene words and slang acronyms frequently used on the internet, such as recognizing "LOL" as "laughing out loud" and "n00b" as "noob". The AFINN dictionary has been applied in multiple studies, including a comparative analysis of twitter sentiments between shrinking cities and growing cites (Hollander & Renski, 2017), an analysis of Facebook fan page ranking based on user opinion (Ngoc & Yoo, 2014), sentiment analysis of online restaurant reviews (Gan, Ferns, Yu, & Jin, 2017), and identification of character-to-character sentiment in Shakespeare's Plays (Nalisnick & Baird, 2013).

After running New York City' tweets through Python, initial statistics were summarized in terms of five valuables for each tweet including number of positive words, number of negative words, positive score, negative score and overall sentiment score. The positive score

of each tweet is obtained by adding up the score of every single positive word, and negative score is the total score of every single negative word. The overall sentiment score of each tweet is the sum of positive score and negative score. Figure 3.3 shows a small set sample of sentiment analysis result.

0	1	2	3	4	5	positive_words	negative_words	positive_score	negative_score	overall_score
85549638	28447705	Want to work at CVS Health? We're #hiring in #NewYork NY	-74.00594	40.71278	2017-04-2	1	0	1	0	1
85549638	38118086	This #job might be a great fit for you: Credit Risk Intern - htt	-73.19517	41.18655	2017-04-2	2	1	4	-2	2
85549641	1.61E+08	「Black is Beautiful in every shade」 #メンズ#ファッショ	-73.93764	40.71154	2017-04-2	2	0	5	0	5
85549643	5.74E+08	Want to work in #GardenCity NY? View our latest opening:	-73.63429	40.72677	2017-04-2	1	0	1	0	1
85549656	43909158	Throwback Fail Friday from 2011. #fail #oops #failfriday #b	-74.07111	40.96192	2017-04-2	0	2	0	-4	-4
85549661	4.88E+09	This #job might be a great fit for you: Senior Accountant - ht	-74.16375	40.85843	2017-04-2	2	0	4	0	4
85549665	24220809	See our latest #NewYork NY #job and click to apply: Financ	-74.00594	40.71278	2017-04-2	0	1	0	-2	-2
85549666	61431068	Sweet California @ CDMX https://t.co/iXs93AX8Z5	-74.00076	40.72076	2017-04-2	1	0	2	0	2
85549679	5E+08	Can you recommend anyone for this #job? Regional Sales M	-74.15020	40.57953	2017-04-2	1	0	2	0	2
85549680	5.7E+08	Can you recommend anyone for this #job in #Stamford CT?	-73.53879	41.05296	2017-04-2	1	0	2	0	2
85549684	75719630	Packaging is exciting part of my work✨ #lovefactoryny	-74.0064	40.7142	2017-04-2	1	0	3	0	3
85549684	52272942	Update: Construction on #MosholuParkway WB from Henr	-73.88741	40.90093	2017-04-2	0	1	0	-1	-1
85549684	52272942	Cleared: Construction on #OysterBayBranch Both direction	-73.80803	40.70054	2017-04-2	1	0	1	0	1
85549685	51149096	Update: Construction on #NJ21 Both directions from CR 5(	-74.16724	40.74446	2017-04-2	1	0	3	0	3
85549685	50728535	Cleared: Construction on #SawMillRiverParkway NB from E	-73.75890	41.1802	2017-04-2	1	0	1	0	1

Figure 3.3. Sample result of tweet sentiment analysis

After finishing the initial sentiment analysis of the original tweets data, 869118 tweets which contain sentiment words are selected and prepared for processing in GIS. The tweets are geolocated by using the latitude and longitude coordinates. I clipped the tweets data with the boundary of NYC to exclude the unnecessary data and then use Spatial Join to affix sentiment data to block group shapefile. The tweets that cluster into the same block group are summarized into five outcome variables including positive word frequency per tweet, negative word frequency per tweet, average positive score, average negative score, and average sentiment score. The new sentiment variables are calculated as the sum of original value divided by the total tweets within the block group (see Figure 3.4). The distribution of sentiment-related tweets does not cover every block group in NYC. Therefore, I fulfill the missing values through Kriging interpolation and then use Zonal Statistic tool to join the table to block group shapefile. In this step, both sentiment data and Environmental Justice score are in block group level.

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	GEOID	NAMELSAD	MTFCC	FUNCSTAT	ALAND	AWATER	INTPTLAT	INTPTLON	Count_	Avg_X_lon	Avg_Y_lat	positive_frec	negative_fre	positive_ave	negative_av	sentiment_a
2	360470000000	Block Group	G5030	S	37441	0	40.6688898	-73.986582	1	-73.98602	40.66886	0	3	0	-12	-12
3	360050000000	Block Group	G5030	S	40963	0	40.8974045	-73.853314	1	-73.8533	40.89775	0	3	0	-12	-12
4	360050000000	Block Group	G5030	S	20673	0	40.8527041	-73.883724	1	-73.883472	40.8537807	2	2	2	-10	-8
5	360811000000	Block Group	G5030	S	143816	0	40.7093658	-73.902273	1	-73.90443	40.7092109	0	3	0	-9	-9
6	360610000000	Block Group	G5030	S	18989	0	40.8617869	-73.92755	1	-73.927931	40.8621262	0	3	0	-9	-9
7	360471000000	Block Group	G5030	S	144613	0	40.6457434	-73.932316	1	-73.933325	40.6446986	0	2	0	-8	-8
8	360470000000	Block Group	G5030	S	39835	0	40.6839002	-73.960946	1	-73.960167	40.6834412	0	2	0	-8	-8
9	360811000000	Block Group	G5030	S	37802	0	40.7020275	-73.907896	1	-73.9077	40.7028	0	2	0	-8	-8
10	360470000000	Block Group	G5030	S	39072	0	40.6787023	-73.978975	1	-73.9791	40.67866	1	2	1	-7	-6
11	360050000000	Block Group	G5030	S	33542	0	40.8439928	-73.887117	1	-73.888016	40.8440978	0	3	0	-7	-7
12	360050000000	Block Group	G5030	S	53858	0	40.8063159	-73.911951	1	-73.911219	40.8054272	1	2	3	-7	-4
13	360050000000	Block Group	G5030	S	46331	0	40.8353139	-73.921658	1	-73.921999	40.834742	2	2	4	-7	-3
14	360810000000	Block Group	G5030	S	126114	0	40.7573672	-73.942468	1	-73.94264	40.7581	0	2	0	-7	-7
15	360470000000	Block Group	G5030	S	46165	0	40.683458	-73.948513	1	-73.949327	40.6828493	0	2	0	-6	-6
16	360610000000	Block Group	G5030	S	44785	0	40.7546268	-73.997254	1	-73.99868	40.7549412	0	2	0	-6	-6
17	360470000000	Block Group	G5030	S	106834	0	40.6814326	-73.904644	1	-73.904436	40.682088	1	2	2	-6	-4
18	360470000000	Block Group	G5030	S	42080	0	40.6897911	-73.958657	1	-73.959506	40.6902645	1	2	1	-6	-5

Figure 3.4. sample result of tweet sentiment at block group level

3.3 Statistical Analysis

Statistical analysis implemented in this study with descriptive statistics, Spearman correlation coefficient, and multiple regression analysis. Descriptive statistics is used to summarize the general characteristics of the Environmental Justice and Tweet sentiment. Associations were assessed with Spearman correlation coefficient and multiple regression analysis. In order to explore whether the sentiment variables are associated with EJ indicators, I analyzed the Spearman correlation coefficient and its significance test between them. The Spearman correlation coefficient measures the degree of association between two sets of data. The positive and negative words frequency, positive and negative score per tweet, and overall sentiment score per tweet are the dependent variables, and independent variables are the scores of proximity to hazards, health and risk exposure, population vulnerability and Environmental Justice. To further verify the influence of EJ levels on the volume and intensity of tweet sentiment, I decided to conduct a multiple linear regression model.

3.4 Spatial Autocorrelation Analysis

Spatial autocorrelation refers to potential interdependence between some variables in the same distribution area. Tobler (1970) pointed out “the first law of geography: everything is

related to everything else, but the near things are more related than the distant things."

Spatial autocorrelation is an optimal method to detect spatial clusters of Environmental Justice inequities and sentiment strength. In this thesis, global and local indices of Moran's I were used to measure spatial autocorrelation. Moran's I was developed by Patrick Alfred Moran in 1950. Global Moran's I quantifies the spatial autocorrelation as whole, while Local Moran's I detect the degree of spatial autocorrelation at every spatial unit and identifies local spatial cluster patterns and explicit outliers.

Global Moran's I has been used to test the null hypothesis that the spatial autocorrelation of a variable is zero. If the null hypothesis is rejected, the variable can be considered spatially autocorrelated (Cliff & Ord, 1981). To further assess the presence of the clusters, Anselin Local Moran's I has been used to test the null hypothesis that the there is no local spatial association between the value observed at a location and the values observed at nearby sites (Anselin, 1995). If the null hypothesis is rejected, the variable can be considered local spatially autocorrelated. I will also use Bivariate Moran's I to analyze the spatial autocorrelation between EJ variables and sentiment variables. The area of where the two variables are related in map can show how their relationship change across NYC. At last, I will run Ordinary Least Square (OLS), Spatial Lag Model (SLM) and Spatial Error Model (SEM) to analyze the spatial association between the EJ and tweet sentiment.

Chapter 4: Results and Analysis

In this chapter, I will introduce the findings of this study with maps and tables.

4.1 Correlations between EJ and tweet sentiment

The relationship between EJ and tweet sentiment was analyzed at the block group unit level in NYC for 2017. There are 6302 block groups in NYC, and 6302 groups of data used in the correlation analysis. Before analyzing the association between sentiment and EJ variables, raw data is restructured into descriptive statistics. Since the negative score is negative number, the strength of negative sentiment becomes stronger when the negative score becomes smaller. This negative association will make the final result of correlation difficult to understand. Thus, I use the absolute value of negative score.

Table 4.1 lists statistical results of EJ variables and sentiment variables. The mean and standard deviation of population vulnerability, proximity to hazard, and health risk and exposure are very close to each other, which indicates the central tendency and the degree of dispersion of EJ valuables are very similar. The mean of positive word frequency and positive score per tweet are both higher than that of negative side, which means the volume and intensity of negative sentiment are larger than positive sentiment. The mean of overall sentiment score per tweet is 1.29, indicating the overall tweet sentiment in NYC is slightly positive.

	N	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
Population vulnerability	6302	1	5	21065	3.34	1.38	1.91
Proximity to hazard	6302	1	5	19559	3.10	1.35	1.83
Health risk and exposure	6302	1	5	20208	3.21	1.38	1.91
Total EJ score	6302	3	15	60832	9.65	2.74	7.49
Positive word frequency	6302	0	1.59	5374.03	0.85	0.24	0.06
Negative word frequency	6302	0	0.80	1602.60	0.25	0.10	0.01
Positive score	6302	0	3.75	11193.07	1.78	0.56	0.31
Negative score	6302	0	1.73	3044.29	0.48	0.22	0.05
Overall Sentiment score	6302	-0.73	3.52	8147.18	1.29	0.56	0.31

Table 4.1. Descriptive statistics of EJ and sentiment variables at the census block group level

As I mentioned in the Chapter 1, one of my research questions is about how the level of Environmental Justice indicator affect the tweet sentiment volume and intensity. The null hypothesis is “there is no association between the sentiment variables and EJ variables”. The null hypothesis is rejected when $p<0.05$. The outcome of the cross-tabulation in Table 4.2 allows us to see if the associations exist. The p-values between the following variable are all smaller than 0.05 except the “negative word frequency” and “population vulnerability”, which indicates that there is insufficient evidence to conclude a significant relationship between “negative word frequency” and “population vulnerability” because the correlation coefficient is not significantly different from 0. Hazard proximity, heath risk and exposure and overall EJ scores all show a weak positive correlation ($R=0.30, 0.20, 0.23$) with negative

word frequency, as well as negative sentiment score (R=0.29, 0.19, 0.24). However, the above three EJ variables also show weak positive correlation with positive word frequency (R=0.43, 0.38, 0.28), positive sentiment score (R=0.43, 0.37, 0.27), and overall sentiment score (R=0.33, 0.31, 0.19). It means people in the area of higher proximity to hazard or higher health risk produce tweets with more positive sentiment words or stronger positive sentiment, which contradicts my assumption drew from the previous studies that they are negatively correlated. Only population vulnerability shows a weak negative correlation (R=-0.24, -0.26, -0.26) with positive words frequency, positive score and sentiment score, which indicates that increasing in population vulnerability is associated with decreasing level of positive word frequency, positive score and overall sentiment score.

