Developing a Data Mining Framework to Identify a Sense of Gentrification through Social Media Data: A Case Study Using Instagram Posts in Salt Lake City, Utah

---

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Geography with Concentration in Geographic Information Science

---

by

Cheng-Chia Huang

Fall 2017

SAN DIEGO STATE UNIVERSITY

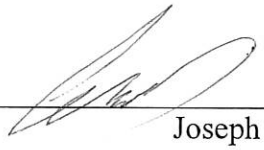
The Undersigned Faculty Committee Approves the

Thesis of Cheng-Chia Huang:

Developing a Data Mining Framework to Identify a Sense of Gentrification through Social Media Data: A Case Study Using Instagram Posts in Salt Lake City, Utah

Atsushi Nara (Chair)
Department of Geography

Ming-Hsiang Tsou
Department of Geography



Joseph Gibbons
Department of Sociology

7/11/2017

Approval Date

ABSTRACT OF THE THESIS

Developing a Data Mining Framework to Identify a Sense of Gentrification through Social Media Data: A Case Study Using Instagram Posts in Salt Lake City, Utah
by
Cheng-Chia Huang
Master of Science in Geography with a Concentration in Geographic Information Science
San Diego State University, 2017

Gentrification is a transformation of a working-class or abandoned area of a city under the influence of redevelopment and influx of higher-income residents, which involves economic upgrading and replacement of long-term residents who were often of lower social status. Researchers utilize various quantitative and qualitative methodologies to measure the gentrification dynamics; however, incorporating human perceptions of neighborhoods into a large-scale measurement has not been much explored. Moreover, there is a lack of research considering gentrification dynamics at a finer spatio-temporal scale across a large area. This thesis fills these gaps by introducing an innovative social media based data mining framework, which utilizes both qualitative and quantitative approaches by analyzing social media data to explore human activities and perceptions related to gentrification. As a case study, the framework was applied to Instagram data collected from Salt Lake City, UT to investigate gentrification dynamics by capturing the yearly change of a sense of gentrification at a spatial scale of census block groups.

There are two studies in this thesis. The first is a comparative study to examine gentrifying dynamics identified by quantitative and qualitative measures, which are based on five existing census-based gentrification typologies and social media driven *human-perceived gentrifying areas*. As a result, five typologies showed inconsistency in delineating gentrifying areas. Furthermore, they did not match with the gentrifying areas identified based on human perception. The second study introduces a novel data mining framework to examine gentrification dynamics utilizing social media data. Specifically, this study developed two gentrification indicators, *nightlife activities* and *gentrification ambience* to characterize the sense of gentrification. They were extracted from geotagged Instagram posts by employing text processing and text clustering techniques. The spatial distribution of those two indicators over years revealed that areas observed with a sense of gentrification closely correspond to the *human-perceived gentrifying areas*. Furthermore, clustering results delineated two types of gentrifications, residential-driven and commercial-driven gentrifications. Finally, the result portrayed the yearly changes of a sense of gentrification illustrating the potential gentrification expansion. This research demonstrates the framework's capability of incorporating human perception and social media data to explore spatially and temporally fine-grained gentrification dynamics.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

ACKNOWLEDGEMENTS

The writing of this thesis is a fantastic journey that I never could have imagined. I would not have been able to accomplish this research without those people who supported me throughout these two years. First, I would like to express my sincere gratitude to my advisor Dr. Atsushi Nara for his patience, encouragement, guidance, help, and all the learning opportunities he provided me. Also, I would like to express my appreciation to my committee members, Dr. Ming-Hsiang Tsou and Dr. Joseph Gibbons for their advice and revision. Finally, I would like to thank my parents, friends, the members of HDMA for their love and support.

Chapter 1: Introduction

1.1 Introduction

Gentrification, a term which was coined by Ruth Glass (1964), originally refers to an urban process that involves property price increase and replacement of working-class residents by the incoming middle classes. Literally, Glass used “gentry-fication” to describe a phenomenon when a new “urban gentry” moves back to inner cities and replaces the existing residents. The term, new “urban gentry” explains a group of people who are different from the conventional middle-class. Unlike affluent middle-class households who prefer to live in the suburbs, urban areas are more attractive to a new “urban gentry” because of the shorter commute, the higher return on housing investment, or the different lifestyle from the mundane suburban (Lees, Slater, & Wyly, 2008). Researchers also use different terms to refer the “urban gentry” such as “gentrifiers,” “yuppies” (Short, 1989), “hipsters” (Hae, 2011), or “creative workers (Florida, 2002).

Today, although different interprelations exist in the gentrification literature and various types of gentrification has been reported (Barton, 2016; Lees et al., 2008) generally gentrification is a transformation of a working-class or abandoned area of a city under the influence of redevelopment and influx of higher-income residents, which involves economic upgrading and replacement of long-term residents who were often of lower social status. Furthermore, the gentrification-related displacement involves both residential and cultural facets (Gibbons & Barton, 2016). Resident displacement refers to low-income residents of a neighborhood pushed out due to increasing housing costs (Anderson, 1990; J. Betancur, 2011; Lees et al., 2008), whereas the cultural displacement refers to the change of culture in the neighborhoods (Freeman, 2006; Zukin, 2011). Researchers also emphasized that this transformation is spatial-temporal dynamic since the spatial patterns and characteristics of neighborhoods continue to change during a gentrification process (Gale, 1980; Kerstein, 1990; Liu & O’Sullivan, 2016; O’Sullivan, 2002; Pattison, 1983; Torrens & Nara, 2007). Despite the rich gentrification literatures and their significant contributions to examine gentrification phenomena, previous studies were often focused on either qualitative or quantitative approach and not many employed a mixed method to take the advantages from both. Incorporating human perception is one major advantage from a qualitative perspective to detect the cultural shift resulted in gentrification. Barton (2016) pointed out that qualitative approaches can confirm not only the identified areas experienced a natural form of improvement such as incumbent economic upgrading but also a cultural change. Nevertheless, qualitative approaches are typically labor-intensive and time consuming when they are applied to a large area as compared to quantitative approaches. On the other hand, quantitative approaches utilize secondary data and systematic statistical methods to study gentrification, which can be applicable to a larger area but possibly overlook subtle cultural shifts.

In recent years, researchers have sought to bridge the gap between qualitative and quantitative methods. For example, Papachristos et al. (2011) used the number of coffee shops as the cultural measurement of gentrification in their quantitative study. While their method incorporated human perception in a quantitative way, the counts of neighborhood coffee shops only reflect the certain changing consumption patterns (Barton, 2016). In other words, the attempt to quantitatively measure cultural shifts did not examine other cultural characteristics of gentrification such as artists (Zukin, 1989), yoga (Kern, 2012), nightlife (Hae, 2011, 2011).

Another issue in previous gentrification studies is that spatial and temporal scales were often either confined or relatively coarse due to the limited data availability. For instance, plenty qualitative studies examined gentrification dynamics at a fine temporal scale, but only focused on a single or small gentrifying areas (Boyd, 2008; Hamnett & Whitelegg, 2007). Quantitative studies looked into a large area and calculated gentrification indices; nevertheless, the temporal scales of those indices were based on multi-year aggregated data (Ley, 1986; Ley & Dobson, 2008). Therefore, there is a lack of research considering gentrification dynamics at a finer spatio-temporal scale across a large area. Without examining the fine spatial and temporal scale across a large area, those studies overlooked the variance of gentrification dynamics during a short period in a large region.

The emergence of location-based social media now provides an opportunity to address these issues in gentrification studies. Social media has both “big” and “deep” characteristics (Sui & Goodchild, 2011). In the 20th century, social science relied on two types of data: “surface data” about lots of people and “deep data” about few individuals. “Surface data” is a big volume of data used in the disciplines that adapted quantitative methods like sociology, economics, political science, and computational geography to explore a phenomenon on a macro level. In contrast, “deep data” recording detail information about a small group of people have been used in the disciplines that adapted qualitative approaches such as psychology, anthropology, ethnography, and art history to explore a phenomenon on a micro level (Manovich, 2011). In other words, researchers had to sacrifice the details for constructing a macro picture, and vice versa. However, today, we can follow opinions, ideas, and feelings of hundreds of millions of people through social media (Manovich, 2011). As a consequence, we no longer need to make a choice from either data volume or data depth (Sui & Goodchild, 2011). Yet, in order to take the benefit of “big” and “deep” social media data, there is a need to develop an analytical framework to examine the fine-grained gentrification dynamics and to quantify human perception for identifying gentrifying areas across a large area.

1.2 Goals and Objectives

With the publicly available social media data, this thesis develops an effective and efficient data mining framework that utilizes both qualitative and quantitative approaches to identify and characterize the gentrifying areas. The framework incorporates human perception to recognize a sense of gentrification at the finer spatial and temporal scale,

and it is easily applied to a large region. This innovative application will add to the current literature on gentrification, social media, urban geography, and space-time GIS. In addition, although discussing consequences of gentrification often lead to controversial debates, the analytical results about fine-grained gentrification dynamics acquired from the framework can support decisions for local residents, communities, and governments. For example, the gained knowledge can inform them gentrifying areas at the very early stage, which could further help preventing from or minimizing those potential negative impacts such as displacement, segregation, and discrimination (Atkinson, 2000; J. J. Betancur, 2002; Butler & Robson, 2003; Davidson, 2010; Hae, 2011; Powell & Spencer, 2002; Slater, 2006; Wyly & Hammel, 2004).

1.3 Research Questions

To achieve the research goals, this thesis designed two research questions and conducted two studies to answer them.

(1) How accurate are existing measurements to identify gentrifying areas?

To answer the first research question, this study reviewed and evaluated the existing measurements. Specifically, it compared gentrifying dynamics identified by quantitative and qualitative measures. As a quantitative approach, it replicated five existing census-based gentrification typologies to quantify and map the spatial distribution of gentrifying areas. This study compared the results with those human-perceived gentrifying areas, which were determined by qualitatively analyzing online forums and news reports.

(2) Can social media data approach incorporating human perception better identify gentrifying areas as compared to conventional census-based typologies?

The second study first introduced a novel data mining framework that utilizes social media data to examine gentrification dynamics. To answer this research question, this study collected Instagram data in Salt Lake City from 2013/1/1-2015/12/31. Then, it constructed two gentrification indicators—nightlife activities and gentrification ambience— that facilitate capturing the sense of gentrification. Nightlife activities and gentrification ambience were quantified by applying text processing and text clustering techniques to the Instagram data. Two indicators of a sense of gentrification were then mapped to visually investigate their spatial distribution over the years and compared with human-perceived gentrifying areas.

Chapter 2: Literature Review

2.1 The Background of Gentrification Studies

The term, “Gentrification,” was coined by British sociologist Ruth Glass in 1964:

“One by one, many of the working class quarters of London have been invaded by middle-classes—upper and lower. Shabby, modest mews, and cottages—two rooms up

and two down—have been taken over, when their leases have expired, and have become more elegant, expensive residences" (Glass, 1964: 17-19)

Glass described gentrification from the aspect of class succession and pointed out the two main characteristics of gentrification: (1) sharp increase in house prices and rent in inner cities and (2) displacement, which involves an upper middle-class population displacing a working class population. Today, although researchers defined gentrification from various aspects and lack of consensus concerning the conceptualization of gentrification, most researchers agree that it is a transformation of a working-class or vacant area of a city under the influence of redevelopment and influx of higher-income residents, which involves economic upgrading and replacement of long-term residents who were often of lower social status (Barton, 2016; Lees et al., 2008; Torrens & Nara, 2007).

One of the main gentrification research topics is the cause of gentrification. For a long time, researchers have focused on urban dynamics including *urbanization*, *suburbanization and urban sprawl*. *Urbanization* refers to the process of population concentration in the dense cities. *Suburbanization* is the process of expansion when population move away from densely city areas to low-density areas, which leads to *urban sprawl* (Docampo, 2014). Yet, in the mid 20 century, Glass found a new urban phenomenon which involves the middle class moved to inner cities, housing value increase, and potential displacement. Over years, researchers tried to find the causes and usually explained it from the economic perspective or the cultural perspective.

Another research topic is the process of gentrification. Researchers found that gentrification is a dynamic process, and evaluated the characteristics of gentrification in different stages (Clay, 1979; Gale, 1980; Kerstein, 1990; National Association of Neighborhoods, 1980; National Urban Coalition, 1978; Pattison, 1983). Some scholars even developed models to simulate gentrification process (Liu & O'Sullivan, 2016; O'Sullivan, 2002; Torrens & Nara, 2007).

Debating on the gentrification consequence is another research area. Some researchers thought gentrification leads to positive consequences including tolerance, social mix, and stronger competitive capacity (Butler, 1997; Castells, 1983; Caulfield, 2016; Freeman & Braconi, 2004; Ley, 1994; Rose, 1984). Some researchers yet thought gentrification cause negative consequences such as displacement, discrimination, and segregation (Atkinson, 2000; J. J. Betancur, 2002; Butler & Robson, 2003; Davidson, 2010; Hae, 2011; Lees et al., 2008; Powell & Spencer, 2002; Slater, 2006; Wyly & Hammel, 2004).

The following subsections in Section 2.1 provide literature reviews on the three major topics including the cause of gentrification, the process of gentrification, and the consequences of gentrification.

2.1.1 The Causes

(1) Economic perspective

Two mainstream explanations of gentrification predominate the literature: the cultural perspective and the economic perspective. *The bid rent model* revised by Schill and Nathan (1983) and *the rent gap theory* created by Smith (1979) are based on the economic perspective. In the traditional *bid rent theory*, every resident only considers accessibility and space. High income or wealthy residents can afford the cost of transportation, so they usually choose to live in the suburbs for spacious areas. However, low-income residents live in the inner cities for accessibility at the expense of spacious houses. In their revised *bit rent model*, Schill and Nathan stated that gentrification is caused by a new type of people. They are so rich that they can afford spacious houses in the inner city and don't have to sacrifice both space and accessibility. Smith's rent gap theory is another common explanation for gentrification. The rent gap theory is based on the assumption that consumers desire to achieve a reasonable rate of return on their financial investment. The capitalized ground rent and the potential ground rent are two key concepts in this theory. The capitalized ground rent describes the actual economic return from those who have the right to use the land given the present land use. The potential ground rent represents the potential economic return from those who have the right to use the land if the land is put to its "highest and best" use. Once the gap between potential ground rent and capitalized ground rent is so large that the developers can afford the cost of purchase and renovation, redevelopment and gentrification will occur.

Many researchers accepted Smith's rent gap theory and proved the theory with their own operational definition of rent gap. For example, Clark (1988) studied Malmo, Sweden and showed that the largest rent gap caused the initial redevelopment of the land (gentrification). Hammel (1999) also provided a historical view of the rent gap in the 1960s in Minneapolis, MN. He combined data from tax assessments and deeds of sale to measure the capitalized ground rent and the potential ground rent for each property from the 1870s to 1970s. Then, he pointed out that the general pattern conformed to rent gap theory.

(2) Cultural Perspective

Instead of adopting the economic perspective, many humanist and cultural geographers believe that gentrification is associated with cultural factors. They consider gentrification as the consequence of the preference of gentrifiers who appeared in post-industrial society (Hamnett, 2003; Lees et al., 2008; Ley, 1994; Lipton, 1977). Post-industrial society here means the stage of social development which appears in late the 20th century, when the society oriented toward advanced services and a white-collar employment structure (Daniel. Bell, 1973). Gentrifiers, a new kind of middle-class appearing in post-industrial society or the people who confirm their lifestyle, are willing to minimize their commuting time and prefer the lifestyle of the inner cities, and

consequently their decision to move to inner cities leads to gentrification. Gentrifiers are also called new “urban gentry” (Glass, 1964), “yuppie” (Short, 1989), “hipster” (Hae, 2010) or “creative workers.” Although it is highly controversial, Florida’s (2002) definition is extensively used by researchers. He used “creative workers” to describe “gentrifiers,” and stated that they are composed of the *Super-creative core* and the *Creative professionals.* The *Super-creative core* includes researchers, engineers, artists, designers, and media workers since they fully engage in the production that involves intense knowledge and creativity (Florida, 2002). The *Creative professionals* refer to those highly educated people who use their knowledge to solve specific problems such as people work in the legal sector, healthcare, business, and finance.

For the cultural perspective, the cultural characters that appeals gentrifiers and incubates gentrification is a distinctive ambience of a gentrifying area. Ley (2001) pointed out that gentrifiers love living in inner cities because these places represent a social distinction that separates them from the mundane suburbs. Jager (1986) further attributed the social distinction to a distinctive ambience of gentrifying areas and called it *gentrification aesthetic*. Researchers after Jager thought that *gentrification aesthetic* is embodied in a specific art style such as historical or Victorian buildings (Carpenter & Lees, 1995; Jager, 1986), “postmodern landscape” (Mills, 1988), and marginal and bohemian art (Zukin, 1989). According to these discussions, art is one of the main components of gentrification ambience. Yet, not only the art style but also artists and *aestheticisation* process comprise the special “feel” in a gentrifying place (Ley, 2003).

Specific demographic population and business also contribute to the distinctive gentrification ambiance. Those populations include young people (Ley, 1986; Van Criekingen, 2009), hipsters (Vermeulen, 2016), and gay and lesbian communities (Castells, 1983; Knopp, 1990; Lees et al., 2008; Rothenberg, 1995). For instance, Florida (2002) developed the "gay index" for measuring the tolerance and creative atmosphere of a city. The certain business is related to leisure activities such as trendy restaurants, galleries, cafés, coffee shops, bars, and yoga (Kern, 2012; Ley, 1986; Papachristos et al., 2011; Zukin, 2014).

Additionally, vibrant nightlife plays an important role in gentrification (Currid, 2009; Florida, 2002; Hae, 2011; Zukin, 2011). Hae (2011) stated that nightlife activities generate an image of lively and diverse urban sociality which attracts *the creative class (gentrifiers)*, and rekindles depressed property markets in derelict neighborhoods. Hae (2011) further discussed nightlife activities became an important strategy for repositioning a city as competitive in global/regional markets and being able to attract quality human capital.

The strong relationship between gentrification and the special taste of gentrifiers was proved by researchers. For example, Ley (1986), one of most influential cultural geographers, studied the whole inner cities of 22 Census Metropolitan Areas (CMAs) in the Canadian urban system. He implemented Pearson correlation analysis in the gentrification indices and the 35 independent variables in each inner city. His study indicates that the *post-industrialism metropolitan economy* and the aesthetically

pleasing landscapes are strongly correlated with gentrification. In other words, his research supports the argument that gentrification takes place because of the emergence of gentrifiers in post-industrialism society and the gentrification aesthetics which cater to gentrifiers' taste. Ley and Dobson (2008) further explored what factors impede gentrification. Using the City of Vancouver as their study area, they applied Pearson correlation to test for associations between the gentrification index and locational, land use and population characteristics. Their results point out that the factors against gentrifiers' taste impede gentrification such as local poverty cultures and industrial land use. In sum, the reason why gentrification happens and the reason why gentrification doesn't happen associate with gentrifiers' preference and their culture.

2.1.2 The Process

Different from figuring out the factors which trigger gentrification, some researchers focused on the process and the dynamics of gentrification. Gale (1980), Clay (1979), Pattison (1983), the National Urban Coalition (1978), the National Association of Neighborhoods (1980) and Kerstein (1990) used interviewing and questionnaire to analyze the characteristics of different stages of gentrification. For example, Kerstein (1990) interviewed 347 homeowners in South Hyde Park and North Hyde Park in Tampa, FL to examine the demographic and attitudinal attributes of the residents at different gentrification stages. In his research, the characteristics of residents at each stage are consistent with the prediction of other models; however, the residents at the earlier stage don't value the diversity in the gentrifying areas like what Clay's stage models predict.

Instead of focusing on residents at different stages of gentrification, Van Criekingen (2009) examined the human migratory characteristics by analyzing the in-movers and out-movers in the process of gentrification. He combined the three-year Population Register (1996, 2000 and 2002) with the 2001 census data to analyze the demographic and socioeconomic characteristics of in-movers, out-movers, and stable residents in Pentagon, Brussels. The study identified that the in-movers and out-movers to or from Brussels' historical core were young adults, and renting in a gentrifying area were usually associated with a transitional step in their housing career.

Unlike stage models and migration dynamics which are a series of snapshots of a gentrifying area, researchers also established simulation models by using a geocomputational technique (O'Sullivan, 2002; Torrens & Nara, 2007). O'Sullivan (2002) used graph-based cellular automata based on Smith's *rent gap theory* to simulate the house price dynamics which occurred in Hoxton in inner East London in the UK. Based on O'Sullivan's simulation model, Torrens and Nara (2007) introduced a hybrid automata model and applied it to a real world example: Salt Lake City, Utah. Their simulation model predicted different scenarios, and pointed out that gentrification will happen or will not happen in certain conditions.

2.1.3 The Consequence

Besides debating over the cause and the process of gentrification, researchers argued about the impacts of gentrification. Lees (2000) used *emancipatory city thesis* and *revanchist city thesis* to explain why studies hold opposite opinions, i.e., positive and negative gentrification consequences. Those studies that argue gentrification is a positive process are based on *emancipatory city thesis* (Butler, 1997; Castells, 1983; Caulfield, 2016; Freeman & Braconi, 2004; Rose, 1984). They considered the process of gentrification as a practice when the new middle class who love social mixing escapes from mundane suburban areas. For example, by interviewing the gentrifiers in Toronto, Caulfield (2016) argued that the inner city offers a place for gentrifiers to subvert the dominance of hegemonic culture. Ley (1994) interviewed Canadian pioneer gentrifiers in three largest Canadian cities and points out that they welcome the difference in inner cities. Butler (1997) interviewed the gentrifiers in Hackney, inner London, and stated that those gentrifiers valued the social mixing. In addition, researchers like Freeman and Braconi (2004) argued that there is no significant displacement during the process of gentrification. They examined the triennial New York City housing and vacancy survey and found that lower income and lesser educated residents were less likely to move out of gentrifying areas than non-gentrifying areas between 1996 and 1999.

In contrast to *emancipatory city thesis*, Slater (2006) harshly criticized that those studies lacking in critical perspective and paid too much attention to gentrifiers rather than the working class. Thus, they tended to sugarcoat the process of gentrification. Slater's view belongs to *revanchist city thesis* which considered gentrification as a process that the middle class wants to reform inner cities, and takes inner cities back from the working class (Lees et al., 2008). Studies based on *revanchist city thesis* stated that gentrification is a negative process including displacement, segregation, and discrimination (Atkinson, 2000; J. J. Betancur, 2002; Butler & Robson, 2003; Davidson, 2010; Hae, 2011; Powell & Spencer, 2002; Slater, 2006; Wyly & Hammel, 2004). For example, Butler and Robson (2003) interviewed the middle-class gentrifiers in London and pointed out that they engaged in little social mixing with local low-income groups. Through analyzing the regulation and laws with regard to nightlife business in New York City, Hae (2011) criticized that the original subcultural nightlife business was gentrified in the gentrification process. Those trendy lounges and carpeted live music venues displaced the subcultural pubs and poor artists. Some researchers took quantitative approaches to prove the negative consequences resulted from gentrification. For example, Davidson (2010) choose three neighborhoods undergoing gentrification in London, the UK as his study areas. He used measures of social cohesion to assess neighborhood-based social mixing including behavioral mixing and perceived mixing and noticed that there was little mixing between middle-class development residents and working-class residents in surrounding neighborhoods. Wyly and Hammel (2004) created four models to test their segregation and discrimination of hypothesis, in which model parameters include mortgage trends of inner cities of 23 metropolitan areas in U.S. They proved their hypotheses that there are intensified segregation and discrimination in gentrifying areas.

2.2 Gentrification measurements

Gentrification has been intensively discussed from mainly three viewpoints mentioned above. However, there are challenges of gentrification research. First, due to the limited data availability, there is a lack of fine-grained gentrification study across a broad area. Specifically, researchers examined gentrification dynamics on a large scale need to sacrifice the spatio-temporal details, while the studies explored the process of gentrification on a fine spatio-temporal scale can only focused on a small area. For instance, Ley (1986) evaluated the inner cities' gentrification in Canada. He selected the whole inner cities of 22 Census Metropolitan Areas (CMAs) in the Canadian urban system, and calculated their gentrification indices derived from the social index. Yet, since census data is an aggregated data published decennially, the gentrification indexes were generated by subtracting social indices in 1971 from the social indices in 1981 in each census tract. In other words, the gentrification index used in this research was unable to explain gentrification within a census tract, and the gentrification dynamics during the ten years were overlooked. Ley and Dobson's (2008) study examined the gentrification in Vancouver from 1971 to 2001. However, they calculated a gentrification index by subtracting the social index of 1971 from 2001. In other words, they aggregated 30-year data into a gentrification index per census tract. Their gentrification index only showed the change from 1971 to 2001 but didn't show the variance in this period of time. Other studies looked into the fine-grained gentrification dynamics; however, those works were confined to a relatively small region. Take Clark's (1988) research as an example, it examined the shift of the yearly potential ground rent and the yearly capitalized ground rent. The temporal resolution is finer than Ley's (1986) and Ley and Dobson's (2008) works, but the study area is smaller. The study area consists of 6 areas in the inner-city of Malmo, Sweden. It is almost equal to the size of four blocks which are located in different sections of Malmo's inner city. Hammel (1999) also provided a detailed historical view of gentrification in Minneapolis, MN. He combined data from tax assessments and deeds of sale to measure the capitalized ground rent and the potential ground rent for each property from the 1870s to 1970s. The study area is nine groups of parcels which were assembled and redeveloped in the 1960s in Minneapolis. Most qualitative research also focused on small groups of gentrifying neighborhoods. Hamnett and Whitelegg (2007) examined the process of change from industrial to postindustrial land use and argued that this change is a process of gentrification involving economic, financial, and planning factors. They provided a temporally fine-grained study and they discussed the gentrification dynamics from the 1970s to the 1990s. However, this study only looked into Clerkenwell, London which is a small area of central London. Boyd (2008) investigated the racial conflict resulted from gentrification during the 1980s to the 1990s. Although she recorded the yearly details about how American African community reacted to redevelopment, she focused on a confined neighborhood, Douglas/Grand Boulevard, Chicago. There are the simulation models which are able to offer a fine-grained view of individual houses across large spatial scale (O'Sullivan, 2002; Torrens & Nara, 2007). However, in order to further validate, improve, and apply their models, more exploratory gentrification studies on fine spatial-temporal scale across the macro area are needed.

