Noise Levels Associated with Sentiment Analysis on Twitter: A Case Study of New York City

A thesis submitted by

Xiang Yu

In partial fulfillment of the requirements of the degree of Master of Arts

in

Urban and Environmental Policy and Planning

TUFTS UNIVERSITY

August 2016

Advisor: Justin Hollander

Reader: Mary Davis

ABSTRACT

Resulting from the urbanization and motorization, urban noise has become one of the most severe hazards in current society. This thesis firstly investigated the background of noise-control regulations and laws, how noise varies in different urban environment, and the health impacts of noise pollution. After that, the author made several maps to illustrate the information such as distribution of 311 noise complaints and population density in New York City via ArcGIS. Finally, this thesis explored how to use social media data (Tweets) to determine the correlations among different types of variables including population density, average noise complaints, and average sentiment scores in New York City, and two methods, sentiment analysis and statistical test were applied in this process. The results of this study demonstrated that noise complaint and population density were significantly correlated and moved together positively. Meanwhile, the relationship between noise complaint and sentiment score were beyond the author's expectation. It had been approved in this study that when noise complaints go up the sentiment of tweets becomes more negative than positive in most of the community boards in New York City. In addition, this thesis also provided some suggestions and recommendations on applying social media data in urban planning and strategies of noise control for policy makers and future studies.

ACKNOWLEDGEMENTS

This research would not be possible without the support of many people. I would like to express my sincere thanks to the members of my thesis committee: my advisor Prof. Justin Hollander who always enlightens and encourages me by his extensive knowledge and professional attitudes, and my reader Prof. Mary Davis whose invaluable suggestions and supports helped me to overcome the difficulties and improved this thesis successfully.

I also wish to extend my gratitude to my academic advisor, Prof. Weiping Wu who offered numerous assistance, guidance and care to me throughout my two years at Tufts University. I hope she has a wonderful time in her new chapter at Columbia University.

I would also like to express my appreciations to all other faculties and staffs at UEP for their unselfish assistances and consistent inspirations, and to all my good friends who have worked, struggled and shared joy together. I am so glad that I could have your love and encouragement in my journey here.

Last but not the least, I wish to express my special thanks to my beloved family for their endless love, support and understanding throughout the graduate program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Urbanization is proceeding rapidly in many places of the world, and more than half of the global population are now living in cities. Accompanied by this large-scale urbanization, concerns about improving urban environments have become increasingly important for both policy makers and urban citizens (Rehan 2015). Besides air pollution, urban noise can be considered as another main source of pollution affecting public health in cities. The World Health Organization (WHO) has declared noise as a pollutant since 1972 (WHO 1997). In that same decade, the negative effects of noise pollution and the importance of urban soundscape has been noticed by several organizations, such as the World Soundscape Project founded in Canada (Westerkamp 1991).

Nowadays, the quality of the acoustical environment in urban areas, particularly in mega-cities, is facing great threats. We can find various audible noise sources in the urban environment: traffic noise created from vehicles, trains, and aircraft; noise hazards from industrial facilities and municipal constructions; and even social activities like loud parties and open air markets. Although these audible sources may not be viewed as serious impacts by some of the urban residents, they all contribute to the conversion of the soundscape in urban areas (Vianna, Cardoso, and Rodrigues 2015).

For this thesis, I conducted a noise-related study of New York City (NYC). NYC, as one of the biggest and busiest city in the world, is creating a tremendous volume of noise every single day, and its residents have been annoyed from different kinds of noise pollution for a long time. In 2007, under Local Law 113 for the year 2005, the City updated the Noise Code for the first time in 30 years to reflect the changing urban landscape and advances in acoustic technology. According to this updating, different

noise sources were categorized more detailedly and reasonably, which allows the government to control and manage the noise pollution more efficiently.

311 is one of the online/call-in services running by the city since 2003, whose mission is to provide the public with a quick, easy access to all the government services and information in NYC. Today, it fields more than 55,000 calls per day, offering information about more than 3,600 topics. As we can see from Figure 1, Noise (the largest pink area in the middle of the chart) is the topics with the most complaints during one week in September 2010 (Johnson 2010). Noise complaints are reported, recorded, and answered according to NYC Noise Code. In order to enforce its objective, the Department of Environmental Protection (DEP) and the New York City Police Department (NYPD) share duties based on the type of noise complaints.

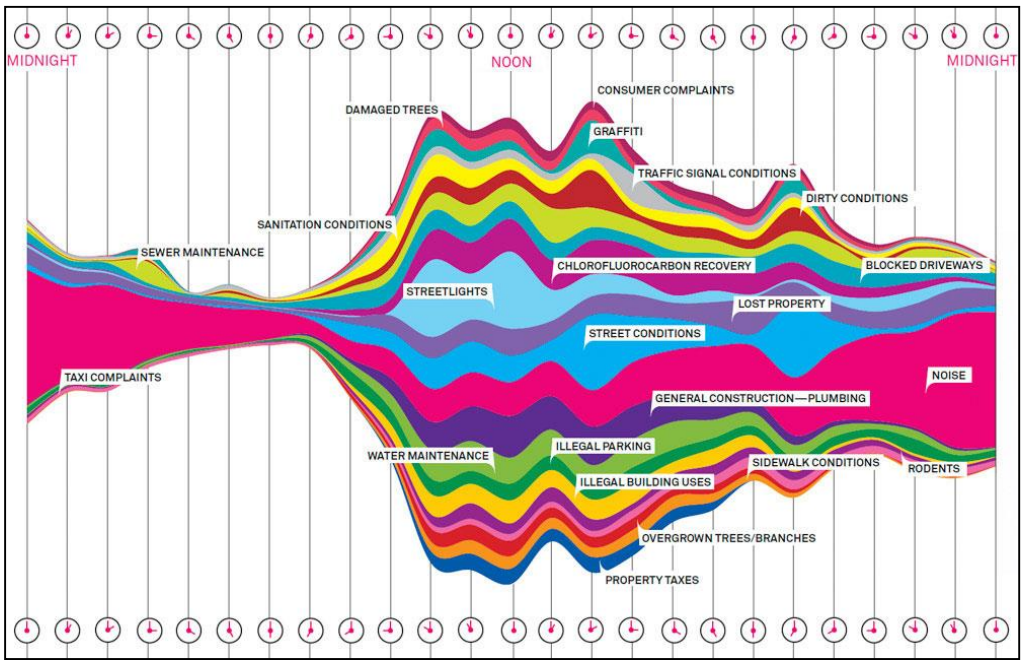

Figure 1: Complains called into 311 between Sept 8 and 15, 2010
(Source: Steven Johnson, Wired)

So, the first part of this thesis is a review of noise pollution and its side effects on human health. It covers a range of urban areas both domestically and internationally, and

especially focuses on NYC. It also included the processes by which I managed the 311 service data such as data collecting, geographic categorizing, and basic mapping, which helped me to get a better understanding of the general condition of NYC's noise pollution.

The second field is an analysis what we called "Big Data Mining" in social media, which was applied in this study through tweets collection and sentiment analysis. Due to wide spreading of smartphones and advanced digital technology, social media is gaining an increasing number of users and becoming more and more popular across all generations. Meanwhile, its academic potential and capability of researching also draw wide attention from researchers in various fields. Recently, urban study has experienced a growing interest in using digital data from different sources to understand the city and its dwellers. Particularly, with the increasing availability of those location-based data, researcher are able to obtain more valuable information from the social media database (Ciuccarelli, Lupi, and Simeone 2014).

Since social media has already become a popular place for people to share their life and connect to others, those data is able to show how people live and experience the city where they live, work, or even visit, and tells us about their feelings and thoughts. For example, researchers could determine the subjective well-being of individuals based upon Facebook status updates (Kim and Lee 2011), tweets sent by users of Twitter might provide useful information about land use and social inequality (Frias-Martinez and Frias-Martinez 2014), and geo-located reviews from TripAdvisor might show what the most popular attraction looks like (Ciuccarelli, Lupi, and Simeone 2014). Moreover, short post from microblog such as Twitter also become a good source for sentiment

analysis, which is also called opinion mining, refers to the process of identifying and determining whether opinions expressed in a piece of writing is positive, negative, or neutral. It has been widely applied to social media to discover how people feel about a particular topic (Lexalytics 2016).

In this study, I investigated the relationship between two superficially unrelated datasets: 311 noise complaints and sentiment trend of a group of tweets in NYC during a two-week period of time. At the same time, I hoped to find some evidence showing whether there are interactions between population density and noise complaints, and between noise complaints and tweet sentiment. With all of the above in mind, I come up with the following research questions:

1. How does urban noise influence public health, particularly people's mental health and sentiment?
2. Does the distribution of noise complaints match up with population density?
3. How can the sentiment of tweets reflect this kind of noise impact in NYC?

To answer the first question, I conducted a literature review in chapter 2 after reading through over twenty previous studies and government documents, in which I introduced a brief history of noise-related legislation and standards in both the US and EU, explained the bad effects of urban noise, and summarized some case studies world widely.

For the last two questions, two sets of data, briefly mentioned above, were applied in this study. The first one is a group of noise complaints from the 311 online services collected in a two-week period between January 11 and 25, 2016. These data are available on NYC's official website called 311 Online Service Request Map. More

details such as location, request time, and solutions of each complaint can be found on the website of NYC Open Data (NYC 2016). The process of managing the 311 data can be seen in the first part of Chapter 3. Generally, three noise level including high, medium, and low were defined in this study based on the complaint number of each community board, and the distribution of these categories was mapped and illustrated via ArcGIS.

Figure 2: Tweet Output Area

The second set of data is a bunch of tweets collected in exactly the same period of time with the 311 complaint data. With the help from my thesis advisor Justin Hollander, the director of Tufts Urban Attitudes Lab (UAL), the tweets were extracted from the Twitter service within the area shown in Figure 2. Then, I clipped and geo-located those tweets based on their longitude and latitude coordinates via GIS. The second part of

chapter 3 will introduce the whole process of data preparation, as well as the sentiment analysis conducting via the software: "*Urban Attitudes*", developed by the Tufts UAL.

Then I did a series of statistical tests to verify the relationships among those factors. The process and findings will be described in chapter 4 and chapter 5. Additionally, chapter 5 will also point out the limitations of this thesis and raise some recommendations for future researchers.

CHAPTER 2: LITERATURE REVIEW

Resulting from rapid population growth and unprecedented motorization in big cities worldwide, urban noise is having an increasingly significant impact on urban dwellers' health and life quality. Evidence from epidemiological studies suggests that daily noise exposure could cause stress, sleep disturbance, depression, and increased risk of getting hypertension and cardiovascular disease (Ising and Kruppa 2004). Meanwhile, exposure to loud industrial noise has long been recognized as an important occupational health hazard, which may have serious impacts on employees' health and productivity (Rabinowitz et al. 2007). Moreover, excessive noise in residential areas has been viewed as a prevalent urban environmental hazard associated with adverse psychosocial and physiological health effects (Moudon 2009) and a cause of impaired cognitive performance of children (Clark et al. 2006).

Background of Noise-Control Laws and Regulations

The negative health effects of noise were recognized more than 50 years ago. Physically, exposure to high sound levels can cause direct harm to human ears and lead to noise-induced hearing loss. Therefore, regulations for protecting people from occupational noise hazards have already been implemented for over half a century. Here is a brief history of how the system has been developed. According to Moudon (2009) study, hearing loss from occupational exposure to hazardous noise was identified as a compensable disability by the US courts in 1948 to 1959, and with the development of the jet engine and other modern motorized technologies, occupational noise could bring more damages to individuals. In order to minimize those side effects, a series of noise-control laws including the Occupational Safety and Health Act (1970) and Noise Control

Act (1972) were implemented. Then the US Environmental Protection Agency (EPA) created the Office of Noise Abatement and Control (ONAC) for abating noise pollution by following the guidelines enacted in the Noise Control Act.

In short, the legislation for industrial noise protection came a little bit earlier, and then the regulations for residential areas were taken into action by the federal government a few years later. One possible reason could be that the people at that time might not have had too much awareness of their health and the impacts of noise pollution. The EPA determined that outdoor noise levels would affect public health and welfare, and released several standards of noise exposure in 1974. According to those standards, exposure to noise equal to 70 decibels (dB) for 24 hours might cause hearing loss over a lifetime. Likewise, 55dB and 45dB were set as the ceilings of acceptable outdoor and indoor sound volume, respectively, for preventing activity interference and annoyance (EPA 1972). In 1978, the Quiet Communities Act, which was most recently updated in 2015, authorized the EPA to provide grants to the state and municipal governments for noise abatement (Congress 2015). However, in 1981, the White House concluded that noise issues should be best handled by the state or local government. Consequently, the responsibility for noise control started to switch from the federal level to the state and local government level. During that period, ONAC played an important role in noise abatement until its funding was phased out in 1992. After that, the Noise Control Act of 1972 and the Quiet Communities Act of 1978 still remain in effect today, though essentially unfunded.

From an international perspective, regulations related to noise control or protection were updated more recently than in the US. According to the World Health

Organization (WHO)'s Community Noise Guidelines (1999), continuous outdoor noise in residential areas should not exceed an average level of 55dB within 16 hours. In 2002, the European Union (EU) passed Directive 2002/49/EC, also known as the Environmental Noise Directive, whose aim was to seek a common approach intended to avoid, prevent, or reduce the harmful effects due to exposure to environmental noise. It has provided a lot of information on environmental noise, such as the causes, the subcategories, and the assessment and protection methods for the EU member states. In addition, the applications and the suggestions for future policy were also emphasized in this document. More information on permissible exposure time under different dB can be found in Figure 3, and any over-time exposure might cause temporary or permanent damage to our hearing system.