Spearman Correlations									
variables	**vulnerability**	**hazard**	**exposure**	**EJ**	**pos word**	**neg word**	**pos score**	**neg score**	**sentiment score**
vulnerability	1.00	-0.05**	-0.04**	0.45**	-0.24**	-0.02	-0.26**	0.01	-0.26**
hazard	-0.05**	1.00	0.58**	0.76**	0.43**	0.30**	0.43**	0.29**	0.33**
exposure	-0.04**	0.58**	1.00	0.77**	0.38**	0.20**	0.37**	0.19**	0.31**
EJ	0.45**	0.76**	0.77**	1.00	0.28**	0.23**	0.27**	0.24**	0.19**
pos word	-0.24**	0.43**	0.38**	0.28**	1.00	0.22**	0.94**	0.23**	0.88**
neg word	-0.02	0.30**	0.20**	0.23**	0.22**	1.00	0.25**	0.92**	-0.07**
pos score	-0.26**	0.43**	0.37**	0.27**	0.94**	0.25**	1.00	0.24**	0.93**
neg score	0.01	0.29**	0.19**	0.24**	0.23**	0.92**	0.24**	1.00	-0.10**
sentiment score	-0.26**	0.33**	0.31**	0.19**	0.88**	-0.07**	0.93**	-0.10**	1.00
**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).									

Table 4.2. Results of Spearman correlation coefficient at the census block group level

The result of the multiple regression analysis is shown in Table 4.3. According to Spearman correlation analysis, the correlations between population vulnerability, proximity to hazard and health risk exposure are not strong, so they could all be used for multiple linear regression analysis. Therefore, the multiple linear regression analysis with a dependent variable of overall sentiment score and independent variables of three EJ variables were used to predict tweet sentiment. The prediction model was statistically significant (F = 416.947, $p < 0.01$) and accounted for approximately 16.6% of the variance of tweet sentiment score (R=0.41, adjust R^2=0.167). Multiple regression analysis shows that, among the three predictor variables, population vulnerability is found to be the variable with the highest value of coefficient (b=-0.10, p<0.01), following by proximity to hazard (b=0.08, p<0.01) and health risk exposure (b=0.07, p<0.01). The result indicates that people's sentiments are more sensitive to their vulnerability of environmental stressors.

	Unstandardized B	**Coefficients Std. Error**	**Standardized Coefficients Beta**	**t**	**Sig**	**Tolerance**	**VIF**
(Constant)	1.14	0.02		46.99	0.00		
Population vulnerability	-0.10	0.01	-0.24	-20.36	0.00	0.99	1.00
proximity to hazard	0.08	0.01	0.20	14.42	0.00	0.67	1.49
Exposure to health risk	0.07	0.01	0.16	11.38	0.00	0.67	1.49

Table 4.3. Multiple linear regression analysis of overall sentiment score and EJ variables. dependent variable: overall sentiment score; N=6302. R^2 = 0.166, F = 416.947, and $p < 0.01$.

4.2 Spatial analysis of EJ and tweet sentiment

GIS mapping technique helps to identify spatial pattern of Environmental Justice and tweet sentiment visually, and Geoda helps to determine the underlying spatial autocorrelation

statistically described by Global Moran's I and Local Moran's I. In this thesis, the null hypothesis is defined as the lack of clustering pattern in EJ and tweet sentiment. Table 4.4 shows the Z-score and p-values for Global Moran's I. All p-values are all smaller than 0.05, and thus, the null hypothesis is rejected with 95% confidence level and 5% error, which reveals all variables have significant clustering.

Variables	Global Moran's I	Z-value	P-value
Population vulnerability	0.70	2.57	**
Proximity to hazard	0.85	2.83	*
Health risk exposure	0.89	2.88	**
Total EJ score	0.85	2.83	**
Positive word frequency	0.85	2.82	**
Negative word frequency	0.74	2.60	*
Positive sentiment score	0.85	2.85	**
Negative sentiment score	0.77	2.67	*
Overall sentiment score	0.81	2.77	**
**. P-value is significant at the 0.01 level (2-tailed). *. P-value is significant at the 0.05 level (2-tailed).			

Table 4.4. Global Moran's I indices of EJ and sentiment variables

The values of Local Moran's I for all EJ and sentiment variables are higher than 0.7, indicating they have strong spatial correlation. There are very few high-low and low-high outliers for both EJ and tweet sentiment variables. Block groups with higher cumulative impacts from environmental and social stressors, more tweet sentiment words, or stronger sentiment intensity are surrounded by block groups of similar situations.

According to overall EJ cluster map (see Figure 4.1), large and clear high-high spatial clusters are observed in upper Manhattan and north-western part of Queens and Brooklyn. The low-low spatial clusters are mainly located in Staten Island and south-eastern area along

the city boundary. In terms of population vulnerability (see Figure 4.2), the high-high clusters are mainly located in upper Manhattan and Bronx, while low-low clusters are mainly found in Staten Island and lower Manhattan. In terms of health risk exposure and proximity to hazard cluster map (see Figure 4.3 and Figure 4.4), the distribution of high-high clusters and low-low clusters are very similar. Manhattan and norther-western part of Queens and Brooklyn are dominated by high-high clusters, and the low-low cluster are mainly located in suburb area.

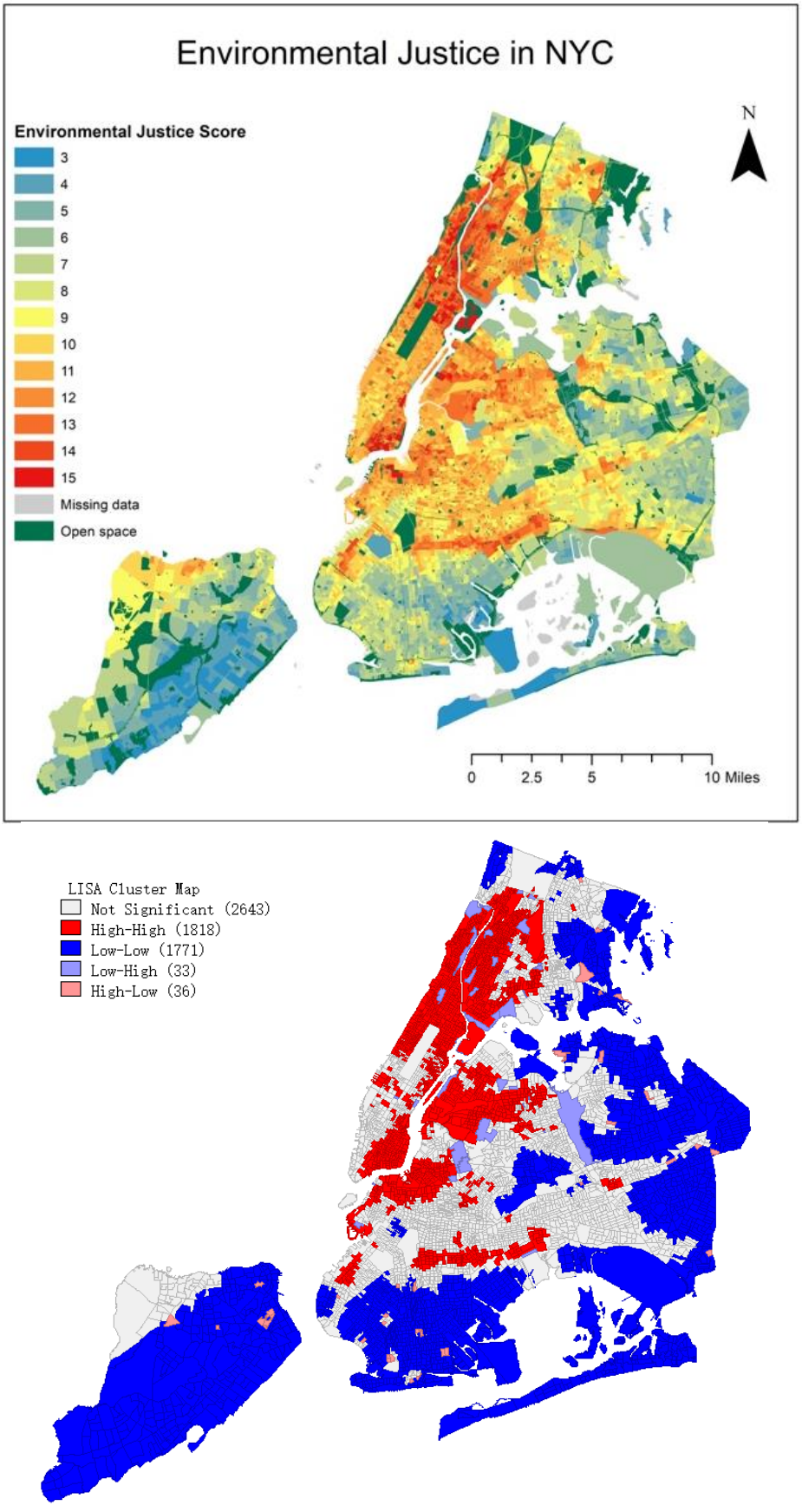


Figure 4.1. Spatial distribution map and cluster map of Environmental Justice by block group