Another research challenge is how to incorporate human perception for detecting the gentrifying areas across a large area. Researchers use qualitative approaches or quantitative approaches to identify gentrifying areas. One major advantage of qualitative strategies is incorporating human perception, which helps us detect the change of culture in a gentrifying area. Like Gibbons and Barton's (2016) argument, gentrification involves not only the change of demographic characters but also an emergence of distinctive culture. By recognizing the cultural factors, qualitative strategies can confirm that the identified areas not just experienced a natural form of improvement such as incumbent economic upgrading (Barton, 2016), but also a change in culture. However, various form of qualitative approaches such as content analysis, interview, and fieldwork are labor-intensive and difficult to be applied to a large region. Because of the easy data accessibility, most of quantitative approached take advantage of census data, and these methods are called *census-based strategies* (Barton, 2016). Quantitative strategies relying on census data are easier to be operated across a large area, while they can only examine gentrification dynamics on a relatively coarse spatio-temporal scale due to the decennial availability of census data. Another drawback on most quantitative strategies is not incorporating human perception for measuring the shift of culture. Quantitative approaches are often based on threshold strategies, which divide neighborhoods into non-gentrifiable, gentrifiable, and gentrifying if they meet certain economic and demographic criteria (Bostic & Martin, 2003; Ding, Hwang, & Divringi, 2016; Freeman, 2005; McKinnish, Walsh, & Kirk White, 2010; Wyly & Hammel, 1998). These strategies are unable to capture the cultural characteristics because they rely only on census indicators of neighborhood changes. More recently, some studies have attempted to incorporate human perception or sought to other cultural measurements with census data (Hwang & Sampson, 2014; Wyly & Hammel, 1998). Wyly and Hammel conducted fieldwork to evaluate the quality of housing structures. Hwang and Sampson took advantage of Google Street View to manually evaluate the appearance of buildings. These new methods involve the manual assessing process on every parcel in neighborhoods either physically or digitally, which is relatively time-consuming. Papachristos et al. (2011) used a non-census data source. They utilized the number of coffee shops listed in annual Chicago Business Directories as the cultural measurement of gentrification. Coffee shops are regarded as a special place which is integral to the leisure and lifestyle amenities package and so attractive to gentrifiers according to the literature. However, the counts of coffee shops in the neighborhoods only reflect a certain changing consumption pattern relating to gentrification. Namely, the number of coffee shops is only one of cultural indicators, and it evaluates the cultural change from a limited aspect. Those earlier attempts to incorporate human perception or evaluate the cultural measurement are limited in their research scope and/or scale; thus, more effective and efficient approaches should be explored.

In terms of quantitative measurements, Table 1 illustrates the details of every census-based strategies. All of these methods are threshold strategies, which means that census block groups were identified as gentrifying if they had some economic or demographic characteristics at the beginning of a period of time, and the characteristics changed at the end in a particular way. The first strategy is created by Hammel and Wyly (1998). This method uses income level and housing structure to define non-

gentrifiable, gentrifiable and gentrifying areas, and gentrifying areas including poor, fringe and core gentrifying areas. The second strategy is created by Bostic and Martin (2003). They modified Hammel and Wyly's strategy, and took two steps to identify non-gentrifiable, gentrifiable, and gentrified areas. First, it uses the same method as Hammel and Wyly's to distinguish gentrifiable areas from non-gentrifiable areas; then, two methods were developed to determine gentrifying and gentrifiable areas. One method regards an area where was gentrifiable at the earlier time point and had changed to non-gentrifiable at the later time point as a gentrifying area. The other method is based on the 8 demographic and economic factors proposed by Hammel and Wyly to identify a gentrifying area. While Bostic and Martin's (2003) strategy can identify gentrification to some extent, it doesn't take new construction into account which is also an important identifier of gentrification. Therefore, Freeman (2005) developed another strategy. He created four criteria for identifying gentrifiable areas in a city including population, income, housing market, and age of buildings. Like previous strategies, Ding, Hwang, and Divringi (2016) strategy also take economic factors and housing market into account. In addition, it uses the change of residents' education status to measure the residential displacement. Among the existing gentrification typology strategies, McKinnish's method (2010) is the simplest. Average family income is the only considered factor in this method. In contrast to those strategies that identify non-gentrifiable, gentrifiable, and gentrifying areas, Hwang, and Sampson (2014) regarded neighborhoods are in different stages of gentrification including early, middle, and late stage. These systematic and statistical quantitative measurements focus on demographic and socio-economic perspectives based on Census survey data, which are limited to capture subtle cultural characteristics of gentrification such as gentrification aesthetics discussed in 2.1.1.2 and their changes over time.

Developing a method to identify different types of gentrifying areas is another challenge. Although gentrification has been studied for more than 50 years, researchers today haven't reached a consensus about its definition. One reason is that there are many types of gentrification. They can be divided into two main categories: *residential gentrification* and *commercial gentrification* (Lees et al., 2008). *Residential gentrification* includes *rural gentrification* (Parsons, 1980), *new-build gentrification* (Zukin, 2000)*, super-gentrification* (Lees, 2003), and *Studentification* (D. Smith, 2005; D. Smith & Holt, 2007) . Apart from locating in inner cities, Parsons pointed out that *rural gentrification* happens in rural areas. Zukin argued that *new-build gentrification* takes place in brownfield rather than residential areas. Lees stated that super-gentrification happens in areas that have already been gentrified. Smith used *Studentification* to refer to the process of gentrification resulted from students but not middle-class adults. *Commercial gentrification* includes *tourism gentrification* (Gotham, 2005) and *provincial gentrification* (Dutton, 2003). Gotham created the term, *tourism gentrification*, to describe the gentrification in French Quarter, New Orleans. It became an enclave in which corporate entertainment and tourism venues have proliferated, and resident-oriented business has been evicted. Leeds observed *Provincial gentrification* at Dutton, UK, which was different from other types of gentrification because it occurred in provincial cities rather than major cities. As the increasing different types of gentrification, we need a method to differentiate them.

Table 1. Gentrification Typology Strategies

<table>
<tr><th colspan="5">Census-Based Gentrification Strategies</th></tr>
<tr><td rowspan="3">Hammel and Wyly (1998)</td><td>Non-gentrified</td><td>Poor
(Gentrifiable)</td><td colspan="2">Gentrified</td></tr>
<tr><td>1. A tract in the inner city.
2. And the median income of a census tract is more than 50% of the median income in the city; the census tract is a non-gentrifiable</td><td>1. A tract In the central city
2. If the median income of a census tract is less than 50% of the median income in the city.</td><td>Fringe</td><td>Core</td></tr>
<tr><td>1. The gentrifiable tract has at least one improved
structure on a majority of the blocks, while improved units must comprise 1/3 of all structures on at least one block.

* To classify structures as improved, they considered the quality and style of repainting, ornamentation, signage, and renovations to apartment buildings, and other signs of reinvestment.</td><td>1. The gentrifiable tract has at least one improved housing structure
on each block

2. And the improved structures should comprise at least 1/3 of all housing units in the tract.</td></tr>
<tr><td>Bostic and Martin (2003)</td><td colspan="4">Their method to identify a non-gentrifiable and gentrifiable tract is the same as Hammel and Wyly’s (1996) method. However, they develop more complicated two methods to identify a gentrifying tract.
Method 1: A census tract was considered as gentrifiable at the earlier time point and had changed to non-gentrifiable at the later time point
Method 2: They use the 9 descriptive gentrification factors proposed by (Hammel and Wyly 1996; 1999) to identify gentrifying areas:
(1) the t+1 population share of persons 25 and older with some college education.
(2) the ratio of median family income at time t+1 and median family income at time t
(3) the home-ownership rate at time t+1
(4) the change in population share of the cohort that is aged between 30 and 44 at time t to that at t+1
(5) the t+1 poverty rate
(6) the t+1 population share of White non-family households
(7) the t+1 Black population share
(8) managerial and administrative workers as a share of the total workforce at t+1

The average rank of the tracts across the nine measures is used as a score and the tracts with the lowest average rank are identified as gentrifying.</td></tr>
</table>

<table>
<tr><td>Freeman (2005)</td><td>Not potential gentrifying</td><td>Potential gentrifying</td><td>Gentrifying</td><td>Non-gentrifying</td></tr>
<tr><td></td><td>1. A tract is not in an MSA area
2. Or a census tract doesn’t meet the criteria of potential gentrifying</td><td>1. A tract is In an MSA area.
2. And a census tract with a median income that is at or less than the median in their respective metropolitan areas.
3.
have a proportion of housing built within the past 20 years lower than the proportion
found at the median (40th percentile) for the respective metropolitan area (MSA).

And a census tract with the proportion
of its housing stock built within the past 20 years falling below the the median for their respective metropolitan areas (MSA).</td><td>A census tract meets the three criteria on the left column, and also meets the following two criteria:
4. And have a percentage increase in educational attainment greater than the median increase in educational attainment for that metropolitan area.

*The education attainment refers to the percentage of those 25 years and older with at least four years of college
5. And The census tract has an increase in real housing prices during t1-t2.</td><td>A census tract meets the 1, 2 and 3 criteria, but not 4 and 5 criteria.</td></tr>
<tr><td>McKinnish et al. (2010)</td><td colspan="2">Low-income neighborhood sample</td><td colspan="2">Gentrifying</td></tr>
<tr><td></td><td colspan="2">At t1, the tract is in the bottom quantile of national average family income.</td><td colspan="2">1. At t1, the tract is a low-income neighborhood
2. The tract experiences an increase in the average family income by at least 10,000 during t1-t2</td></tr>
</table>

<table>
<tr><td rowspan="3">Voorhees Center (2014)</td><td></td><td>Type 1
No Change, Upper</td><td>Type 2
No Change, Middle Class</td><td>Type 3
No Change,</td><td>Type 4
No Change, Extreme Poverty</td><td>Type 5
Increase, Not Gentrification</td><td>Type 6
Increase, Gentrification</td><td>Type 7
Decrease, Mild</td><td>Type 8
Decrease, Moderate</td><td>Type 9
Decrease, Severe</td></tr>
<tr><td>Overall Average Scores</td><td>>7</td><td>0~7</td><td>-1~-7</td><td><7</td><td><=7</td><td>>7</td><td colspan="3">13~-13</td></tr>
<tr><td>Socioeconomic change from T1-T2</td><td colspan="4">4~-4</td><td colspan="2">>4</td><td>-5~-7</td><td>-8~-9</td><td><=-10</td></tr>
</table>

<table>
<tr><td></td><td colspan="4">Socioeconomic status index:
1. If the percentage of white people (non-Hispanic) of the tract above city average, the tract get score +1
2. If the percentage of African-Americans of the tract above city average, the tract get score -1
3. If the percentage of Latino of the tract above city average, the tract get score -1
4. If the percentage of Elderly (Age 65+) of the tract above city average, the tract get score -1
5. If the percentage of children (Age 5-19) of the tract above city average, the tract get score -1
6. If the percentage of college education (Bachelor’s degree or higher) of the tract above city average, the tract get score +1
7. If the median family income of the tract above city average, the tract get score +1
8. If the percentage of owner-occupied of the tract above city average, the tract get score +1
9. If median house value of the tract above city average, the tract get score +1
10. If the percentage of family below poverty of the tract above city average, the tract get score -1
11. If the percentage of manager occupation of the tract above city average, the tract get score +1
12. If the percentage of family with children of the tract above city average, the tract get score -1
13. If the percentage of private school attendance (pre-K through 12) of the tract above city average, the tract get score +1
* According to the 13 criteria of the socioeconomic index, each tract has one socioeconomic index one year. Based on the change of socioeconomic index, census tracts are divided into nine types. Type 6 means the area is undergoing the process of gentrification.</td></tr>
<tr><td rowspan="2">Ding, Hwang, and Divringi (2016)</td><td>Non-gentrifiable
(old gentrification)</td><td>Gentrifiable</td><td>Gentrifying
(continued gentrification)</td><td>Non-gentrifying
(stalled gentrification)</td></tr>
<tr><td>A tract has a median household income above the citywide median at the beginning of the period of analysis.</td><td>A tract has a median household income below the citywide median at the beginning of the period of analysis.</td><td>1. A census tract was gentrifiable at the beginning of the time period.
2. And the tract has experienced an above citywide median percentage increase in either its median gross rent or median home value.
3. And the tract has experienced an above citywide median increase in its share of college-educated residents.</td><td>A tract was gentrifiable at the beginning of the period of analysis but doesn’t meet the criteria of gentrifying.</td></tr>
<tr><td rowspan="2">Hwang, and Sampson (2014)</td><td colspan="3">Stages</td></tr>
<tr><td>Early</td><td>Middle</td><td>Late</td></tr>
<tr><td></td><td>Below 0.5 scores:
1. Low-middle composite structural mix score
2.Few types of beautification
3. Many types of disorder/decay</td><td>0.50 ~ 0.65 scores:
1. Middle composite structural mix score
2.Some types of beautification
3. Some types of disorder/decay</td><td>0.65~0.80
Scores:
1. Middle-high composite structural mix score
2.Many types of beautification
3. Few types of disorder/decay</td></tr>
</table>

	1. Composite structure mix: a tract in the combined condition of older structures, and the degree of new structures. * New structures: the amount of new or rehabilitated building structures, new traffic signs/structures, new public courtesies, new large developments, and new construction for sale. 2. Visible beautification efforts: if it efforts discouraging disorder (e.g., painting over graffiti) if it has personal frontage beautification, and if it has vacant/public space beautification. 3. Lack of disorder and decay: if it is a lack of physical disorder, if it is a lack of unkempt vacant/public space, and if it is a lack of decaying structures.

2.3 Social Media Research and Gentrification Studies

Social media now bring geographers an opportunity to address those challenges of gentrification research. Social media includes, for instance, MySpace, Facebook, Twitter, and LinkedIn which have been described as one of the characteristics of Web 2.0 technologies (Sui & Goodchild, 2011). Sui and Goodchild pointed out that one major character of social media during recent years is the increased use of mapping and location-based features. This location-based social media shift social media from cyberspace to real place, and interconnect cyberspace and real space (M.-H. Tsou, 2011). In other words, social media reflects our real world to some extent and can be the valuable research sources.

Social media has been used in various GIScience applications, such as tracing, monitoring, and mapping the spread of social movements, disease outbreaks prediction, elections, and political campaigns (M.-H. Tsou, 2011). For example, Nagel et al. (2013) analysed the tweets which are the text Twitter users post to express their opinions and sharing the information. They collected tweets with flu keywords and evaluated which keyword is more appropriate for predicting flu outbreak. The work of Aslam et al. (2014) is another example of social media research applied to public health. They tested whether tweets contain the keyword “flu” can help people predict influenza. Although the correlation between tweets and Influenza-like illness (ILI) varied by city, this study demonstrated increased accuracy in using Twitter as a supplementary surveillance tool for influenza as better filtering. In terms of political research, Tsou et al. (2013) exploited tweets and web pages with the keyword “Obama” and “Romney” from top thirty US cities to analyze the public attention these two US presidential candidates get before and after Hurricane Sandy.

The advantage of social media is the both “deep” and “big” volume of data (Manovich, 2011). “Deep” means the data contains fine spatio-temporal details, and the information about human imagination, opinions, ideas, and feeling; “big” means its big volume. Manovich stated that with social media data, for the first time we don’t have to sacrifice the details for the volume or vice versa. For example, before the emergence of social media data, GIScientists who adopted quantitative approaches often used aggregated data for analyzing the phenomenon across a large region. On the other hand, cultural geographers who adopted qualitative methods can only focus on a small area for examining the details of the phenomenon. By applying social media to geographic research, geographers now have a great opportunity to reconcile the perspective of quantitative research and the perspective of qualitative research, and further reconcile the world of space (traditional GIS and Marxism geography) and the world of place (cultural geography) (Sui & Goodchild, 2011). Therefore, taking advantage of social media data in gentrification research, it is capable of designing a data analytics framework that combines both quantitative and qualitative strategies for evaluating the gentrification dynamics. Moreover, such a big data approach has a possibility to measure human perception to the neighborhoods and analyze gentrification at a fine spatial-temporal scale across a large region.

Some researchers have used social media data for their gentrification studies. For example, Zukin, Lindeman, and Hurson (2015) studied restaurant reviews on the website *Yelp.com*. Business reviews on Yelp has been becoming increasingly popular with more than 61 million reviews and 138 million unique visitors per month from more than 20 countries (Yelp, 2014). Restaurants' reviews are the most popular among all review categories and create cultural and financial value for individual restaurants, which would result in the increase of economic investment and thus accelerate the process of gentrification. At the same time, those restaurant reviews would also change the cultural landscape of an area. However, Zukin et al. only analyzed those Yelp reviews in two gentrifying areas and do not focus on how to capture the signal of gentrification with Social Media data.

Hristova et al. (2016) used Twitter and Foursquare data to explore the social diversity of urban locations. Their study introduced a novel network perspective on the interconnected nature of people and places, which shows a capability of social media data for identifying gentrifying areas. Specifically, they defined four metrics of the social diversity of places: (1) Brokerage, (2) Serendipity, (3) Entropy, and (4) Homogeneity. Brokerage includes *brokerage of a person* and *brokerage of a place*. The first expresses a person's ability to connect or disconnected others; the latter one means a place's ability to bring visitors together or disconnect individuals in physical space. Serendipity is for measuring the extent to which a place can induce chance encounters between its visitors. Entropy refers to the extent to which a place is diverse with respect to visits. Homogeneity describes the extent to which a place's visitors are homogeneous in their characteristics. With these four metrics of social diversity, they found that neighborhoods in deprived areas with high entropy and brokerage are currently undergoing processes of gentrification. This finding indicates that social media data is capable of detecting a signal of gentrification.

While geographers are increasingly using different social media as their data sources, Instagram, one of the most popular photo sharing and social networking application, is rarely used. Most Instagram studies were conducted by computer science researchers, and they mainly focused on data exploration and analytics. For instance, Hu, Manikonda, and Kambhampati (2014) took advantage of computer vision and clustering technique aiming to find out the types of photos posted on Instagram. They grouped Instagram photos into eight categories: friend, food, gadget, captioned photos, pet, activity, selfie, and fashion. Their research revealed that friend, selfie, and activities are the most popular topic, whereas pet and fashion are least popular categories. This kind of research provides other researchers a fundamental understanding of Instagram; however, few research used Instagram data to explore geographic phenomena.

Some studies show the value of Instagram for geographic research, but they haven't demonstrated how to use Instagram for studying geographic questions. For example, Silva et al. (2013) compared Instagram with Foursquare in order to see whether data from one system could complement the other, or if they are compatible or not. They conducted CDF (Cumulative Density Function) and correlation analysis, and found that

Instagram and Foursquare are compatible generally; however, Instagram user behavior is slightly different from Foursquare. For example, Instagram users tended to share more content in the same place than Foursquare users. Additionally, by examining the number of posts each hour during a day in different cities, they pointed out that the sharing behavior in Instagram in different city is more distinct to each other than the sharing patterns observed for Foursquare, meaning that Instagram provides a more distinguishable "cultural signature." Their research showed that the posting pattern of Instagram is a proper source which helps researchers explore the cultural characteristic of a place.

Hochman's (2013; 2012) works used Instagram data and image analysis techniques to characterize cities. Unlike computer scientists, Hochman, as an artist and digital humanity researcher, is interested in different cities' characteristics and lives of individuals. In 2012, by taking image processing technique, Hochman and Schwartz's compared "Vision Rhythm" in New York City with Tokyo. They found that the hue, brightness, other image attributes, and the amount of images generated by users during a specific time period are significantly different in two cities. For example, people in New York City preferred to use blue-grey tones, while people in Tokyo preferred to use red-yellow tones. Their findings conform to Silva et al.'s (2013) statement that Instagram posts have great value for studying the characteristic of a place. Hochman (2013) also demonstrated how Instagram data helps us to explore geographic phenomenon on dynamic scale. Unlike traditional data like census decennial data or other government data in which the individual and the particular are sacrificed for the sake data aggregation, Hochman (2013) presented individual photos in 13 cities by visualizing all photos sorted by time or attributes. In other words, this study illustrates that taking advantage of Instagram or other social media data, researchers can explore urban dynamics on a finer scale.

In sum, research by Hochman and Schwartz's (2012) and Silva et al. (2013) provides a case study that investigates how spatial and cultural patterns can be traced with Instagram posts. In additional, Hochman (2013) demonstrated how fine-scale urban study can be done by analyzing Instagram data.

2.4 Text Mining Technique and Research

2.4.1 Data Mining and Text Mining

Data mining is an essential approach to quantitatively process and analyze the big volume of social media data. It is a process of discovering useful knowledge from raw data, or simply Knowledge Discovery from Data (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Tan, 1999). In order to stress the characteristics of social media data, Tsou and Leitner (2013) proposed the concept KDC (Knowledge Discovery in Cyberspace) to discuss the data mining process to handle the very large human messages collected from cyberspace and social media.

In general, data mining algorithms can be grouped into supervised, unsupervised, and semi-supervised learning algorithms (Gundecha & Liu, 2012). Whether the dataset has

known class labels is the major difference between supervised and unsupervised learning algorithms. Namely, researchers took supervised learning algorithms on the dataset with known labels and use unsupervised learning algorithms on the dataset without known labels. Supervised learning methods include decision tree induction, k-nearest neighbors, naive Bayes classification, and support vector machines. Clustering based on similarity or dissimilarity between data object is the typical method of unsupervised learning. Data scientists develop several ways to measure similarity or dissimilarity such as Euclidean distance, Cosine distance, Minkowski distance, and Mahalanobis distance. Other statistical methods like simple matching coefficient, Jaccard coefficient, and Pearson's correlation can also be used to measure the proximity between the pair datasets. K-mean clustering, hierarchical clustering, and density-based clustering are the three common methods for clustering data after the similarity or dissimilarity has been calculated (Gundecha & Liu, 2012). Semi-supervised learning algorithms can be used when there are small amounts of labeled data and large amounts of unlabeled data.

Text mining also known as *text data mining* or *knowledge discovery from textual databases* is an extension of data mining (Hotho, Nürnberger, & Paaß, 2005; Tan, 1999). Text mining is used to discover useful models, general trends, and patterns, interesting information, or rules from large quantities of text contents such as unstructured textual data in text files, HTML files, chat messages and emails (He, Zha, & Li, 2013).

2.4.2 The Concepts and Methods of Text Mining

(1) Bag of Words

In text mining, documents containing text data are typically transformed into a set of words, or *bag of words*. Created by a linguist, Zellig Harris (1954), the *bag of words* model regards the text as a bag which contains many words, and it disregards grammar and word order. After transforming the documents into *bag of words*, researchers used different ways to characterize these documents. The most common way is *term frequency*. For example, if a word appears in a document twice, the word will be represented by 2. Using term frequency, bag of words from a document is summarized as a frequency vector for further analyses.

(2) Text Preprocessing

Before transforming a document into a *bag of words*, there are several text data preprocessing including *tokenization*, *filtering*, *lemmatization*, and *stemming*. A word or other meaningful symbols in a document are *tokens*. *Tokenization* is an initial process to remove the punctuation marks, the white spaces, and other non-text characters. *Filtering*, *lemmatization*, and *stemming a*re techniques to reduce the size of the bag of words for improving calculating speed and results (Hotho et al., 2005).

A standard *filtering* method is *stop word filtering*. The idea of *stop word filtering* is to remove the words which have little or no content information, or the words that we know are not pertain to our research topics (Hotho et al., 2005). *Lemmatization* is a process to standardize verb and noun forms. It transforms all verb forms into the infinite tense and transforms all nouns to the singular forms. However, for conducting *lemmatization*, we need to recognize the word forms first. Namely, we need to achieve the other task, to tag every word with its *part of the speech*. Since this tagging process is time-consuming and still error-prone, researchers usually apply the other text preprocessing method called *stemming* (Hotho et al., 2005). *Stemming* is simpler because it transforms every word into its stem. A stem is a natural group of words with similar meaning. Comparing with *lemmatization*, *stemming* is a crude heuristic process, which only strips the plural "s" from nouns and "ing" from verbs, and removes the derivational affixes.

(3) Vector Space Model

Most text mining studies used *vector space model* to represent a document. In this model, a document is an *m*-dimensional vector.

The simplest way to represent a document as a vector is to use binary term vector. For example, if a word appears in a document (no matter how many times), we treat it as one. If a word doesn't appear in a document, it will be regarded as zero in the vector. However, this method considers all words appear in a document as the same, and it doesn't take how important the word is into account. In order to address this problem, *term weighting schema* was developed. With this method, the importance of a word can be represented by its weight. *TF-IDF* (*Term Frequency - Inverse Document Frequency*) is the most common way to weight the words. *TF* means how frequent a word appears in a document, and *IDF* refers to the adjustment of its weight based on the frequency of a word in the entire *corpus,* the collection of text or documents. For instance, a word will get a larger weight if it appears more frequently in one document, but rarely in the entire corpus (Salton & Buckley, 1988).

(4) Data Mining Methods for Text—Classification, Clustering, and Topic Modelling

Text classification is to classify documents into pre-designed classes. In the classification process, pre-labeled data must be divided into two groups: a training dataset and a test dataset. The training dataset is for training the classifiers such as Naive Bayes Classifier, Nearest Neighbor Classifier, Decision Trees, and Support Vector Machine (SVM) (Hotho et al., 2005). To evaluate the performance of the classification model, we can classify the test dataset with the trained classifier, and compare the estimated labels with the true labels. One of common text classification methods is sentiment analysis which determines whether the attitude of the document is positive or negative (Pang & Lee, 2006).