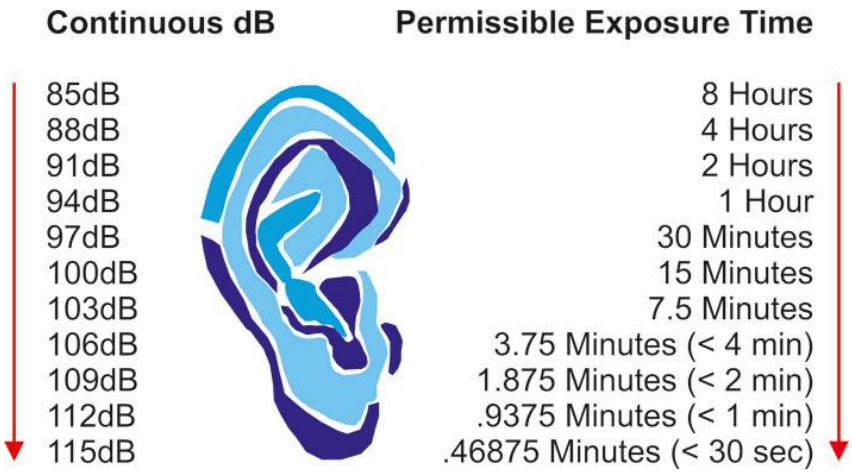

Figure 3: Exposure Time under Different Continuous dB
(Source: The Hearing Company)

In order to show how our ears feel at different sound levels based on the standards that I mentioned in last paragraph, I found a graph (Figure 4) that shows how our acoustic feeling changes from low sound levels to high. From this figure, I also noticed that common things like city traffic, hair dryers, and lawn mowers could easily create very

loud noise. It is no wonder that the noise limitations are often exceeded in built-up areas, both in the EU and the US (Lee et al. 2014). For example, a recent study in Fulton County, Georgia (Seong et al. 2011), showed that 48% of the population in the county were exposed to noise levels over 55dB during the daytime, and 32% of the population were constantly exposed to the noise level of 50dB. The condition of Europe could be even worse. Lee et al. (2014) also revealed that approximately one fifth of the European people were exposed to daytime levels exceeding 65dB, and nearly one third were exposed to nighttime levels beyond 55dB.

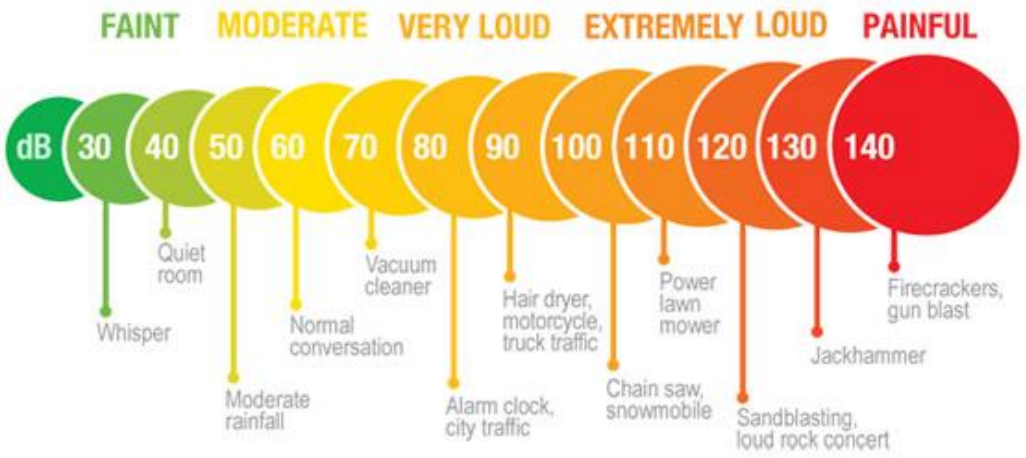

Figure 4: Acoustic Feelings of Different Sound Levels
(Source: Ototronix Diagnostics)

Relationship between Noise Pollution and Building Environment

Many studies have demonstrated that a noticeable percentage of the urban population is being affected by urban noise worldwide. I think the question of what kind of condition and environment will cause people to become more vulnerable. Vianna, Cardoso, and Rodrigues (2015) did related research on the urban soundscape and its perceptual effect on urban citizens in 2015. They evaluated the noise level and the effects among six different scenarios in which people work, relax, or stay with family, and the

goal was to find which scenario contained the loudest noise and had the most serious bad effects. After interviewing 180 individuals in Porto, Portugal, and doing the statistical analysis, the results showed that interviewees had the greatest perception of noise sources when they were at home, and nearly half of them reported annoyance due to the noise in their workplace. Although no significant annoyance was detected when people were at leisure, 65% of the interviewees said that they had perceived noise sources in urban parks.

The situation in the US was also discouraging. San Francisco is one example. A study in 2007 estimated that 17% of the city's population was at risk of annoyance from the traffic noise alone, and this percentage could possibly be higher when other noise sources were considered (Seto et al. 2007). In NYC, the large number of noise-related calls from the 311 service also reflected how annoying the noise pollution could be. According to the 311 data, 111,730 noise-related complaints were recorded in the system in 2009 (Blasio and Lloyd 2007).

Before moving into noise-controlling strategies, I think it is important to figure out in what kind of building environment citizens can be easily bothered by noise pollution. As we know, urban elements vary in different building environments. Resulting from the rapid motorization in recent decades, particularly in developing countries, air pollution is no longer the only threat brought by modern transportation, while traffic-related noise is becoming the main portion of urban noise pollution. One recent study by Lee et al. (2014) showed that traffic noise has attracted growing attention as a public health concern in the US. They investigated the traffic noise in three US cities—Atlanta, Los Angeles, and NYC—and collected data in those cities' downtown

areas via noise survey and noise detector. The results showed that urban noise in those three cities was quite high, and some of the areas even exceeded the WHO's thresholds associated with adverse health impacts. In addition, it was clear that traffic noise level was tightly associated with traffic density, and NYC had the greatest number of vehicles and highest noise level among those three cities.

In terms of traffic noise, it is hard for urban dwellers to escape being annoyed nowadays, because communities of big cities are usually surrounded by well-developed roads. Seto et al. (2007) did a further investigation on the relationship between indoor noise and traffic volume in San Francisco. It was found in this study that urban noise tended to increase by 6.7dB with every 10-times increase in street traffic, and the high frequency of buses and heavy trucks on roads would make the situation even worse. The researchers also found that residents of high-density areas like Chinatown, downtown, and fast-growing neighborhoods always faced a higher risk of noise exposure and annoyance.

A more recent study of NYC was focused on how traffic noise varies at different times of day (Kheirbek et al. 2014). In this study, researchers set several monitors at 56 different sites around the city that have diverse traffic volume and building density. The purpose was to measure the sound level over a one-week period. This study found that the average daytime noise level was significantly higher than the nighttime one, and the noise on weekdays was also higher than on weekends. Unfortunately, the record from all the detectors went beyond 55dB, which was the EPA's limitation of activity interference and annoyance, and half of them even exceeded the line that might cause hearing loss based on the EPA's noise guidelines. Since mass traffic can bring both noise and air

pollution, it was not odd to see that the researchers also found that these two kinds of pollution shared the similar distribution among the study sites (Kheirbek et al. 2014).

From the perspective of urban transportation, planners often speak highly of public transportation for its sustainability, effectiveness, and energy saving. But how does public transportation perform in terms of noise control? Perhaps people will think that public transit can help with controlling the number of private vehicles, and thus contribute to reducing traffic noise. However, the facts about public transportation in NYC may not be that ideal, based on a pilot survey in 2006. In that study, over 90 sensors were set by Gershon et al. (2006) for collecting the noise level of the NYC transit system. Their findings were quite surprising. The average noise level of the subway platforms was around 86dB, and the maximum noise level inside subway cars could go up to 112dB, which both exceeded the guidelines of the EPA and WHO for an exposure duration of no more than 45 minutes. The noise level of bus stops was a little lower than subway platforms, but the maximum sound was still as high as 89dB. On the basis of my own experience, the subway trains there could be very loud when they were running on the tracks, especially the elevated ones. I could barely hear people talking when there was a train running above my head.

Health Impacts of Noise Pollution

After I went through the studies above, I came to the conclusion that a large number of urban residents were being affected by various sources of urban noise, including industrial noise, traffic noise, loud activities, and so on. Consequently, the health impacts of noise pollution should not be ignored here. Noise can give people unpleasant experiences because it not only disturbs people's daily life, but also interferes

with social activities and family gatherings (Babisch 2002). A number of epidemiological studies have reported an association between noise exposure and different types of health problems, including high blood pressure, heart disease, and noise-introduced sleep disturbance (Lee et al. 2014). Besides those bad impacts, noise pollution also contributes to substantial social losses, such as loss in property values, community decline, and the abandonment of civic facilities (Ising and Kruppa 2004). After investigating some related studies, Seto et al. (2007) believed that environmental hazards were more likely to annoy the low-income population in urban areas, and they found that higher noise was more likely to be detected in those low-income neighborhoods. The US situation is quite the same, at least in terms of traffic noise. We know that interstate highways or rapid transit often cut through the low-income and minority neighborhoods.

From the perspective of public health, the association between noise and health is easy to understand (Babisch 2002). Generally, noise has been widely considered as a kind of stressor, which can change the stress hormones of people and affect metabolism. The basic process can be found in the flowchart below (Figure 5).

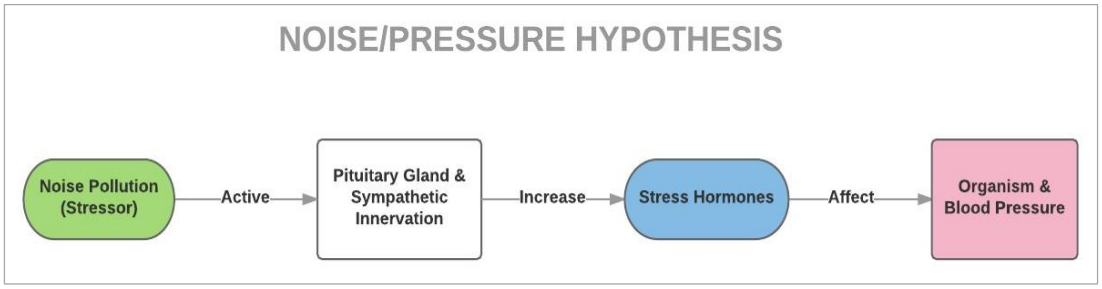

Figure 5: Noise/Pressure Hypothesis

Noise-related health problems can be summarized in three aspects. First, it has been recognized that consistent exposure to noise can cause both acute and chronic changes of stress hormones, and then result in a high possibility of getting high blood

pressure and cardiovascular disease. Moreover, it can also affect the equilibrium of vital body functions (Ising and Kruppa 2004).

Second, continuous noise at nighttime can reduce sleep quality by increasing the time to fall asleep or inducing frequent awakenings (Sygna et al. 2014). Although it has not been proven that noise-introduced sleep disorders will immediately cause mental health problems, Ising and Kruppa (2004) said in their study that the bad health influences of long-term sleep disorders should not be neglected. They also listed several tests done in the 1990s, which confirmed that even low noise could probably increase the concentration of stress hormones when people were sleeping. Later, a Norwegian study (Sygna et al. 2014) also revealed the weak but positive indications of an association between traffic noise exposure and mental health, particularly for people who have poor sleep quality.

Third, side effects on schoolchildren need a specific concern. Due to the lack of self-control and concentration, children are more likely to be distracted by different kinds of noise in the surrounding area. Clark et al. (2006) did a study in three different countries, including Netherlands, Spain, and the UK. According to their findings, aircraft noise was more harmful to children's reading comprehension ability when compared with road traffic noise. Another study in Germany found the similar result that aircraft noise was significantly associated with a lower rating of children's mental and physical well-being when they were at school (Schreckenberg et al. 2011). This study also explained the reason why aircraft had worse impacts than vehicles. One possible reason was that aircraft noise was more intense and less predictable, which cause interferences more easily than any other source of noise. However, from the perspective of long-term

impacts, road traffic noise is still a great threat, especially for the people who live close to highways or arteries.

Social Media Data and Its Research Value

The second part of the literature review was focused on how to apply social media to social analysis and urban planning. Resulting from the rapid development of smartphones and the expanding coverage of the Internet, social media has become one of the most prosperous industries nowadays. Meanwhile, it has experienced an increasingly tight connection with different fields and businesses. Decision makers, consultants, and researchers have all started to identify ways in which firms like YouTube, Facebook, Instagram, and Twitter make profits and affect the world (Kaplan and Haenlein 2010). Even though social media has already become very popular among people of different ages, some still have not noticed how powerful it can be and what it means to our lives.

Firstly, let us take a look at the numbers of active users of Facebook and Twitter. According to Facebook's statistical data, there were 1.09 billion daily active users in March 2016 on average, and over 90% of them logged in to their account via mobile devices(Facebook 2016). On the other side, Twitter seems to have fewer active users than Facebook. Based on its report at the end of 2015, there were 320 million active users who used their Twitter every month, and 80% of them preferred to post tweets through their mobile app (Twitter 2015). If we put those numbers into perspective, surprisingly, we can see that nearly one seventh of the world population (2013) were using Facebook frequently every day, and the number of Twitter's active users nearly equals the population of the US (2016).

Therefore, with the help of its numerous users, social media can connect people and affect the world much more easily than any traditional media. Since the current social media applications are developed as the combination of previous Internet products like bulletin board systems and blogs, as well as being updated very frequently, they can be extremely user-friendly and always able to be improved in terms of user experience. In addition, social media users are able to create and share information more easily and effectively, and this information has become what we call "Big Data."