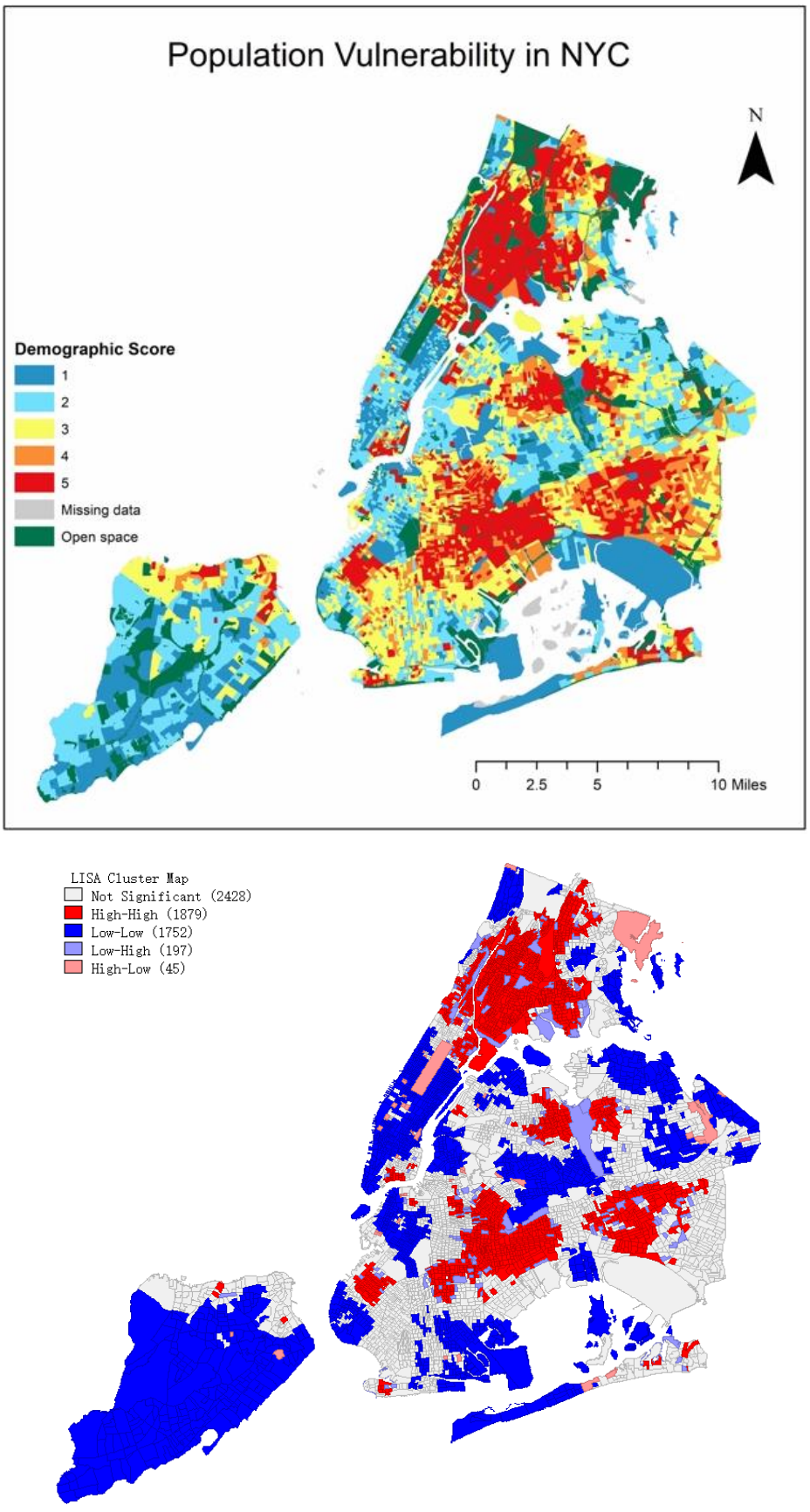
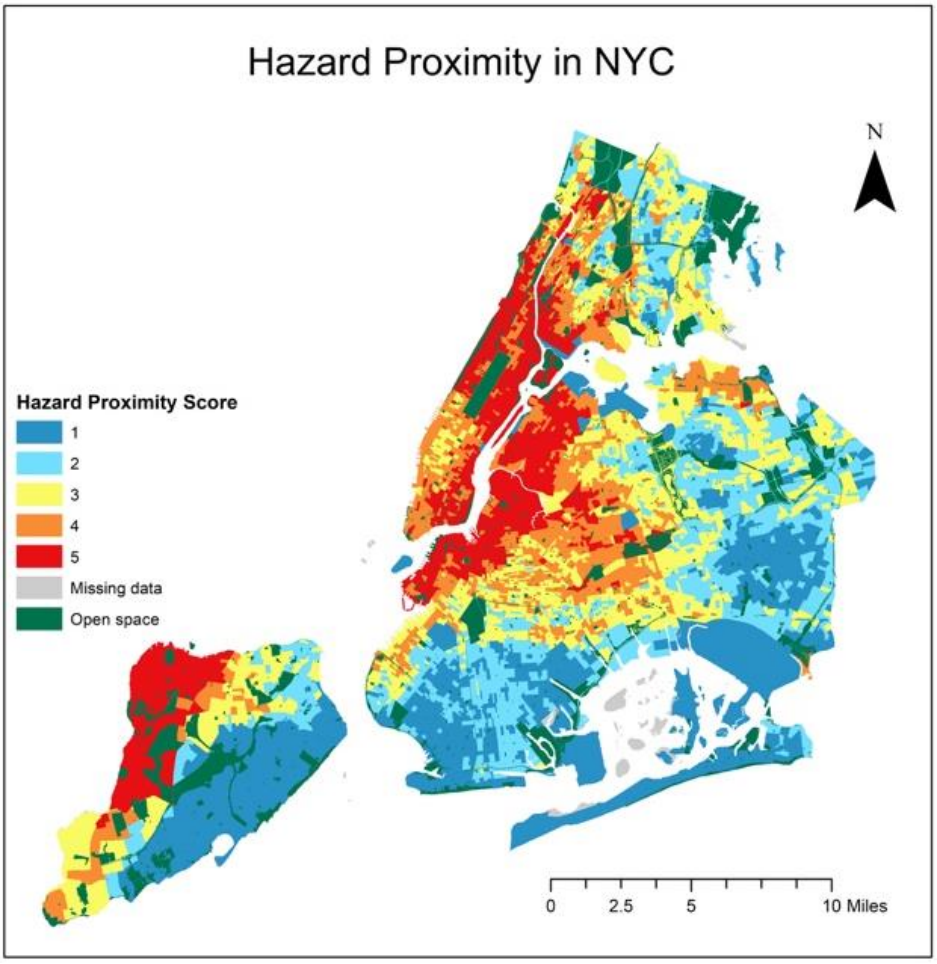


Figure 4.2. Spatial distribution map and cluster map of population vulnerability by block group



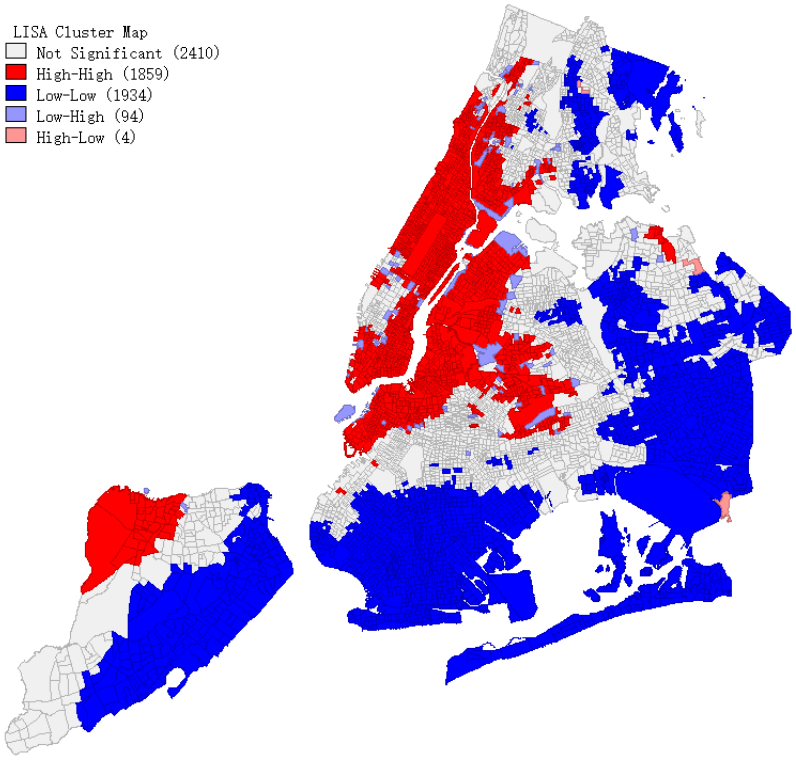
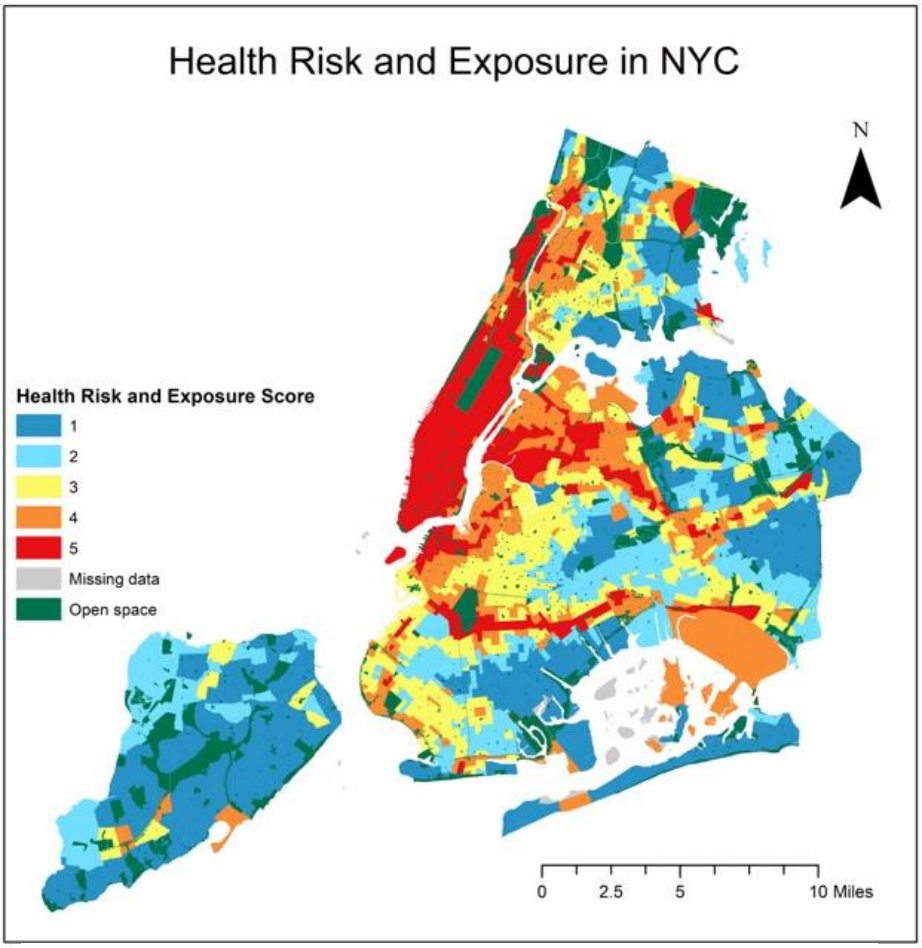


Figure 4.3. Spatial distribution map and cluster map of proximity to hazard by block group



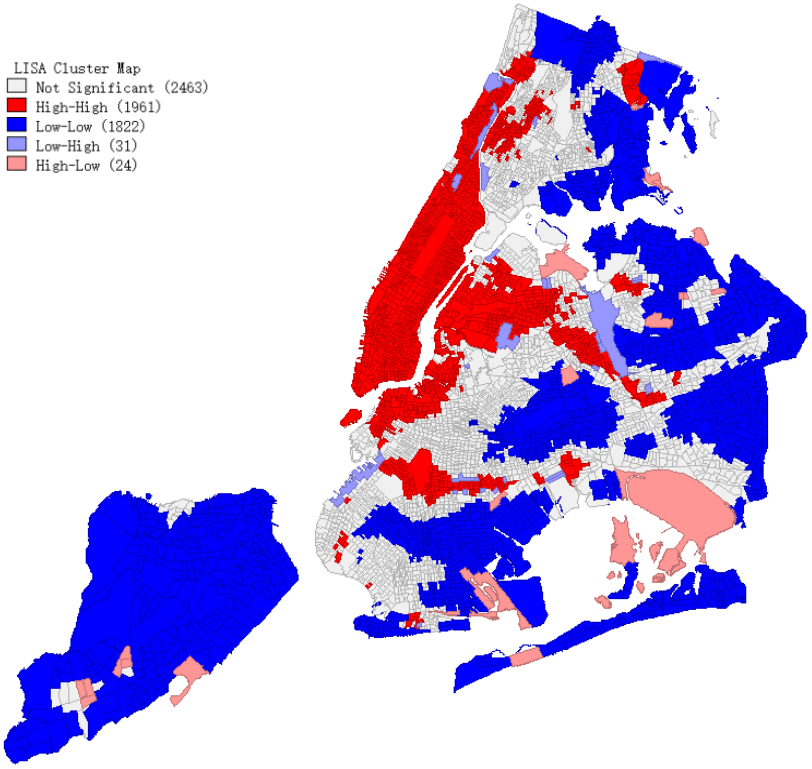
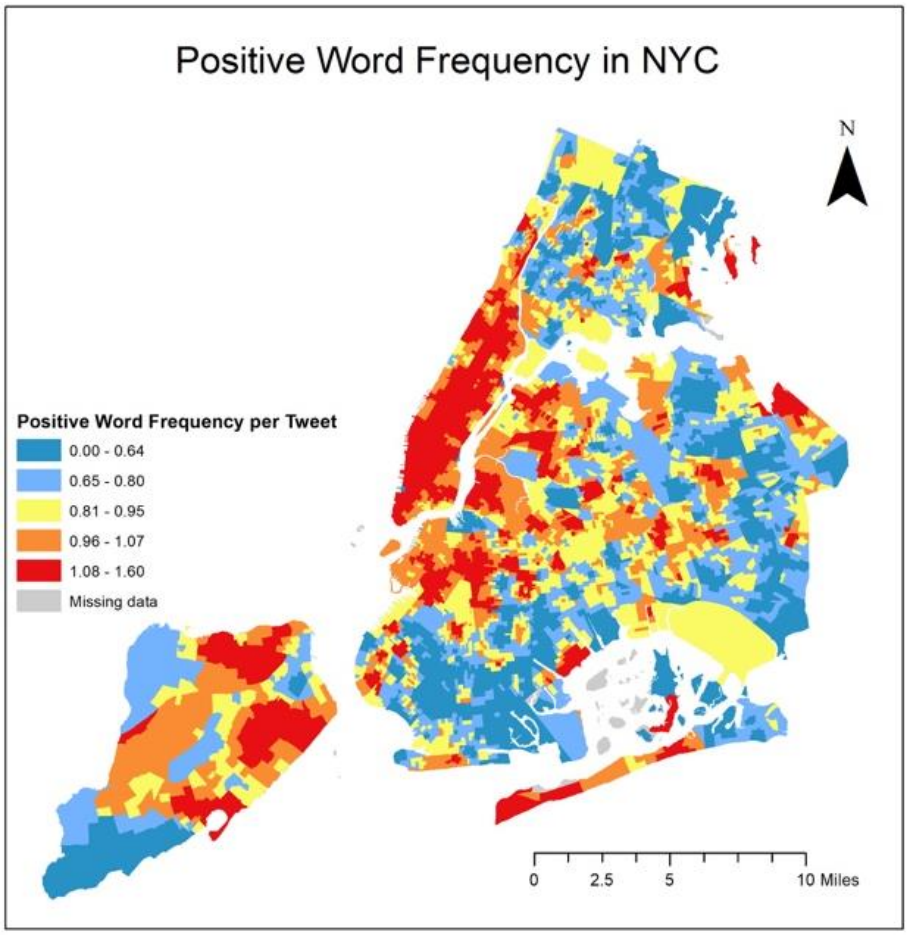