Text clustering is for finding groups of documents with similar content. In a good clustering result, the documents within their cluster are more similar and between the clusters more dissimilar. Most text mining techniques are based on distance-based

clustering algorithms. For distance-based clustering algorithms, the similarity and dissimilarity between documents can be measured by the distance between documents. The most well-known function of calculating the distance in the text domain is the cosine similarity function (Aggarwal & Zhai, 2012). Comparing with Euclidean similarity function that should only be used for normalized vectors, cosine function can better process the documents with different length (Hotho et al., 2005). Clustering algorithms, such as hierarchical clustering and k-means clustering, are employed to the computed text distance to divide documents into groups with similar text contents. Other clustering algorithms include bi-section-k-means, self-organizing map (SOM), co-clustering, and fuzzy clustering.

Topic modeling, another text mining technique, find some groups of words that represent a document, or topics, to summarize a large text dataset. Like text clustering, topic modeling is for unlabeled documents. It assumes that (1) the n documents in the corpus are assumed to have a probability of belonging to one of k topics. Namely, a given document may have a probability of belonging to multiple topics; (2) each topic consists of a probability vector, and this vector quantify the probability of the different terms for the topic (Aggarwal & Zhai, 2012). Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA) are most common methods used for topic modelling. However, they have three known limitations. First, the order of word is ignored in these methods. Second, these method do not take temporal order of the documents into account. Third, the number of topic is assumed known and fixed. Because of these disadvantages of the traditional topic modelling methods, there were some new methods developed such as dynamic topic model (Blei & Lafferty, 2006) and the Bayesian nonparametric topic model (Gershman & Blei, 2012).

2.4.3 Text Mining and Geography Research

Text mining has been implemented in various disciplines, such as business and marketing (He et al., 2013; Ingvaldsen & Gulla, 2012), public health and biomedical domains (Cohen & Hersh, 2005; Collier, 2012; Lu, Zhang, Liu, Li, & Deng, 2013; Pereira, Rijo, Silva, & Martinho, 2015; Popowich, Vx, & Va, 2005), and education (Grobelnik, Mladenić, & Jermol, 2017). Like other disciplines, geographic research has been incorporating text mining technique during recent years.

Bertrand et al. (2009) took advantage of basic text processing and demonstrated how twitter data help us to monitor the location of wildfire. They extracted the place names posted during the incident, and found that the origin of the fire was mentioned in early tweets and was mentioned most frequently. Other place names mentioned in later tweets coincided with the locations of the spread of the fire.

Besides text processing, text classification is also used in geographic research. Xu et al. (2013) explored how geographical characteristics of people affect their geographical awareness. They applied Named-entity recognition (NER) to classify the atomic elements in tweets, and then used the GeoNames Gazetteer to recognize the place names and their coordinates in those elements. By mapping those places mentioned by

each Twitter users, their study revealed the spatial association between users' home locations and places mentioned in their Tweets. As another example, Longley and Adnan's (2016) estimated the flow of people with certain characteristics, night-time residence, and the interactions between different groups. The text classification technique including *Onomap name classification* and *enhanced version of CACI's Monica system* was used to extract more detail information from Twitter users' name such as their age, and ethnicity. Sentiment analysis also becomes common in geographic studies. For example, Mitchell et al. (2013) analyzed tweets with sentiment analysis and investigated how geographic place correlates with the levels of happiness.

Topic modeling, a technique that assists researchers to look into the contents of their text dataset, is incorporated in geographic research. Take Wang et al. (2012) for instance, they aggregated tweets by day and utilized *semantic role labeling (SRL)* to classify the semantic event content of the tweets, so that they had multiple events are associated with each day. Then, they further extracted 10 topics from each day using LDA topic modelling. With prior criminal incidents and tweet topics derived via LDA, they trained their generalized linear regression model (GLM) and developed a model that is able to predict the probability of hit-and-run incident based on the new tweets.

Ghosh and Guha (2013)'s research is another example. After the basic text preprocessing procedures such as removing punctuation and stop words, and stemming, they discovered three themes from obesity-related tweets with LDA topic modelling. Those themes include 'childhood obesity and schools,' 'obesity prevention,' and 'obesity and food habits.' Taking advantage of GIS analysis, they identified that those three topics show distinct spatial patterns between rural and urban areas, northern and southern states, and between coasts and inland states. Lansley and Longley (2016) also used LDA topic modelling for extracting common topics from London daily geo-tagged tweets. In their research, they found that since the characteristics of place, tweets topics were significantly varied across entire London.

Chapter 3: Methodology

3.1 Research Design

This research addresses two research questions: (1) how accurate are existing measurements to identify gentrifying areas, and (2) can media data approach incorporating human perception better identify gentrifying areas as compared to conventional census-based typologies? To answer these two questions, there are two studies in this thesis. The first study is a comparative study which compared five existing census-based gentrification typologies with the gentrifying areas identified based on human perception. A novel data mining framework to examine gentrification dynamics utilizing social media data was introduced in the second study. By applying text processing and text clustering techniques to Instagram posts, the second study examined the spatial-temporal distribution of two gentrification indicators, nightlife activities and gentrification ambience to characterize the sense of gentrification.

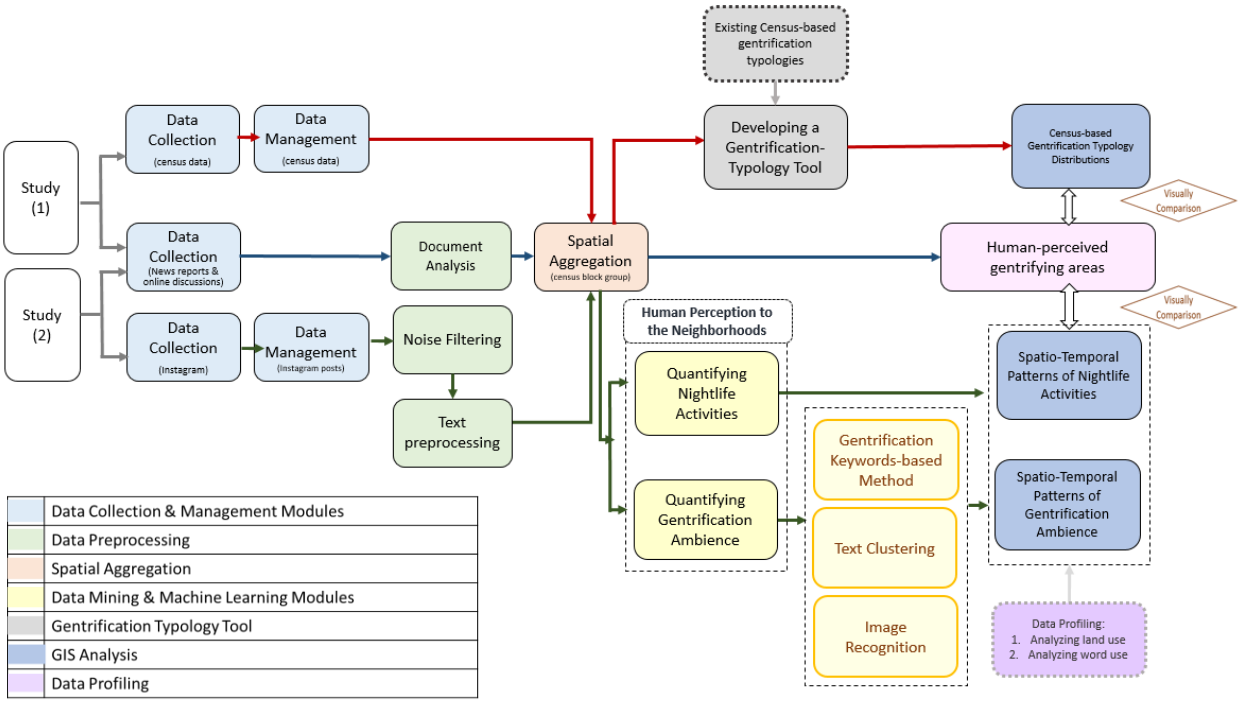


Figure 1. Research workflow

Figure 1 represents the research workflow of this thesis research. The workflow consists of seven key modules including (1) data collection/management, (2) data processing, (3) spatial aggregation, (4) data mining/machine learning, (5) gentrification typology tool, (6) GIS analysis, and (7) data profiling. These seven modules, which complete different categories of tasks, compose two studies in this research.

The first study developed data collection module first. This module was written in Python scripts worked with Census FTP server to collect 2000 decennial census data and 2014 ACS 5-year estimate data, and stored the census data in a PostgreSQL database. Then, a spatial aggregation module was developed to aggregate the census data based on the 2010 census block group's boundaries. With the aggregated census data, the gentrification typology tool module was developed to generate five gentrification typologies based on existing census-based gentrification typologies methods. GIS analysis module was used to map the results. Finally, the results were compared with those human-perceived gentrifying areas, which were determined by qualitatively analyzing online forums and news reports. The procedure of this qualitative analysis is illustrated in the middle of the workflow.

The second study developed another data collection module written in Python scripts. It worked with Instagram API to automatically collect Instagram data from 2013 to 2015 and stored the data in a Mongo database. The data preprocessing module was used to filter noise and clean the Instagram text. Then, the spatial aggregation module grouped all collected Instagram posts by census block groups. The aggregated Instagram posts were analyzed by the data mining/machine learning module. This module written in Python and R quantified nightlife activities and gentrification ambience. The GIS analysis module based on ArcGIS mapped the patterns of nightlife activities and the patterns of gentrification ambience. These patterns were compared with those human-perceived gentrifying areas, and further explored by the data profiling module.

In the data mining framework, two gentrification indicators were quantified including nightlife activities and gentrification ambience. The strategy used to measure nightlife activities is calculating the number of night Instagram posts in a census block group during a year. This strategy assumes that the more nightlife activities the more night posts there are. Three strategies were developed to quantify gentrification ambience including gentrification keywords-based method, text clustering, and image recognition method. These three ways measured gentrification ambience from different angles. Gentrification keywords-based method identified a place with strong gentrification ambience by counting how often gentrification keywords were used at a place. Text clustering method identified a place with strong gentrification ambience by looking for the place which used similar words to the gentrifying place. Image recognition method identified a place with strong gentrification ambience based on how often the Instagram photos' topics were associated with gentrification keywords. Among these three strategies, this thesis research has not conducted image recognition method because of the difficulty to obtain the permission for applying computer vision technique to Instagram photos granted by Instagram.

3.2 List of Main Software and Packages

The main software and programming packages used in this research are listed in Table 2. MongoDB and PostgreSQL are databases to store and manage the dataset of this

research. MongoDB is a NoSQL database used for unstructured social media data, Instagram data; PostgreSQL is a SQL database for structured data such as census decennial data, ACS data, and parcel data. PyMongo and Psycopg2 are two Python libraries used for connecting and interacting with Mongo database and PostgreSQL database in Python respectively. NLTK (Nature Language Toolkit) and Graphlab Create are two Python libraries for applying text processing and text clustering techniques to Instagram text. In addition to Python libraries, some R packages are used. For example, NbClust is an R packages used to decide the optimum number of clustering for Instagram posts. Wordcloud is another R package to generate word clouds from Instagram text. Besides, PostGIS, a spatial database extender for PostgreSQL, was used for performing spatial queries; ArcMap, a GIS software, was used for mapping the spatial patterns.

Table 2 The main software and tools

Names	Description	Used For
MongoDB	A NoSQL database	Managing and querying Instagram data
PostgreSQL	A SQL database	Managing and querying census data
PyMongo	A python library	Interacting with MongoDB
Psycopg2	A python library	Interacting with Postgres database
NLTK, Graphlab Create	A Python library	Text processing and text clustering
NbClust	An R package	Finding out the optimum number of clustering
wordcloud	An R package	Generating word clouds
PostGIS	A spatial database extender for PostgreSQL	Spatial querying
ArcMap	A GIS software	Mapping

3.3 Study Area

The study area is Salt Lake City, UT, where has experienced rapid population growth and urban development. According to the report on U.S. Census website[1], Utah is the fastest-growing state. Its population increased 2.03 percent from 2015 to 2016. The growth of population is associated with its urban development, and the rapid growth can result in a significant gentrification phenomenon in its city areas. Geographically, Salt Lake City and its suburban neighborhoods are located within a topographically confined Salt Lake Valley. In terms of urban structure, Salt Lake City can be considered as a simple mono-centric city (i.e., single core Central Business District in the Valley).

[1] https://www.census.gov/newsroom/press-releases/2016/cb16-214.html

Therefore, urban dynamics are simpler than other topographically complex and interconnected cities like the San Francisco Bay Area.

The dashed line shown below (Figure 2) is the boundary of Salt Lake City based on the Salt Lake City government. However, the boundaries of census block groups in the city do not completely match the boundary of Salt Lake City. The study area only includes those census block groups, which have more than 50% area within the city's boundary. Additionally, those census block groups with the population density less than 5 percentile (446.02 people/km^2) are not discussed in this research. Therefore, instead of the actual city area, the pink area in Figure 2 is the study area in this thesis.

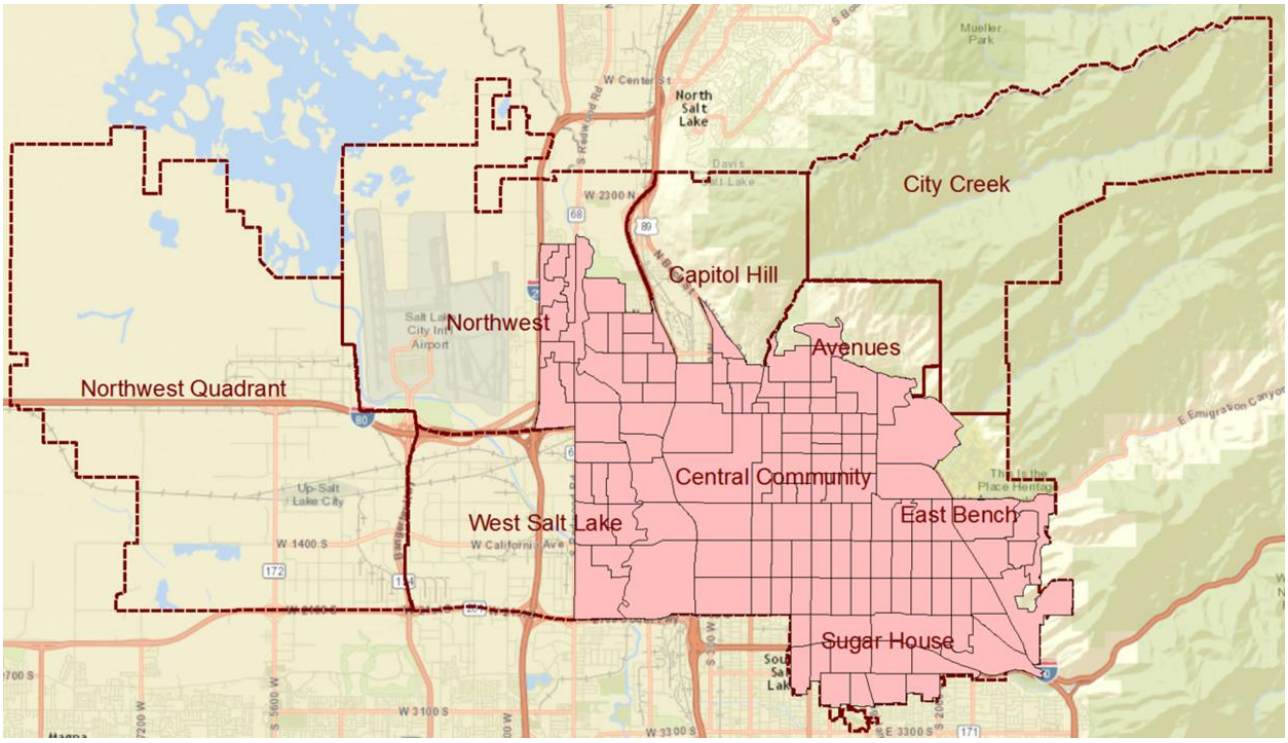


Figure 2. Study area

3.3 Data Collection and Data Description

In this research, there are four major data sources as described in Table 3. The first data source is U.S. Census survey data. The census 2000 decennial data and 2014 ACS 5-year estimate data in Salt Lake County at the census block group level were collected from *Census FactorFinder* and *Census FTP servers*, and they were stored in a PostgreSQL database.

The second data source is from online news reports and online forums. With *Google Search Engine,* this research used two keywords, "gentrification" and "Salt Lake City" to search and collect those discussions on online forums and the News report. From 2000 to 2015, a total 11 articles were found. The sources include The Great American

Country, Salt Lake Tribune, Summit Sotheby's International Realty, Utah Stories, Reddit, and, Zillow.

The third data source is social media posts from Instagram, a mobile photo, text, and video sharing service. Only Instagram text data was used, which includes captions and hashtags. Instagram captions are the text for describing the pictures and videos. Instagram hashtags are the text using the # symbol to identify messages on a specific topic (Hu, Manikonda, and Kambhampati, 2014). Comparing to other social media, Instagram data is more suitable for this research because it is the most popular social media in Utah[2] except for Facebook, in which most posts are not publicly accessible through API. An Instagram retrieval engine was developed to collect Instagram posts from 2013/1/1~2015/12/31 in Salt Lake County through Instagram API. *MongoDB,* a NoSQL database*,* was used to temporally store and manage the collected Instagram data. Figure 3 shows a map of collected Instagram posts. The black dots indicate that most Instagram posts were present in the Central Community.

Parcel data based on Salt Lake City assessment is fourth data source, which was gathered through the Assessor's comprehensive appraisal process. In this research, parcel data was used for exploring the land use in gentrifying areas.

Table 3. Data used in this research

Category	Source	Data Collection
Census Data	1. 2000 decennial data (154 fields) 2. 2014 ACS 5-year estimate data (73 fields)	1. Census FactFinder 2. Census FTP servers
News reports and discussion on online forums	1. The Great American Country 2. Salt Lake Tribune 3. Summit Sotheby's International Realty 4. Utah Stories 5. Reddit 6. Zillow	Search Keywords: *gentrification* and *Salt Lake City* Duration: 2000-2015
Social Media Data	Instagram	Instagram API Duration: 2013/1/1~2015/12/31
2015 Parcel Data	The Assessor's 2017 CAMA database	2015 – 1st Quarter

[2] https://www.similarweb.com/blog/second-most-popular-social-network-by-state

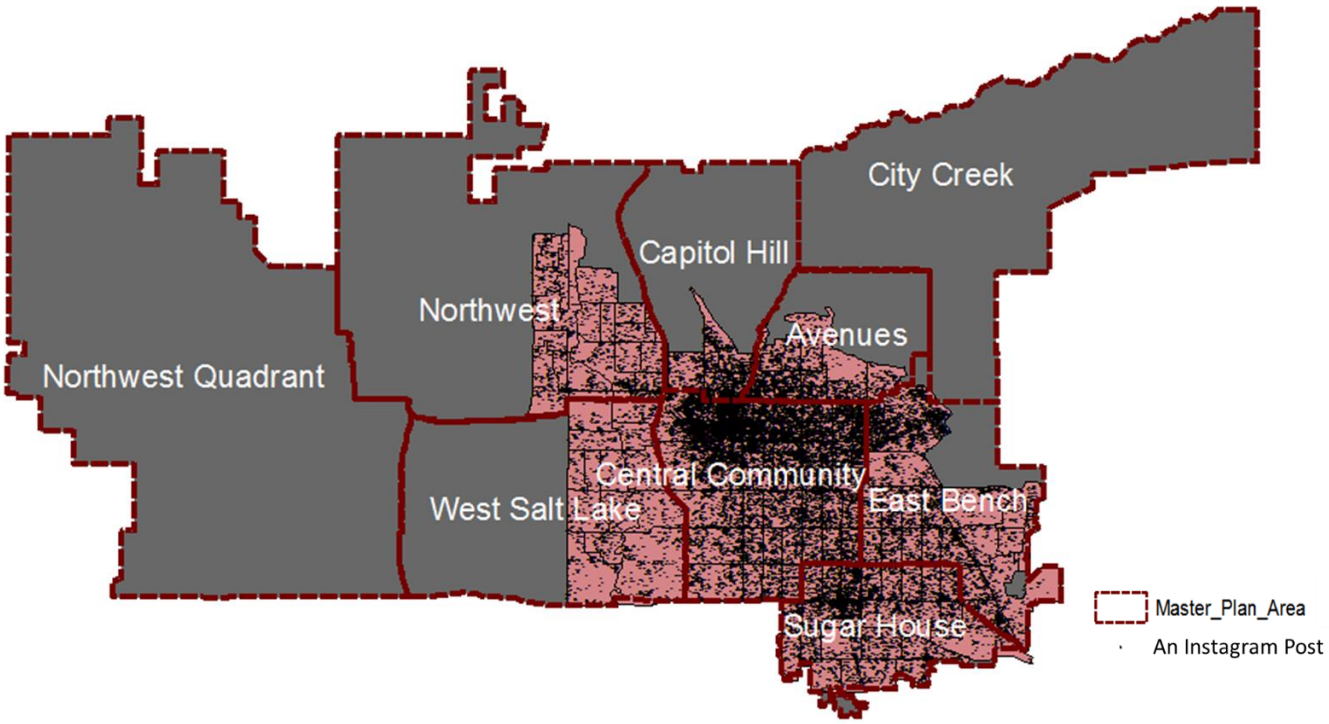


Figure 3. The distribution of Instagram posts between 2013 and 2015

3.4 Data Preprocessing

Before analyzing these posts, this research filtered the posts defined as noise. The sources of noise in the social media include advertisements, marketing, messages, robots, and non-relevant conversations (M.-H. Tsou, 2015). The extreme users were defined as the noise in this research because Instagram posts were used to measure the posting patterns of all users, and a large amount of Instagram posted by a tiny group of users seriously influences the research outcome. From an overview of the raw dataset presented in Figure 4, the pattern of Instagram posts per user matches a power law distribution. It indicates there are extreme users who contributed to a much higher number of Instagram posts. In this research, the method to filter extreme users is to remove users with post counts above the 90th percentile of posting frequencies. The 90th percentile of posting frequencies is 576.6, which means that the users with more than 577 posts during 2013 to 2015 were removed from the raw dataset. Figure 5 represents the dataset after filtering extreme users depicting a steady decrease in user counts by Instagram posts. Table 4 shows the counts of Instagram data before and after the filtering process.

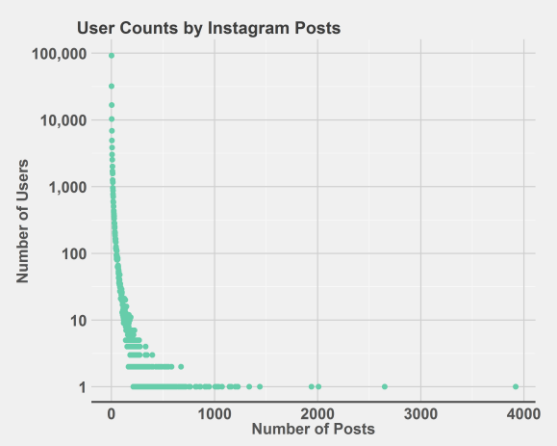


Figure 4. User counts by Instagram posts (before)

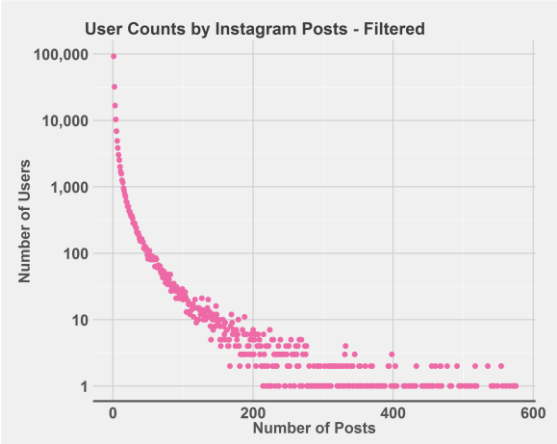


Figure 5. User counts by Instagram posts (after filtering)

Table 4. Data before and after filtering

Dates	Total Counts in Entire County (before filtering)	Total Counts in the Study Area (before filtering)	Total Counts in the Study Area (after filtering)
2013/1/1 ~ 2015/12/31	1,236,070	652,190	640,353

3.5 Data Analysis

3.5.1 Human Perceived Gentrifying Areas

The human-perceived gentrifying areas in this research means the places that people regard as gentrifying based on human perception. People notice the gentrification happens in the neighborhoods because of the increase of house value or rent, the changing in appearance of buildings and environment, and the cultural characteristics associated with gentrification. To qualitatively determine the human-perceived gentrifying areas, this research conducted a simple content analysis on those online news reports and online discussions containing “Salt Lake City” and “gentrification,” recorded the place names mentioned in the content on the map.

3.5.2 Generating the Gentrification Typology Distributions in Salt Lake County

To evaluate how accurate those existing census-based gentrification typology strategies to identify gentrifying areas, this research developed a tool for replicating five census-based strategies, and generating the gentrification typologies distribution in the study area. The five census-based strategies include (1) Bostic and Martin’s method, (2) Freeman’s method, (3) Ding, Hwang and Divringi’s method, (4) McKinnish, Walsh, and White’s method, and (5) The Voorhees Center’s method. The details of these five methods are introduced in the chapter 2. The tool developed was written in Python scripts and worked with a PostgreSQL database. Calculating typologies requires comparing census data in the same census boundary in two different time points. Since the boundaries of census block group in 2000 and 2010 are different, these census data were harmonized to 2010 census block group boundaries using areal weighting method.

Different typology strategies choose different comparative areas. For example, while Bostic and Martin's method uses metropolitan statistical area (MSA) for comparison, Ding's method chooses a city for being the contrast. Specifically, Bostic and Martin's method classifies a MSA as a gentrifiable area if the median income of a census tract is less than 50 percent of the median income in entire MSA. Here, an MSA is the comparative area in Bostic and Martin's method. Ding's method, in contrast, determines a gentrifiable tract if it's median household income is below the citywide median. In Ding's method, a city is a comparative area. For consistency, this research defines Salt Lake City is a comparative area.