The term "Big Data" is usually used for describing data sets that are too large or too complex to be easily captured, managed, and analyzed by commonly used software (Snijders, Matzat, and Reips 2012). Since the term was created by NASA in 1997, it has been used in many different fields, and its size and coverage are growing constantly by receiving data from mobile devices, media software, smartphones, and various sensor networks. According to the statistics for 2015, 2.5 quintillion bytes of data were created daily, and the stored data grew four times faster than the world economy (Walker 2015). Meanwhile, individuals' capacity for information storing had roughly doubled every forty months since the 1980s, thanks to the information and technical revolution (Hilbert and López 2011).

Therefore, Big Data fits perfectly with Gartner's widely used concept of "3Vs" (volume, velocity, and variety) because of its increasing volume, fast velocity, and high variety of information assets. By benefiting from those advantages, Big Data has been applied to improve the decision-making process in different fields such as economic development, health care, business analysis, natural resource management, and academic research (McAfee et al. 2012). For example, in order to explore how Big Data can help

the government to address critical problems, the Obama administration announced the Big Data Research and Development Initiative, which is composed of 84 different Big Data programs spread across six departments in 2012 (President 2012).

Urban planners have also already started to apply social media data in urban planning for improving urban functions and enhancing decision-making processes. There are some natural characteristics of social media data that are well adapted for use by planners. First of all, most of the current social media apps have a location-based check-in function, by which the users can share their daily life experiences and activity-related choices wherever they like from their laptops, tablets, or smartphones (Hasan, Zhan, and Ukkusuri 2013). From this perspective, these people act like urban or social sensors, constantly providing their diverse thoughts and observations from the physical world to the online network. This huge body of information offers researchers many new forms of data set and easier access for doing urban studies, and these data can hardly be obtained via any traditional data-collecting methods (Silva et al. 2014).

As Ciuccarelli, Lupi, and Simeone (2014) described in their book *Data City*, decision-making processes normally include two types of knowledge, institutional knowledge and local knowledge. Institutional knowledge is usually collected and created by institutions in forms like documents, reports, and plans. The information is usually collected via traditional data collection methods like interviews, surveys, questionnaires, and long-term observations, so that decision makers will be able to get a sense of how their policies or the public services work. On the other hand, local knowledge is the information that comes from local residents. This knowledge distributes more randomly and varies dramatically by different groups of people, because they view the cities from

different perspectives and backgrounds. Interactions between these two kinds of knowledge partially decide how successful the decision-making process will be.

With the rising awareness of the concept of Participatory Urbanism (Paulos, Smith, and Honicky 2016), public participation is playing an increasingly important role in urban planning. Things like community meetings, co-design sessions, and digital collaborative platforms have become more and more popular in decision-making processes. Meanwhile, different levels of government have become more welcoming to residents' involvement and have been willing to listen to local demands (Hanzl 2007). The reasons could be twofold. On one hand, public opinion is quite essential for the problem-solving process in most of the current urban complex. On the other hand, bringing in the public can certainly narrow the conceptual gap between decision makers and citizens and enhance the efficiency and practicality of the policy-making process. Therefore, based on its features, social media can be a perfect tool for gluing those two kinds of knowledge that I mentioned above.

The most common method is to capture urban dynamics and social activities patterns by using location-based data from social media (Silva et al. 2014). Hasan, Zhan, and Ukkusuri (2013) did a study on how social media data reflected people's travel destinations in 2013. The data set they used consisted of several groups of tweets in NYC, Chicago, and Los Angeles. Here is some information about Twitter and the way it works. Twitter allows its users to post short messages of up to 140 characters, which is also sometimes called status updating. A tweet can be posted with the location information if the user wants, and its content can be a short text, a picture, a short video, or a GIF (graphic interchange format). Some third-party apps like Foursquare also can

post in Twitter when people link them together. So Hasan and his colleagues aggregated this location-based information to analyze people's destination choice among six different kinds of activity. They noticed that people's destination of activity was not selected randomly. Yet the spatiotemporal distribution of it could be reflected by the concentration of tweets distinctly in these cities. So they believed that Twitter could influence urban dwellers' activity and destination choices in those cities, which might help improve urban function.

A similar method was also applied in later studies. For example, Silva et al. (2014) considered every social media user as the participatory sensor of a huge network, which they called "city image," that could measure city dynamics differently than the traditional ways. In the raster image they created, the different color of each cell represented the strength of people's willingness to shift activities. In addition, that image could be overlapped with other layers such as traffic condition, weather condition, or place of interest. So it could help city planners to understand overall urban dynamics and to use those commonly invisible resources more effectively (Silva et al. 2014). In addition, this geolocated information was capable of driving smart growth and improving the urban environment, because these data reflected the demands of urban inhabitants and showed the pattern of various activities in a 3D urban complex (Sagl, Resch, and Blaschke 2015).

Sentiment Analysis and Social Media

Besides making use of the check-in function, the sentiment value contained in social media can also be applied to social analysis. For example, sentiment analysis, also described as opinion mining, usually represents studies analyzing people's general

attitude or sentiment trend based on their opinions, emotions, and appraisals reflected in written language. It has been increasingly used in social media analysis, and also widely applied in data and text mining (Liu 2015).

Since short sentences or summarized information are more suitable for sentiment analysis, microblogs like Twitter are more commonly analyzed in this approach (Nakov et al. 2013). Usually, researchers set several sentiment classifications for judging the sentimental attribute of the written text. For example, emoji were chosen as the sentimental symbols in a computer science study done by Go, Bhayani, and Huang (2009). Their approach was that they considered the tweets that ended with "☺" as positive, and those that ended with "☹" as negative. Then the samples were divided into two groups representing the positive sentiment and the negative sentiment. A similar method also could be applied to written text. One of the applications is the AFINN Dictionary mentioned in chapter 1. In addition, Agarwal et al. (2011) introduced an extension of WordNet in their study which gave pleasantness scores (from 1 to 3) to about 8000 English words.

In terms of applying those dictionaries to Twitter, the most common way is to give tweets containing positive words like "good" or "nice" a positive score and give tweets containing negative words like "bad" or "sad" a negative score. The number varies based on the sentimental strength of each word. Additionally, symbols like a hashtag (#) can also help with locating the hot topics and collecting sentiment data on Twitter (Kouloumpis, Wilson, and Moore 2011).

Since social media software is developed on the basis of computer technology, a big part of the social media studies is related to computer science. However, approaches

like sentiment analysis should be introduced to the field of urban planning as well. The reasons are twofold. Firstly, as Liu (2015) describes in his book, "opinions are central to almost all human activities and are key influencers of our behaviors" (p. 5). The attitude of urban inhabitants is so important that it should not be neglected by the decision makers, since it is the goal of public policies to serve the public. So how can decision makers know whether their strategies are successful or not without knowing people's feelings about those policies? Since the current urban planning concept contains growing humane considerations, social media will play an increasingly important role in connecting decision makers and citizens.

Secondly, sentiment data can be collected more easily from social media than by using traditional measures. It usually takes quite a long time to obtain data on a group of people's attitudes by doing interviews or telephone calls, and it is hard to guarantee that every interviewee is willing to provide enough information. But for social media, all the data are provided voluntarily by the users. The characters of these data, such as ubiquity, diversity, and ease of access, make social media an ideal tool for evaluating urban context and helping with the following improvements.

Take a recent feasibility study as an example: it explored the correlation between public transit and stigma on Twitter (Schweitzer 2014). From the result, it was surprising to see that nearly all those mass public transit providers studied, such as NY MTA, MBTA, and BART, had a negative sentiment score based on the tweets related to their services, which indicated that there were a lot of complaints about their services from the Twitter users. Among those complaints, complaints about the quality of services such as on-time performance, facilities, driver and staff conduct, and security were roughly

equivalent to those about social justice concerns such as access for subpopulations, race equality, and gender issues. From this perspective, social media not only provides suggestions for physical urban context, but also helps with enhancing social justice. Another interesting thing that I found was that this paper had already been retweeted 40 times by 34 individual users, with an upper bound of 144,791 followers after being published two years ago (Schweitzer 2014). So we can get a sense of how fast a piece of information can be spread through social media.

Some of the studies that I mentioned above have illuminated ways of applying social media in urban planning scenarios, while the methods will vary or be upgraded corresponding to different kinds of social media and their future updating. Nonetheless, social media analysis still has limitations that affect the study's accuracy. To begin with, the size of those data sets could be very large, with valuable and useless information mixed together. So it is necessary to filter and manage the data via some particular software (Snijders, Matzat, and Reips 2012). Moreover, cultural difference could have impacts on people's behavior and attitude among different countries, which should be noticed in further investigations (Silva et al. 2014). For instance, two geographically separated cities with a similar culture should be compared together, and two closed areas having distinct cultures should be considered differently. The next chapter starts describing the process of my own research, which gained inspiration from this literature review.

CHAPTER 3: METHOD

As I mentioned in the first chapter, two data sets were used in this study, as well as two analyzing tools, ArcGIS and STATA. The first data set is the noise complaint data from NYC's 311 Online Service. The city has a website called "311 Request Map" showing 311 request spots all over the city. On that website, all the requests and complaints are set into different categories such as air and water quality, noise, transit, parking, and so on. Meanwhile, the complaints are recorded and counted at the community board (CB) level. The users can know the total complaints of a specific category or a specific period of time by clicking and filtering at the drop-down menu.

GIS Mapping

The data used in this paper were collected for two weeks between Jan. 11 and Jan. 25, 2016, with a total number of 10,885. Every complaint was geolocated based on the provider's address, and counted within its CB. Here is some basic information on the administrative division of NYC: the city is now composed of five boroughs, which are Manhattan, the Bronx, Queens, Brooklyn, and Staten Island; among these boroughs, there are currently 59 CBs. In order to distinguish those CBs more clearly, I gave each CB a sample name. For example, the 12 CBs in Manhattan were named M1 to M12, and the ones in the Bronx were named BX1 to BX12. Brooklyn is the borough with the most CBs, and Queens is the runner-up with 14 CBs. Staten Island has the smallest number, which is only three. Figure 6 is a basic map which I drew using ArcGIS (GIS), showing where the boroughs and the CBs are.

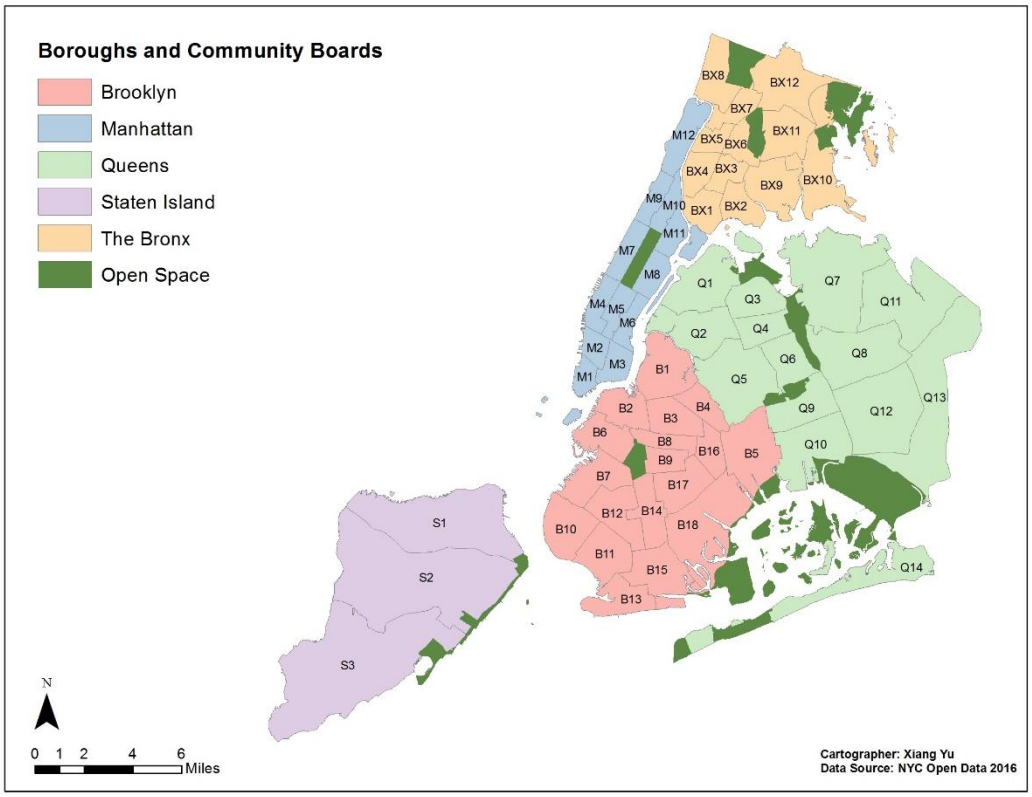

Figure 6: Map of Boroughs and Community Boards in New York City

The next step was adding the complaint data to that basic map. First, I recorded the total number of noise complaints of each CB and then typed those numbers manually in an Excel table. After that, I joined the map with the table via GIS. It should be noticed that the number of complaints varies significantly among different CBs, from over 600 to only 45, indicating that the acoustical environment differs between CBs in NYC. For the sake of illustrating how those noise complaints are distributed in the city, three different levels—low, medium and high—were created by using the "Quantile Cut-Point" in GIS, and the related map can be found in Figure 7. The Quantile Cut-Point divided all the variables equally into different groups, and each group had the same number of variables.