Figure 4.4. Spatial distribution map and cluster map of health risk exposure by block group

According to overall sentiment score cluster map (see Figure 4.9), large high-high clusters are mainly located in lower Manhattan and Staten Island, and some small high-high clusters are located in central areas of Queens and Brooklyn. The low-low clusters are mainly found in Staten Island, Bronx and Queens. The cluster maps of positive sentiment score and positive word frequency show very similar local spatial patterns (see Figure 4.5 and Figure 4.7). The location of high-high clusters and low-low clusters of positive word frequency coincide with those of positive sentiment score. This kind of spatial similarity is also found between cluster maps of negative word frequency and negative sentiment score, indicating sentiment intensity and sentiment volume have a similar tendency (see Figure 4.6 and Figure 4.8). People in stronger emotion will be more likely to produce more sentiment words in a tweet. Positive sentiment variables have stronger spatial correlation than negative sentiment variables based on Local Moran's I. I examined the local cluster and outliers of each sentiment variable and found negative sentiment clusters have a different local spatial pattern compared with positive sentiment clusters. The high-high clusters and low-low cluster of positive word frequency and positive sentiment scores are large and are located in some specific areas, while the size and distribution of negative words frequency clusters and negative sentiment score clusters are small and more scattered. It's hard to identify the district with a higher rate of negative sentiment clusters.



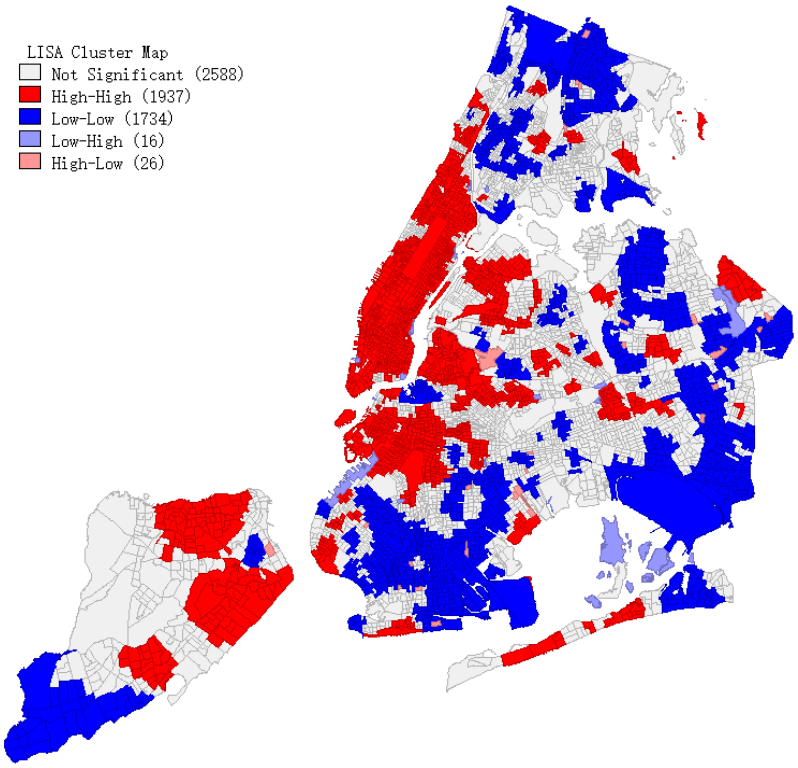


Figure 4.5. Spatial distribution map and cluster map of positive word frequency by block group

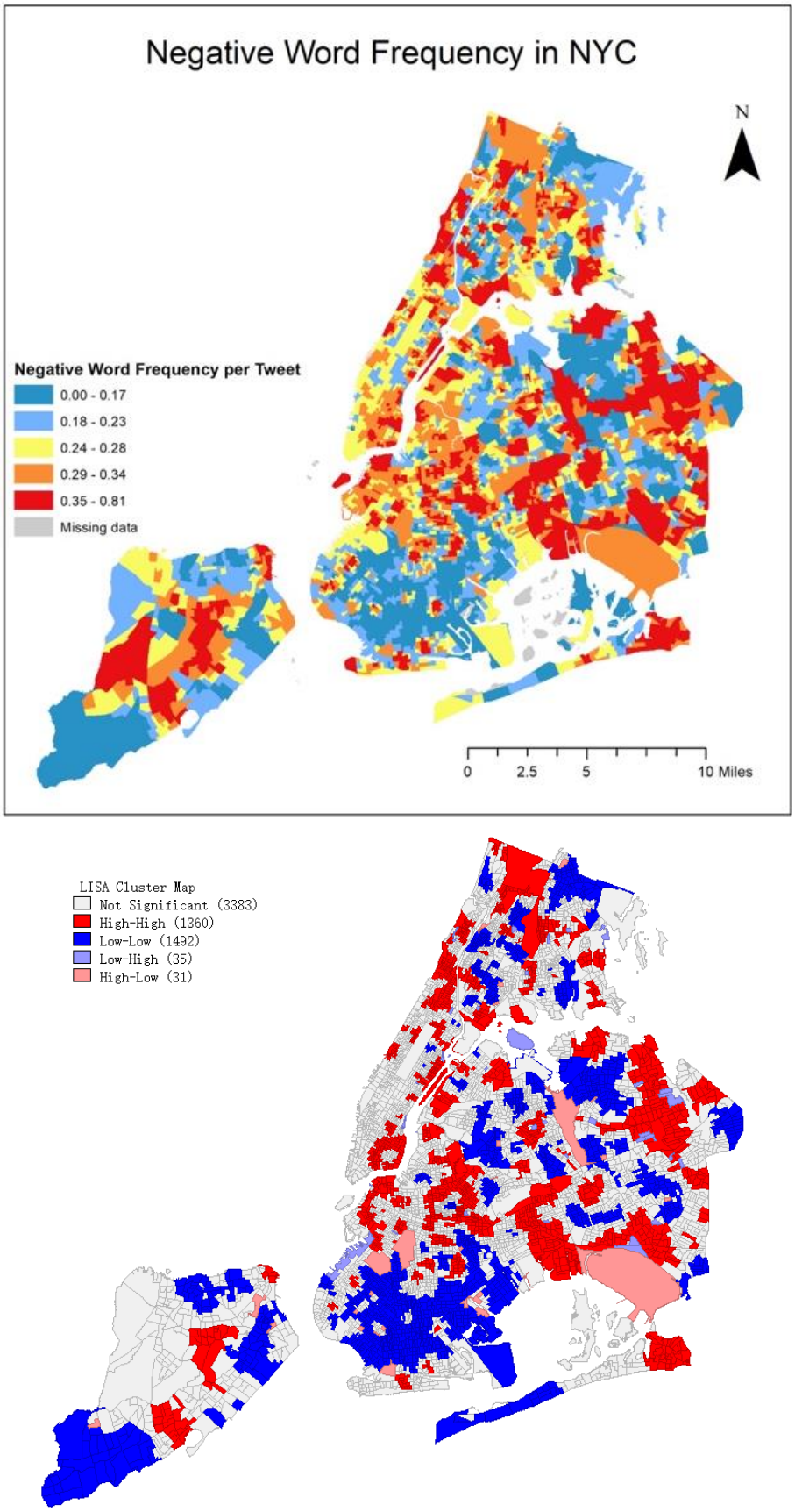


Figure 4.6. Spatial distribution map and cluster map of negative word frequency by block group

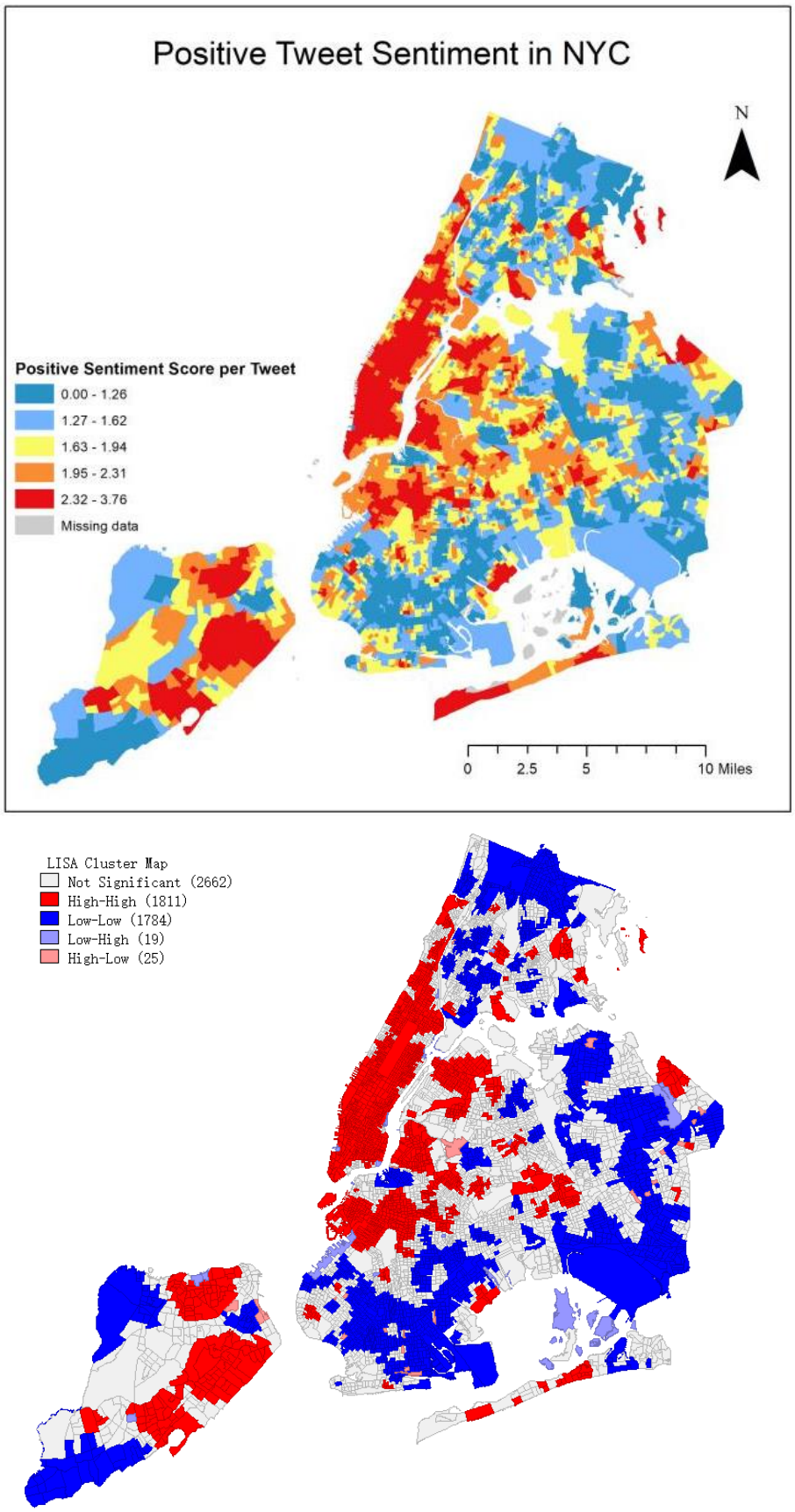


Figure 4.7. Spatial distribution map and cluster map of positive sentiment score by block group