Additionally, since the spatial scale of this research is census block group, the demographic or economic attributes from Census data used are based on census block group level, not census tract level that the original methods used.

3.5.3 Quantifying the Nightlife Activities

According to the analysis of Instagram from previous research (Hu et al., 2014), selfies, friends, and activities are the top three topics posted on Instagram. It indicates that most people post Instagram when they are accompanied with friends or in some places where activities happen. Therefore, it is very likely that a person posts an Instagram content because an activity is happening. Given on this finding, this study assumes that the more nightlife activities, the more Instagram posted at night time. The night time in this research means the time during 6 pm to 1 am . Because most nightlife business such as bars and nightclubs opens before or at 6 pm, and they must close at 1 am due to Utah Liquor Law[3][4], Instagram posted during 6 pm to 1 am better capture the nightlife activities. In order to analyze the spatial-temporal patterns of nightlife activities in Salt Lake City, the counts of Instagram night posts per year were grouped by the census block groups, and they were normalized with the population of the place. The normalization is to reduce the influence of the number of residents (Eq. 1). Specifically, the equation is to avoid from the situation that lots of night posts in a census block group are just because many people live there and it has strong chance of getting Instagram night posts. *n* is the normalized count of night Instagram posts in a census block group during a certain year, N is the total posts at night time in a census block group during a year, and P is the total population in a census block group in a year.

$$n = \frac{N}{P} \tag{1}$$

3.5.4 Quantifying the Gentrification Ambience

(1) Gentrification Keywords

This research took two approaches to measure gentrification ambience. One approach is based on the assumption that gentrifying areas have more Instagram posts including text and photos that are related to gentrification keywords. In other words, the more

[3] https://abc.utah.gov/license/licenses_club.html
[4] https://abc.utah.gov/license/licenses_restaurant_beer.html

often gentrification keywords appear in a place, the stronger gentrification ambience it is. A place with stronger gentrification ambience is more likely undergoing gentrification. The second approach assumes that the Instagram posts posted in gentrifying areas are similar and build resembling ambiences. Therefore, grouping Instagram posts for each area by their similarity, gentrifying areas can be distinguishable from other areas.

In this research, gentrification keywords were determined according to existing literatures (Beauregard, 1986; Castells, 1983; Hae, 2011, 2011; Holt, 2008; Jager, 1986; Kern, 2012; Knopp, 1990; Lees et al., 2008; Ley, 1986; Rothenberg, 1995; Van Criekingen, 2009; Zukin, 1989), which were further revised by domain experts. Specifically, there are 26 keywords: *tonight, night, gentrification, gentrifier(s), gentrifying, gentrified, bar(s), coffee, café, restaurant(s), gallery/galleries, young, youth, trendy, aesthetic(s), art, hipster(s), beer, gay(s), lesbian(s), Victorian, bohemian, yoga, yuppie, expensive, and pricey*. After stemming process with *porter stemmer* (Porter, 1980, 2001), the gentrification keywords are converted into*: tonight, night, gentrif, gentry, bar, coffe, cafe, restaur, galleri, young, youth, trendi, aesthet, art, hipster, beer, gay, lesbian, victorian, bohemian, yoga, expens*, and *pricey*.

(2) Gentrification-Keyword Based Method

This approach assumes that there are more Instagram posts associated with gentrification keywords when the place experiences gentrification. Therefore, the patterns of gentrification ambience were measured by calculating the frequency of Instagram posts containing gentrification keywords per year in each census block group. The patterns then were mapped and compared with the five gentrification typology distribution as well as human-perceived gentrifying areas.

Like quantifying nightlife activities, in order to avoid from the influence of the number of residents, the frequency of Instagram containing gentrification keywords was normalized by the population in a block group during a certain year (see Eq. 2). g is the normalized counts of Instagram posts containing gentrification keywords, G is the total posts with gentrification keywords in a census block group during a year, and P is the total population in a census block group in a year.

$$g = \frac{G}{P} \tag{2}$$

(3) Text Clustering Method

The other approach is text clustering method which was exploited to distinguish gentrifying areas by measuring the similarity of Instagram text contents and grouping areas based on the text similarity. The assumption of this approach is that gentrifying areas have similar gentrification ambience, so their Instagram contents are similar. Combining all Instagram text including captions and hashtags in a census block group during a year into a document, and clustering these documents with the text clustering technique, so that the gentrifying block groups could be distinguished from other block groups. In this research, one-year Instagram text in a census block group are called a

document; entire Instagram text in the study area in a year is called a *corpus*; “bag of word” means all words in the corpus, and the term frequency vector is a vector that represents how many times each word appears in a document.

Before applying text clustering, this research applied text pre-processing to each document. It includes removing punctuations, stop words, emoji, and non-English words as well as the stemming process. Stemming methods is to build the basic forms of words. It strips the plural ’s’ from nouns, and the ’ing’ from verbs, or other affixes. (Hotho et al., 2005) In this research, the stemming process was accomplished with *Porter stemmer* (Porter, 1980, 2001) built in NLTK python library. When finishing text pre-processing, each *document* was transformed into a *term frequency vector*. To reflect how important a word is to a document in the entire corpus, every *term frequency vector* was converted into a *TF-IDF* (Term Frequency–Inverse Document Frequency) vector. In particular, through *TF-IDF* technique, larger weights were given to the important words and smaller weights were given to common words. Important words are defined based on two characters: (1) appear frequently in a document (common locally), and (2) appear rarely in entire corpus (rare globally). In other words, if a frequent word in a document appears in many other documents, its *TF-IDF* will be very small. By contrast, if a frequent word in a document appears rarely in other documents, its *TF-IDF* will be very large. The *TF-IDF* equation is shown below (Eq. 3). *W* is the number of this word in this document, and *D* here is the number of documents using this word within the whole corpus.

$$TF - IDF = W \times Log\,(W/(1 + D)) \quad (3)$$

In order to reduce the influence of increasing users over the years and the different size of documents, the *TF-IDF* of each word in a vector was normalized by the number of unique users in a census block group during a year. The normalized *TF-IDF* is shown below (Eq. 4). *U* represents the number of unique users in each census block group during a year.

$$TF - IDF = W \times Log\,(W/(1 + D))/U \quad (4)$$

After normalization, the normalized *TF-IDF* vectors were converted into distance matrixes based on cosine distance for the clustering. The method of clustering used is *hierarchical clustering* which assigned all of the documents to a single cluster and then partitions the cluster into two least similar clusters. It proceeded recursively on each cluster until there is one cluster for every document. The procedure of text pre-processing and text clustering was performed by Python libraries including NLTK

(Natural Language Toolkit) and Graphlab Create*, and R* packages such as NbClust. A statistic method and manual try and error method were mixed for seeking the optimum number of clusters. *Elbow Method* (Ketchen & Shook, 1996)was used to calculate the total WSS (within-cluster sum of the square) for each number (k) of clustering, and plotted the curve of WSS according to the number of clusters k. The significant turning point of the graph indicates the optimum number of clusters. According to *Elbow Method*, the suggested number of the clusters is 3 (Figure 6). Furthermore, this study manually examined cluster results from 3 to 20 groups to determine the appropriate number of clusters that represent gentrifying areas. As a result, this research determined 7 groups.

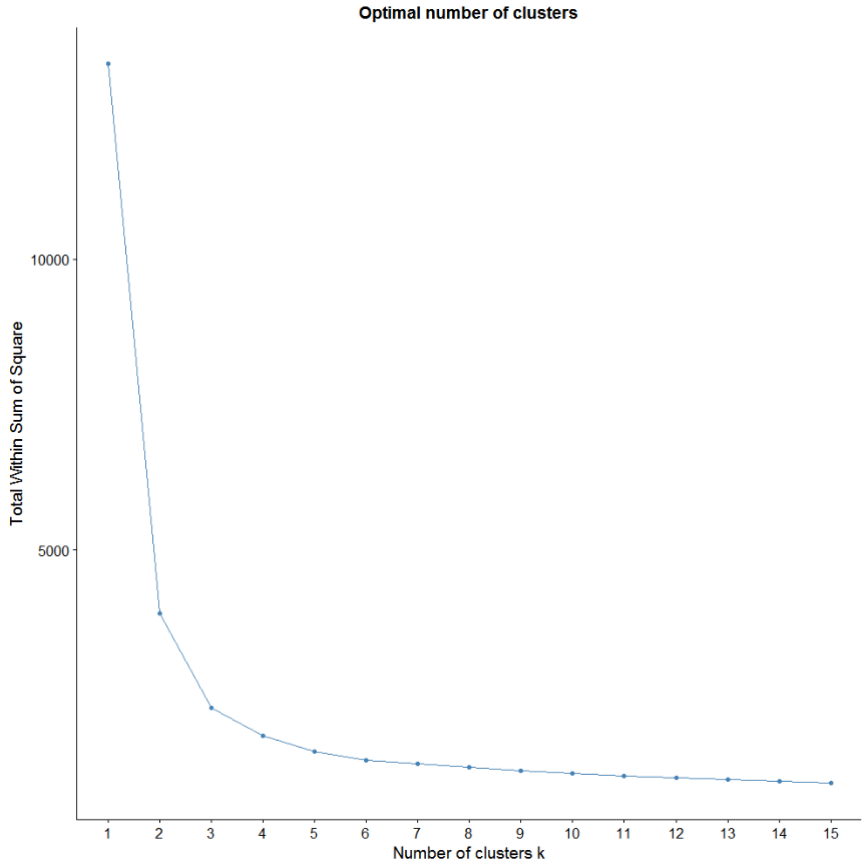


Figure 6. Elbow method

3.5.5 Data Profiling

To discuss the text clustering results, this research examined the land use and word use in each group. The land use information is from 2013, 2014, and 2015 parcel data. Through spatial joining the land use of parcels with text clustering results, total area of

different types of land use in each group was calculated and visualized with *100% Stack Bar Graphs.* These graphs visually present the percentage of different types of land use in each group. For analyzing the word use of each group, the Instagram document of each group was converted to a 36,581-dimential vector (there are 36,581 words in the study area from 2013-2015). Every number in the vector is a normalized *TF-IDF* which represents the importance of a word. Then, all vector was visualized with a bar chart and a word cloud. The bar charts indicate how people use words in the group. Bars in the bar chart evenly distributed, which means that people use various words in the group. By contrast, the bar chart with uneven bars and some obvious peaks, which shows that the group is dominated by some words and these words are relatively rarely appear in other groups. To visualize the peaks in the bar charts, which are the important words in each group, this research generated word clouds for each group based on normalized *TF-IDF* .

Chapter 4: Results

4.1 Human-perceived Gentrifying Areas in Salt Lake City

This research evaluated the social media-based framework by comparing it with the human-perceived gentrifying areas, places people regarded as gentrifying. The Human-perceived gentrifying areas were determined by analyzing qualitatively discussions on online forums and in news reports. Using the Google Search Engine, materials were selected if they contained two keywords, “gentrification” and “Salt Lake City,” and if they were posted during 2000–2015. After analyzing the 11 collected articles, it was determined that 9 places in the study area have experienced gentrification based on the descriptions. Those 9 places are, The Avenues, Sugar House, Downtown Salt Lake City, South of Downtown, Trolley Square, The Marmalade District, West of 300 West, 9th & 9th, and 15th & 15th as shown in Figure 7.

The first is the Avenues. A report on the Great American Country website listed five great neighborhoods that experienced gentrification, including the Avenues. The report described the Avenues as having the typical landscape of gentrification, stating, “The Avenues is an affluent, walkable, outdoor-centric section of town…The Avenues has more diversity in the style of home, with bungalows, Victorians, ramblers/ranches, and two-story houses (Cutler, 2015).” The Visit Salt Lake website mentioned the aesthetics and leisure activities available in the Avenues, which are the distinct cultural factors in a gentrifying area, stating, “Perhaps the quirkiest and artsiest neighborhood of Salt Lake, The Avenues combines beautiful historic residential neighborhoods with hip contemporary features like yoga and Pilates studios, spas and bed-and-breakfasts (Kukura, 2016)”. The Avenues was also referred to as a gentrifying area by the Zillow community in 2011 (Bcgallo et al., 2011).

Sugar House is another neighborhood listed on the Great American Country report. It was described as a “walkable/bike-able” area where “a mix of young professionals and young families” lived (Cutler, 2015). According to this description, the walkable streets and the diversity of young professionals are the characteristics of gentrification. There are other reports about Sugar House. In 2013, a Utah Stories report mentioned that gentrification is happening in Sugar House, stating, “I fear what is happening in Sugar House isn’t the rebirth of a neighborhood. It’s surpassing gentrification and heading straight towards homogenization (“A Local’s Perspective on the Sugar House Development,” 2013).” In 2014, the Reddit community also discussed the gentrification of Sugar House (Kimberlyjo et al., 2014).

Downtown Salt Lake City is another place that has experienced gentrification. In 2011, a real estate broker mentioned the gentrification of the eastern part of downtown on Zillow (Bcgallo et al., 2011). Another Zillow user in the discussion supported this opinion by describing the new condominiums popping up and the refurbishment going on in the south of downtown. The gentrification of downtown is also supported by the Great American Country report, which stated that a large amount of renovations, walkability, and a vibrant nightlife are found in downtown. Additionally, the Artspace project might contribute to the gentrification of downtown: “Artspace has long been involved in Salt Lake City's redevelopment. It revitalizes former industrial sites or abandoned buildings and builds spaces that appeal to tenants who help transform the community (Markosian, 2007).” According to Markosian’s report on the Utah Stories, while it is a nonprofit organization and it aims to provide affordable places for artists to prevent increases in rent, the strategy for redeveloping run-down areas is highly similar to the other cases of gentrification in such neighborhoods as SoHo, New York City. Besides downtown, the south of downtown, Trolley Square, and its surrounding areas were also referred to as gentrifying areas (Bcgallo et al., 2011).

The Marmalade District on the western side of the Capitol was regarded as another gentrifying area. The real estate broker who mentioned the gentrification in downtown Salt Lake City also mentioned the Marmalade District: “The Marmalade area has been seeing a lot of remodels and upgrades in quality; I recently sold a home there that is currently being upgraded. Even West of 300 West has seen some gentrification. (Bcgallo et al., 2011)” The Summit Sotheby’s International Realty (“Living in Salt Lake City,” 2012) and Great American Country reports (Cutler, 2015) have described the gentrification there. In the reports, The Marmalade District was regarded as an “up-and-coming” section of town. It was “a blighted industrial area,” but “rapidly becoming an artsy community with a bohemian feel.” In the same manner, the Marmalade District experienced a transformation from an industrial or disinvested area to a place full of gentrification ambience.

According to the report of Summit Sotheby's International Realty (“Living in Salt Lake City,” 2012) and Great American Country reports (Cutler, 2015), the landscape in the 9th and 9th (at the intersection of 900 East and 900 South Streets) and 15th and 15th (at

the intersection of 1500 East and 1500 South Streets) areas also match the characteristics of gentrification. They were delineated as a place known as its "foot-traffic friendly", "culturally diversity," and "the gay community" full of "quirky shops", "independent movie theaters", "art galleries" and "coffee shops." Based on the online reports, 9th and 9th and 15th and 15th are walkable and culturally diverse places. The coffee shops and art galleries there are the key cultural factors associated with gentrification.

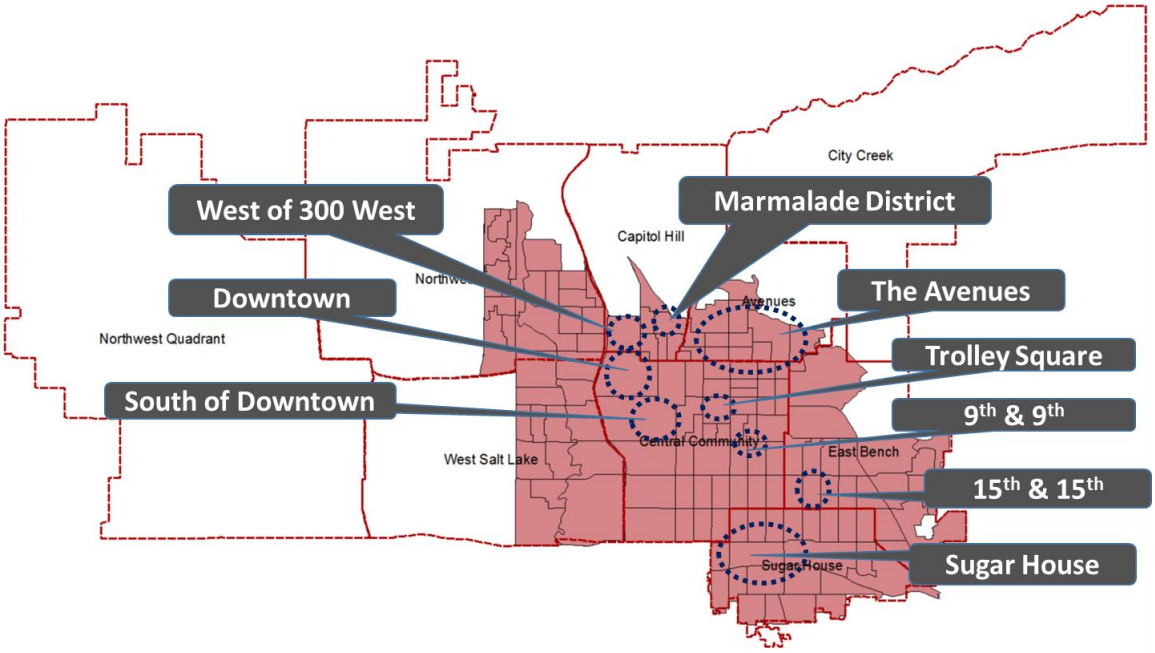


Figure 7. Areas of human-perceived gentrification in Salt Lake City

4.2 Five Gentrification Typology Distributions in Salt Lake County

To examine how applicable the conventional census-based strategies are, this research replicated five census-based gentrification typology strategies, including (1) Bostic and Martin's strategy (2003); (2) Freeman's strategy (2005); (3) Ding et al.' strategy (2010); (4) McKinnish's strategy (2014); and (5) Voorhees Center's method (2015)[5]. Then, the gentrification typology distributions were mapped and compared with the human-perceived gentrifying areas.

The tool used to replicate these gentrification typology distributions was written in Python, which interacted with a PostgreSQL database containing block group-level 2000 decennial census data and 2014 ACS five-year estimates data. Based on the five census-based gentrification typology strategies, census block groups were identified as gentrifying if they had some economic or demographic characteristics in 2000 and if the

[5] The details of these census-based methods are listed in Table 1.

characteristics changed in 2014 in a particular way. Figures 8 through 12 show the gentrification typology distributions mapped with ArcGIS.

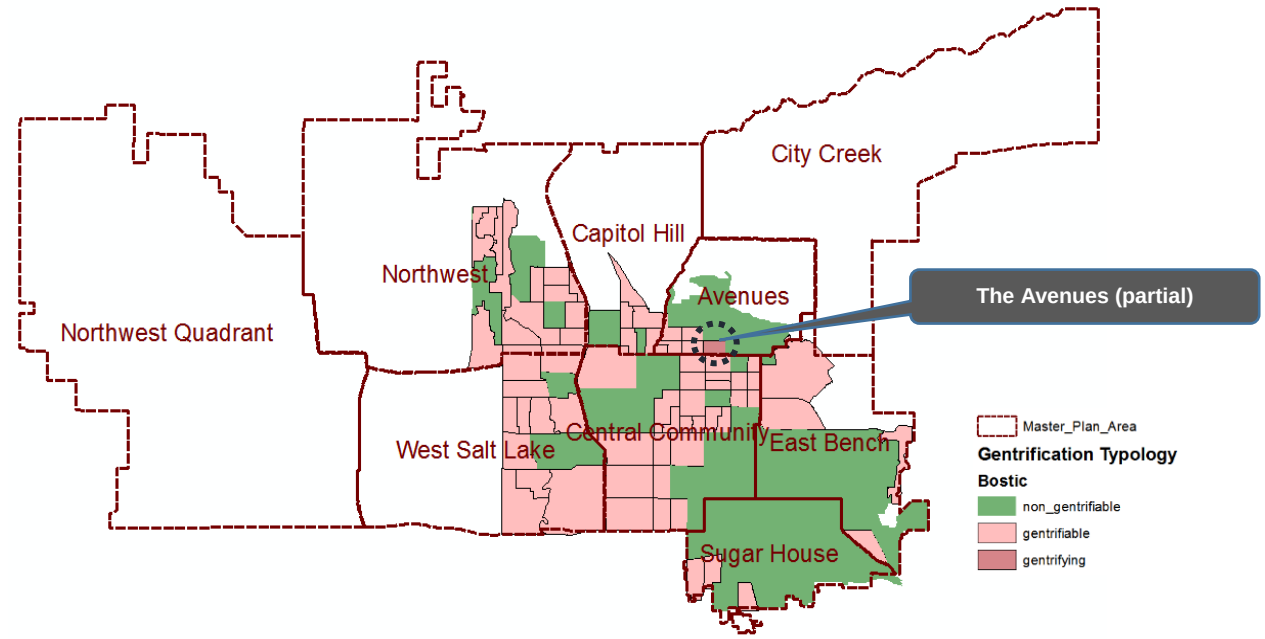


Figure 8. Gentrification typology distribution (Bostic and Martin's method)

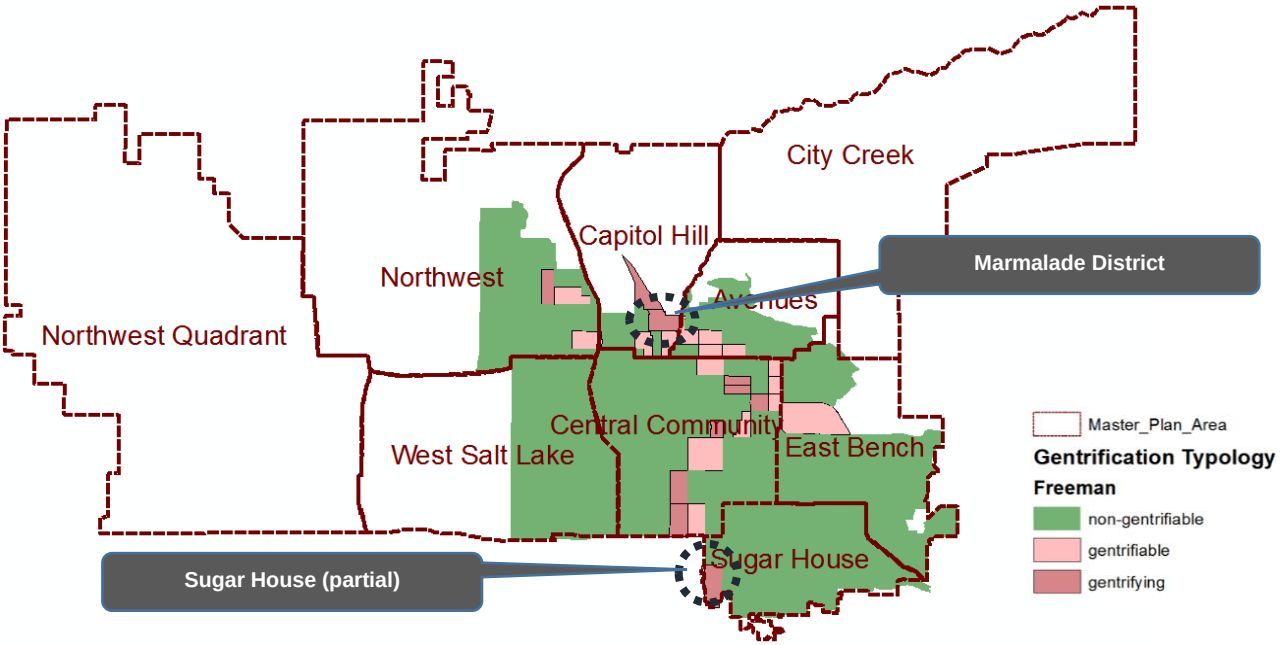


Figure 9. Gentrification typology distribution (Freeman's method)

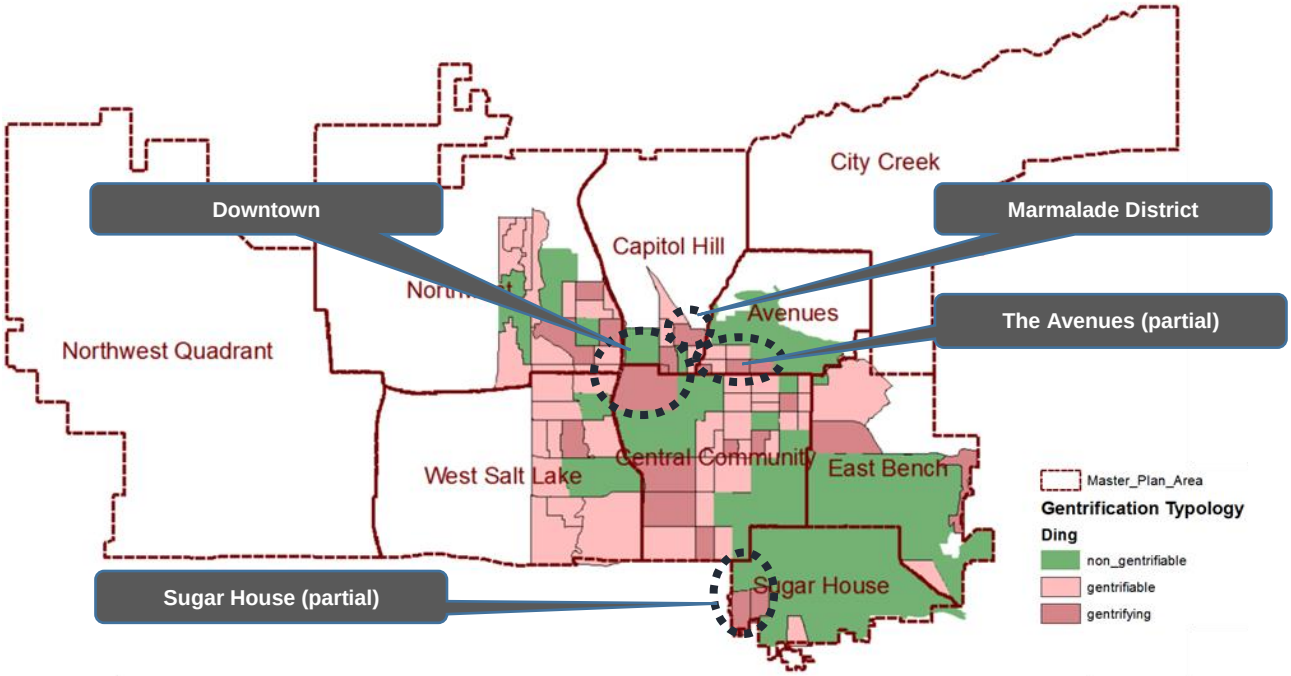


Figure 10. Gentrification typology distribution (Ding et al.'s method)