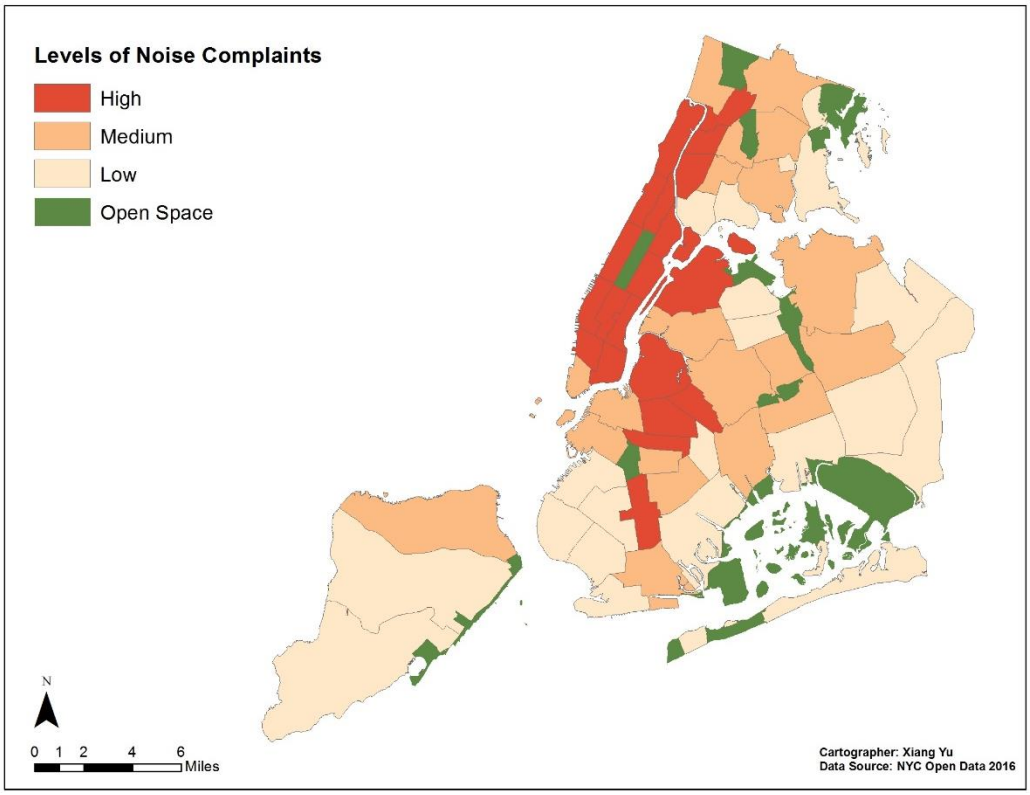

Figure 7: Map of Noise-Related Complaints in New York City

From Figure 7, we can see the distribution of noise-related complaints clearly. In general, the metropolitan area, including all of Manhattan and its surrounding places in the Bronx, Queens, and Brooklyn, had the largest number of noise complaints. The condition becomes better gradually from downtown to the suburbs. However, we can also find some abnormal spots whose noise levels are much different than those of the CBs surrounding them. For example, Q3 and Q4 are surrounded by several areas with medium or high levels of noise complaints, but they are the only two CBs with low noise levels in this area. On the contrary, B14, Brooklyn, is supposed to lie in the area with decreasing noise complaints, yet its color is darker than the surrounding boards. When I "walk" into this CB through Google Map, I found that this area is highly covered by either residential buildings or commercial concentrations, and the only open space that I found was a

community part called Kolbert Park. In addition, arteries like Coney Island Ave and Avenue J are full of different types of shops and restaurants, and most of them are streetfront commercial properties. This kind of commercial concentration can definitely make that area more attractive and prosperous, but it might create more noise than the other types like shopping mall. All the CBs in Manhattan except M1 are covering by the high level of noise complaints, and CBs in the Bronx, Queens, and Brooklyn areas close to Manhattan are also under a high risk of noise annoyance. Staten Island has the best conditions when compared with other boroughs in this scenario. As I expected, the metropolitan area of NYC seems to concentrate louder urban noise because of its high density of population and social activities. By contrast, the condition of the suburbs is better because of lower density and more open spaces (Raimbault and Dubois 2005).

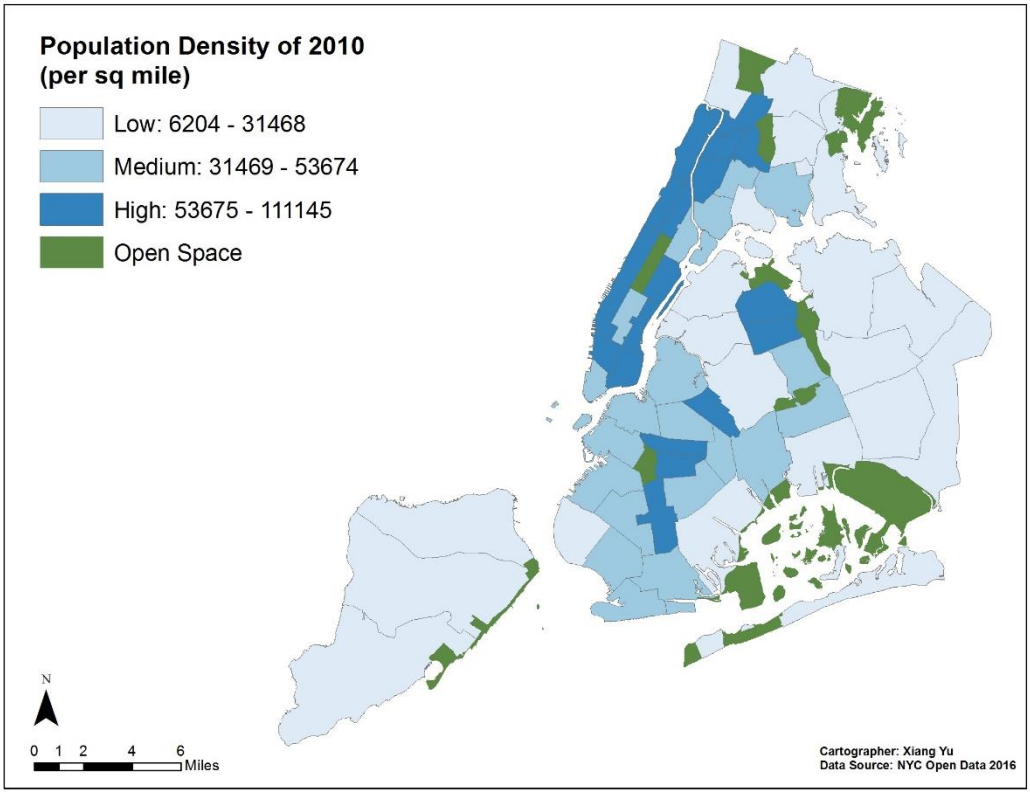

Figure 8: Map of Population Density, New York City, 2010

Therefore, population density is always one of the most important leverages for the building environment, such as the soundscape in urban areas. To address the question of how population density affects noise complaints in my study, I added a layer showing the population density of each CB in GIS. By doing this, I hoped to see whether I could find some correlations between population distribution and noise complaints. Using the same method that I applied to Figure 7, I also divided the CBs into three groups from low to high based on their population density, as shown in Figure 8.

In order to find the correlations between the two variables of population density and noise complaints, I made a bivariate map, shown in Figure 9, by using the population density of CB as the x-axis and the level of noise complaints as the y-axis, in which I overlaid the layer of population density and the layer of noise-related complaints together. There were three steps for doing this bivariate map. First, each category of population density and noise complaints was given a score. For example, I numbered the low, medium, and high population density as 1, 2 and 3, respectively, and numbered the low, medium, and high levels of noise complaint as 10, 20, and 30.

Next, I added a new column for summing those two scores together, so that the new numbers varied in the range of 11, 12, 13, 21, 22, 23, 31, 32, and 33. The number in the tens place, ranging from 1 to 3, represents how high this CB's noise complaints are. The number in the ones place represents how high the population density of a CB is, which also varies from 1 to 3. For example, CBs with the number 13 have a low level of noise complaint but a high population density.

Last, I renamed those numbers by some corresponding abbreviations in case the readers might get confused. In those abbreviations, the letters L, M, and H mean low,

medium, and high, respectively; the letter N stands for the level of noise complaints; and the letter P stands for population density. For example, the abbreviation of number 12 would be LNMP, indicating that the CBs with this number have a low level of noise complaint and a medium population density. Then I colored those nine different cells by a sequential bivariate color scheme suggested by Josh Stevens (Stevens 2015).

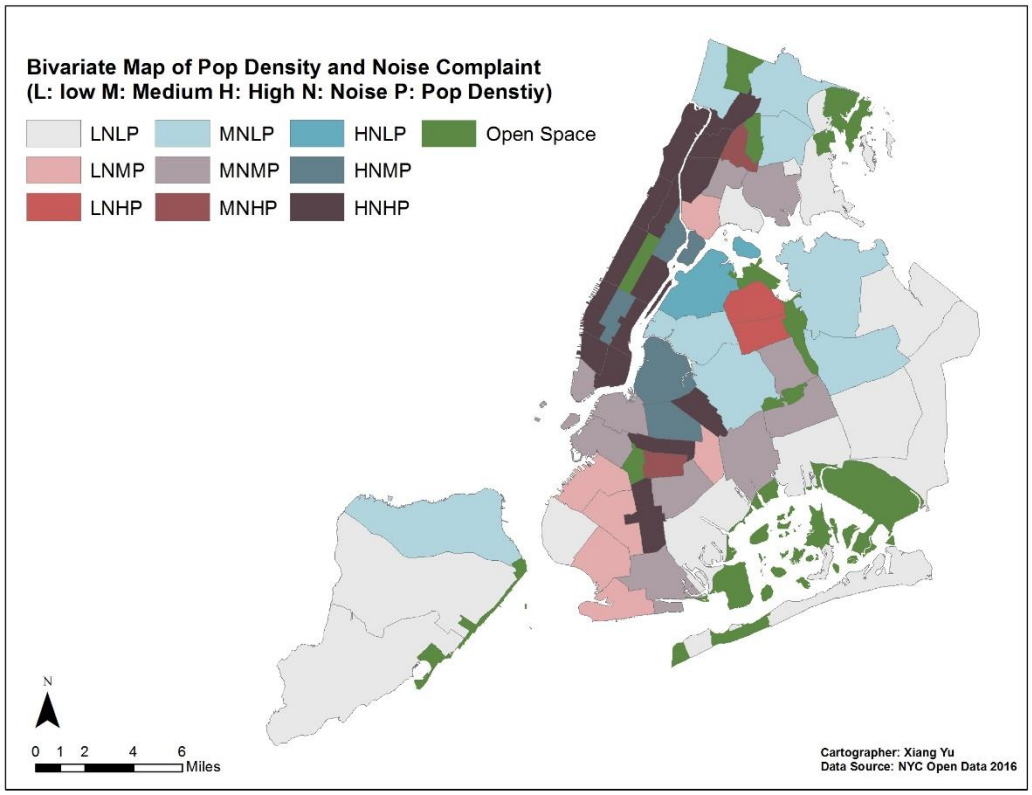

Figure 9: Bivariate Map of Population Density and Noise Complaint in New York City

When we look at Figure 9, we can see that noise complaints and population density are chasing each other in NYC, which matches the common knowledge that highly dense areas are usually noisier. There are 20 CBs with a high level of noise, and 15 of them have a high population density. The last three are quite remarkable because they are not close to the downtown area. In particular, B14, the one that we discussed in the last paragraph, is located near the middle of Brooklyn, quite far away from the other

"double high" CBs, suggesting that population density is one of the dominant factors in the noise pollution there. At the same time, a considerable comparison can be found between Q1 and either Q3 or Q4. Q1 is the only CB that has a high level of noise complaint and low population density. By contrast, the number of noise complaints of Q3 and Q4 are both low, but they are highly dense. Therefore, population density does not seem to be that powerful in these cases.

Based on my research, in general, Manhattan has the most severe noise pollution among all the boroughs, and some of the CBs in Brooklyn and the Bronx have higher risk of being affected by noise pollution. On the other hand, the situations of Queens and Staten Island are much better, and Staten Island seems to be the quietest borough in NYC based on the data that I studied. Since the elements of urban environment and society are very complicated and changing all the time, the urban soundscape can be influenced by many factors. Obviously population density is one of the most important factors, but not always the dominant one.

Tweet Collection and Clipping

The second data set is a group of 159,949 tweets collected in a two-week-long period. All the tweets were obtained via software developed by the Tufts UAL. In order to make the results more reliable, unnecessary tweets such as advertisements had already been filtered before use. The tweets were extracted from the Twitter application program interface (API) "Decahose" and saved as a .csv file that can be opened by Microsoft Excel. From Figure 10 we can see some of the sample: column B shows a series of identifying numbers for the users instead of their actual user name, and the text part of

each tweets can be seen in column C. Columns D and E contain the information on latitude and longitude coordinates, and column F tells the time when a tweet was posted.

	A	B	C	D	E	F
1	6.89E+17	3.21E+08	We're all on a different journey. #streetdreamsmag @ New York New York	-74.0064	40.7142	2016-01-18 00:00:58
2	6.89E+17	4.8E+08	TONIGHT????TONIGHT????NYC BIGGEST TS EVENT Tranny Strip NYC Party Su	-73.9997	40.75798	2016-01-18 00:01:04
3	6.89E+17	94039059	#Repost @greenlightradio with repostapp. ????????? Chopping it up with o	-73.9421	40.71499	2016-01-18 00:01:19
4	6.89E+17	19002016	#DUMBO ???????????????????????????? @ DUMBO Brooklyn https://t.co/;	-73.9899	40.70305	2016-01-18 00:01:22
5	6.89E+17	2.63E+08	Great to meet up and catch up with my friend and fellow #operasinger @am	-73.9789	40.75882	2016-01-18 00:01:24
6	6.89E+17	1.17E+09	Make the right call! Delight your guest without the work. Let @solacebar1 #	-73.9514	40.82508	2016-01-18 00:01:30
7	6.89E+17	14827129	Crowded for @salesforce presentation at suits_supply #nrf16 #retail #GeoM	-74.0009	40.72218	2016-01-18 00:01:32
8	6.89E+17	48211232	Matt's I can't believe I'm here during NFL playoffs face #truelove #marriage	-73.9709	40.76456	2016-01-18 00:01:39

Figure 10: Sample Tweets Collected in New York City

Since Twitter has a function that allows its users to post a tweet with the location information, I was able to geolocate all the tweets that I needed by using their geographical coordinates in ArcGIS. Considering that the rectangular area from which the original tweets were extracted is much larger than the actual area of NYC, I clipped the tweets with the land boundary of NYC to get rid of those unnecessary areas.