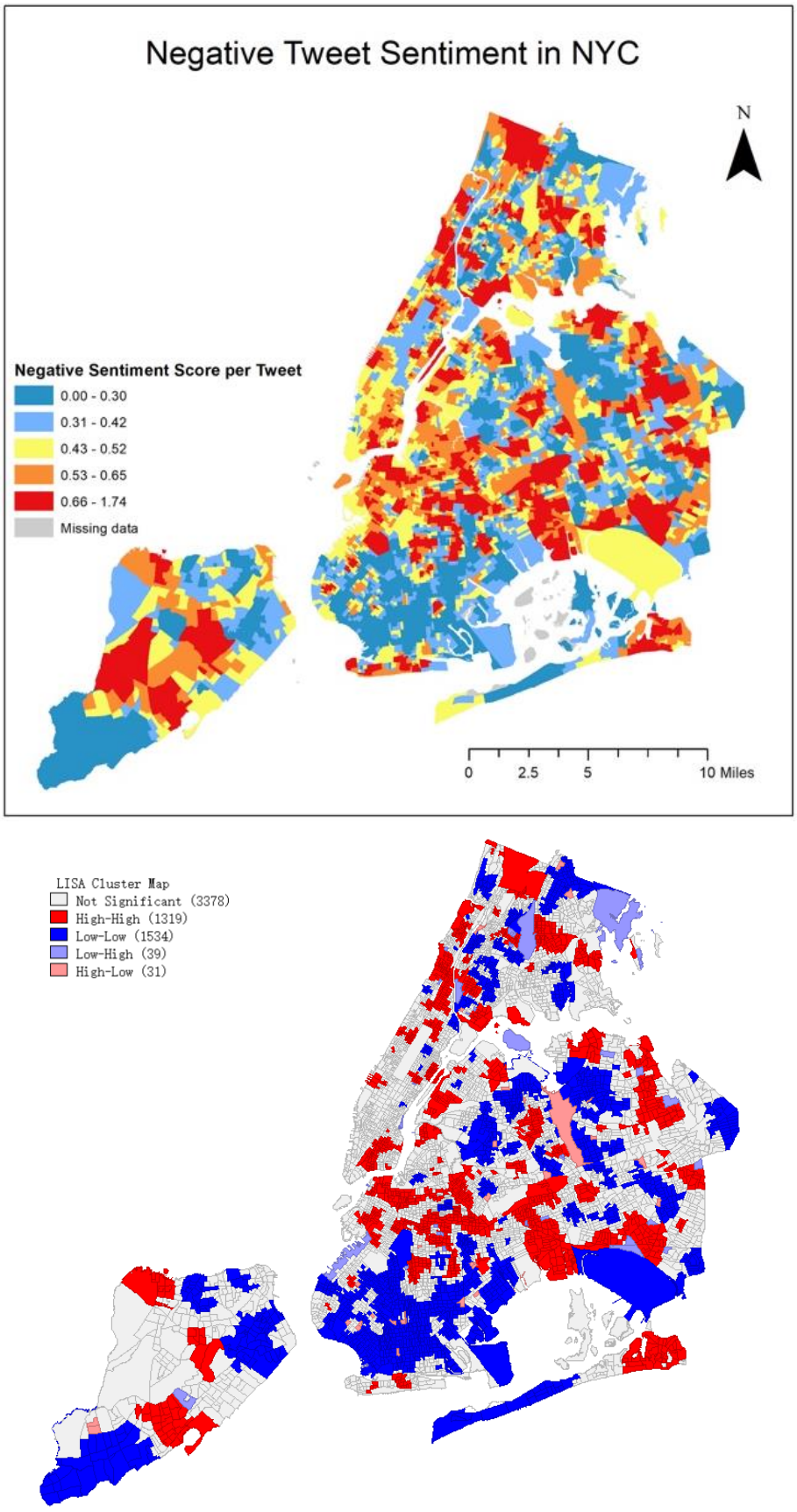


Figure 4.8. Spatial distribution map and cluster map of negative tweet sentiment by block group

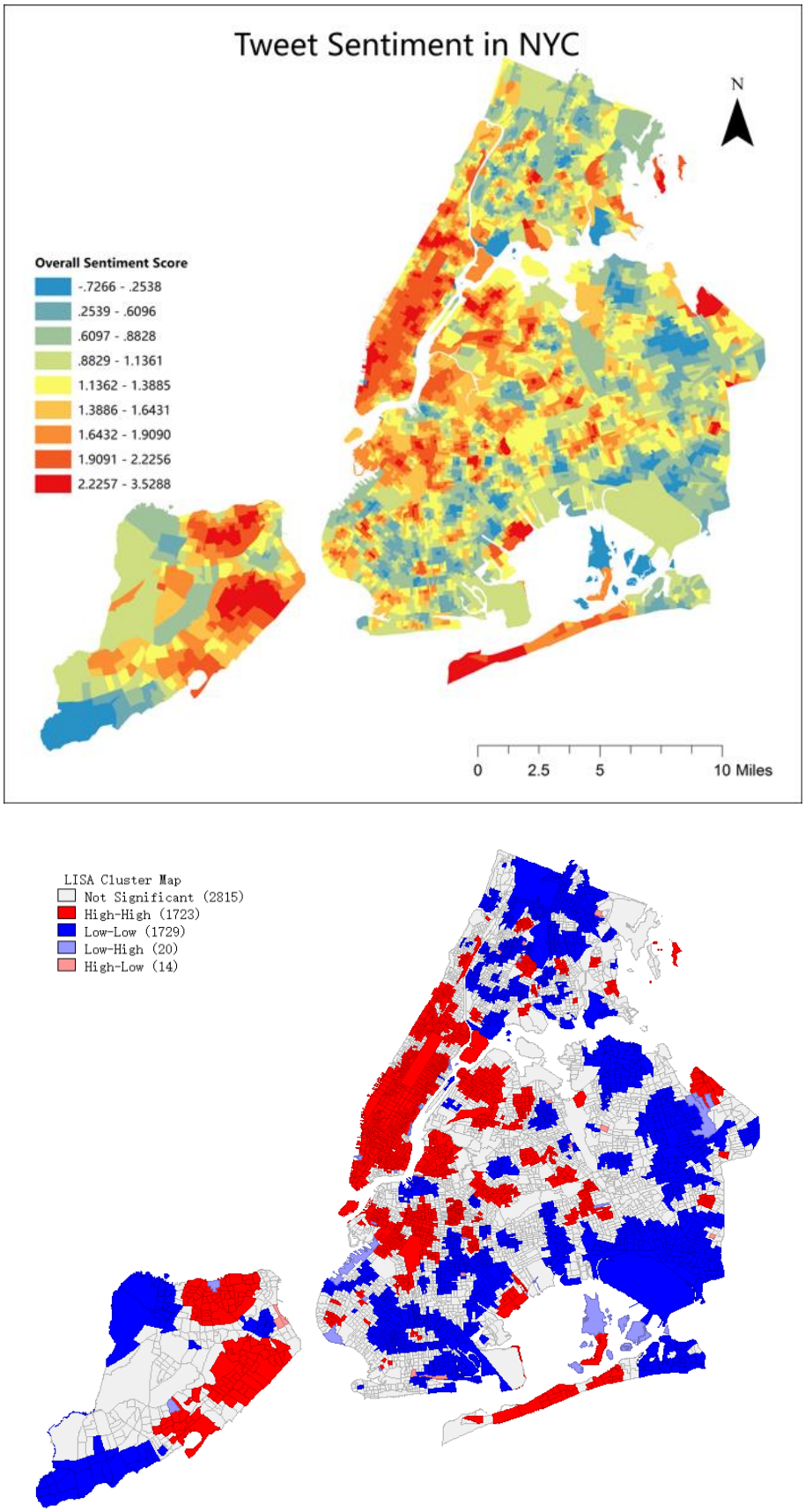


Figure 4.9. Spatial distribution map and cluster map of overall sentiment score by block group

In correlation analysis between EJ and tweet sentiment, I mentioned an interesting phenomenon that the overall EJ score is positively correlated with positive word frequency, positive sentiment score and overall sentiment score. I decided to use Bivariate Moran's I to further identify the block groups that have the positive correlations. Figure 4.10, Figure 4.11 and Figure 4.12 are the Bivariate cluster maps of EJ with positive word frequency, positive sentiment score and overall sentiment score. High-high cluster refers to clustering of a high EJ score surrounded by clustering of high positive word frequency, positive sentiment score and overall sentiment score, whereas low-low cluster refers to the clustering of a low EJ score surrounded by the clustering of low positive word frequency, positive sentiment score and overall sentiment score.

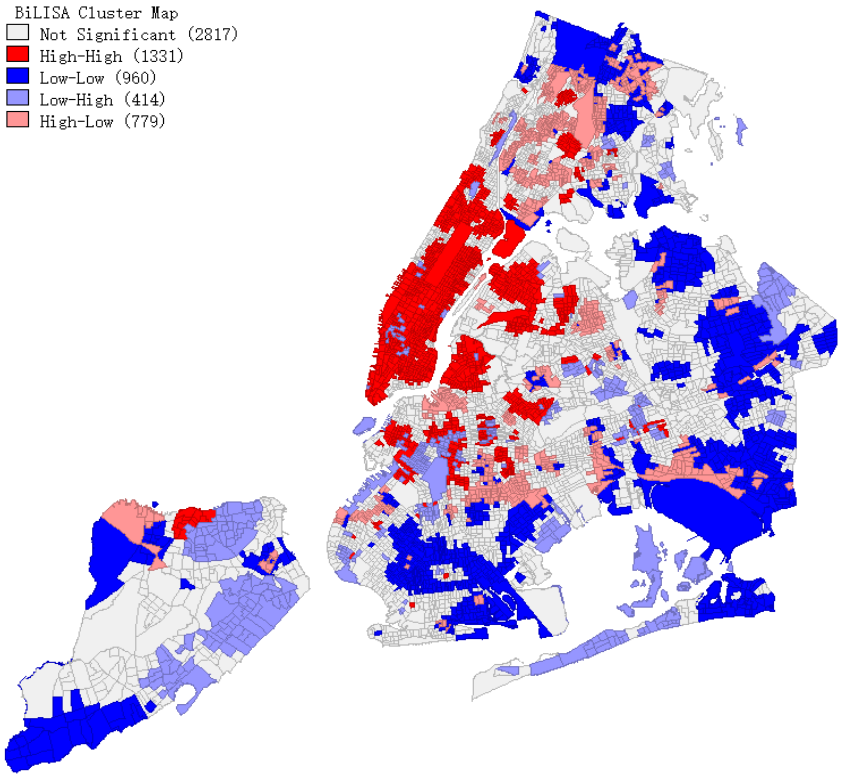


Figure 4.10. Map of bivariate association between EJ and positive word frequency by block group