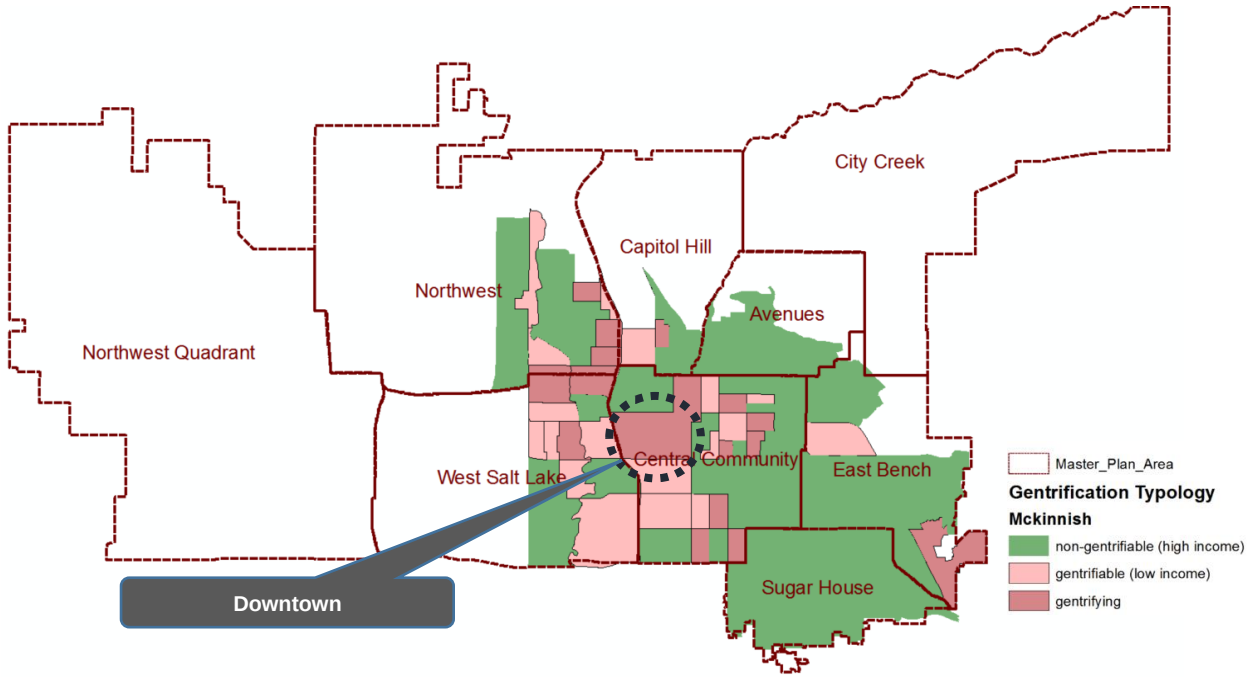


Figure 11. Gentrification typology distribution (McKinnish's method)

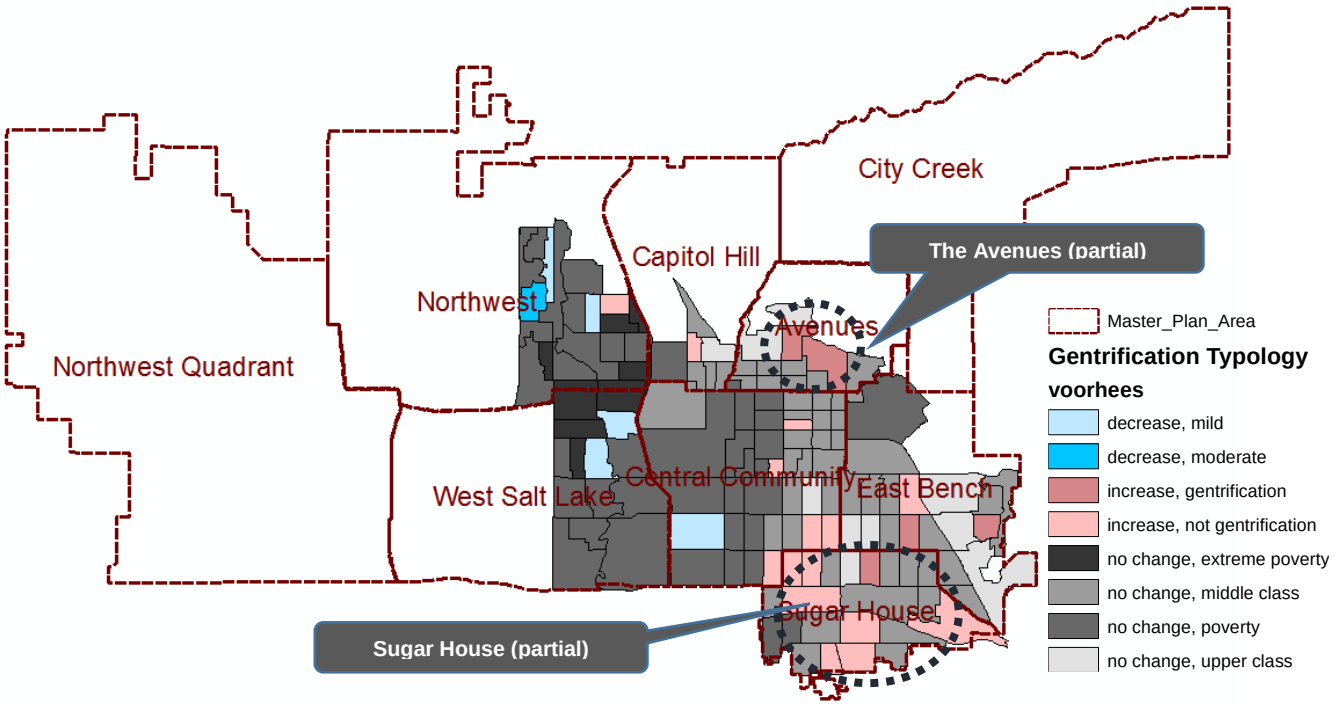


Figure 12. Gentrification typology distribution (Voorhees Center's method)

Surprisingly, while several census block groups were identified as gentrifying by these census-based strategies (Figures 8 to 12), most do not match the human-perceived gentrifying areas (Figure 7). Bostic and Martin's method only identified one gentrifying block group. While their method captured a part of The Avenues, one of the human-perceived gentrifying areas, the other eight gentrifying areas were not identified. Freeman's method identified 10 gentrifying census block groups, and two of them matched the human-perceived gentrifying areas, including the Marmalade District and part of Sugar House. Yet, the other seven areas of human-perceived gentrification were overlooked, including The Avenues, West of 300 West, downtown, south of downtown, Trolley Square, 9th and 9th, and 15th and 15th. Ding et al.'s method identified 21 gentrifying census block groups, seven of which were in areas of human-perceived gentrification. One was the Marmalade District, another was downtown, two were within Sugar House, and the other three were in the Avenues. Namely, Ding et al.'s method captured the Marmalade district, downtown, part of Sugar House, and part of the Avenues. However, the other five areas of human-perceived gentrification were not recognized. McKinnish's method selected 17 gentrifying block groups, but only south of downtown matched the human-perceived gentrifying areas. Voorhees Center's method identified 20 census block groups as areas of socio-economic increases, 10 of which belong to areas of human-perceived gentrification. As shown in Figure 12, two census

block groups identified as increasing in gentrification are in the Avenues and one is in Sugar House. Additionally, 10 census block groups in Sugar House or near Sugar House are classified as not increasing in gentrification. This indicates that while Voorhees Center's method did not regard these places as gentrifying areas, it recognized that the socio-economic index of these areas increased from 2000 to 2014. In other words, the growth of the socio-economic status in these places was highly similar to gentrifying areas, but their overall average socio-economic indices have not reached seven[6]. Nevertheless, the other seven areas of human-perceived gentrification were not identified by this method.

In this case study, the gentrifying areas selected based on Bostic and Martin's method do not match the human-perceived gentrifying areas well because of the method's rigorous standard of comparison with other measurements. Although it uses the same criteria to determine a gentrified area as that of Ding et al.'s method, it only selects one area with the highest average rank among nine gentrification factors to identify a gentrifying neighborhood. In other words, no matter how many possible gentrifying areas there are, this method only recognizes one. If we look to the ranking shown in Table 5, the census block group with the second average rank is in Sugar House, which was also identified by the Voorhees Center's method. The census block group with the third average rank is in the Avenues. This finding indicates that many gentrifying areas were not identified, as Bostic and Martin's method only identified the top ranking area.

[6] The details of Voorhees' Center method are available in Table 1.

Table 5. The top three census block groups based on Bostic and Martin's Method

Bostic and Martin's Method												Name of the place
Census block group ID	Median Income in 2000 (USD)	College-educated people in 2014 (Rank)	Change of family income (Rank)	Homeownership rate in 2014 (Rank)	Population aged between 30–44 in 2014 (Rank)	Poverty rate in 2014 (Inversed Rank)	White non-family households in 2014 (Rank)	Black population in 2014 (Rank)	Managerial workers in 2014 (Rank)	People with some college education in 2014 (Rank)	Ranking	
490351012004	30,368	6	25	32	4	8	21	1	5	21	1	The gentrifying block group based on Bostic and Martins' method
490351048003	29,590.99	18	26	8	24	1	25	1	21	2	2	A block group in Sugar House
490351011022	29,537	21	12	34	62	4	4	1	4	13	3	A block group in the Avenues
* Median household income: 35,192 USD												

Table 6. The two gentrifiable census block groups based on McKinnish's method

	McKinnish's method		Ding et al.'s method		
	Average family income in 2000 (USD)	Average family income increase (USD)	Median percentage increase in median home value (%)	Median percentage increase in median gross rent (%)	Increased number of college-educated people (person)
490351029002	19,731.99	-13,178.42	59	65	118
490351029001	27,499.99	4607.21	68	3	244

	*The bottom quantile of average family income in 2000: 27,464.50 *The median percentage increase in median home value: 58% *The median percentage increase in median gross rent: 46% *The median increase in the share of college-educated residents: 77 people

Table 7. A census block group downtown

	The conditions of a gentrifiable area (Freeman's method)		The conditions of a gentrifiable area (Ding et al.'s method)	Other socioeconomic factors	
Block group ID	Median household income in 2000 (USD)	The proportion of housing built within the past 20 years (%)	Median household income in 2000 (USD)	The percentage increase in median gross rent (%)	The percentage increase in college-educated people (%)
490351025001	22,786	74	22,786	63	381
* Median household income: 35,192 USD * The median proportion of housing built within the past 20 years: 0.05					

Table 8. A census block group in the Avenues

Ding et al.'s method				Freeman's method		
Block group ID	Median percentage increase in median home value (%)	Median percentage increase in median gross rent (%)	Increased number of college-educated people (person)	Percentage increase in college-educated people (%)	The change of median home value (USD)	
490351011012	-40	57	118	41	-80,600	
*The median percentage increase in median home value: 58% *The median percentage increase in median in median gross rent: 46% *The median increase in the share of college-educated residents: 77 people						

Table 9. Two census block groups in the Avenues and Marmalade District

Voorhees Center's Method			Ding et al.'s method		
Block group ID	Overall socioeconomic score	Socioeconomic change from 2000–2014	Median percentage increase in median home value (%)	Median percentage increase in median gross rent (%)	Increased number of college-educated people (person)
490351025001	1	-1	-16	63	751
490351007002	6	0	62	32	234
* No increase in gentrification: overall socioeconomic score must be <7 and the socioeconomic change from 2000–2014 must be >4 * Increase in gentrification: overall socioeconomic score must be >7 and the socioeconomic change from 2000–2014 must be >4					

Table 10. Census block groups in the Upper Avenues and the core of Sugar House

Ding et al.'s method (the criteria of a gentrifiable area)		
Block group ID	Median household income in 2000 (USD)	The places
490351010001	46,793	The block group in the Upper Avenues
490351012001	45,603	The block group in the Upper Avenues
490351141001	38,000	The block group in the core of Sugar House
* Median household income: 35,192 USD		

Table 11. Census block groups in western Sugar House

Ding et al.'s method (the criteria of a gentrifying area)				Voorhees Center's method	
Block group ID	Median percentage increase in median home value (%)	Median percentage increase in median gross rent (%)	Increased number of college-educated people (person)	Overall socioeconomic score	Socioeconomic change from 2000–2014
490351049003	88	53	151	6	0
490351049002	75	43	194	2	0
*the median percentage increase in median home value: 58% *the median percentage increase in median in median gross rent: 46% *the median increase in the share of college-educated residents: 77 people				* **No increase in gentrification**: overall socioeconomic score must be <7 and the socioeconomic change from 2000–2014 must be >4 * **Increase in gentrification**: overall socioeconomic score must be >7 and the socioeconomic change from 2000–2014 must be >4	

The gentrifying areas identified by McKinnish's method also do not match the human-perceived gentrifying areas well, as this method only recognized one area of human-perceived gentrification and overlooked the other eight areas. The major reason is its strict definition of a gentrifiable area. A place identified as gentrifying must meet the criteria of a gentrifiable area at a prior time point. Thus, the more rigorous the definition is of a gentrifiable area, the fewer the number of areas that can be regarded as gentrifying at a later time point. According to this method, a gentrifying area must have an average family income in the bottom quantile of the entire city at the beginning of a certain period. Compared to other methods, such as Bostic and Martin, Freeman, and Ding et al.'s methods, using the median of the median household income at the prior time point as the threshold of being a gentrifiable area, McKinnish's criteria are more rigorous. In other words, if a place has a high median income the prior year, no matter how much its median income increases after a period, it will not be recognized as gentrifying. Therefore, many middle-class neighborhoods that have experienced gentrification were not recognized. Except for south of downtown, none of the other areas of human-perceived gentrification meet McKinnish's criteria. The other reason is that this method only relies on economic change; it does not consider demographic factors. Table 6 shows the two examples, wherein two census block groups are gentrifiable but not gentrifying based on McKinnish's method, but they were identified as gentrifying by Ding et al.'s method. Specifically, their average incomes were lower than 10,000 USD, which does not meet the criteria of a gentrifying area in McKinnish's method. However, their increase in the percentage of either the median home value or the gross rent is greater than the citywide median percentage increase. The increased number of college-educated people is also greater than the median of the city. Thus, the two census block groups in Table 6 were regarded as gentrifying using Ding et al.'s definition. Therefore, McKinnish's method ended up identifying areas that were low-income neighborhoods at a prior time point and experienced rapid economic upgrading during a period, but it did not consider whether displacement or a culture of gentrification appeared in this place.

Freeman's method is highly similar to Ding et al.'s method; however, Ding et al.'s method identified more areas of human-perceived gentrification, including the Marmalade District, downtown, some areas in the Avenues, and a part of Sugar House. One of the reasons is that Freeman's method defines a gentrifiable area as not only low income but also disinvested in the prior year, whereas Ding et al.'s method considers any low-income area as a gentrifiable area. In particular, a gentrifiable area for Freeman must meet two criteria. The first is that the median income of the area in the prior year must be lower than the median of the city, and the second is that the area's proportion of housing built within the past 20 years must be lower than the median proportion of the entire city. Yet, based on Ding et al.'s method, having a median income in the prior year that was lower than the median of the city is the only criterion to determine a gentrifiable area. Therefore, in Freeman's method, places with a high percentage of new construction in the past 20 years are not considered gentrifiable, even though there are significant housing value or rent increases and increases in the number of college-

educated residents. Table 7 shows one example in downtown[7], where the proportion of housing built within the past 20 years is more than 70%, so it does not match the definition of a gentrifiable area according to Freeman's method. However, its number of college-educated residents increased by more than 300% and the rent increased by more than 60%, which reflect the change in the quality of amenities, the increasing housing costs, and the increase in young professionals. These changes reveal that downtown experienced gentrification from 2000 to 2014. The other reason is that when determining whether a place is gentrifying, Ding et al.'s method evaluates the increase in either home values or gross rent, but Freeman's method only uses an increase in housing prices as an indicator of neighborhood upgrades. One example is the census block group in the Avenues[8] (Table 8). Although the house prices decreased in the area, the gross rent increased significantly. Therefore, it was identified as gentrifying by Ding et al.'s method, but as not gentrifying by Freeman's measurement.

Compared to other measurements, Ding et al.'s method and Voorhees Center's method are more effective. The former method recognized the greatest number of different areas of human-perceived gentrification, including the Marmalade District, the Avenues, downtown, and Sugar House. In total, four areas of human-perceived gentrification were identified. Yet, Voorhees Center's method identified the greatest number of gentrifying census block groups in areas of human-perceived gentrification. In other words, although only two areas of human-perceived gentrification were identified by Voorhees Center's method, including the Avenues and Sugar House, there were 10 identified census block groups in these two areas, including three areas with an increase in gentrification and seven areas with no increase in gentrification. There are two reasons that Voorhees Center's method performs well. First, it is based on a change in the comprehensive socio-economic status. Namely, without a median household income as a threshold in a gentrifiable area, Voorhees Center's method is the only strategy that considers upper middle-class areas when evaluating gentrification. Different from other methods that only regard low-income neighborhoods as gentrifiable, Voorhees Center's method is able to detect middle-income or upper middle-income neighborhoods that have experienced gentrification, which is why Voorhees Center's method identified 10 gentrifying block groups in the affluent Avenues and middle-income Sugar House districts (Figure 12). Second, the socio-economic status includes a detailed evaluation of demographic characteristics that reflect better the displacement of residents through gentrification. Different from McKinnish's method, which ignores resident displacement, or Freeman's and Ding et al.'s methods, which only use the increase in college-educated residents as an indicator of resident displacement, socio-economic status, used by Voorhees Center's method, is a detailed evaluation of demographic characteristics. These characteristics include the percentage of white people, African Americans, Latinos, the elderly, children, college-educated people, managers, families with children, and private school attendees in the area. Although Bostic and Martin's

[7] See the details in the appendix. Block Group ID: 490351025001
[8] See the details in the appendix. Block Group ID: 490351011012

method also considers nine demographic characteristics, only the top-ranked area is identified as gentrifying, which makes it less effective than Voorhees Center's method.

In general, the gentrifying areas identified by both Ding et al.'s method and Voorhees Center's method better match the human-perceived gentrifying areas. However, the gentrifying block groups identified by these two methods do not coincide. Ding et al.'s method recognized downtown[9] and the Marmalade District[10], but Voorhees Center's method did not, as shown in Table 9. The reason for this is that house values or gross rent and the number of college-educated people increased dramatically in these two places, and these factors met the criteria of Ding et al.'s method. Yet, Voorhees Center's method identified a gentrifying area based on its change in socio-economic status. If there was no significant change in an area's socio-economic status, it was not identified by Voorhees Center's method. Additionally, although both strategies recognized partial areas of gentrification in the Avenues and Sugar House districts, those areas did not match each other. Ding et al.'s method recognized two block groups[11] in the western part of Sugar House and three block groups[12] in the southern part of the Avenues. Meanwhile, Voorhees Center's method identified two block groups[13] in the Upper Avenues and a couple of block groups[14], including those with an increase in gentrification and with no increase in gentrification, in Sugar House, but not the two identified by Ding et al.'s method. As shown in Table 10, Ding et al.'s method could not identify the Upper Avenues and the core of Sugar House because these places had high household incomes in 2010, which excludes them from being gentrifiable areas based on Ding et al.'s threshold of a gentrifiable area. Table 11 indicates that the two western block groups in Sugar House have experienced a rapid increase in house prices and educated residents, making them identifiable by Ding et al.'s method. Nevertheless, their static socio-economic status explains why Voorhees Center's method could not recognize them.

In conclusion, there is inconsistency among these five existing gentrification typologies. This result corresponds with Barton's (2016) findings in New York, which were that the gentrifying neighborhoods selected by Bostic and Martin's method and Freeman's method varied greatly. Yet, these methods capture areas of human-perceived gentrification to some extent. Among these five strategies, Ding et al.'s method and Voorhees Center's method identified more areas of human-perceived gentrification. However, even Ding et al.'s method can only identify four areas of human-perceived gentrification out of nine gentrifying areas, and no gentrification typology strategy is able to recognize the Trolley Station, 9th and 9th, and 15th and 15th neighborhoods. This

[9] See the details in the appendix. Block Group ID: 490351025001
[10] See the details in the appendix. Block Group ID: 490351007002
[11] See the details in the appendix. Block Group ID: 490351049003, 490351049002
[12] See the details in the appendix. Block Group ID: 490351011021, 490351011012, 490351012004
[13] See the details in the appendix. Block Group ID: 490351010001, 490351012001
[14] See the details in the appendix. Block Group ID: 490351033002, 490351033003, 490351038001, 490351141001, 490351047002, 490351048003, 490351048002, 490351044001

result suggests that some gentrifying areas observed by humans cannot be detected by census-based methods that only rely on economic growth and resident displacement. One reason is that gentrification involves not only changes in economic and demographic characteristics, but also the emergence of a distinct culture usually detected by human perception. Thus, to examine gentrification dynamics better, human perceptions of the neighborhoods and the measurements of culture should be incorporated when identifying gentrifying areas.

In the following sections, the study introduces a novel social media-based framework that identifies gentrifying areas by recognizing the cultural characteristics of gentrification.

4.3 Nightlife Activity and Gentrification

To detect the cultural characteristics of gentrification, this research used two gentrification indicators. One is nightlife activities and the other is gentrification ambience. Nightlife activities are the distinct culture of gentrifying places, as discussed in the literature (Currid, 2009; Florida, 2002; Hae, 2011, 2011).

Nightlife activities in the study area during 2013 to 2015 were quantified by calculating and normalizing the number of Instagram night posts in 2013, 2014, and 2015 per census block group. Figures 13 to 15 present maps that visualize the results. On these maps, normalized Instagram night posts were classified into seven groups using Jenks Natural Breaks algorithm (Jenks, 1967). The brightness represents the frequency of Instagram night posts. The darker the census block group, the more Instagram night posts it had.