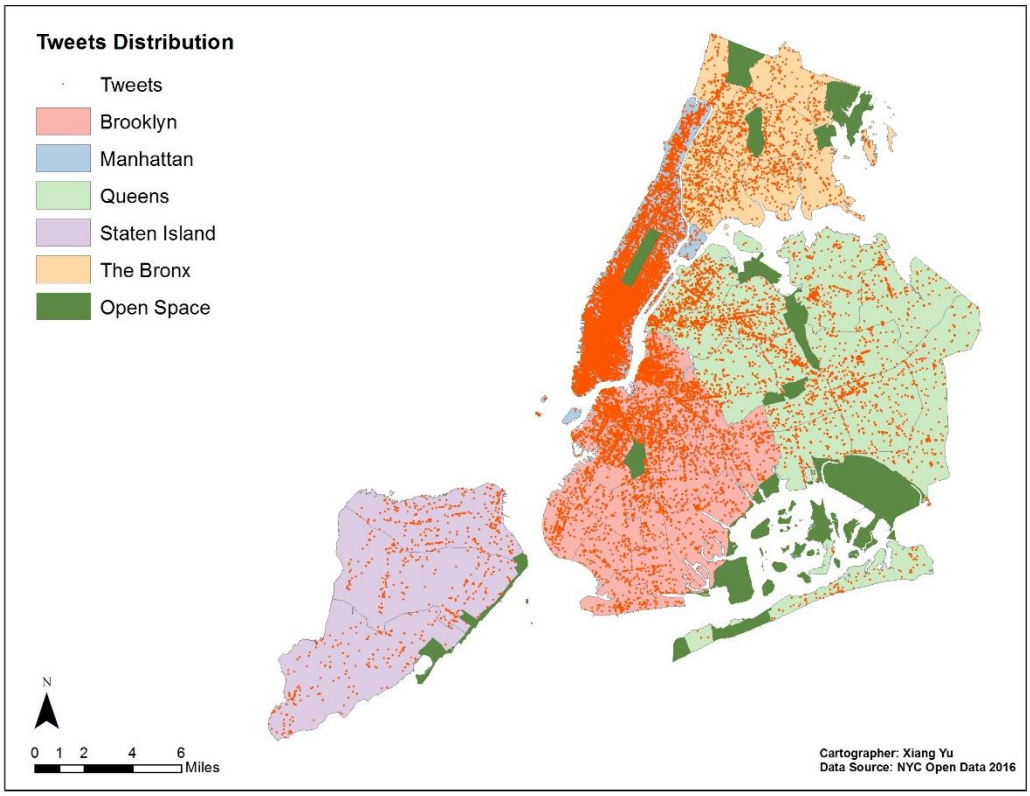

Figure 11: Map of Tweets Distribution

In Figure 11, we can see how the tweets were distributed in NYC, and each tiny dot stands for a tweet with geolocation information. The thickly dotted areas of Manhattan indicate that numerous tweets were posted there at that time, and the dots gradually faded away from the metropolitan area to the suburbs. Since the noise complaints were collected by CB, I did another clipping by using the CB boundary and then saved each CB's data respectively for the following sentiment analysis. Figure 12 shows what the tweets of Q1 and Q2 look like after the clipping.

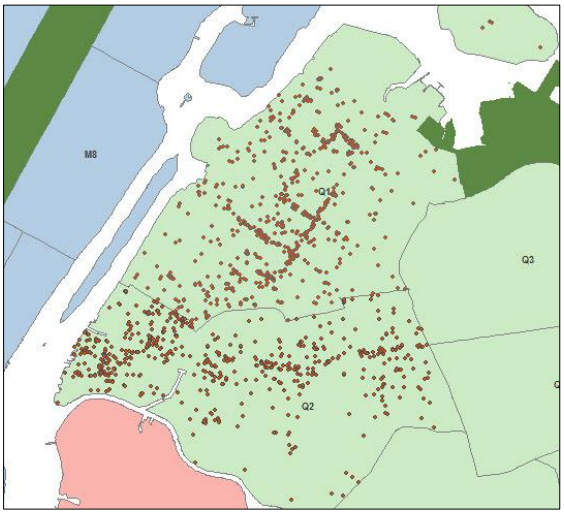

Figure 12: Tweets Clipped by Community Board

Sentiment Analysis

The sentiment analysis was conducted via the "Urban Attitude" software. This program was developed by my thesis advisor, Justin Hollander, and a Tufts student, Dibyendu Das, for aggregating public tweets and processing sentiment analysis. In Urban Attitude, there are two tools that can be used for sentiment analysis, which are called Tweet Analyzer and Text Analyzer. With Tweet Analyzer, researchers can analyze a group of tweets or a subset of tweets by setting either keywords or dates or both. In addition, it is possible for researchers to filter tweets with specific words using this tool.

By comparison, Text Analyzer has some of the functions that Tweet Analyzer has, but it can do the sentiment analysis much more quickly than Tweet Analyzer. Last but not least, both of the tools need to be linked up with a dictionary that can identify the sentiment score of different words before doing the sentiment analysis.

Here I applied the AFINN Dictionary in this study. It was developed by Finn Årup Nielsen, and the newest version of it contains a list of 2477 English words and phrases. Each of the words was evaluated with a sentiment score based on an ordinal scale between -5 (negative) and +5 (positive). For example, the word "catastrophic" was given a score of -4, whereas "brilliant" was given a score of +4. In addition, the dictionary was capable of capturing frequently used Internet slang, such as recognizing "LOL" as "laughing out loud" and "loooove" as "love." This dictionary has been applied to over 100 studies among various fields since it was developed. Examples include a comparative analysis of the sentiment of tweets before and after the presidential elections both in the US and France in the year 2012 (Nooralahzadeh, Arunachalam, and Chiru 2013), a study addressing the sentiment expression on Twitter during the San Bruno Gas Explosion in September 2010 (Nagy and Stamberger 2012), and the introduction of a neural network for classifying commercial advertisements based on their semantic and sentiment content (Abrahams et al. 2013).

I divided the process of sentiment analysis into two steps. The first step has already been talked briefly in last section: that I first joined the .csv file of all the tweets and geolocated them by using the latitude and longitude coordinates. Then I clipped those tweets by CBs and outputted them respectively for every CBs via an ArcGIS extension tool called "Split by Attribute" provided by USGS. The second step was to calculate each

CB's sentiment score by using the tool Text Analyzer. There are six pieces of valuable information summarized in the result: the total number of sentiment words and the number of sentiment-containing tweets found, the total positive and negative score, and the total number of positive and negative tweets or paragraphs. The positive score is obtained by summing up every single word's positive scores, and a negative score is the total of all the negative words' score. The rule of how to calculate the numbers of positive and negative tweets can be described as follows. Those tweets that contain words with positive sentiment scores ranging from +1 to +5 are considered as positive tweets/paragraphs. Similarly, the negative tweets/paragraphs are the ones containing words or phrases with a negative sentiment score. In addition, tweets containing both positive and negative words are counted by both sides. So the sum of positive and negative tweets is sometimes bigger than the total number of sentiment tweets.

The results varied quite significantly among different CBs. One possible reason could be that in a certain CB more tweets mean a higher sentiment score because of the way the sentiment is calculated. For example, M1 has the highest positive score of 27,466 calculated from 11,449 total sentimental tweets, but Q14 only has a positive score equal to 184 based on 89 sentiment tweets. Figure 13 shows the results of M1 as a sample of the sentiment analysis.

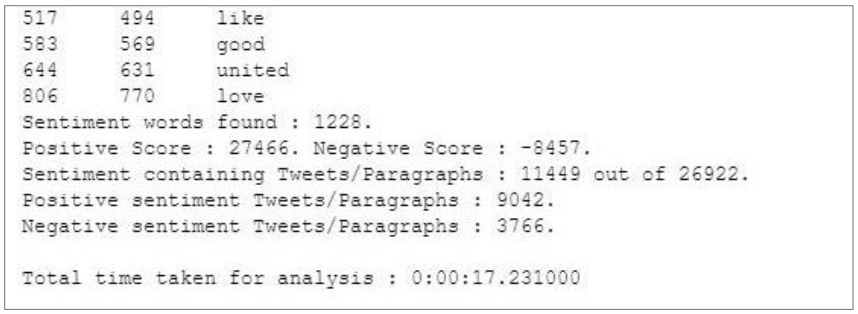

Figure 13: Sample Result of Sentiment Analysis

After that, I created a table to summarize all the data that I got including the noise-related complaints from 311 and the results of the sentiment analysis of the tweets. Since the table is too big to list here, I just put the data on Brooklyn as a sample in Table 1. For one thing, this summary table makes it easier and more straightforward to compare different variables. For another, this summary table is necessary for the following statistical analysis. As I mentioned in chapter 1, one of the research questions is about how the sentiment trend of tweets can reflect the impact of urban noise on people's health in NYC. In order to seek the answer for this question, I conducted a statistical analysis to see whether the tweets of those areas with a high level of noise complaints are more negative than in areas with a low level of noise complaints. If so, I could assume that living with high urban noise would give people more negative impacts, such as pressure and anxiety, which are reflected in their posts from the social media.

Community Board	Population	Complaints per 1000	Sentiment Tweets	Positive Score	Negative Score	Avg. Positive	Avg. Negative
Brooklyn							
B1	173083	2.05	2641	6559	-1906	2.48	-0.72
B2	99617	1.94	1830	4461	-1421	2.44	-0.78
B3	152985	1.59	718	1932	-565	2.69	-0.79
B4	112634	1.99	731	1775	-590	2.43	-0.81
B5	182896	0.83	560	1227	-646	2.19	-1.15
B6	104709	1.76	1436	3218	-1159	2.24	-0.81
B7	126230	0.88	1049	2072	-565	1.98	-0.54
B8	96317	2.65	523	1179	-385	2.25	-0.74
B9	98429	1.67	199	498	-105	2.50	-0.53
B10	124491	0.70	254	605	-246	2.38	-0.97
B11	181981	0.54	161	336	-165	2.09	-1.02
B12	191382	0.47	195	507	-284	2.60	-1.46

Table 1: Sample Summary Table: information of the CBs in Brooklyn

Correlation Coefficient Test

Besides illustrating the variables such as population density, noise complaints, and tweet sentiment score in GIS, a statistical analysis was also conducted via STATA in this thesis, in which two common correlation analyses, Pearson's correlation and Spearman's correlation, were applied. Both of these two correlation analyses are able to measure the relationship among variables, but there are some differences between them. In terms of their diagrams, Spearman's correlation usually measures the rank order of the points without caring exactly where the points are, while Pearson's coefficient usually measures the linear relationship between the variables. In addition, Spearman's correlation measures the statistical dependence between two variables by assessing how well their relationship can be presented by a monotone function, while Pearson's shows how well a straight line can describe the relationship between the variables. Therefore, Spearman's correlation is more suitable for analyses of non-normally distributed data.

Before doing the statistical analysis, it was very important for me to find the appropriate variables. As we discussed before, population density has a great impact on the number of noise complaints. For example, more densely populated CBs usually have more noise complaints than the ones in the suburbs. Thus, the CB's total noise complaints cannot represent the actual acoustic environment appropriately and objectively. Similarly, the total sentiment score can barely reflect the average sentimental trend of a CB because of its algorithm. Urban Attitude software has the capability to capture every single word and phrase included in the AFINN dictionary and then assign the sentiment score to it. After that, the software can track how many times the word appears in the entire text and multiply this number with the sentiment score given by the AFINN. So the total

sentiment score is decided in part by how many the tweets are in a certain area. For instance, the total positive score of M1 is over 20 times bigger than Q14.

In order to minimize impact from those interference factors such as population density and uneven tweet distribution, I made some conversions based on some variables that I collected before. First of all, I divided each CB's total complaints by its population of 2010 and then multiplied by 1,000, and I got the average noise complaints per thousand people. Second, I divided both the total positive score and the total negative score by the number of sentiment tweets and got the average positive and negative score, respectively, of each sentiment tweet. Third, since the negative sentiment scores are all negative, which might make the final result complicated to read, I decided to use the absolute value of those average negative scores rather than the original one. So in this case, the higher a number is, the stronger negation it contains. The mean of these three variables can be found in Table 2. Please notice that "ComPer1000" stands for the average noise complaints per 1,000 capita; "AvgPos" and "AvgNeg" represent the average positive score and negative score of each sentiment tweet, respectively; and "absAvgNeg" is the absolute value of AvgNeg.

	Mean	Std. Err
ComPer1000	1.45	0.12
AvgPos	2.29	0.03
AvgNeg	-0.94	0.03
absAvgNeg	0.94	0.03

Table 2: Mean and Standard Error of the Variables Used in the Correlation Tests

At first glance of the mean of ComPer1000, it seems not that bad to have fewer than two complaints for every 1,000 people, typically in a huge city like NYC. However, the actual conditions might be not as good as that number showed. As we can see from

Figure 7, nearly all NYC's metropolitan area is covered with a high level of noise complaints, while the suburbs have much fewer complaints. So the average number could dilute that huge gap of noise distribution between the downtown and the suburbs. According to the average sentiment score in Table 2, it seems that the tweets used in this study are generally more positive.