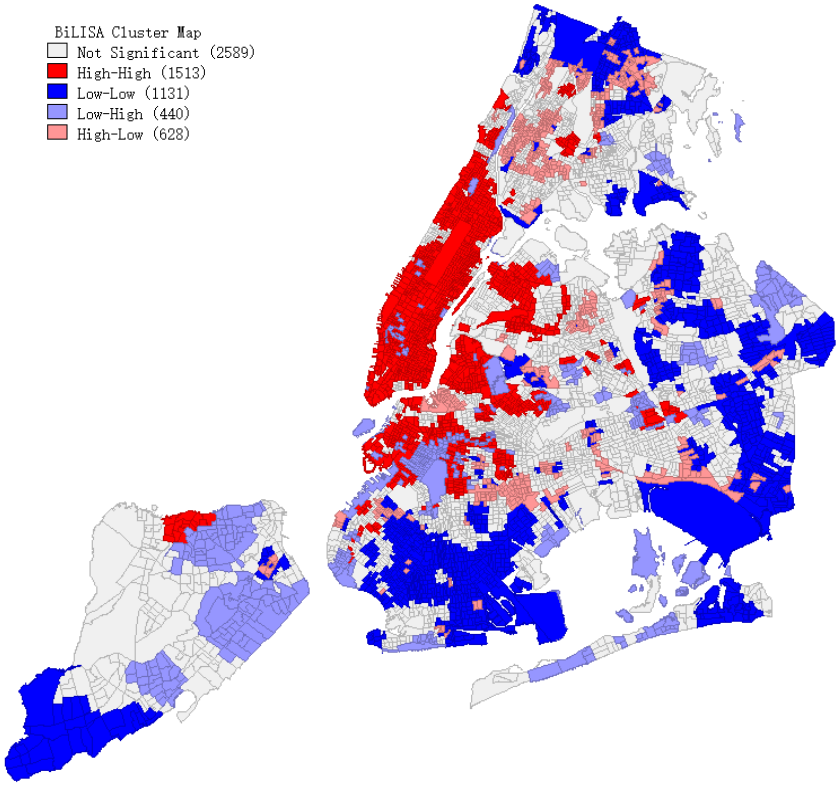


Figure 4.11. Map of bivariate association between EJ and positive sentiment score by block group

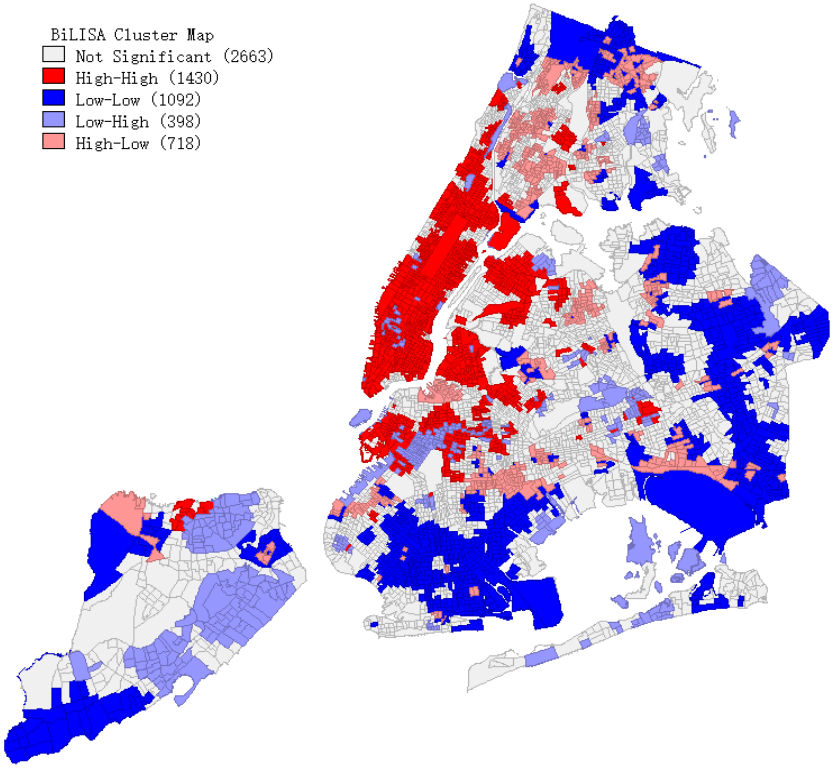


Figure 4.12. Map of bivariate association between EJ and overall sentiment score by block group

The three Bivariate cluster maps have very similar spatial patterns. The high-high clusters are mainly located in downtown and midtown Manhattan, while the low-low clusters are mainly located along the northeast and southern city boundaries. By comparing the Bivariate cluster maps to the community board map and Google street map of NYC, the landmarks and features in high-high clusters are determined. The polarity and intensity of tweet sentiment are affected by various factors including stimulating, physiological and cognitive factors (Helliwell & Wang, 2015). Environmental Justice indicators are identified as stimulating factors which may induce negative sentiment, but other factors also affect people's emotions. Downtown and midtown Manhattan have many tourist destinations, leading a large number of positive sentiment tweets produced by temporary visitors. They enjoyed the beautiful views and visited tourist destinations such as Times Square, Museum of Modern Art and the Empire State Building, and then posted positive sentiment tweets to share their experience. Central Park, as the most significant green space in NYC, is also associated with positive sentiment according to previous studies (Garvin& Brands, 2011; Chiesura, 2004).

4.3 Spatial Regression Analysis

According to spatial autocorrelation analysis, the volume and intensity of tweet sentiment show a similar spatial tendency. Thus, positive sentiment score, negative sentiment score and overall sentiment score are selected as dependent variables in order to avoid repeating analysis. The independent variables are proximity to hazard, health risk exposure and population vulnerability. Firstly, I assessed the influence of three EJ variables on the volume and intensity of tweet sentiment respectively using ordinary least square (OLS) regression.

The summary of OLS models are provided in Appendix A. Table 4.5 shows the diagnostics of OLS models.

Variable	Positive sentiment		Negative sentiment		Overall sentiment	
	Value	Probability	Value	Probability	Value	Probability
MCN	8.52		8.52		8.52	
Jarque-Bera	125.88	0.00	1549.75	0.00	39.32	0.00
Breusch-Pagan	149.97	0.00	103.70	0.00	42.12	0.00
Koenker-Bassett	115.89	0.00	54.75	0.00	35.56	0.00
LM-Lag	13778.37	0.00	13434.41	0.00	13406.91	0.00
R-LM-Lag	265.97	0.00	68.70	0.00	231.42	0.00
LM-Error	14231.65	0.00	13578.20	0.00	13645.59	0.00
R-LM-Error	719.26	0.00	212.49	0.00	470.10	0.00

Table 4.5. Summary of diagnostics for OLS models at the census block group level

The Multicollinearity Condition Number (MCN) is smaller than 30, indicating the EJ variables are not correlated with each other. The Jarque-Bera is statistically significant ($p<0.01$), which means the residuals are not normally distributed. The model is biased to predict tweet sentiment due to the missing of some key variables. The results of Breusch-Bera test and Koenker-Pagan test are both statistically significant ($p<0.01$), indicating the relationships between EJ variables and tweet sentiment variables are non-stationary. Thus, the spatial regression models (SLM and SEM) that control the spatial dependency are used to predict tweet sentiment.

According to the Lagrange Multiplier test and Robust LM diagnostic, LM-error and LM-lag are statistically significant ($p<0.01$), as well as the Robust LM-error and Robust LM-lag ($p<0.01$). Then both of the spatial error model (SEM) and spatial lag model (SLM) are conducted. The results of OLS, SLM and SEM analysis of the association between the tweet sentiment and the neighborhood EJ characteristics are given in the Appendix. By comparing the values of R^2, log-likelihood, Akaike information criterion (AIC), and Schwarz criterion (SC), SLM is most suitable for positive sentiment score and overall sentiment score, while SEM performs better interpretation of negative sentiment score. A lower AIC value and SC value and higher log likelihood value means the model is a better fit for the data. Therefore, SLM is adopted as the final model for positive sentiment score and overall sentiment score, and SEM is chosen for negative sentiment score.

Table 4.6 displays the summary of spatial regression model for tweet sentiment variables. In general, three independent EJ variables are all statistically significant to the positive, negative and overall sentiment scores. With respect to overall sentiment score and positive sentiment score, R^2 values of SLM are both 0.27, indicating that the EJ variables explain of 27% of the variance in the data. For the negative sentiment score, R^2 value of SEM is lower at 0.12. The potential reason for the low fit might be that the overall polarity of tweet sentiment is positive. Although the SLM and SEM have better performance than the original OLS model, the heteroskedasticity test and spatial dependence test remain significant ($p<0.01$). The spatial effects of the EJ variables are not eliminated in the spatial regression models.

Model	**SLM**				**SEM**	
Variable	**Positive sentiment**		**Overall sentiment**		**Negative sentiment**	
	Coefficient	Probability	Coefficient	Probability	Coefficient	Probability
Constant	0.48	0.00	0.51	0.00	0.32	0.00
Population vulnerability	-0.04	0.00	-0.05	0.00	0.01	0.02
Proximity to hazard	0.04	0.00	0.03	0.00	0.03	0.00
Health risk exposure	0.02	0.00	0.02	0.00	0.01	0.00
Log likelihood	-3858.91		-4352.21		1041.10	
R^2	0.27		0.27		0.12	
AIC	7727.82		8714.41		-2074.20	
SC	7761.56		8748.15		-2047.21	
N	6302		6302		6302	

Table 4.6. Summary of spatial regression models for sentiment variables at the census block group level

When the population vulnerability increases by 1 unit, the score of negative sentiment increases by 0.01, holding other variables constant. On the other hand, when population vulnerability increases by 1 unit, positive sentiment and overall sentiment decrease by 0.04 and 0.05, holding the other variables constant. Population vulnerability in this thesis is evaluated by race (% residents of color) and poverty (% resident living below twice national poverty level). The result is consistent with the study of well-being measurement based on tweet sentiment (Jaidka et al.,2020), finding that communities of color and low-income families are negatively associated with positive tweet sentiment. When proximity to hazard

increase by 1 unit, the scores of positive sentiment, negative sentiment and overall sentiment increase by 0.04, 0.03 and 0.03 on average respectively, holding the other variables constant. When health risk exposure increases by 1 unit, the scores of positive sentiment, negative sentiment and overall sentiment increase by 0.02, 0.01 and 0.02 on average respectively, holding the other variables constant. The coefficient values of proximity to hazard are higher than those of health risk exposure in all regression models, indicating it has a stronger influence on tweet sentiment than health risk exposure. The result suggests that it is important in predicting sentiment compared to the other factors.

Chapter 5: Conclusion, Limitation, and Recommendation

In this chapter, I will discuss the results of this study by answering my research questions. I will also come up with some limitations from both data and methodology aspects, and then put forward some recommendations for future relevant studies.

5.1 Conclusion

This thesis demonstrated the spatial patterns of Environmental Justice and tweet sentiment, and then investigated the association between them.

The first research question asked, "How is Environmental Justice spatially distributed in NYC?" To address the question, I adopted the EJSCREEN indicators and EJSM scoring methodology to quantify EJ. EJSCREEN indicators are grouped into three variables including population vulnerability, proximity to hazard and health risk exposure. Each variable was given a score from 1 to 5 based on quintile ranking, and the overall EJ score is the sum of scores for the three categories, ranging from 3 to 15. In order to identify the clustering area of EJ, I calculated Global and Local Moran's I. The cluster map shows that high-level EJ block groups are mainly clustered in Manhattan and the norther-western part of Queens and Brooklyn, while low-level EJ block groups are mainly clustered in suburban area.

The second research question asked, "How is tweet sentiment spatially distributed in NYC?" In order to answer this question, I performed lexicon-based sentiment analysis using AFINN dictionary in Python. Five sentiment variables including positive word frequency, negative word frequency, positive sentiment score, negative sentiment score and overall sentiment score are summarized. Global and Local Moran's I are calculated and indicated strong local spatial correlation. The spatial pattern of positive word frequency, positive

sentiment score and overall sentiment score are very similar. The high value clusters are mainly located in midtown and downtown Manhattan. The spatial pattern of negative word frequency and negative word score are similar too. The high value clusters of negative sentiment are more scattering than that of positive sentiment.