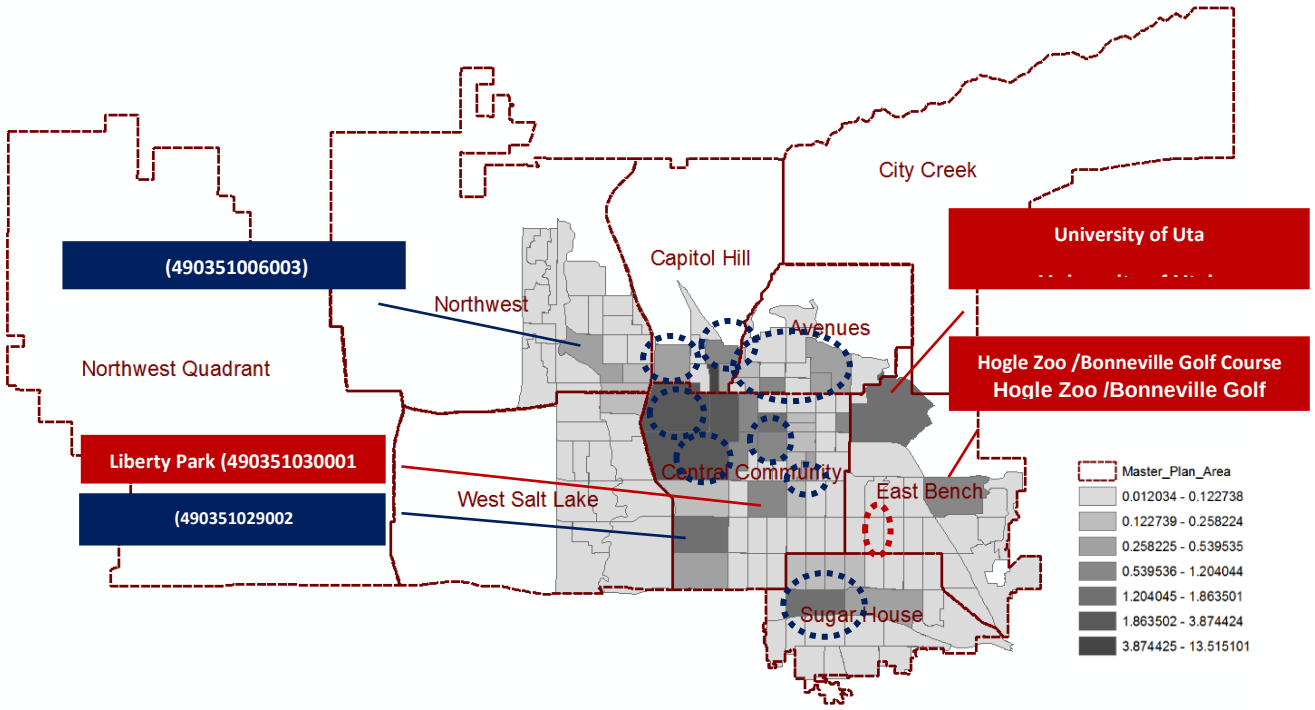


Figure 13. Patterns of Instagram night postings in 2013

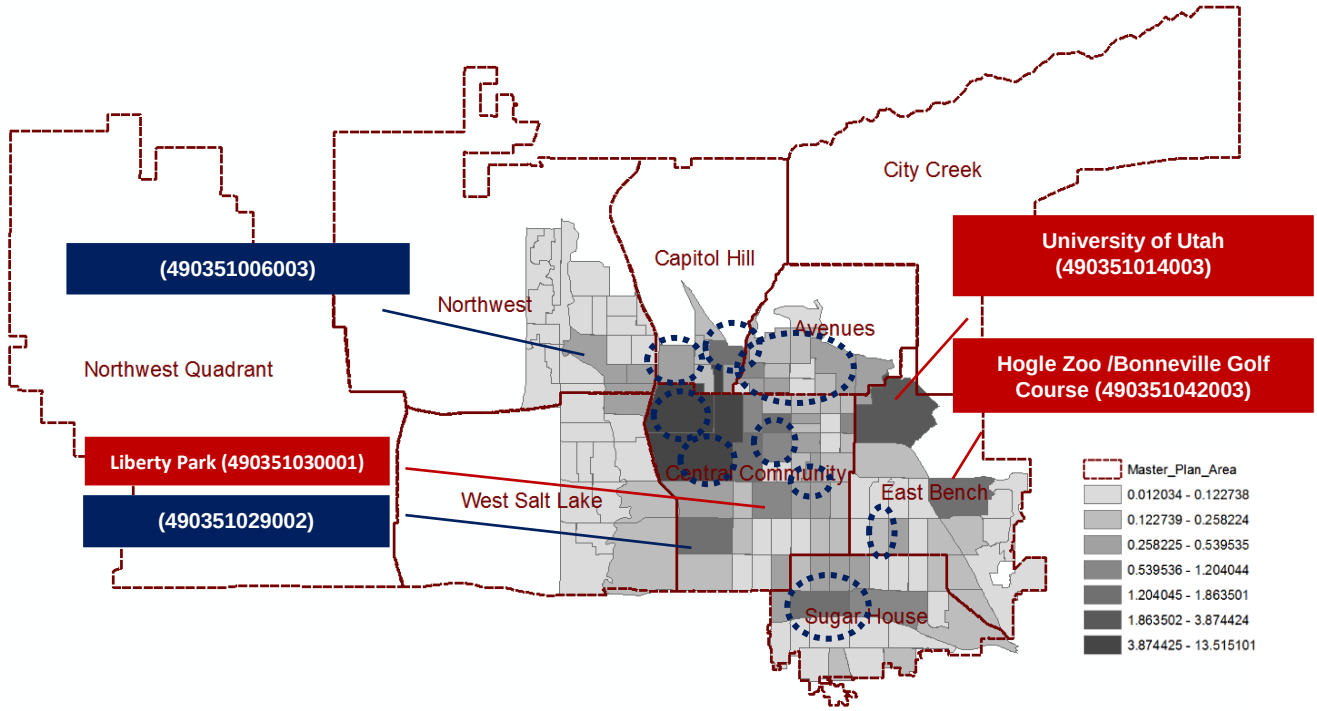


Figure 14. Patterns of Instagram night postings in 2014

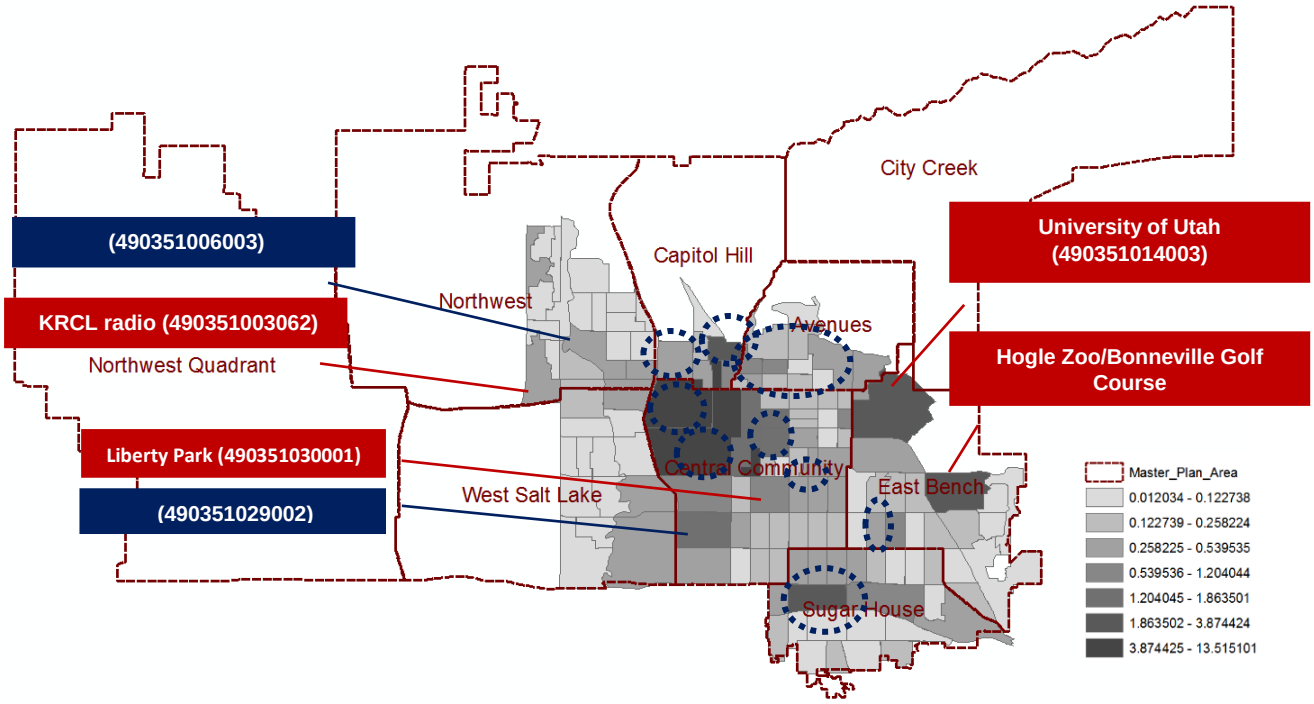


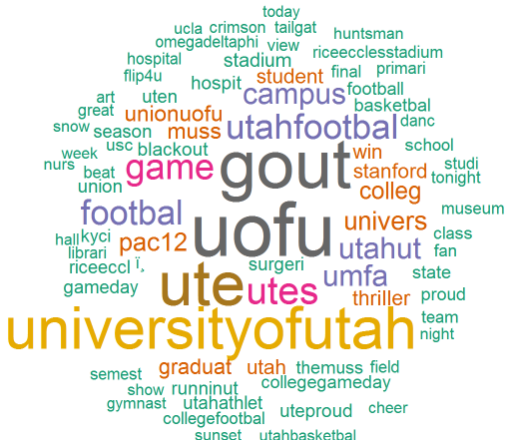
Figure 15. Patterns of Instagram night postings in 2015

The circled areas on these maps are the human-perceived gentrifying areas. As shown by the blue circles, most areas of human-perceived gentrification are located in those census block groups with the highest or a higher frequency of night posts. Among the human-perceived gentrifying areas, downtown and south of downtown had the most Instagram posts across three years, and their normalized numbers of night posts were between 3.87 and 13.52. Additionally, the frequency of night posts in the Marmalade District, Trolley Station, and Sugar House was between 1.86 and 3.87, indicating significantly more night posts. However, the patterns of nightlife activities do not perfectly coincide with the human-perceived gentrifying areas. West of 500 West and 9th and 9th only had slightly more night posts. As the red circles show in Figure 13, 15th and 15th had the lowest frequency of night posts in 2013. This result reveals that nightlife activities quantified with the number of Instagram night posts can help us identify gentrifying areas to some extent.

While there is a relationship between the number of Instagram night posts and areas of human-perceived gentrification, some census block groups with a high frequency of night posts do not match the human-perceived gentrifying areas. As indicated by the blue text boxes in Figures 13 to 15, the block group IDs 490351029002 and 490351006003 are darker on these maps, but they are not areas of human-perceived gentrification, as derived from news reports and online forums. Instead, they were identified as gentrifying by Ding et al.'s method. This result implies that night posting patterns on Instagram can further help us to identify places that experience gentrification but that are not mentioned in the mass media. This finding also agrees with Sampson (2013) and Barton's (2016) argument, which states that media sources may emphasize the changes in particular areas and thus overlook the changes in other areas. Specifically, they posit that when discussing gentrification in New York City, the media usually mentioned Brooklyn and Manhattan.

However, as indicated by the red text boxes on these maps, the inconsistency between night posting patterns and areas of human-perceived gentrification might also be due to some popular venues. To explore further the venues and the activities in such areas as block group IDs 490351030001, 490351014003, 490351042003, and 490351003062, a word cloud analysis was applied to their Instagram night posts (Figures 16A– Figures 16D).

A.



B.

C.

D.

Figure 16. Word clouds of four census block groups. (A) is the word cloud based on the Instagram text posted in block group ID 490351030001; (B) is the word cloud based on the Instagram text posted in block group ID 490351014003; (C) is the word cloud based on the Instagram text posted in block group ID 490351042003; (D) is the word cloud based on the Instagram text posted in block group ID 490351003062.

A word cloud is a visual representation technique used to depict keywords from a text-based data source by highlighting frequent keywords with larger text sizes[15]. Figure 16A depicts “libertypark,” “liberty,” and “park” as frequently posted Instagram captions in Block Group ID 490351030001. This indicates that many visitors to Liberty Park contributed to the night posts. As shown in Figure 16B, the words people used most frequently in block group ID 490351014003 were “gout,” “ute,” and “uofu,” where “gout” means “goutes” after the stemming process. All these words refer to the University of Utah, which shows that the Instagram night posts here were made by many college students. “Zoo” and “hoglezoo” are the largest words in Figure 16C, indicating that the high frequency of night posts in block group ID 490351042003 resulted from visits to Hogle Zoo. “Krclradio” is one of the biggest words in Figure 16D, indicating KRCL radio is located here. “Prowrestl” and “ucwzero” are the other two biggest words in Figure 19, where “prowrestl” means “prowrestler” after the stemming process. These two words refer to an American independent professional wrestling promotion based in Salt Lake City, Utah. According to this word cloud, many Instagram night posts were made because of events hosted by KRCL radio or because of wrestling competitions.

Thanks to the flexibility of the Instagram data, the dynamics of nightlife activities over three years were visualized. Based on Figures 13 to 15, nightlife activities were concentrated in the same places during 2013 to 2015, but the nightlife became more vibrant over the years. Take the Marmalade District for example, where the number of Instagram night posts was between 0.54 and 1.20 in 2013, but which increased to 1.20–1.86 in 2014. In 2015, it rose to 3.83–13.52. From 2013 to 2015, nightlife activities also expanded along the Avenues and Sugar House surrounding areas. Although, without a fine-grained census-based strategy or conducting fieldwork, it is difficult to prove whether these expanding areas were gentrifying during these three years, as the spatio-temporal shift suggests the nightlife patterns changed, and the areas with increased nightlife activities shared a similar cultural characteristic to a gentrifying area.

In sum, there is a relationship between Instagram night posting patterns and areas of human-perceived gentrification. However, without a statistical analysis, the result does not show how accurate the Instagram night posting approach is in identifying gentrifying areas, but suggests that the patterns of night posts can help us to capture the distribution of a sense of gentrification and its yearly shift. In addition, two gentrifying areas identified by a census-based strategy match the places with a high frequency of night posts, even though they were not mentioned on news reports or in online forums. This suggests that the social media-based framework identified some gentrifying areas that cannot be detected by human perception. However, the evaluation of gentrification dynamics using Instagram night posts has flaws. On the one hand, although 15th and 15th was regarded as a gentrifying area based on a media source, the number of Instagram night posts from this district was not significantly higher than that from other non-gentrifying areas. On the other hand, the results were influenced by some venues

[15] https://commons.wikimedia.org/wiki/Category:Word_clouds

that were irrelevant to nightlife, including the park, the zoo, the college, and the wrestling promotion. Furthermore, the change in night posts might also be influenced by the increase in Instagram users.

4.4 Gentrification Ambience and Gentrification

Gentrification ambience is a distinct cultural factor that appeals to gentrifiers, as discussed in previous studies (Carpenter & Lees, 1995; Jager, 1986; Ley, 1994; Mills, 1988; Zukin, 1989). Therefore, gentrification ambience was used as the other gentrification indicator to detect the cultural characteristics of gentrification in the following sections. Two approaches were taken to quantify gentrification ambience: (1) a gentrification keyword-based method and (2) a text clustering method. Both these two approaches measure how people perceive a place by analyzing what people talk about in that place. The gentrification keyword-based method assumes that gentrification ambience is embodied in the frequency of gentrification keywords; a gentrifying place should have more gentrification keywords than a non-gentrifying place. Meanwhile, the text clustering method assumes the ambience of a place is embodied in Instagram posts, so gentrifying areas should have similar Instagram texts because they have similar ambience. The results were examined by comparing gentrification ambience patterns with areas of human-perceived gentrification.

4.4.1 Gentrification Keyword-based Approach

The frequency of Instagram posts containing gentrification keywords in 2013, 2014, and 2015 per census block group was calculated and normalized by the population (Eq. 2). Based on the literature review and as revised by domain experts, the gentrification keywords include *night, tonight, gentrification, gentrifier(s), gentrifying, gentrified, bar(s), coffee, café, restaurant(s), gallery/galleries, young, youth, trendy, aesthetic(s), art, hipster(s), beer, gay(s), lesbian(s), Victorian, bohemian, yoga, yuppie, expensive*, and *pricey*. However, because the words *gentrification, gentrifier(s), gentrifying, gentrified, yuppie*, and *pricy* do not appear in the Instagram dataset, after the stemming process, there were only 20 gentrification keywords used in this research, including *night, tonight, bar, coffe, café, restur, gelleri, young, youth, trendi, aesthet, art, hipster, beer, gay, lesbian, Victorian, bohemian, yoga*, and *expens*.

The patterns of gentrification keywords were mapped and they are shown in Figures 17 to 19. Like the maps of nightlife activities, the normalized Instagram posts are classified into seven groups using Jenks Natural Breaks algorithm (Jenks, 1967). The brightness represents the frequency of Instagram posts, so the darker the census block group, the more Instagram posts the area generated containing gentrification keywords. Additionally, blue circles refer to areas of human-perceived gentrification.

According to Figures 17 to 19, areas of human-perceived gentrification are related to areas with frequent gentrification keywords. Specifically, all the human-perceived

gentrifying areas have a higher or the highest frequency of gentrification keywords. Over three years, downtown and south of downtown always have the most gentrification keywords; the number of normalized Instagram posts containing gentrification keywords was 0.45 to 3.72. The Marmalade District, Trolley Square, and Sugar House also have a significantly higher frequency of gentrification keywords. Especially, in 2015, Instagram posts generated in these areas that contained gentrification keywords rose to 0.45–3.72, equal to the numbers found in downtown and south of downtown. Other areas of human-perceived gentrification have a slightly higher frequency of gentrification keywords, including West of 500 West, the Avenues, 9th and 9th, and 15th and 15th. In 2013, there were 0.03 to 0.06 keywords per capita in West of 500 West and 9th and 9th and 0.001 to 0.03 keywords per capita in 15th and 15th. These numbers were slightly higher than most areas with only 0 to 0.001 keywords. In 2015, the numbers of keywords in West of 500 West, the Avenues, 9th and 9th, and 15th and 15th were 0.06–0.13 and 0.13–0.27 per capita. This result indicates that the patterns of gentrification ambience quantified by calculating the number of Instagram posts with keywords help us identify gentrifying areas to some extent.

However, there is inconsistency between the patterns of keywords and the human-perceived gentrifying areas. One of the reasons is that the patterns of keywords reference some gentrifying census block groups that cannot be identified by human perception. The blue text boxes in Figures 17 to 19 show two examples. They were identified by Ding et al.'s method but they are not located in areas of human-perceived gentrification. This result was also found when analyzing night posting patterns. The influence of certain venues is another reason. Corresponding with the findings in the last section, Liberty Park, the University of Utah, and Hogle Zoo are in block group IDs 490351030001, 490351014003, and 490351042003 (the red text boxes), respectively, and the frequent keywords used in Instagram posts from these areas are generated by the many visitors and students in some venues not relevant to gentrification.

While the overall patterns of gentrification keywords and the patterns of night posts are similar, the patterns in block group ID 490351028022 are different. There were many gentrification keywords posted in this area, but the frequency of night posts was not significantly higher than in other places. This study examined block group ID 490351028022 further by generating a word cloud based on the Instagram text in this location (see Figure 20), and it was found that many keywords were generated due to many Instagram posts from a café there, which indicates that gentrification keyword patterns were influenced by only one shop.

Furthermore, the spatio-temporal dynamics of gentrification keywords are visualized in Figures 17 to 19. The gentrification keywords are concentrated in the same places during 2013 to 2015, but the number of posts containing gentrification keywords per capita increased over the years. Take Sugar House for example, whose normalized number of keywords was between 0.06 and 0.13 in 2013, but that number increased to 0.27–0.45 in 2014 and to 0.45–3.72 in 2015. Besides, the areas with a higher frequency

of keywords expanded each year. Although the association between this expansion and the gentrifying areas needs further evaluation, the spatio-temporal shift suggests the gentrification keywords were posted in more places and the sense of gentrification was expanding.

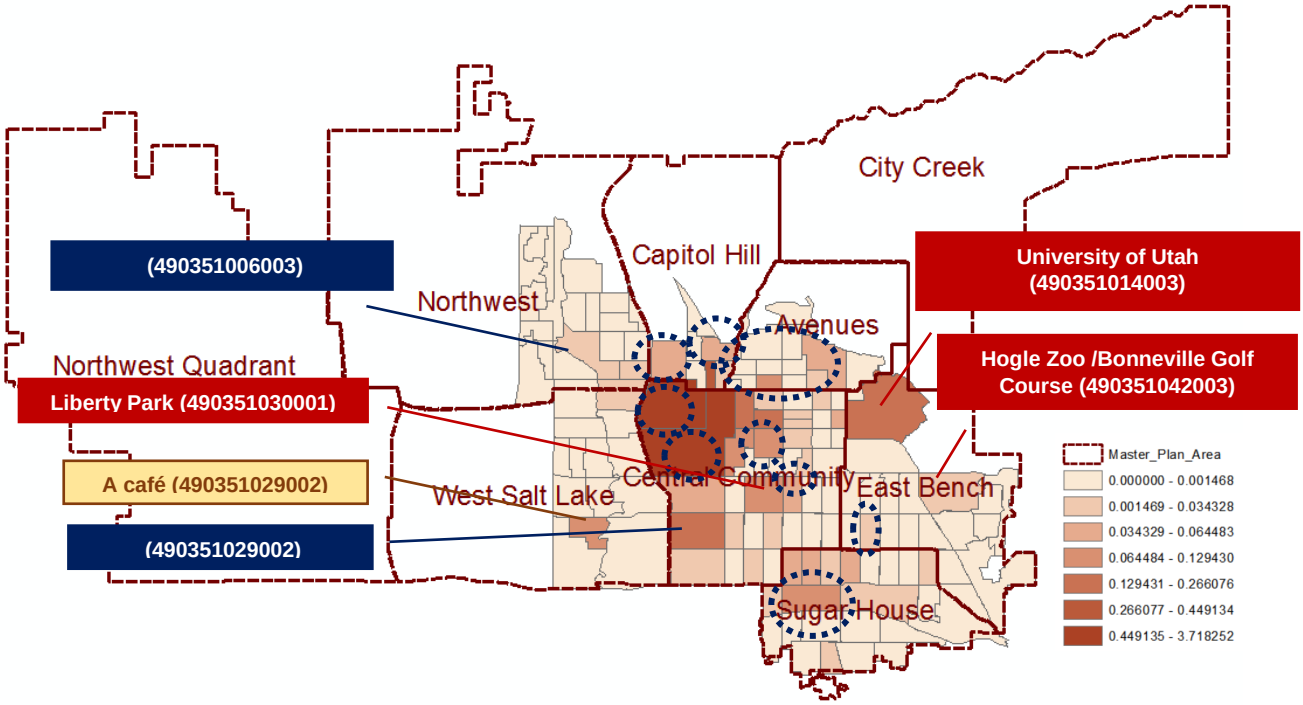


Figure 17. Patterns of gentrification keywords in 2013

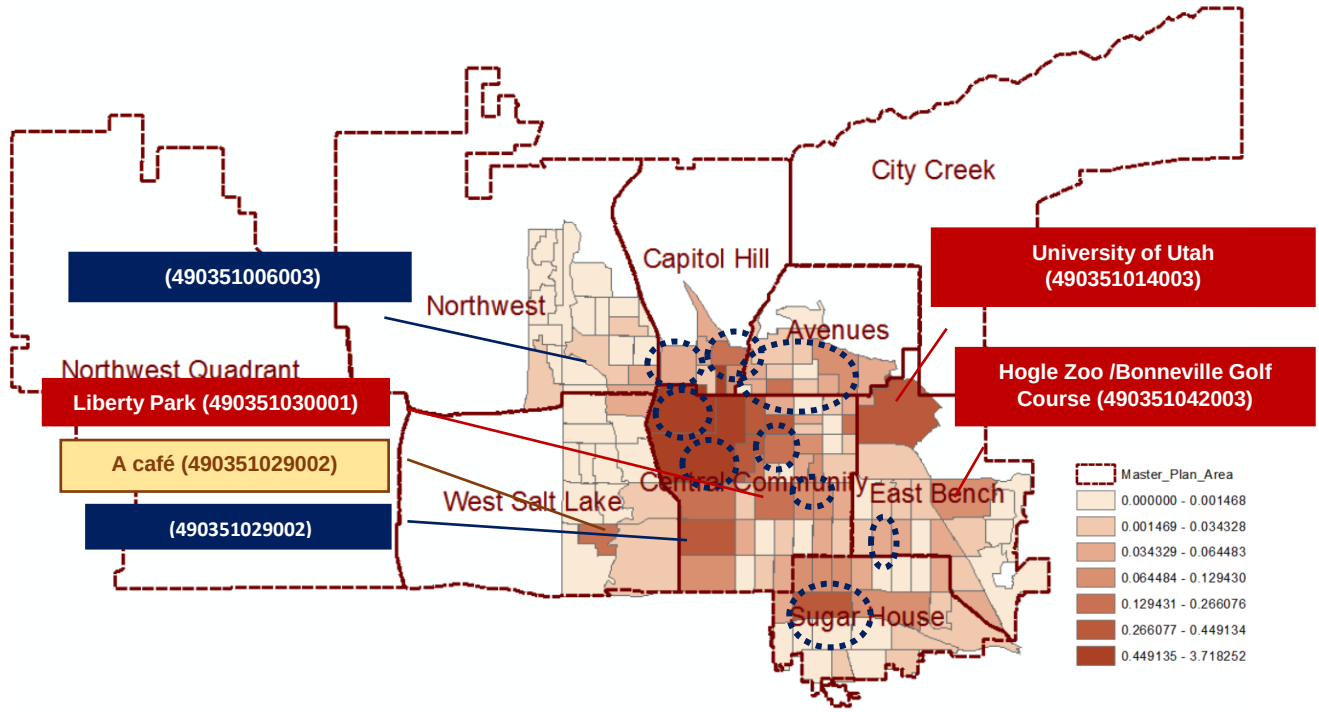


Figure 18. Patterns of gentrification keywords in 2014

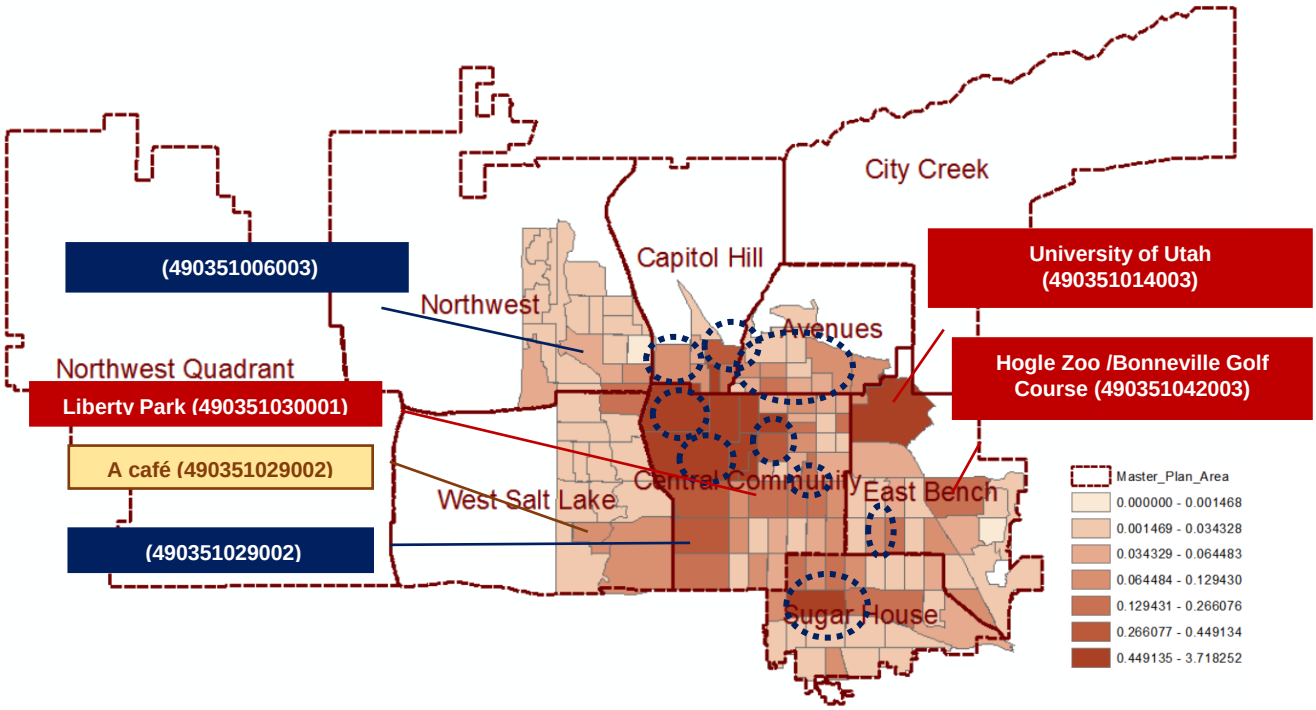


Figure 19. Patterns of gentrification keywords in 2015

Figure 20. Word cloud based on block group ID 490351028022

To conclude, there is a relationship between the patterns of gentrification keywords and areas of human-perceived gentrification. Additionally, the result suggests the keyword-based approach can help us to recognize some gentrifying areas using the census-based strategy that cannot be identified by human perception. Because of the flexibility of social media sources, this approach also helps us explore fine-grained gentrification dynamics. Nevertheless, this approach has limitations. The numbers of gentrification keywords were influenced by a few venues. For example, many posts contained the word “coffee” but only because of one shop there. Moreover, the patterns of keywords were also affected by the increasing number of Instagram users.