From my perspective, those newly created variables are capable of showing the average intensity of noise complaints and sentiment of tweets more objectively, and thus contribute to the accuracy of the statistical analysis. In Figure 14, I tested the normality of these variables via STATA. We can see that all the variables that might have an impact on the result of the statistical analysis are not normally distributed.

For the statistical analysis, I raised two hypotheses, and the cutoff for the statistical significance was 0.05. The first hypothesis was that the average noise complaint and population density should be positively correlated, indicating that the increasing population density should create more noise and bring more related complaints. For the second hypothesis, I assumed that the average noise complaint should be correlated with the average sentiment score. Furthermore, the noise complaint should be correlated negatively with the average positive sentiment score and have the opposite correlation with the average negative sentiment score, meaning that urban noise has bad effects on people's mental health and makes their tweets more negative.

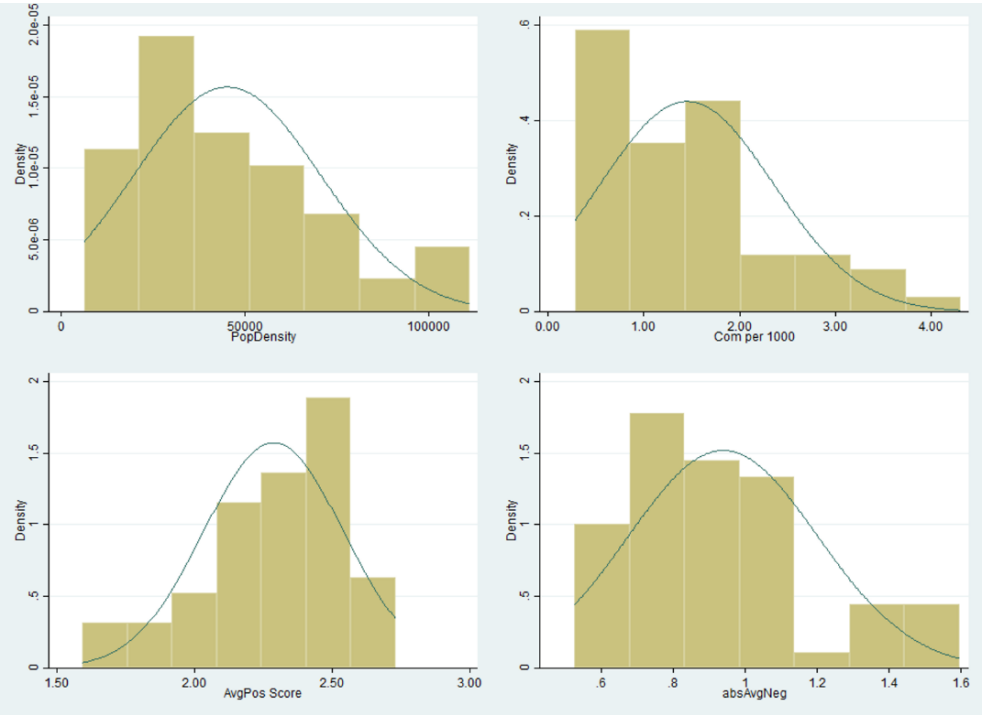

Figure 14: Histograms of the Variables Used in the Correlation Tests

CHAPTER 4: RESULTS

In this chapter, I will introduce the findings of this study. First I will look at the relationship between population density and noise complaint. From Figure 9 in chapter 3, the bivariate map, we can see that most of the highly dense CBs also attract more noise complaints, indicating that these two variables, noise complaints and population density, should have similar distributions and be correlated with each other. However, it was hard for me to prove the strength of the correlation using only that map. As I mentioned before, the Quantile Cut-Point in ArcGIS is more or less controversial because of the lack of classification accuracy. The quantile cut-points are the numbers that divide the entire sample pool equally into different categories based on the ascending order. So it is possible to set two variables with tiny differences into two adjacent different groups when they just stay at the edges beside the cut-point. For example, M1 is the CB that has the fewest complaints (121) among the CBs with a medium level of noise complaints, while B18 was classified as a CB with low noise complaints even though the difference between these two CBs was only four complaints.

Therefore, I did a further statistical test for those two variables for determining whether the bivariate map is accurate. First, the result of the Pearson's analysis showed that those two variables are significantly correlated with an r-value of 0.50 and a nil p-value. Then the Spearman's analysis revealed the same result by showing similar numbers to those of the Pearson's analysis ($p = 0.00$, $r = 0.61$). So my first hypothesis that noise complaint and population density are positively correlated has been confirmed, which means that when a CB's population density increases, the noise complaints there will goes up as well.

After that, the same statistical tests were applied to test the relationships between the average noise complaint and the tweets' average sentiment scores. As I introduced in chapter 4, three variables, ComPer1000, AvgPos, and absAvgNeg, were used for this analysis. According to my previous hypotheses, I suggested that the average noise complaint should be negatively correlated with the average positive score but be positively correlated with the average positive score, namely that people's tweets should become more negative when they experience a lot of loud noise.

	ComPer1000	**AvgPos**	**absAvgNeg**
Pearson's			
ComPer1000	r = 1; p = N/A		
AvgPos	r = 0.33; p = 0.01**	r = 1; p = N/A	
absAvgNeg	r = -0.48; p = 0.00**	N/A	r = 1; p = N/A
Spearman's			
ComPer1000	r = 1; p = N/A		
AvgPos	r = 0.34; p = 0.01**	r = 1; p = N/A	
absAvgNeg	r = -0.55; p = 0.00**	N/A	r = 1; p = N/A

*Table 3: Results of Correlation Analysis of Average Complaints and Sentiment Score (**: P < 0.05)*

Surprisingly, the results (Table 3) of the statistical test were totally inverse with my assumption. For the Pearson's analysis, the correlation between ComPer1000 and AvgPos is statistically significant, and the positive r-value (0.33) suggests that the tweets tend to be more positive when the average noise complaints go up. On the contrary, the r-value between ComPer1000 and absAvgNeg is negative (-0.48), meaning the sentiment of the tweets becomes less negative when the average noise complaints go up. Meanwhile, a similar result was also found in the Spearman's test: ComPer1000 is positively correlated with AvgPos (r = 0.34, p = 0.01) while negatively correlated with absAvgNeg (r = -0.55, p = 0.00), and both of them are statistically significant. As a

result, my hypothesis of the correlation between average noise complaints per 1,000 capita and average sentiment score of the tweets was clearly rejected.

To sum up, among my three hypotheses, only the first one was approved, and the other two were rejected based on the results of statistical tests. From the results, we saw that the CBs' population density and average noise complaints moved correspondingly in the same direction, and it was quite surprising to see that the correlation between average noise complaints and sentiment score was much beyond my expectation. Frankly speaking, it is not easy to believe that the increasing number of noise complaints can bring the happier sentiment of the tweets people posted. In order to get a sense of how the tweets' sentiment varies from one CB to another, I made two other GIS maps using the same method that I introduced in chapter 3.

To some degree, the results of the statistical analysis were supported by the maps presented in Figures 15 and 16. As we can see from Figure 15, nearly all the CBs in Manhattan are covered by a high level of average positive score, and some CBs in Queens and Brooklyn, known as the extended metropolitan areas, also have the same level of noise. So the positive sentiment trend just followed the distribution of noise complaints, which went in the opposite direction from my previous assumption.

However, some special CBs also drew my attention. S3 and B7 were two examples. S3 is at the bottom left corner of Figure 15, with a high level of positive scores. According to Figure 7, the noise complaints map, S3 was one of the quietest CBs of the city. If we look at this CB from a satellite map, we can see that the communities there are quite low-density, surrounded by several large open spaces. In addition, S3 stays far away from the metropolitan area, and thus becomes much quieter, but it also has a

high positive score. On the contrary, B7 is a quiet CB near the Upper Bay to the south of Manhattan, and its sentiment score tends to be more negative than the surrounding areas. In terms of building environment, I didn't find many differences between this CB and the other surrounding ones. More than that, it has a good view of the ocean, and the community there looks pretty tidy from the street view of Google Map. So it is hard to determine why the positive score there is lower than in the surrounding areas. However, as I learnt from the City-Data website, the percentages of Asian and Hispanic residents of B7 are much higher than the surrounding CBs. So I am wondering whether this demographic difference can be a possible reason.

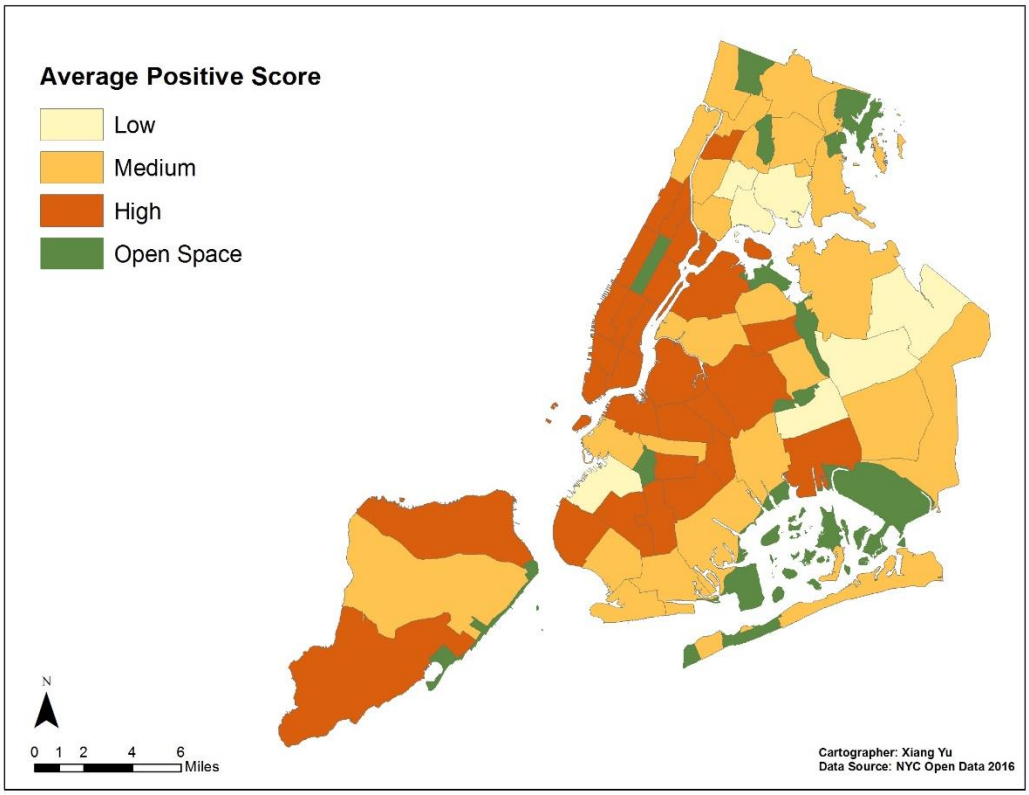

Figure 15: Map of Average Positive Sentiment Score in New York City

Meanwhile, as we can see from Figure 16, the distribution of the average negative score is clearly opposite to the average positive score. CBs at the eastern and southern

edge of the city tend to be covered with the lightest color, meaning that the general sentiment score there is more negative than the other CBs. At the same time, the color gets darker and darker from the suburb to the downtown area, which matches the result that I got from the statistical test that the sentiment scores of those loud areas are less negative. Nonetheless, it should be pointed out that some of the CBs don't follow this general rule. First, CBs at the bottom of Brooklyn including B13, B15 and B18 have the less negative scores than the CBs just above them. As we know from Figure 7, these CBs are considered as quiet ones based on the number of noise complaints that they have. If the negative sentiment score is positively correlated with noise complaint, then these CBs could be the opposite examples. When I looked at those CBs' features, I noticed that they were closed to beaches and parks that probably makes them good places for leisure. So people might post something happy when they were relaxing themselves there. Second, I took B7 and S3 as examples again. As these two maps show, both B7's positive and negative score are at the low level, indicating the people there might like to post tweets in a relatively neutral sentiment. By contrast, S3' negative score remains the high level as same as its positive score. Since this CB's population density is relatively low and it also has fewer noise complaints. I believe that noise may not be an important factor in this case, and the reason why S3' has the high score in both positive and negative scores may need future investigation.

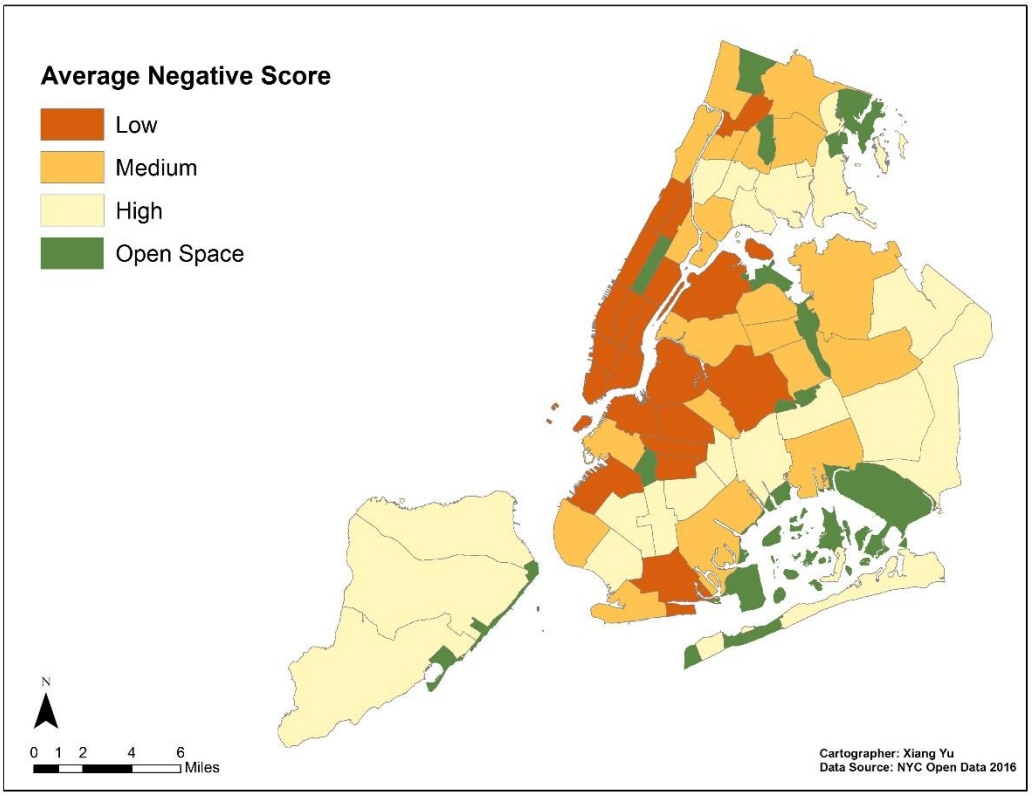

Figure 16: Map of Average Negative Sentiment Score in New York City

CHAPTER 5: CONCLUSION

Thanks to the help from my advisor and reader, most of my original ideas were applied to this study, and the flow of this study moved quite smoothly. In general, this thesis has applied a sentiment analysis of tweets and demonstrated that urban noise in NYC was correlated with population density and tweet sentiment.