The third research question asked, "Is there a relationship between EJ and tweet sentiment and how does EJ affect people's sentiment?" I used both statistical and spatial correlation analysis to answer this question, including Spearman correlation, multiple regression, and spatial regression model such as SLM and SEM. According to the results of the above analysis, the overall EJ is significantly correlated with tweet sentiment. Among three EJ variables, proximity to hazard has the strongest correlation with sentiment intensity and volume. Population vulnerability is negatively associated with positive sentiment score, positive word frequency and overall sentiment score, while positively associated with negative sentiment score and negative word frequency. The result is consistent with a previous study that happy tweets were less frequent in area with higher poverty levels and higher proportions of racial minorities (Nguyen et al., 2016). Mitchell et al. (2013) also found that higher socioeconomic status has a positive relationship with higher sentiment scores. Proximity to hazard as well as health risk exposure have positive relationships with all of the five sentiment variables in this study. The result is not entirely consistent with previous studies, some of which found environmental hazard and health risk exposure will decrease subject well-being (HHS, 2012; CDC, 2011; Galea et al., 2005). However, it is also argued that the contradictory results are not surprising since younger and more educated people are

more likely to post tweets, thus the results underestimate the negative sentiment from elder and less-educated people who are more vulnerable to hazard (Zheng, 2018).

Given the above conclusions, some suggestions can be put forward for residents, city planners and political interest groups to seize the framework of distributive justice and procedural justice to address the Environmental Justice problems and promote relevant public policy (Pearsall & Pierce, 2010).

Distributive justice refers to how environmental burdens are distributed across the city. The GIS technique helps to identify the spatial patterns of demographic characteristic and environmental hazard and determine areas that suffers from disproportionate cumulative environmental burden. The development of an EJ screening tool is a key step in this process. In California, strong effort has been made to encourage CalEPA to develop CalEnvironScreen, a state level EJ screening tool which updates its indicators and methodology continuously. I think other states could learn from California to develop their own screening tool with more specific indicators taken into EJ evaluation.

Procedural justice focuses on the right of all people to participate in decision-making process related to Environmental Justice. Based on the result of this study, people's sentiment is more sensitive to population vulnerability than proximity to hazard and health risk exposure. It should be noted that city planners are not aiming to change the demographic characteristics, but to promote Environmental Justice through community participation. Minkler (2008) constructed a dynamic model with ten dimensions of community and partnership capacity relevant to Environmental Justice, which could serve as a theoretical

framework for legislators. The ten dimensions include leadership, participation, skills, resources, social and organization networks, sense of community and partnership identity, understanding of community and partnership history, community and partnership power, and shared values and critical reflection. In brief, it aims to educate communities in high environmental burden areas about the knowledge of Environmental Justice and encourages them to participate in the development and implementation of Environmental Justice policies and activities to build healthy communities and reduce health disparities related to environmental factors.

5.2 Limitation and Recommendation

This thesis has several important limitations in data collection. Firstly, I'm not able to capture all the environmental and socioeconomic variables or risk factors which may be related to Environmental Justice. The EJSCREEN indicators, selected by EPA, were used for assessing patterns of cumulative impacts from environmental and social stressors in this thesis. Other variables such as sensitive land use type and climate change are also taken into consideration in some recent EJ studies. Secondly, the tweet data analyzed in this thesis only covered three months of a single year in 2017 (from April to July). Therefore, the characteristics of tweet sentiment are not comprehensive enough to perform stable spatiotemporal analysis. For instance, if we have more tweet data that covers different seasons throughout a year, we can compare the differences of sentiment spatial pattern between winter and summer. NYC tends to have fewer tourists during winter due to the cold weather, which helps to filter out some sentiment-related tweets produced by temporary city

visitors. Environmental Justice, unlike other environmental factors such as air pollutants could be felt and described immediately, it's a more tangible cumulative environmental burden that influences our life unconsciously. Thus, the sentiment-related tweets from the local residents would be more significant when analyzing the association between tweet sentiment and Environmental Justice. Given the above limitations of data, I would recommend that future relevant studies acquire tweets covering at least one whole year and include the mobility patterns of Twitter users into their analysis. In this way, the differences of sentiment scores between residents and city commuters could be determined, which would also help discover other factors that affect tweet sentiment.

This thesis also has some limitations in methodology. For the sentiment analysis section, I did not filter the data by setting some "key words" related to Environmental Justice. As I mentioned above, Environmental Justice includes a wide range of factors related to cumulative environmental burden and health disparities. Identifying the key words for Environmental Justice will be too complicated in this study. I have to keep the original data although this might increase some variance and noise in the result. Also, the lexicon-based sentiment analysis method is not as accurate as machine learning method. Sentiment words are assigned a single score using the sentiment dictionary regardless of the topic and context. But sometimes the same sentiment word could have opposite sentiment polarity. Take AFINN dictionary as an example, a tweet saying "Boston is a breathtakingly beautiful city" and another tweet saying "The movie is breathtakingly scary" appear in different contexts. The polarity of the sentiment word "breathtakingly" appears to be positive for the first tweet

and negative for the second tweet. But in a lexicon-based method, they will be both given a positive score. Based on the above limitations in methodology, it should be noted that future urban and environmental studies will not only require the application of traditional tools such as GIS and Geoda, but also use programming software such as Python and R, especially when dealing with large data sets.

Appendix A: Summary of Spatial Regression Models

Overall sentiment Score						
Variable	OLS		SLM		SEM	
	Coefficient	Probability	Coefficient	Probability	Coefficient	Probability
Constant	1.14	0.00	0.51	0.00	1.41	0.00
Population vulnerability	-0.10	0.00	-0.05	0.00	-0.03	0.00
Proximity to hazard	0.08	0.00	0.03	0.00	0.07	0.38
Health risk exposure	0.06	0.00	0.02	0.00	-0.01	0.54
Log likelihood	-4720.20		-4352.21		-4380.10	
R^2	0.17		0.27		0.26	
AIC	9448.39		8714.41		8768.21	
SC	9475.39		8748.15		8795.20	

Table 1. Summary of spatial regression models for overall sentiment score

Positive sentiment Score						
Variable	OLS		SLM		SEM	
	Coefficient	Probability	Coefficient	Probability	Coefficient	Probability
Constant	1.46	0.00	0.48	0.00	1.80	0.00
Population vulnerability	-0.09	0.00	-0.04	0.00	-0.02	0.00
Proximity to hazard	0.12	0.00	0.04	0.00	0.02	0.02
Health risk exposure	0.07	0.00	0.02	0.00	0.00	0.90
Log likelihood	-4476.59		-3858.91		-4380.10	

R^2	0.23	0.27	0.26
AIC	8961.17	7727.82	8768.21
SC	8988.17	7761.56	8795.20

Table 2. Summary of spatial regression models for positive sentiment score

Negative sentiment Score						
Variable	OLS		SLM		SEM	
	Coefficient	Probability	Coefficient	Probability	Coefficient	Probability
Constant	0.32	0.00	0.16	0.00	0.32	0.00
Population vulnerability	0.01	0.00	0.01	0.01	0.01	0.02
Proximity to hazard	0.04	0.00	0.03	0.00	0.03	0.00
Health risk exposure	0.01	0.04	0.01	0.03	0.01	0.00
Log likelihood	886.297		1025.01		1041.10	
R^2	0.07		0.12		0.12	
AIC	-1764.59		-2040.02		-2074.20	
SC	-1737.60		-2006.27		-2047.21	

Table 3. Summary of spatial regression models for negative sentiment score

Appendix B: Python Code for Tweet Sentiment Analysis

```
import pandas as pd
import numpy as np
import sys
import os
import nltk

file="D:\\0pp\\A1010104\\"
file1=file+"NewYork\\"
file2=file+"result\\"
emotions= pd.read_csv(file+"afinn.csv")
#%%%
neg=emotions.loc[emotions['value']<0,"word"]
pos=emotions.loc[emotions['value']>0,"word"]

files= os.listdir(file1)
n=0
for i in files:
    positive_words=[]
    negative_words=[]
    positive_score=[]
    negative_score=[]
    trainDF=0
    n=n+1
    print(i)

    print(n)
    trainDF= pd.read_csv(file1+i,header=None)
    trainDF['total_content'] = trainDF.iloc[:,2].str.lower()
    lines=trainDF.total_content.values.tolist()
    nn=1
    for line in lines:
        nn=nn+1
        if nn%5000==0:
            print(nn)
        words = nltk.word_tokenize(text=str(line))
        neg_list=list(set(words).intersection(set(neg)))
        pos_list=list(set(words).intersection(set(pos)))
        neg_words=[i for i in words if i in neg_list]
        pos_words=[i for i in words if i in pos_list]
        positive_words.append(len(pos_words))
```

```
            negative_words.append(len(neg_words))
            s=0
            for w1 in pos_words:
                temp=emotions.loc[emotions['word']==w1,:].value.tolist()
                s=s+temp[0]
            positive_score.append(s)

            s=0
            for w1 in neg_words:
                temp=emotions.loc[emotions['word']==w1,:].value.tolist()
                s=s+temp[0]
            negative_score.append(s)

        trainDF['positive_words']=positive_words
        trainDF['negative_words']=negative_words
        trainDF['positive_score']=positive_score
        trainDF['negative_score']=negative_score
        trainDF['overall_score']=trainDF['positive_score']+trainDF['negative_score']
        print("----------")
        trainDF=trainDF.drop(['total_content'], axis=1)
        print(trainDF.columns)

        trainDF.to_excel(file2+i.split('.')[0]+".xlsx",index=False,encoding="utf_8_sig")

files= os.listdir(file2)
n=0
for i in files:
    trainDF=0
    n=n+1
    print(i)
    print(n)
    trainDF= pd.read_excel(file2+i)
    print(trainDF.shape)
    if n==1:
        df=trainDF.loc[(trainDF['positive_words']>0)|
                (trainDF['negative_words']>0)
                ,:]
    else:
        temp=trainDF.loc[(trainDF['positive_words']>0)|
                (trainDF['negative_words']>0)
                ,:]
        df=pd.concat([df,temp])
```

```
    print(df.shape)
#%%
print(df['negative_score'].value_counts)
#%%
df=df.drop(['total_content'], axis=1)
df.to_excel(file+"result_ALL.xlsx",index=False,encoding="utf_8_sig")
```

References

Aubree Driver, Crystal Mehdizadeh, Samuel Bara-Garcia, Coline Bodenreider, Jessica Lewis, & Sacoby Wilson. (2019). Utilization of the Maryland Environmental Justice Screening Tool: A Bladensburg, Maryland Case Study. International Journal of Environmental Research and Public Health, 16(3), 348.

Bertrand, K., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in New York City: A High Resolution Spatial and Temporal View.

Bao, J., Liu, P., Yu, H., & Xu, C. (2017). Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. *Accident Analysis & Prevention*, *106*, 358-369.

Bonorris, S. (2010) Environmental justice for all: A fifty state survey of legislation, policies and cases (4th ed.). San Francisco: University of California Hastings School of Law. Retrieved February 7, 2012, Retrieved from p0http://www.uchastings.edu/centers/public-law/environmental-justice.html.

Burger J., Gochfeld M. Conceptual environmental justice model for evaluating chemical pathways of exposure in low-income, minority, native American, and other unique exposure populations. Am. J. Public Health. 2011 doi: 10.2105/AJPH.2010.300077.

Brainard J.S., Jones A.P. Modelling environmental equity: Access to air quality in Birmingham. Environ. Plan. A. 2002; 34:695–716.

Centers for Disease Control and Prevention (CDC). Health Disparities and Inequalities Report – United States, 2011. MMWR Supplement 60(2011):1-113

Chiesura, A. (2004). The role of urban parks for the sustainable city. Landscape Urban Plan, 68, 129-138. http://dx.doi.org/10.1016/j.landurbplan.2003.08.003

Cutter, S.L. 1995. Race, class and environmental justice. Progress in Human Geography 19(1):111-122

Cutter, S. L., J.T. Mitchell and M.S. Scott. (2000) Revealing the Vulnerability of People and Places: A Case Study of Georgetown County, South Carolina. Annals of the Association of American Geographers, 90(4), p. 713-737.

Charles Camille Z. The Dynamics of Racial Residential Segregation. American Review of Sociology. 2003; 29:167–207.

Christopher G. Boone (2013) An Assessment and Explanation of Environmental Inequity in Baltimore, Urban Geography, 23:6, 581-595, DOI: 10.2747/0272-3638.23.6.581

D. Azar, D. Rain. Identifying population vulnerable to hydrological hazards in San Juan, Puerto Rico. GeoJournal, 69 (1–2) (2007), pp. 23-43

Dhaoui, C., Webster, C., & Tan, L. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing*, *34*(6), 480-488.