4.4.2 Text Clustering Approach

The text clustering approach is the other method used to measure gentrification ambience. The purpose of the text clustering method is to identify areas with similar Instagram texts as gentrifying areas. Different from the keyword-based method, this approach considers all words posted in the study area and it avoids subjective keyword selection.

(1) The Advantages of Text Clustering

Figures 21 to 23 show the text clustering results from 2013 to 2015, respectively. The circles on the maps represent the human-perceived gentrifying areas. The results indicate that the text clustering method grouped most areas of human-perceived gentrification into two groups, Groups 5 and 7, and they were distinguished from other areas. As the map shows in Figure 21, in 2013, some areas of human-perceived gentrification, including downtown, south of downtown, and the core of Sugar House,

were clustered into Group 5; the other areas of human-perceived gentrification, including West of 300 West, the Marmalade District, a part of the Avenues, and 9th and 9th, were clustered into Group 7. Only Trolley Square and 15th and 15th were clustered with other groups. In 2014 and 2015, Trolley Square was also clustered into Group 5. In other words, the text clustering method identified seven areas of human-perceived gentrification in 2013 (Figure 21) and eight areas of human-perceived gentrification in 2014 and 2015. Only one area of human-perceived gentrification, 15th and 15th, was overlooked in 2014 and 2015.

Moreover, the influence of certain venues is mitigated by the text clustering approach. Many venues, such as the college, the park, and the wrestling promotion, seriously influenced the number of night posts and the number of gentrification keywords, but these issues were not obvious when using the text clustering technique. For example, there were many Instagram night posts and gentrification keywords collected from the University of Utah and Hogle Zoo; however, these venues are not relevant to gentrification. With the text clustering method, the census block group wherein the University of Utah and Hogle Zoo are located was differentiated from Groups 5 and 7.

In addition, the text clustering approach is capable of exploring gentrification dynamics on a fine-grained temporal scale. As the patterns show in Figures 21, 22, and 23, Group 7 continued to expand over the years. This expansion is similar to what we found from the patterns of night posts and the patterns of gentrification keywords, indicating that more and more census block groups have similar Instagram texts, which might result from the expansion of gentrification ambience. Although we cannot be sure of whether the expanding areas are really gentrifying, they are more likely to become gentrified because their sense of place is more akin to the gentrifying areas. In addition, it can be noticed that the expanding areas were increasing along the human-perceived gentrifying areas. This finding confirms the arguments of previous studies that pointed out that poor neighborhoods near an area that experiences gentrification are more likely to be gentrified (Guerrieri, Hartley, & Hurst, 2013; Hackworth, 2007).

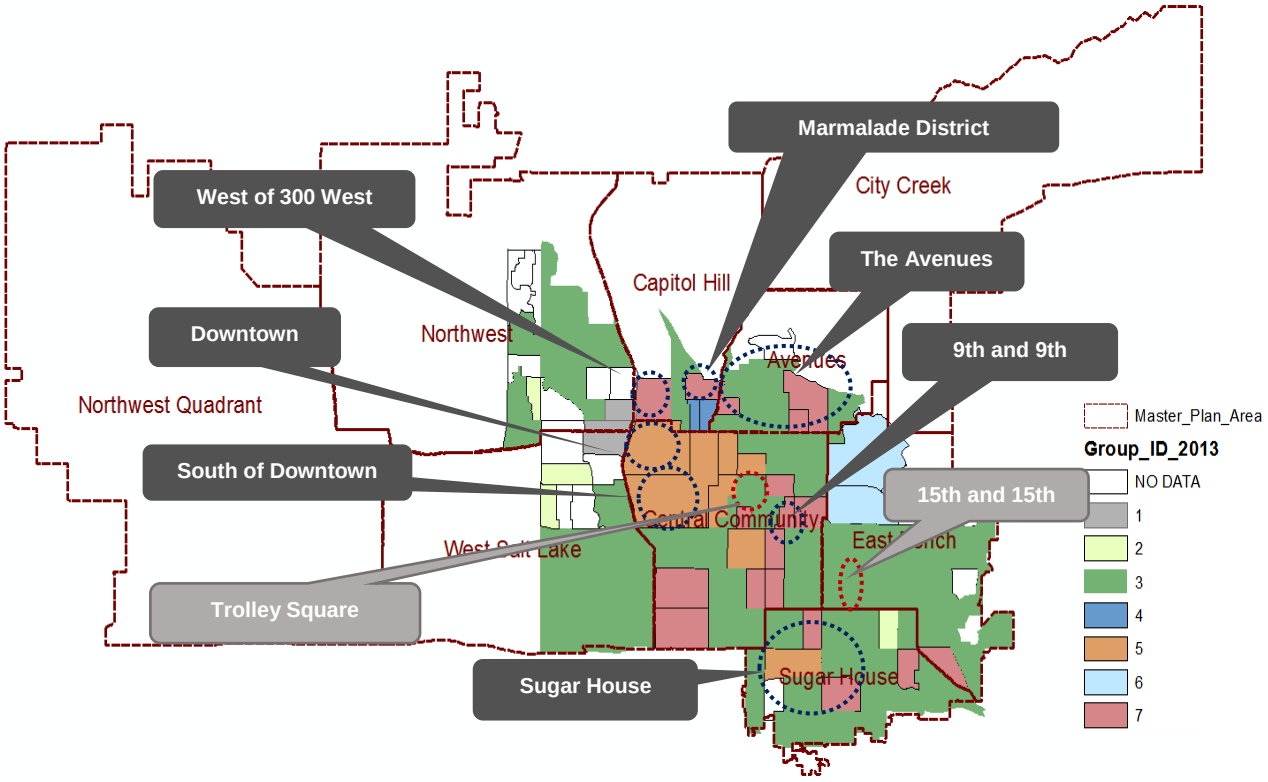


Figure 21. Result of text clustering in 2013

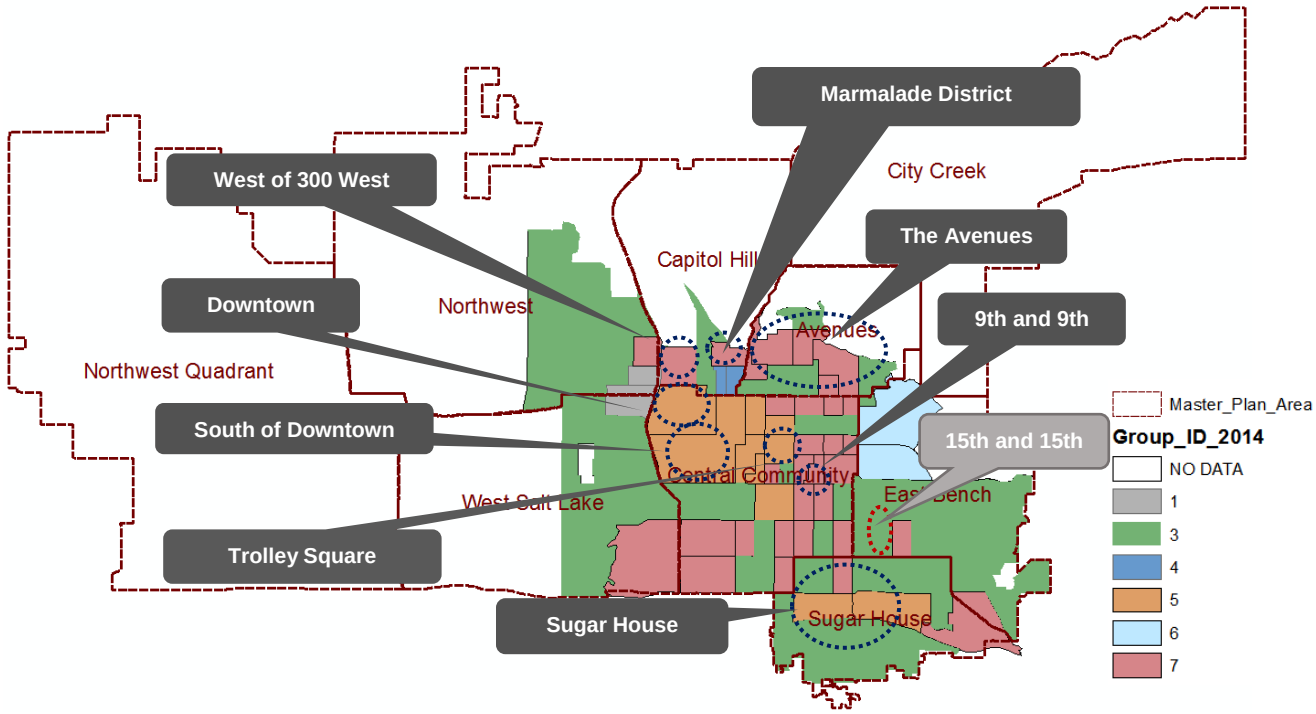


Figure 22. Result of text clustering in 2014

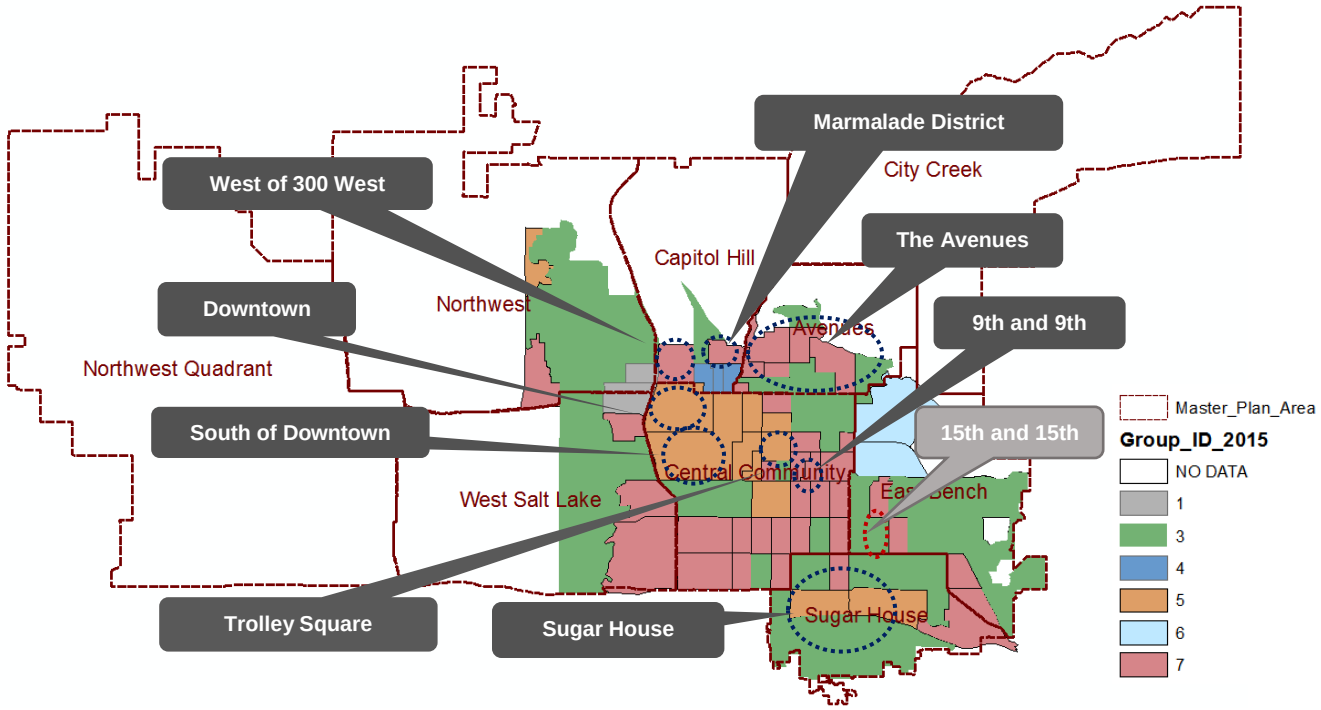


Figure 23. Result of text clustering in 2015

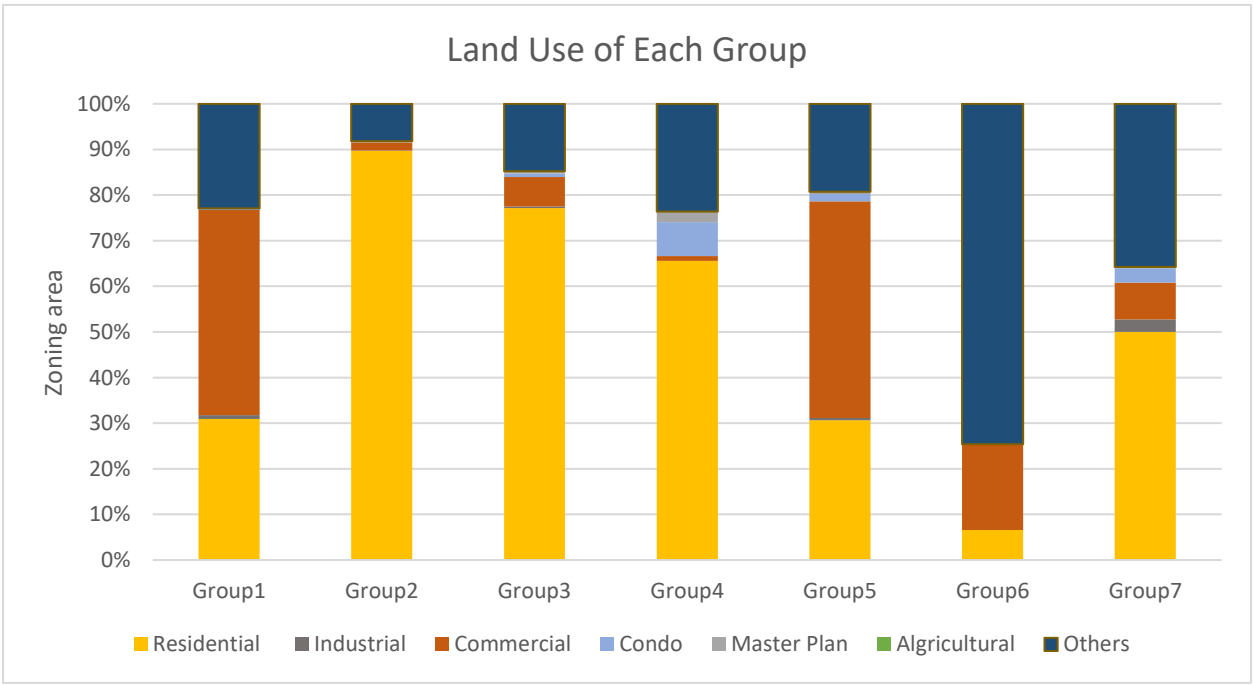


Figure 24. Land use of each group

The text clustering approach also differentiated between the two types of gentrification: commercial gentrification and residential gentrification. As introduced in the literature review, there are several types of gentrification. However, previous studies using the conventional census-based typologies are incapable of distinguishing various types of gentrification. The text clustering approach, on the other hand, revealed two clusters, Group 5 and Group 7, representing commercial and residential gentrification respectively. Neighborhoods in Group 5 including downtown, south of downtown, Trolley Square, and Sugar House are business and shopping districts with large shopping centers and trendy shops, while those in Group 7 consisting of West of 300 West, the Marmalade District, the Avenues, and 9th and 9th are upscale residential areas with diverse architectural styles, historic houses, and small retail shops. The details of the land use of each group are shown in Figure 24. Group 5, representing commercial gentrification, has the largest percentage of commercial areas, making up almost 50% of the entire area, while 30% of the area is residential. This indicates that Group 5 is a mixed-use location characterized by commercial activities. Residential and condominium areas make up more than 50% of Group 7, meaning this area contains large residential neighborhoods, similar to Groups 2, 3, and 4. However, the percentage of commercial land use is larger in Group 7 than in Groups 2, 3, and 4, indicating Group 7 is a residential place with a considerable amount of commercial land use area.

(2) Word Usage in the Gentrifying Groups

To explore what differentiates areas of human-perceived gentrification from other areas, this research examined what words Instagram users used in different groups. First, the bag of word vectors of the census block groups was converted into *TF-IDF* vectors, and all *TF-IDF* vectors were normalized by the number of unique users. Then, the average *TF-IDF* of every word in the same group was calculated and it is visualized in Figure 25.

The horizontal axis refers to all words used in the entire corpus. Specifically, there are 41,205 words on the horizontal axis. If a word was not used in that group, its average *TF-IDF* is zero. Take Group 2 for example, which temporarily appeared in 2013 (see Figure 21), wherein the average *TF-IDF* of each word is zero. This means that almost no words were used in this group. The vertical axis is the average *TF-IDF*. Each bar in the graph represents the mean *TF-IDF* of each word. A significant peak in a group means that the word has high average *TF-IDF*. Namely, it indicates that the word was relatively important for this group because it was used frequently in this group but rarely in other groups. For example, the two peaks in Group 4 represent two significantly important words, “templ” and “templesur,” where “templ” means “temple” and “templesur” means “templesure” after the stemming process. This shows that the location names “temple” and “temple square” are the two most important words in Group 4.

The result shown in Figure 25 differs from the original expectation. Because the mean *TF-IDF* refers to the importance of a word in a group, some gentrification keywords with significant peaks in the gentrifying groups (Group 5 and Group 7) were expected to be

shown in the graph. However, compared with the other groups, the distributions of the bars in Groups 5 and 7 are more even than in the other groups. There are still some peaks, but they are lower than the peaks in the other groups. This result indicates that Instagram users used words that are more diverse in gentrifying areas, and this might result from the greater diversity of activities and events occurring there. This result is similar to what Hristova et al. (2016) found in London. According to their case study, places that experience gentrification have the most diverse venues and visitors. The socio-economic diversity and class diversity in gentrifying areas were also identified in previous studies (Freeman, 2009; McKinnish et al., 2010).

The even distributions of bars do not mean that gentrification keywords play no role in gentrifying groups. By selecting the top 100 words, seven word clouds based on the average *TF-IDF* in each group were generated (Figures 26). As shown in these word clouds, more gentrification keywords appear in gentrifying groups. Specifically, only one gentrification keyword appears each in Group 1 and Group 3: *restaurant* and *coffee*, and there are no gentrification keywords in Groups 2, 4, and 6. Yet, in Group 5, six gentrification keywords appear among the top 100 words, as well as three keywords in Group 7. In other words, the gentrification keywords play a more important role in gentrifying groups than in non-gentrifying groups.

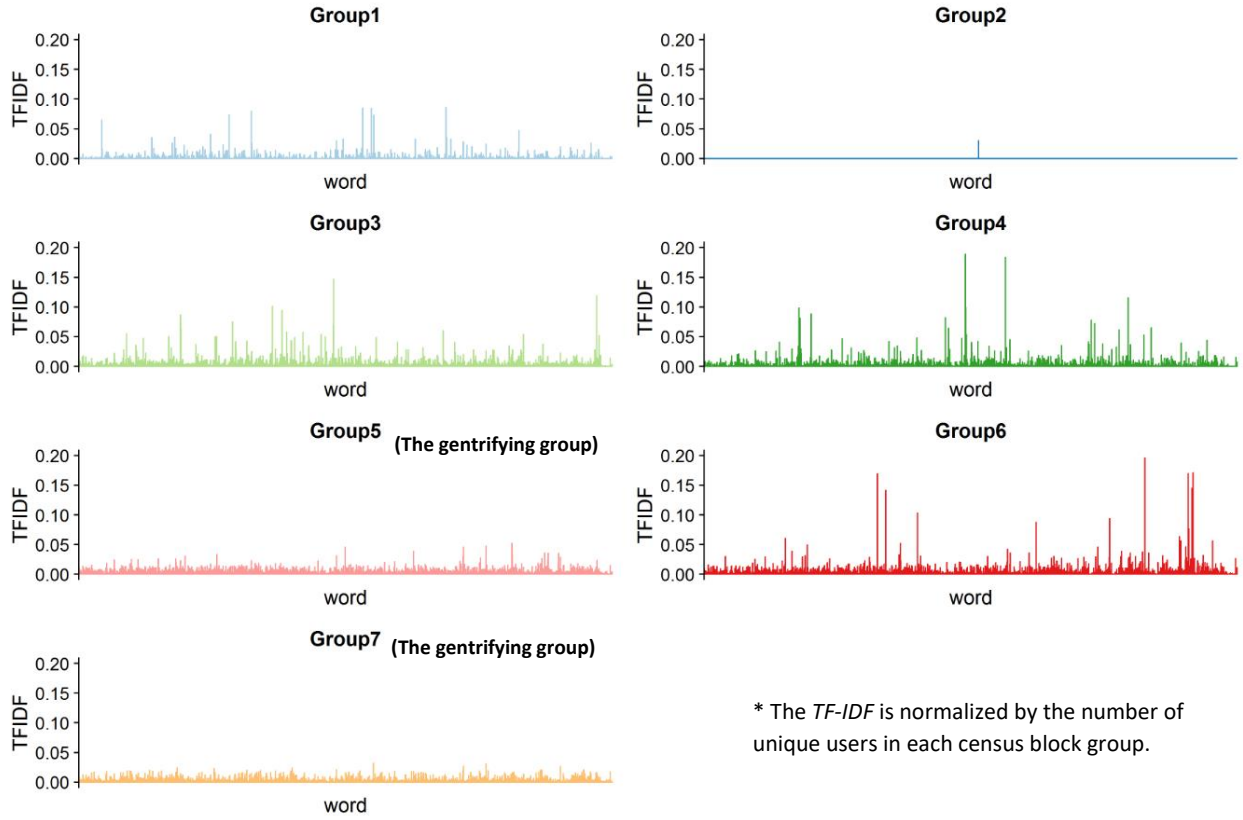
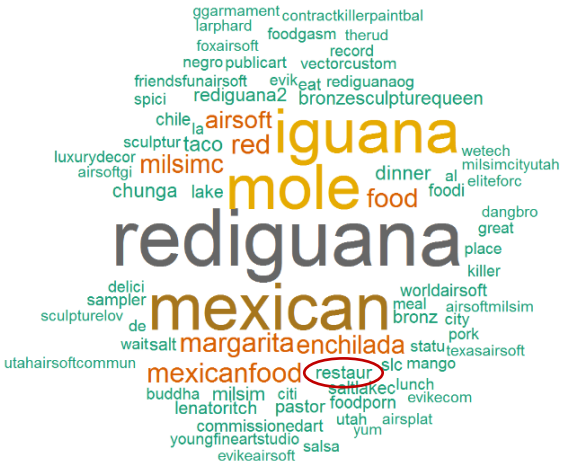


Figure 25. The average *TF-IDF* of each word in different groups


ggarmament
contractkillerpaintbal
larphard
foodgasm
therud
foxairsoft
record
negro
publicart
vectorcustom
friendsfunairsoft
evik
eat
rediguanaog
spici
rediguana2
bronzesculpturequeen
chile
la
airsoft
iguana
sculptur
taco
red
luxurydecor
wetech
milsimcityutah
airsoftgi
milsimc
mole
dinner
al
eliteforc
chunga
lake
food
foodi
dangbro
great
rediguana
place
killer
delici
worldairsoft
sampler
meal
airsoftmilsim
sculpturelov
bronz
city
de
pork
wait
salt
margarita
enchilada
statu
texasairsoft
utahairsoftcommun
mexicanfood
restaur
slc
mango
buddha
milsim
citi
saltlakec
lunch
lenatoritch
pastor
foodporn
evikecom
commissionedart
utah
airsplat
yum
youngfineartstudio
salsa
evikeairsoft

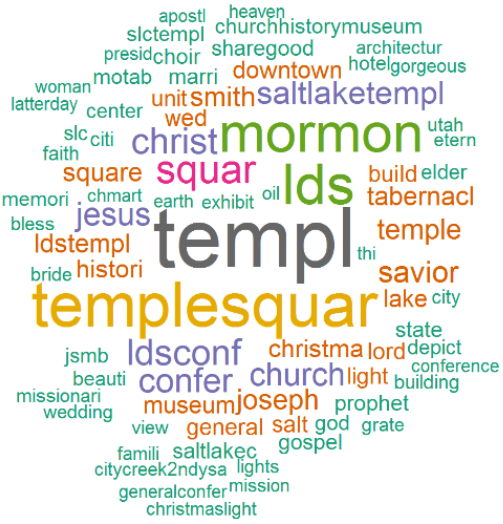
A.


love

B.


latteart
baekhyun
espressoporn
mixandmatch
jinhyeong
hongseok
sm
streetwear
ygstan
coffeeporn
makna
greaser
hallyuwav
yg
coffeeart
bobbi
goodfood
rockabilli
icedcoffe
junhw
ygentertain
espressoart
barista
latt
rapper
ygfamili
ygmakna
exok
kai
bias
fit
oppa
healthi
donghyuk
kpopf4f
xiumin
torani
hiphop
ikon
coffe
instaart
tattoo
exol
luhan
kris
rap
hallyu
kpop
korean
exo
utah
cup
exom
jinhwan
bi
gourmet
chanwoo
foodi
teamb
ot12
instabeauti
slc
espresso
baristam
chen
chanyeol
suho
coffeebean
yunhyeong
bigbang
doubleshot
sehun
rnb
pierc
smtown
nomnomnom
leader
tao
boomboxx
hallyufollow

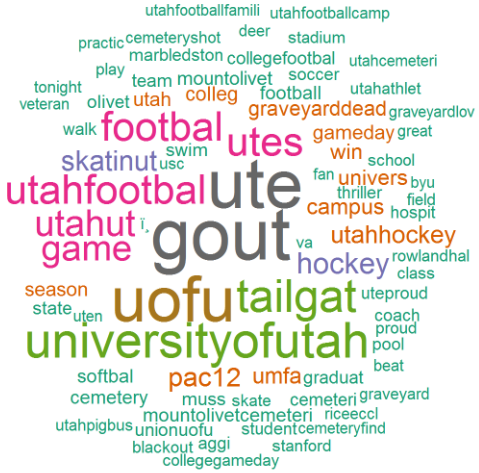
C.


apostl
heaven
churchhistorymuseum
slctempl
sharegood
presid
choir
architectur
hotel
gorgeous
motab
marri
downtown
woman
unit
smith
saltlaketempl
latterday
center
wed
utah
slc
citi
christ
mormon
etern
faith
square
squar
lds
build
elder
memori
chmart
earth
exhibit
oil
tabernacl
bless
jesus
templ
temple
ldstempl
thi
bride
histori
savior
templesquar
lake
city
state
jsmb
ldsconf
christma
lord
depict
conference
beauti
confer
church
light
building
missionari
museum
joseph
prophet
wedding
view
general
salt
god
grate
gospel
famili
saltlakec
citycreek2ndysa
lights
generalconfer
mission
christmaslight

D.