For my first research question, "How does urban noise influence public health, particularly people's mental health and sentiment?" the answer is complicated. According to those studies I have read, loud noise hazard will obviously hurt human ear organs and cause noise-introduced hearing loss. Meanwhile, there are few pieces of evidence showing that urban noise has direct impacts on human mental health, but it does cause sleep disturbance and more stress, and thus contribute to cardiovascular disease or hypertension.

To address my second research question, "Does the distribution of noise complaints match up with population density?" two kinds of approaches, GIS illustration and statistical analysis, were applied, and the answer to this question is "Yes," based on the results. The average noise complaints and population density are positively correlated, and this relationship is statistically significant.

For my last research question, "How can the sentiment of tweets reflect this kind of noise impact in NYC?" I first did the statistical analyses of two pairs of converted variables, and then compared the pattern of their distribution by observing the GIS maps. The answer, which is totally opposite to my expectation, is that tweet sentiment will become more positive when noise complaint goes up and will become more negative when noise complaint drops down.

To sum up, all of my research questions were successfully addressed, and the results were relatively easy to understand. However, there are some limitations that affect the objectivity and authenticity of this study. In the following section, I discussed the limitations that I encountered, as well as the recommendations for future research.

Limitations and Recommendations

Sentiment analysis of social media is still a lately rising form of qualitative research that hardly be considered as perfectly developed. It necessarily involves subjective factors and can always be improved by applying it more widely. The first limitation of my thesis is the period that I used for collecting tweets. Although nearly 16,000 tweets were collected for this study, the two-week-long period was not an ideally long period for predicting the noise pollution of the entire city. As we all know, social activities vary dramatically in different seasons in cities, with a clearly seasonal change in NYC. The tweets for this study were collected in winter, from January 11 to 25, which would be the coldest time of the year. Therefore, people might prefer to stay at home rather than to go out, and there is less ongoing construction in winter. By contrast, there are much more activities and events going on during summertime, which will doubtless bring more people, more traffic, and thus louder noise. At the same time, construction like road maintenance usually takes place in summer, which also destroys the city's soundscape.

An additional limitation is that tweets are not restricted to local residents of NYC, while the noise complaints via calling 311 probably are. Since the "Big Apple" is a well-known destination for tourists, millions of foreign and American tourists come to the city every year. Numerous tweets are posted or retweeted by those tourists while they are

visiting the city, and those tweets will necessarily be added to the tweet pool of sentiment analysis. As a result, we cannot say that the tweets we collected can represent the feeling experienced by local residents. Besides, it is very popular among the young generation, who are the majority of social media active users, to share their travel experience by posting a tweet with the location of that attraction. At the same time, those tweets probably contain positive words and phrases, as people usually feel excited when they visit a new place. For example, a young girl may post a tweet containing words like "fancy" and "beautiful" when she walks in the shops along Fifth Avenue, a couple may post a tweet like "The food here is super great/delicious" when they hit a good restaurant at K-town, or a first-time visitor may speak highly of famous attractions such as the Metropolitan Museum and the Empire State Building. All these behaviors will contribute to the overall sentiment scores of tweets, and thus have an impact on my study. In addition, the 311 data could also have the similar problem that the data might not be able to represent the general residents. For example, some senior citizens may be more sensitive to noise pollution and thus call the 311 service very frequently. Or someone may keep complaining online because he or she have to stay close to an unbearable noise "hotspot".

To address this problem, future researchers might find ways to categorize tweets based on their different characteristics. Taking Andrew Wiley et al. (2015)'s idea as an example, researchers might try to track users who constantly post from the study area and extract tweets from these uses as the local sample. Alternatively, researchers might find a way to check out whether the location of a single tweet matches up with the location on its user's profile. Likewise, tweets from tourists might also be judged by how often a

private account posts from the city being investigated. For future studies, it is worth considering the effect experienced by both local residents and tourists, or perhaps comparing these two groups to see what the differences are.

Furthermore, the software and the dictionary that I used for this study also have some language-based limitations. First of all, the dictionary is still being upgraded and improved, so it is unable to capture and recognize the sentiment words beyond its list and generally cannot cover the full range of slang and vulgarities employed in tweets, especially those with trendy Internet slang words. For example, people might think of the word "dafuq" as a legitimate word when the first time they see it, but actually it has the same meaning as "WTF" on the Internet. Although AFINN does include some slang words and their varieties, as I mentioned in Chapter 3, it is incapable of recognizing words like "dafuq" which express a negative sentiment.

Similarly, as Wiley et al. (2015) described in his thesis, the UAL software would treat a tweet including a phrase like "never good" as a tweet expressing a positive sentiment by only capturing the word "good" and giving it a value of +3. However, the intent of the writer is to describe something negatively. Correspondingly, the overall sentiment score will definitely be misrepresented if this case happens frequently. I also heard from people in the UAL that our software has troubles with recognizing contractions. For example, it might consider the word "won't" to be the word "won" and the letter "t," then give a positive value to it that incorrectly contributes to the total positive score, whereas the original word "won't" usually contains a neutral sentiment. Considering that the word "won't" is commonly used in English, this kind of imperfection would probably leave a significant impact on the total sentiment score in

some cases. Meanwhile, the software is not as smart as humans in recognizing some rhetoric, such as metaphor and sarcasm, and then understanding the actual emotion behind it.

In order to solve these problems, one effective way would be expanding the dictionary's coverage as much as possible updating it more frequently. I noticed that the AFINN dictionary does include some phrases like "not good" or "not working," and gives those phrases a correspondingly negative value, but not recently emerging Internet slang. Additionally, the way of expressing and even the meaning of a single word always vary in different places and countries, though the people there are speaking English as the first language, which makes it not that easy to add all of them into the dictionary. Therefore, a possible solution would be allowing researchers to modify the list manually if it is necessary. If so, they might be able to add words or phrases they find with specific sentiment into the software and make the results more precise. Or program developers could improve the software by adding new algorithms that can evaluate the sentiment of an entire sentence rather than just focusing on single words (Wiley et al. 2015).

Implications for Urban Policy and Planning

As a graduate student majoring in urban policy and planning, I want this thesis to have some inspiration for future urban studies. As the population and the number of vehicles are increasing in most of the mega cities, the living environment of those cities may become worse and worse if there is no appropriate restriction of those uncontrolled developments. With respect to my first research question and the first section of my literature review, the hazard of urban noise should draw more attention from the public. Policy makers and urban planners should view the importance of urban noise control as

the same as other planning elements, like land-use planning or transportation planning. Meanwhile, urban noise control could work together with things like airline planning, traffic management, and land-use control for improving the entire urban environment and the unbalanced urban function.

The GIS part of this study shows how powerful the GIS software could be in converting and illustrating data with geolocation information. With the increase in data access and user-friendly functions, GIS is playing an increasingly important role in urban planning and decision support systems (Yeh 1999). Urban planners can use GIS both for spatial analysis and modelling, and then display the key issue by different types of map. I did a lot of data conversion and management through ArcGIS for this study. But advanced functions like spatial analysis using raster data were not applied here. So, for future studies, researchers may consider using GIS software to deal with the data that they use and making GIS maps to illustrate the issues that concern them.

The rest of this thesis was mainly focused on the sentiment analysis, which could be considered the most important part. As Hollander and Renski (2015) describe, microblogging data has provided rich opportunities for urban social science research, and several studies suggest that sentiment analysis of social media has useful implications for public policy and urban planning. Although the 311 noise complaint data can only partially represent the urban noise of NYC, and a two-week period of tweets is unable to reflect the sentiment of the general public, this study shows how to bridge the gap and make connections between public service and social media datasets, which would be enlightening for future researchers and urban planners.

REFERENCES

Abrahams, Alan S, Eloise Coupey, Eva X Zhong, Reza Barkhi, and Pete S Manasantivongs. 2013. "Audience targeting by B-to-B advertisement classification: A neural network approach." *Expert systems with applications* 40 (8):2777-2791.

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. "Sentiment analysis of twitter data." Proceedings of the workshop on languages in social media.

Babisch, Wolfgang. 2002. "The noise/stress concept, risk assessment and research needs." *Noise and health* 4 (16):1.

Blasio, Bill, and Emily Lloyd. 2007. A Guide to New York City's Noise Code. edited by NYC Environmental Protection.

Ciuccarelli, Paolo, Giorgia Lupi, and Luca Simeone. 2014. *Visualizing the data city: social media as a source of knowledge for urban planning and management*. Cham: Springer.

Clark, Charlotte, Rocio Martin, Elise Van Kempen, Tamuno Alfred, Jenny Head, Hugh W Davies, Mary M Haines, Isabel Lopez Barrio, Mark Matheson, and Stephen A Stansfeld. 2006. "Exposure-effect relations between aircraft and road traffic noise exposure at school and reading comprehension The RANCH project." *American Journal of Epidemiology* 163 (1):27-37.

Congress, The US. 2015. The Quiet Communities Act. edited by US Environmental Protection Agency.

EPA, US. 1972. Noise Abatement and Control. edited by US Environmental Protection Agency.

Facebook. 2016. "Company Info: Stats." http://newsroom.fb.com/company-info/.

Frias-Martinez, V., and E. Frias-Martinez. 2014. "Spectral clustering for sensing urban land use using Twitter activity." *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE* 35:237-245. doi: 10.1016/j.engappai.2014.06.019.

Gershon, Robyn R. M., Richard Neitzel, Marissa A. Barrera, and Muhammad Akram. 2006. "Pilot Survey of Subway and Bus Stop Noise Levels." *Journal of Urban Health* 83 (5):802-812. doi: 10.1007/s11524-006-9080-3.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford* 1:12.

Hanzl, Malgorzata. 2007. "Information technology as a tool for public participation in urban planning: a review of experiments and potentials." *Design Studies* 28 (3):289-307.

Hasan, Samiul, Xianyuan Zhan, and Satish V Ukkusuri. 2013. "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media." Proceedings of the 2nd ACM SIGKDD international workshop on urban computing.

Hilbert, Martin, and Priscila López. 2011. "The world's technological capacity to store, communicate, and compute information." *science* 332 (6025):60-65.

Ising, H 1, and B Kruppa. 2004. "Health effects caused by noise: evidence in the literature from the past 25 years." *Noise and Health* 6 (22):5.

Johnson, Steven. 2010. "What A Hundred Million Calls To 311 Reveal About New York." WIRED, Last Modified November 1 2010.
Kaplan, Andreas M, and Michael Haenlein. 2010. "Users of the world, unite! The challenges and opportunities of Social Media." *Business horizons* 53 (1):59-68.
Kheirbek, Iyad, Kazuhiko Ito, Richard Neitzel, Jung Kim, Sarah Johnson, Zev Ross, Holger Eisl, and Thomas Matte. 2014. "Spatial variation in environmental noise and air pollution in New York City." *Journal of Urban Health* 91 (3):415-431.
Kim, Junghyun, and Jong-Eun Roselyn Lee. 2011. "The Facebook Paths to Happiness: Effects of the Number of Facebook Friends and Self-Presentation on Subjective Well-Being." *Cyberpsychology, Behavior, and Social Networking* 14 (6):359-364. doi: 10.1089/cyber.2010.0374.
Kouloumpis, Efthymios, Theresa Wilson, and Johanna D Moore. 2011. "Twitter sentiment analysis: The good the bad and the omg!" *Icwsm* 11:538-541.
Lee, E. Y., M. Jerrett, Z. Ross, P. F. Coogan, and E. Y. W. Seto. 2014. "Assessment of traffic-related noise in three cities in the United States." *Environmental Research* 132:182-189. doi: 10.1016/j.envres.2014.03.005.
Lexalytics. 2016. "Sentiment Analysis Definition ". https://www.lexalytics.com/technology/sentiment.
Liu, Bing. 2015. *Sentiment analysis: mining opinions, sentiments, and emotions*. New York, NY: Cambridge University Press.
McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. 2012. "Big data." *The management revolution. Harvard Bus Rev* 90 (10):61-67.
Moudon, Anne Vernez. 2009. "Real noise from the urban environment: how ambient community noise affects health and what can be done about it." *American journal of preventive medicine* 37 (2):167-171.
Nagy, Ahmed, and Jeannie Stamberger. 2012. "Crowd sentiment detection during disasters and crises." Proceedings of the 9th International ISCRAM Conference.
Nakov, Preslav, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. "Semeval-2013 task 2: Sentiment analysis in twitter."
Nooralahzadeh, Farhad, Viswanathan Arunachalam, and Costin Chiru. 2013. "2012 Presidential Elections on Twitter--An Analysis of How the US and French Election were Reflected in Tweets." Control Systems and Computer Science (CSCS), 2013 19th International Conference on.
NYC. 2016. "311 Service Requests from 2010 to Present." https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9.
Paulos, Eric, Ian Smith, and RJ Honicky. 2016. "Partcipatory Urbanism." Last Modified January 22, 2016.
President, Executive Office of the. 2012. "Big Data Across the Federal Government."
Rabinowitz, Peter M, Deron Galusha, Christine Dixon-Ernst, Martin D Slade, and Mark R Cullen. 2007. "Do ambient noise exposure levels predict hearing loss in a modern industrial cohort?" *Occupational and environmental medicine* 64 (1):53-59.
Raimbault, Manon, and Daniele Dubois. 2005. "Urban soundscapes: Experiences and knowledge." *Cities* 22 (5):339-350.