Downey, L., & Hawkins, B. (2008). Race, Income, and Environmental Inequities in the United States. *Sociological Perspectives : SP : Official Publication of the Pacific Sociological Association*, *51*(4), 759–781. Retrieved from http://doi.org/10.1525/sop.2008.51.4.759

Dhaoui, C., Webster, C., & Tan, L. (2017). Social media sentiment analysis: lexicon versus machine learning. Journal of Consumer Marketing, 34(6), 480-488.

Das, Dutta, Medina, Minjares-Kyle, & Elgart. (2019). Extracting patterns from Twitter to promote biking. IATSS Research, 43(1), 51-59.

Exec. Order No.12898. (1994) Federal Actions to Address Environmental Justice in Minority Populations and Low-Income Populations, 59 Fed. Reg. 7629.

Gan, Q., Ferns, B., Yu, Y., & Jin, L. (2017). A Text Mining and Multidimensional Sentiment Analysis of Online Restaurant Reviews. Journal of Quality Assurance in Hospitality & Tourism, 18(4), 465-492.

Gee G.C., Payne-Sturges D.C. Environmental health disparities: A framework integrating psychosocial and environmental concepts. Environ. Health Perspect. 2004;112:1645–1653. doi: 10.1289/ehp.7074.

Garvin, A., & Brands, R. (2011). *Public parks: The key to livable communities*. New York: W.W. Norton & Co.

Galea, S., et al. Urban Built Environment and Depression: A Multilevel Analysis. Journal of Epidemiology and Community Health 59(2005): 822-827.

Galster George, Godfrey Erin. Racial Steering by Real Estate Agents in the U.S. in 2000. Journal of the American Planning Association. 2005;71:251–68.

Glickman TS. Measuring environmental equity with geographical information systems. Renewable Resources J. 1994;12:17–21.

Helliwell JF, Wang S (2015) How Was the Weekend? How the Social Context Underlies Weekend Effects in Happiness and Other Emotions for US Workers. PLoS ONE 10(12): e0145123. https://doi.org/10.1371/journal.pone.0145123

Hollander J.B., Shen Y. (2017) Using Social Media Data to Infer Urban Attitudes About Bicycling: An Exploratory Case Study of Washington DC. In: Karakitsiou A., Migdalas A., Rassia S., Pardalos P. (eds) City Networks. Springer Optimization and Its Applications, vol 128. Springer, Cham. https://doi.org/10.1007/978-3-319-65338-9_5

Hollander, J., & Renski, H. (2017). Measuring urban attitudes embedded in microblogging data: Shrinking versus growing cities. The Town Planning Review, 88(4), 465-490.

Hollander, J.B., Potts, R., Hartt, M. *et al.* The Role of Artificial Intelligence in Community Planning. *Int. Journal of Com. WB* **3,** 507–521 (2020). https://doi.org/10.1007/s42413-020-00090-7

Jayajit Chakraborty, Juliana A. Maantay, Jean D. Brender, "Disproportionate Proximity to Environmental Health Hazards: Methods, Models, and Measurement", American Journal of Public Health 101, no. S1 (December 1, 2011): pp. S27-S36. DOI:10.2105/AJPH.2010.300109

Jayajit Chakraborty & Marc P. Armstrong (2013) 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, DOI: 10.1559/152304097782476951

James L. Sadd, Manuel Pastor, Rachel Morello-Frosch, Justin Scoggins, & Bill Jesdale. (2011). Playing It Safe: Assessing Cumulative Impact and Social Vulnerability through an Environmental Justice Screening Method in the South Coast Air Basin,

California. *International Journal of Environmental Research and Public Health*, 8(5), 1441-1459.

Jiang W, Wang Y, Tsou M-H, Fu X. (2015) Using Social Media to Detect Outdoor Air Pollution and Monitor Air Quality Index (AQI): A Geo-Targeted Spatiotemporal Analysis Framework with Sina Weibo (Chinese Twitter). PLoS ONE 10(10): e0141185.

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(19), 10165–10171. https://doi.org/10.1073/pnas.1906364117

J. Chakraborty, G.A. Tobin, B.E. Montz. Population evacuation: assessing spatial variability in geophysical risk and social vulnerability to natural hazards. Natural Hazards Review, 6 (1) (February 2005), pp. 23-3

Katherine S. Nelson, Mark D. Abkowitz, Janey V. Camp. 2015. A method for creating high resolution maps of social vulnerability in the context of environmental hazards, Applied Geography,Volume 63,2015,Pages 89-100,ISSN 0143-6228, Retrieved from https://doi.org/10.1016/j.apgeog.2015.06.011.

Kolchyna, O., Souza, T., Treleaven, P., & Aste, T. (2015). Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. Handbook of Sentiment Analysis in Finance. Mitra, G. and Yu, X. (Eds.). (2016). ISBN 1910571571.

Kuruppuarachchi, L. N., Kumar, A., & Franchetti, M. (2017). A comparison of major environmental justice screening and mapping tools. *Environ Manag Sustain Dev*, *6*(1), 59-71.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*(2010), 627-666.

López-Ornelas, E., Abascal-Mena, R., & Zepeda-Hernández, S. (2017). Social Media Participation in Urban Planning: A New Way to Interact and Take Decisions. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, *42*.

Maantay, J. and A. Maroko (2008). "Mapping urban risk: Flood hazards, race, & environmental justice in New York." Applied Geography (29): 14.

Mah, Alice. (2016) Environmental justice in the age of big data: challenging toxic blind spots of voice, speed, and expertise. Environmental Sociology. Doi:10.1080/23251042.2016.1220849

Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM. The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE.* 2013;8(5):e64417.

Minkler, M., Vásquez, V. B., Tajik, M., & Petersen, D. (2008). Promoting environmental justice through community-based participatory research: the role of community and partnership capacity. *Health education & behavior : the official publication of the Society for Public Health Education*, *35*(1), 119–137. Retrieved from https://doi.org/10.1177/1090198106287692

Meleah Geertsma. (2018, October 25). New Map Shows Chicago Needs Environmental Justice Reforms. Retrieved September 5, 2019, from https://www.nrdc.org/

Muhammad, Aminu ; Wiratunga, Nirmalie ; Lothian, Robert. (2016). Contextual sentiment analysis for social media genres. *Knowledge-Based Systems, 108*(C), 92-101.

Murphy-Greene, C., & Leip, L. (2002). Assessing the Effectiveness of Executive Order 12898: Environmental Justice for All? Public Administration Review, 62(6), 679-687. Retrieved from http://www.jstor.org/stable/3110326

Mitchell R., Popham F. Effect of exposure to natural environment on health inequalities: An observational population study. Lancet. 2008;372:1655–1660. doi: 10.1016/S0140-6736(08)61689-X.

Nalisnick, E. T., & Baird, H. S. (2013, August). Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 479-483).

Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages (2011) 93-98.

Ngoc, P. T., & Yoo, M. (2014, February). The lexicon-based sentiment analysis for fan page ranking in Facebook. In *The International Conference on Information Networking 2014 (ICOIN2014)* (pp. 444-448). IEEE.

Nguyen, Q. C., Kath, S., Meng, H. W., Li, D., Smith, K. R., VanDerslice, J. A., Wen, M., & Li, F. (2016). Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied geography (Sevenoaks, England)*, *73*, 77–88. Retrieved from https://doi.org/10.1016/j.apgeog.2016.06.003

Oliver Melvin L, Shapiro Thomas M. Black Wealth/White Wealth: A New Perspective on Racial Inequality. New York: Routledge; 2006.

Pearsall, H., & Pierce, J. (2010). Urban sustainability and environmental justice: evaluating the linkages in public planning/policy discourse. *Local Environment*, *15*, 569 - 580.

Plunz, Zhou, Carrasco Vintimilla, Mckeown, Yu, Uguccioni, & Sutto. (2019). Twitter sentiment in New York City parks as measure of well-being. Landscape and Urban Planning, 189, 235-246.

Ramsey, M. H. (2009). "Uncerainty in the assessment of hazard, exposure and risk." Environ Geochem Health 31: 13.

Sha, Y., Yan, J., & Cai, G. (2014). Detecting public sentiment over PM2.5 pollution hazards through analysis of Chinese microblog. In *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management* (pp. 722-726). (ISCRAM 2014 Conference Proceedings - 11th

International Conference on Information Systems for Crisis Response and Management). The Pennsylvania State University.

Shan, S., Ju, X., Wei, Y., & Wang, Z. (2021). Effects of $PM_{2.5}$ on People's Emotion: A Case Study of Weibo (Chinese Twitter) in Beijing. *International journal of environmental research and public health, 18*(10), 5422. https://doi.org/10.3390/ijerph18105422

Schroeder, R., K. St. Martin, et al. (2008). "Third World Environmental Justice." Society & Natural Resources 21(7): 547-555.

S.L. Cutter, B.J. Boruff, W. Lynn Shirley. Social vulnerability to environmental hazards. Social Science Quarterly, 84 (2) (2003), pp. 242-261

Susan L. Cutter ; Christina Finch.Temporal and spatial changes in social vulnerability to natural hazards. Proceedings of the National Academy of Sciences, 19 February 2008, Vol.105(7), p.2301

Susan L. Cutter, Christopher T. Emrich & José Luís Zêzere (2014) Application of Social Vulnerability Index (SoVI) and delineation of natural risk zones in Greater Lisbon, Portugal, Journal of Risk Research, 18:5, 651-674, DOI: 10.1080/13669877.2014.910689

Susan L. Cutter, Danika Holm and Lloyd Clark.1995. The role of geographic scale in monitoring environmental justice, Risk Analysis. Vol.16,No.4.1996

Silverman, Zachary S., "Locating Environmental Justice Populations: A Method for Identifying Vulnerable Populations in Massachusetts" (2012). Masters Theses 1911 - February 2014. 836. Retrieved from https://scholarworks.umass.edu/theses/836

Striessnig, E., W. Lutz, and A. G. Patt. 2013. Effects of educational attainment on climate risk vulnerability. Ecology and Society 18(1): 16. Retrieved from http://dx.doi.org/10.5751/ES-05252-180116

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., (2011), "Lexicon-based methods for sentiment analysis", Computational Linguistics, Vol. 37 no. 2, pp. 267-307.

Tanvi Hardeniya, Dilipkumar A. Borikar (2016). Dictionary-Based Approach to Sentiment Analysis - A Review. *International Journal of Advanced Engineering, Management and Science (ISSN: 2454-1311)*,2(5), 317-322.

Taylor, Dorceta E. "Introduction: The Evolution of Environmental Justice Activism, Research, and Scholarship." Environmental Practice 13.04 (2011): 280-301.

Taylor, D.E. 2014. Toxic Communities: Environmental Racism, Industrial Pollution, and Residential Mobility. New York: New York University Press.

U.S. Department of Health and Human Services (HHS), "U.S. Department of Health and Human Services Strategic Plan Fiscal Years 2010 -2015."

U.S. Environmental Protection Agency (EPA), 2017. EJSCREEN Technical Documentation.

Wilson, S. M. (2010). "Environmental Justice Movement: A Review of History, Research, and Public Health Issues." Journal of Public Management & Social Policy 16(1): 19-50.

Wilson SE, Kahn RS, Khoury J, Lanphear BP. The role of air nicotine in explaining racial differences in cotinine among tobacco-exposed children. Chest. 2007;131(3):856–862.

Waller LA, Louis TA, Carlin BP. Environmental justice and statistical summaries of differences in exposure distributions. J Expo Anal Environ Epidemiol. 1999;9(1):56–65. Crossref, Medline

Zandbergen PA, Chakraborty J. Improving environmental exposure analysis using cumulative distribution functions and individual geocoding. Int J Health Geogr. 2006;5:23. Crossref, Medline

Zimmerman R. Issues of classification in environmental equity: how we manage is how we measure. Fordham Urban Law J. 1994;29(3):633–669.

Zheng, S., Wang, J., Sun, C.J., Zhang, X., & Kahn, M.E. (2018). Air pollution lowers Chinese urbanites' expressed happiness on social media. *Nature Human Behaviour, 3*, 237-243.