Rehan, Reeman Mohammed. 2015. "The phonic identity of the city urban soundscape for sustainable spaces." *HBRC Journal*.

Sagl, Günther, Bernd Resch, and Thomas Blaschke. 2015. "Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities." *Sensors* 15 (7):17013-17035.

Schreckenberg, D, T Eikmann, F Faulbaum, E Haufe, C Herr, M Klatte, M Meis, U Möhler, U Müller, and J Schmitt. 2011. "NORAH–Study on Noise-Related Annoyance, Cognition and Health: A transportation noise effects monitoring program in Germany." 10th International Congress on Noise as a Public Health Problem.

Schweitzer, Lisa. 2014. "Planning and social media: a case study of public transit and stigma on Twitter." *Journal of the American Planning Association* 80 (3):218-238.

Seong, Jeong C, Tae H Park, Joon H Ko, Seo I Chang, Minho Kim, James B Holt, and Mohammed R Mehdi. 2011. "Modeling of road traffic noise and estimated human exposure in Fulton County, Georgia, USA." *Environment international* 37 (8):1336-1341.

Seto, Edmund Yet Wah, Ashley Holt, Tom Rivard, and Rajiv Bhatia. 2007. "Spatial distribution of traffic induced noise exposures in a US city: an analytic tool for assessing the health impacts of urban planning decisions." *International journal of health geographics* 6 (1):1.

Silva, Thiago H, Pedro OS Vaz de Melo, Jussara M Almeida, and Antonio AF Loureiro. 2014. "Large-scale study of city dynamics and urban social behavior using participatory sensing." *Wireless Communications, IEEE* 21 (1):42-51.

Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. 2012. "Big data: Big gaps of knowledge in the field of internet science." *International Journal of Internet Science* 7 (1):1-5.

Stevens, Joshua. 2015. "Bivariate Choropleth Maps: A How-to Guide." Accessed February 18, 2015. http://www.joshuastevens.net/cartography/make-a-bivariate-choropleth-map/.

Sygna, Karin, Gunn Marit Aasvang, Geir Aamodt, Bente Oftedal, and Norun Hjertager Krog. 2014. "Road traffic noise, sleep and mental health." *Environmental Research* 131:17-24. doi: http://dx.doi.org/10.1016/j.envres.2014.02.010.

Twitter. 2015. "Twitter Usage & Company Facts." https://about.twitter.com/company.

Vianna, K. M. D., M. R. A. Cardoso, and R. M. C. Rodrigues. 2015. "Noise pollution and annoyance: An urban soundscapes study." *Noise & Health* 17 (76):125-133. doi: 10.4103/1463-1741.155833.

Walker, Ben. 2015. "Every Day Big Data Statistics - 2.5 Quintillion Bytes of Data Created Daily." Last Modified April 5 2015. http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/.

Westerkamp, Hildegard. 1991. "The World Soundscape Project." *The Soundscape Newsletter* No. 01.

WHO. 1997. "Prevention of Noise-induced Hearing Loss." *Report of An Informal Consultation*.

Wiley, Andrew, Urban Tufts University. Department of, Policy Environmental, and Planning. 2015. Shrinking Cities and Subjective Well-Being: An Investigation of

Resident Attitudes and Opinions via Micro-Blog Sentiment Analysis. Ann Arbor: ProQuest Dissertations & Theses.

Yeh, Anthony G-O. 1999. "Urban planning and GIS." *Geographical Information Systems* 2:877-888.

APPENDIX 1: SUMMARY TABLE

Community Board	Population	Complaint Total	Complaints per 1000	Sentiment Tweets	%Positive	%Negative	Positive Score	Negative Score	Avg. Positive	Avg. Negative
Brooklyn										
B1	173083	355	2.05	2641	71.7%	28.3%	6559	-1906	2.48	-0.72
B2	99617	193	1.94	1830	68.9%	31.1%	4461	-1421	2.44	-0.78
B3	152985	243	1.59	718	71.5%	28.5%	1932	-565	2.69	-0.79
B4	112634	224	1.99	731	68.8%	31.2%	1775	-590	2.43	-0.81
B5	182896	151	0.83	560	64.0%	36.0%	1227	-646	2.19	-1.15
B6	104709	184	1.76	1436	69.1%	30.9%	3218	-1159	2.24	-0.81
B7	126230	111	0.88	1049	75.0%	25.0%	2072	-565	1.98	-0.54
B8	96317	255	2.65	523	71.9%	28.1%	1179	-385	2.25	-0.74
B9	98429	164	1.67	199	77.4%	22.6%	498	-105	2.50	-0.53
B10	124491	87	0.70	254	72.0%	28.0%	605	-246	2.38	-0.97
B11	181981	99	0.54	161	66.8%	33.2%	336	-165	2.09	-1.02
B12	191382	90	0.47	195	63.9%	36.1%	507	-284	2.60	-1.46
B13	104278	76	0.73	145	61.7%	38.3%	301	-144	2.08	-0.99
B14	160664	253	1.57	692	65.9%	34.1%	1722	-752	2.49	-1.09
B15	159650	136	0.85	181	70.4%	29.6%	409	-106	2.26	-0.59
B16	86468	100	1.16	278	69.9%	30.1%	739	-292	2.66	-1.05
B17	155252	200	1.29	1792	64.8%	35.2%	4494	-2478	2.51	-1.38
B18	193543	117	0.60	263	63.5%	36.5%	556	-258	2.11	-0.98
Bronx										
BX1	91497	113	1.24	150	64.0%	36.0%	304	-146	2.03	-0.97
BX2	52246	115	2.20	125	52.9%	47.1%	224	-169	1.79	-1.35
BX3	79762	125	1.57	131	53.8%	46.2%	209	-209	1.60	-1.60
BX4	146441	249	1.70	265	64.3%	35.7%	587	-276	2.22	-1.04
BX5	128200	245	1.91	205	70.9%	29.1%	529	-164	2.58	-0.80
BX6	83268	149	1.79	121	66.4%	33.6%	252	-116	2.08	-0.96
BX7	139286	280	2.01	223	66.1%	33.9%	509	-172	2.28	-0.77

BX8	101731	143	1.41	200	64.6%	35.4%	456	-199	2.28	-1.00
BX9	172298	205	1.19	274	51.1%	48.9%	451	-437	1.65	-1.59
BX10	120392	81	0.67	275	66.7%	33.3%	558	-288	2.03	-1.05
BX11	113232	173	1.53	628	69.0%	31.0%	1391	-601	2.21	-0.96
BX12	152344	140	0.92	195	64.7%	35.3%	449	-186	2.30	-0.95
Manhattan										
M1	60978	121	1.98	11449	70.6%	29.4%	27466	-8457	2.40	-0.74
M2	90016	238	2.64	5309	73.4%	26.6%	12792	-3349	2.41	-0.63
M3	163277	510	3.12	3233	72.3%	27.7%	8228	-2157	2.55	-0.67
M4	103245	320	3.10	3679	73.2%	26.8%	9274	-2511	2.52	-0.68
M5	51673	223	4.32	9732	76.2%	23.8%	26381	-5529	2.71	-0.57
M6	142745	264	1.85	1191	75.1%	24.9%	2907	-761	2.44	-0.64
M7	209084	277	1.32	2077	75.6%	24.4%	5023	-1279	2.42	-0.62
M8	219920	336	1.53	1613	70.4%	29.6%	4049	-1285	2.51	-0.80
M9	110193	356	3.23	851	75.7%	24.3%	2128	-503	2.50	-0.59
M10	115723	392	3.39	1117	72.5%	27.5%	2740	-788	2.45	-0.71
M11	120511	254	2.11	396	67.6%	32.4%	960	-377	2.42	-0.95
M12	190020	661	3.48	1041	67.5%	32.5%	2338	-870	2.25	-0.84
Queens										
Q1	191105	293	1.53	1153	72.3%	27.7%	3147	-850	2.73	-0.74
Q2	113200	137	1.21	687	67.3%	32.7%	1509	-611	2.20	-0.89
Q3	171576	94	0.55	202	67.1%	32.9%	461	-178	2.28	-0.88
Q4	172598	105	0.61	277	69.9%	30.1%	680	-230	2.45	-0.83
Q5	169190	136	0.80	267	68.3%	31.7%	630	-208	2.36	-0.78
Q6	113257	124	1.09	271	62.6%	37.4%	590	-266	2.18	-0.98
Q7	247354	121	0.49	364	68.4%	31.6%	783	-303	2.15	-0.83
Q8	151107	134	0.89	274	69.5%	30.5%	524	-256	1.91	-0.93
Q9	143317	128	0.89	151	56.5%	43.5%	265	-171	1.75	-1.13
Q10	122396	57	0.47	238	71.6%	28.4%	588	-232	2.47	-0.97
Q11	116431	45	0.39	197	56.1%	43.9%	372	-303	1.89	-1.54
Q12	225919	92	0.41	354	63.2%	36.8%	763	-474	2.16	-1.34

Q13	188593	59	0.31	226	64.8%	35.2%	521	-254	2.31	-1.12
Q14	114978	84	0.73	89	65.0%	35.0%	184	-101	2.07	-1.13
State Island										
S1	175756	135	0.77	828	76.8%	34.8%	1960	-1092	2.37	-1.32
S2	132003	88	0.67	212	75.5%	33.3%	485	-226	2.29	-1.07
S3	160209	45	0.28	320	79.4%	33.0%	772	-340	2.41	-1.06

APPENDIX 2: RESULTS OF SENTIMENT ANALYSIS

Boroughs	CBs	Sentiment Words	Positive Score	Negative Score	Sentiment Tweets	Positive Tweets	Negative Tweets
Bronx	BX1	117	304	-146	150/411	110	62
	BX2	112	224	-169	125/304	81	72
	BX3	81	209	-209	131/264	84	72
	BX4	140	587	-276	265/679	193	107
	BX5	141	529	-164	205/679	168	69
	BX6	105	252	-116	121/679	89	45
	BX7	155	509	-172	223/688	162	83
	BX8	137	456	-199	200/688	146	80
	BX9	143	451	-437	274/688	162	155
	BX10	176	558	-288	275/629	208	104
	BX11	263	1391	-601	628/1578	489	220
	BX12	135	449	-186	195/1578	143	78
Brooklyn	B1	635	6559	-1906	2641/6467	2102	828
	B2	537	4461	-1421	1830/5079	1424	644
	B3	342	1932	-565	718/1719	579	231
	B4	336	1775	-590	731/1984	568	258
	B5	293	1227	-646	560/1421	412	232
	B6	372	3218	-1159	1436/3797	1118	499
	B7	306	2072	-565	1049/2561	893	289
	B8	261	1179	-382	523/1302	418	163
	B9	126	498	-105	199/716	168	49
	B10	162	605	-246	254/597	203	79
	B11	118	336	-165	161/338	123	61
	B12	158	507	-284	195/473	154	87
	B13	111	301	-144	145/410	103	64
	B14	373	1722	-752	692/1356	538	279
	B15	124	409	-106	181/483	138	58
	B16	198	739	-292	278/1356	221	95
	B17	501	4494	-2478	1792/3661	1400	761
	B18	167	556	-258	263/3661	186	107
Manhattan	M1	1228	27466	-8457	11449/26922	9042	3766
	M2	848	12792	-3349	5309/14660	4283	1553
	M3	657	8228	-2157	3233/8709	2591	991
	M4	683	9274	-2511	3679/9223	2956	1085
	M5	957	26381	-5529	9732/25265	8155	2544
	M6	399	2907	-761	1191/25265	991	329
	M7	521	5023	-1279	2077/25265	1714	553
	M8	509	4049	-1285	1613/3964	1292	544

	M9	345	2182	-503	851/1961	691	222
	M10	382	2740	-788	1117/2621	903	343
	M11	214	960	-377	396/1014	302	145
	M12	347	2338	-870	1041/2963	778	375
Queens	Q1	387	3147	-850	1153/2792	920	353
	Q2	278	1509	-611	687/1853	520	253
	Q3	138	461	-178	202/657	155	76
	Q4	160	680	-230	277/721	221	95
	Q5	186	630	-208	267/703	207	96
	Q6	189	590	-266	271/748	189	113
	Q7	184	783	-303	364/1019	273	126
	Q8	169	524	-256	274/523	216	95
	Q9	118	265	-171	151/467	91	70
	Q10	124	588	-232	238/2792	189	75
	Q11	120	372	-303	197/2792	128	100
	Q12	230	763	-474	354/2792	264	154
	Q13	153	521	-254	226/2792	162	88
	Q14	74	184	-101	89/2792	67	36
State Island	S1	314	1960	-1092	828/1757	6363	340
	S2	155	485	-226	212/536	160	80
	S3	205	772	-340	320/572	254	125