E.

(The gentrifying group)



F.

G.

(The gentrifying group)

Figure 26. The word clouds based on the average TF-IDF in each group. Each word cloud shows the top 100 words with highest average of *TF-IDF*. The word with bigger size represents the word has higher average of TF-IDF in a group. (A) is the word cloud of Group 1; (B) is the word cloud of Group 2; (C) is the word cloud of Group 3; (D) is the word cloud of Group 4; (E) is the word cloud of Group 5; (F) is the word cloud of Group 6; (G) is the word cloud of Group 7. (E) and (G) are the gentrifying groups, and there are more gentrification keywords founded in the word clouds.

Figure 27 helps us to examine further the ranking of gentrification keywords in each group. The horizontal axis represents all gentrification words in the Instagram dataset, and the vertical axis represents the ranking. Different lines refer to different groups, and the dashed lines represent Groups 5 and 7 (gentrifying groups). In the graph, the dashed lines lay above all other lines of most of the gentrification keywords, such as *night, tonight, bar, coffee, café, aesthetics, art, hipster, beer, gay*, and *yoga*. This indicates that when differentiating gentrifying from non-gentrifying areas, these keywords play a more important role than other keywords.

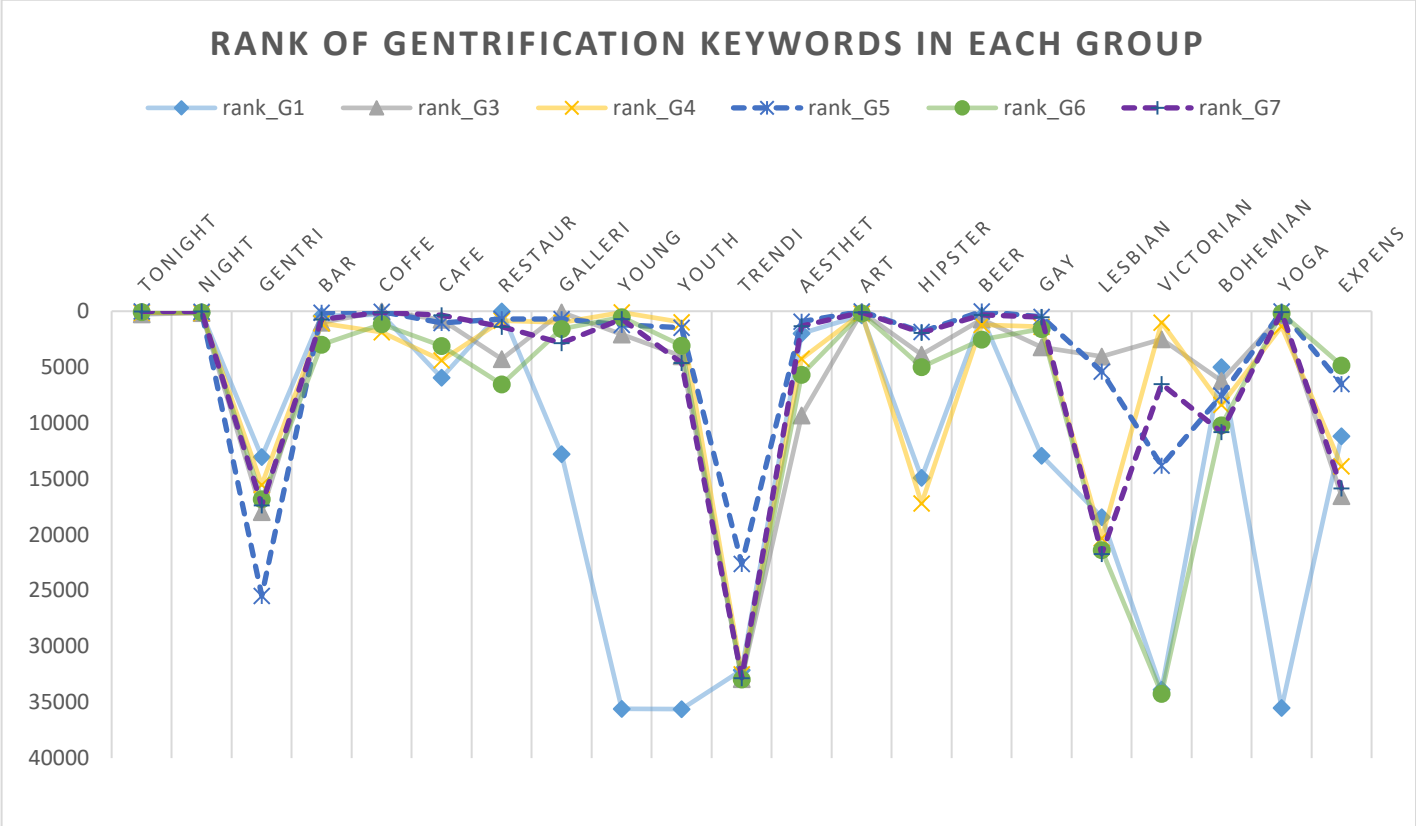


Figure 27. Rank of gentrification keywords in each group

Chapter 5: Conclusion

One of the challenges of previous gentrification studies is taking advantage of the qualitative and quantitative approaches to explore gentrification dynamics. Additionally, the incorporation of human perceptions of neighborhoods into a large-scale measurement of gentrification is underexplored. Another underexplored feature is the lack of research considering gentrification dynamics on a finer spatio-temporal scale across a large area.

To address the issues of gentrification studies, the focus of research is on developing a social media-based framework that incorporates human perceptions of neighborhoods to evaluate gentrification dynamics across a large region on a fine spatio-temporal scale. Nightlife activities and gentrification ambience are the gentrification indicators extracted to capture a sense of gentrification. The case study of Salt Lake City demonstrates how these two indicators can be measured using publically available Instagram data. Additionally, this research shows that most areas of human-perceived gentrification can be identified by this social media-based approach. Furthermore, this new approach also helps us to differentiate between two different types of gentrifying areas, which is highly challenging for conventional census-based methods.

Two studies were conducted to evaluate the social media-based framework. The first is a comparative study between qualitative and quantitative measures of gentrification. As a qualitative approach, areas of human-perceived gentrification were identified based on those areas mentioned in discussions on online forums and in news reports. The human-perceived gentrifying areas were compared with the distribution of gentrification typologies, which were generated according to five census-based gentrification strategies. As a result, the gentrifying areas identified by the five census-based typologies were inconsistent. Furthermore, those areas do not match well with the human-perceived gentrifying areas.

The second study aimed to introduce a novel data-mining framework that identifies a sense of gentrification by utilizing Instagram data. This study constructed two gentrification indicators—nightlife activities and gentrification ambience—that facilitated capturing the cultural characteristics of gentrification. Nightlife activities and gentrification ambience were quantified by applying text processing and text clustering techniques to the Instagram data. By comparing the spatial distributions of these two indicators over years with areas of human-perceived gentrification, the results showed that areas observed as having a sense of gentrification closely correspond to the human-perceived gentrifying areas. Moreover, the text clustering approach differentiated two types of gentrification, namely residential and commercial gentrification, but capturing this differentiation is a challenging task using conventional census-based typologies. Finally, the shifting patterns in relation to a sense of gentrification captured by the text clustering approach illustrate the expansion of gentrification.

This exploratory research suggests that the social media-based data-mining framework combines the advantages of both qualitative and quantitative approaches and it is capable of exploring spatially and temporally fine-grained gentrification dynamics.

5.1 Key Findings

This research has three key findings. First, the social media-based data framework identified areas of human-perceived gentrification that were not recognized with

conventional census-based strategies. The two indicators used in this framework, nightlife activities and gentrification ambience, are the cultural measurements that helped us to detect the nuance of cultural characteristics and to identify successfully most gentrifying areas derived from news reports and online forums.

Second, this social media-based strategy identified different types of gentrification. There are varied forms of gentrification, and the different types of gentrification are grouped into residential gentrification and commercial gentrification (Lees et al., 2008). In this research, a text clustering approach classified gentrifying areas into two groups, where one group matched residential gentrification and the other commercial gentrification. Therefore, the social media-based strategy not only identified accurately most areas of human-perceived gentrification, but also differentiated between residential gentrification and commercial gentrification.

Third, the social media data analysis helped us to explore gentrification dynamics on a finer spatio-temporal scale. Instead of identifying gentrifying areas at the census tract level, this novel framework explored gentrification dynamics at the census block group level. It demonstrated that by using this new approach, there is no need to sacrifice geographic accuracy too much when identifying gentrifying areas. Additionally, instead of aggregating decennial data to generate gentrification typologies, a social media-based approach identified yearly gentrification dynamics. Using the yearly distribution, this case study reveals the expansion of gentrification throughout Salt Lake City from 2013 to 2015.

5.2 Research Limitations and Future Works

While there are several findings in this research, the results are affected by some limitations. First, using nighttime Instagram posts is too simple to represent nightlife activities. Some Instagram night posts might be published from home or from places having nothing to do with the nightlife. To improve this method, one possible method is to check what percentage of nighttime Instagram posts is related to nightlife activities. Then, the frequency of nighttime posts can be adjusted by a proper weight. Another possible method is incorporating Instagram photos to determine whether an Instagram post was published while engaging in nightlife activities. Applying spatio-temporal check-in patterns and a semantic analysis can also help us to check the typology of a place (Cranshaw, Hong, & Sadeh, 1977; McKenzie, Janowicz, Gao, Yang, & Hu, 2015). This technique might be the alternative method of quantifying nightlife activities.

Second, some gentrification keywords might play a less-important role than other keywords. Every keyword is treated equally with the keyword-based method; however, some keywords are too common and frequently appear everywhere. For example, compared to the words *gay* or *gallery*, the words *tonight* and *coffee* are used in daily conversations. Therefore, not only determining the representative gentrification

keywords but also providing the appropriate weights for different keywords is an essential next step.

Third, the shift in Instagram user behaviors might influence the text clustering results. The results of text clustering in 2013 coincide almost perfectly with areas of human-perceived gentrification, but the gentrifying groups increased rapidly in 2014 and 2015. This indicates that more places have Instagram texts similar to those in gentrifying areas. This phenomenon might be due to the expansion of the gentrification ambience, which helps us to identify places likely undergoing gentrification. Yet, this phenomenon could also result from the emergence of certain Instagram user behaviors. In other words, trendy contents (sometimes relate to gentrification) are frequently posted by a large amount of Instagram users, which might lead to homogeneous Instagram text in gentrifying and non-gentrifying places. To address this problem, we need more research to study Instagram user behaviors.

Fourth, the uncertainty of the data and the algorithm also affects the research results. Because social media data are not designed for research purposes, the check-in data can be controlled by users, and it is difficult to know the level of uncertainty of the data (Lazer, Kennedy, King, & Vespignani, 2014). In other words, even though data filtering was applied in this research to reduce the influence of robots and extremely active users, fake Instagram check-in posts might lead to an inaccurate result. Conducting research to explore the level of data uncertainty is the next step in addressing this problem.

The other challenge is the lack of the exact boundaries of areas of human-perceived gentrification. In this research, these areas, derived from news reports and online forums, do not have a clear boundary. They are the names of places with general locations, so they cannot be used to validate statistically the Instagram analysis results. Hence, to validate the results, recruiting residents to draw the boundaries of the gentrifying areas or generating finer spatio-temporal house value dynamics with assessor parcel data is a direction for future work.

Reference

A Local's Perspective on the Sugar House Development. (2013). Retrieved May 1, 2017, from http://utahstories.com/2013/11/a-locals-perspective-on-the-sugar-house-development/

Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Clustering Algorithms. In *Mining Text Data* (pp. 77–128). https://doi.org/10.1007/978-1-4614-3223-4_4

Anderson, E. (1990). The Village Setting. In *Streetwise : race, class, and change in an urban community* (pp. 7–55). Chicago, IL: The Univ. of Chicago Press.

Aslam, A. A., Tsou, M. H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., … Lindsay, S. (2014). The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of Medical Internet Research*, *16*(11). https://doi.org/10.2196/jmir.3532

Atkinson, R. (2000). Measuring gentrification and displacement in greater London. *Urban Studies*, *37*(1), 149–165. https://doi.org/10.1080/0042098002339

Barton, M. (2016). An exploration of the importance of the strategy used to identify gentrification. *Urban Studies*, *53*(1), 92–111. https://doi.org/10.1177/0042098014561723

Bcgallo et al. (2011). Best Area in Salt Lake City for Gentrification Play. Retrieved May 5, 2016, from https://www.zillow.com/advice-thread/Best-Area-in-Salt-Lake-City-for-Gentrification-Play/413670/

Beauregard, R. A. (1986). The chaos and complexity of gentrification. In *Gentrification of the city* (pp. 35–55). https://doi.org/10.4324/9781315889092

Bell, D. (1973). The coming of post-industrial society a venture in social forecasting. New York: Basic Books.

Betancur, J. (2011). Gentrification and Community Fabric in Chicago. *Urban Studies*, *48*(2), 383–406. https://doi.org/10.1177/0042098009360680

Betancur, J. J. (2002). The Politics of Gentrification The Case of West Town in Chicago. *Urban Affairs Review*, *37*(6), 780–814. https://doi.org/10.1177/107874037006002

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning - ICML '06* (pp. 113–120). https://doi.org/10.1145/1143844.1143859

Bostic, R. W., & Martin, R. W. (2003). Black home-owners as a gentrifying force? Neighbourhood dynamics in the context of minority home-ownership. *Urban Studies*, *40*(12), 2427–2449. https://doi.org/10.1080/0042098032000136147

Boyd, M. (2008). Defensive Development: The Role of Racial Conflict in Gentrification. *Urban Affairs Review*, *43*(6), 751–776. https://doi.org/10.1177/1078087407313581

Butler, T. (1997). *Gentrification and the middle classes*. Aldershot [etc.]: Avebury.

Butler, T., & Robson, G. (2003). Plotting the middle classes: Gentrification and circuits of education in London. *Housing Studies*, *18*(1), 5–28.

https://doi.org/10.1080/0267303032000076812

Carpenter, J., & Lees, L. (1995). Gentrification in New York, London and Paris: An International Comparison. *International Journal of Urban and Regional Research*, *19*(2), 286–303. https://doi.org/10.1111/j.1468-2427.1995.tb00505.x

Castells, M. (1983). *The city and the grassroots : a cross-cultural theory of urban social movements*. London: Arnold.

Caulfield, J. (2016). City Form and Everyday Life Toronto's Gentrification and Critical Social Practice. Toronto: University of Toronto Press.

Clark, E. (1988). The Rent Gap and Transformation of the Built Environment: Case Studies in Malmö 1860-1985. *Geografiska Annaler. Series B, Human Geography*, *70*(2), 241–254. https://doi.org/10.2307/490951

Clay, P. L. (1979). *Neighborhood renewal : middle-class resettlement and incumbent upgrading in American neighborhoods.* Lexington, Mass.: D.C. Heath and Co.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/6.1.57

Collier, N. (2012). Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health*, *7*(7), 731–749. https://doi.org/10.1080/17441692.2012.699975

Cranshaw, J., Hong, J. I., & Sadeh, N. (1977). The Livehoods Project : Utilizing Social Media to Understand the Dynamics of a City. *Icwsm*, 58–65. https://doi.org/papers3://publication/uuid/557455DB-AC4A-4C73-968A-31E7A663BC4E

Currid, E. (2009). *The Warhol economy : how fashion, art, and music drive New York City*. Princeton: Princeton University Press.

Cutler, A. (2015). Salt Lake City, Utah: Neighborhoods to Know. Retrieved May 1, 2017, from http://www.greatamericancountry.com/places/local-life/living-in-salt-lake-city--utah

Davidson, M. (2010). Love thy neighbour? social mixing in London's gentrification frontiers. *Environment and Planning A*, *42*(3), 524–544. https://doi.org/10.1068/a41379

De Longueville, B., Smith, R. S., Luraschi, G., Longueville, B. De, Smith, R. S., De Longueville, B., … Luraschi, G. (2009). OMG, from here, I can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. *Proceedings of the 2009 International Workshop on Location Based Social Networks*, (c), 73–80. https://doi.org/10.1145/1629890.1629907

Ding, L., Hwang, J., & Divringi, E. (2016). Gentrification and residential mobility in Philadelphia. *Regional Science and Urban Economics*, *61*, 38–51. https://doi.org/10.1016/j.regsciurbeco.2016.09.004

Docampo, M. G. (2014). Theories of Urban Dynamics. *International Journal of Population Research*, *2014*(Article ID 494871), 1–11. https://doi.org/10.1155/2014/494871

Dutton, P. (2003). Leeds calling: The influence of London on the gentrification of regional cities. *Urban Studies*, *40*(12), 2557–2572. https://doi.org/10.1080/0042098032000136219

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Minin: Towards a Unifying Framework. *Proc 2nd Int Conf on Knowledge Discovery and Data*

Mining Portland OR, *2*, 82–88. https://doi.org/10.1.1.27.363

Florida, R. (2002). Bohemia and economic geography. *Journal of Economic Geography*, *2*(1), 55–71. https://doi.org/10.1093/jeg/2.1.55

Freeman, L. (2005). Displacement or Succession?: Residential Mobility in Gentrifying Neighborhoods. *Urban Affairs Review*, *40*(4), 463–491. https://doi.org/10.1177/1078087404273341

Freeman, L. (2006). *There goes the 'hood : views of gentrification from the ground up*. Philadelphia : Temple University Press. Retrieved from http://www.worldcat.org/title/there-goes-the-hood-views-of-gentrification-from-the-ground-up/oclc/880716083&referer=brief_results

Freeman, L. (2009). Neighbourhood Diversity, Metropolitan Segregation and Gentrification: What Are the Links in the US? *Urban Studies*, *46*(10), 2079–2101. https://doi.org/10.1177/0042098009339426

Freeman, L., & Braconi, F. (2004). Gentrification and displacement new york city in the 1990s. *Journal of the American Planning Association*, *70*(1), 39–52. https://doi.org/10.1080/01944360408976337

Gale, D. E. (1980). Neighborhood resettlement: Washington, D.C. In S. B. Laska & D. Spain (Eds.), *Back to the City* (pp. 95–115). New York: Burlington.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2011.08.004

Ghosh, D., & Guha, R. (2013). What are we "tweeting" about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2), 90–102. https://doi.org/10.1080/15230406.2013.776210

Gibbons, J., & Barton, M. S. (2016). The Association of Minority Self-Rated Health with Black versus White Gentrification. *Journal of Urban Health*, *93*(6), 909–922. https://doi.org/10.1007/s11524-016-0087-0

Glass, R. (1964). *London: aspects of change*. (C. for U. Studies, Ed.) (Vol. Report (Un). London: University College, London.

Gotham, K. F. (2005). Tourism gentrification: The case of New Orleans' Vieux Carre (French Quarter). *Urban Studies*, *42*(7), 1099–1121. https://doi.org/10.1080/00420980500120881

Grobelnik, M., Mladenić, D., & Jermol, M. (2017). *Exploiting Text Mining in Publishing and Education*.

Guerrieri, V., Hartley, D., & Hurst, E. (2013). Endogenous gentrification and housing price dynamics. *Journal of Public Economics*, *100*, 45–60. https://doi.org/10.1016/j.jpubeco.2013.02.001

Gundecha, P., & Liu, H. (2012). Mining Social Media: A Brief Introduction. *Tutorials in Operations Research*, (Dmml), 1–17. https://doi.org/http://dx.doi.org/10.1287/educ.1120.0105

Hackworth, J. (2007). *The neoliberal city : Governance, ideology, and development in American urbanism*. Ithaca, N.Y.: Cornell University Press.

Hae, L. (2011). Dilemmas of the Nightlife Fix: Post-industrialisation and the Gentrification of

Nightlife in New York City. *Urban Studies*, *48*(16), 3449–3465. https://doi.org/10.1177/0042098011400772

Hae, L. (2011). Gentrification and politicization of nightlife in New York city. *ACME*, *10*(3), 564–584.

Hammel, D. J. (1999). Gentrification and land rent: A historical view of the rent gap in Minneapolis. *Urban Geography*, *20*(2), 116–145. https://doi.org/10.2747/0272-3638.20.2.116

Hamnett, C. (2003). Gentrification and the middle-class remaking of inner London, 1961-2001. *Urban Studies*, *40*(12), 2401–2426. https://doi.org/10.1080/0042098032000136138

Hamnett, C., & Whitelegg, D. (2007). Loft conversion and gentrification in London: From industrial to postindustrial land use. *Environment and Planning A*, *39*(1), 106–124. https://doi.org/10.1068/a38474

Harris, Z. S. (1954). Distributional Structure. WORD, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Hochman, N., & Manovich, L. (2013). Zooming into an Instagram City: Reading the local through social media. *First Monday*, *18*(7). https://doi.org/10.5210/fm.v18i7.4711

Hochman, N., & Schwartz, R. (2012). Visualizing Instagram: Tracing Cultural Visual Rhythms. *The Workshop on Social Media Visualization (SocMedVis) in Conjuction with The Sixth International AAAI Conference on Weblogs and Social Media (ICWSM-12)*, 6–9. https://doi.org/10.1080/08838151.2015.1029125

Holt, L. (2008). Embodied social capital and geographic perspectives: performing the habitus. *Progress in Human Geography*, *32*(2), 227–246. https://doi.org/10.1177/0309132507087648

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*, 19–62. https://doi.org/10.1111/j.1365-2621.1978.tb09773.x

Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., & Mascolo, C. (2016). Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16* (pp. 21–30). https://doi.org/10.1145/2872427.2883065

Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we Instagram : a first analysis of Instagram photo content and user types. In *Proceedings of the Eight International AAAI Conference on Weblogs and Social Media* (pp. 595–598).

Hwang, J., & Sampson, R. J. (2014). Divergent Pathways of Gentrification: Racial Inequality and the Social Order of Renewal in Chicago Neighborhoods. *American Sociological Review*, *79*(4), 726–751. https://doi.org/10.1177/0003122414535774

Ingvaldsen, J. E., & Gulla, J. A. (2012). Industrial application of semantic process mining. *Enterprise Information Systems*, *6*(2), 139–163. https://doi.org/10.1080/17517575.2011.593103

Jager, M. (1986). Class definition and the esthetics of gentrification: Victoriana in Melbourne. In *Gentrification of the City* (pp. 78–91). Boston, MA: Allen and Unwin.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography*, *7*(1), 186–190. https://doi.org/citeulike-article-id:8241517

Kern, L. (2012). Connecting embodiment, emotion and gentrification: An exploration through the practice of yoga in Toronto. *Emotion, Space and Society*, *5*(1), 27–35. https://doi.org/10.1016/j.emospa.2011.01.003

Kerstein, R. (1990). Stage Models of Gentrification: An Examination. *Urban Affairs Review*, *25*(4), 620–639. https://doi.org/10.1177/004208169002500406

Ketchen, D., & Shook, C. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, *17*(6), 441–458. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G

Kimberlyjo et al. (2014). Why is everyone here so obsessed with Sugarhouse? Retrieved May 1, 2017, from https://www.reddit.com/r/SaltLakeCity/comments/2cfra0/why_is_everyone_here_so_obsessed_wi%0Ath_sugarhouse/%0A

Knopp, L. (1990). Some theoretical implications of gay involvement in an urban land market. *Political Geography Quarterly*, *9*(4), 337–352. https://doi.org/10.1016/0260-9827(90)90033-7

Kukura, J. (2016). 9 Hippest Neighborhoods of Salt Lake. Retrieved May 1, 2017, from https://www.visitsaltlake.com/blog/post/2016/9/9-Hippest-Neighborhoods-of-Salt-Lake/8367/

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flue: traps in big data analysis. *Science*, *343*, 1203–1205. https://doi.org/10.1126/science.1248506

Lees, L. (2000). A reappraisal of gentrification: towards a “geography of gentrification.” *Progress in Human Geography*, *24*(3), 389–408. https://doi.org/10.1191/030913200701540483

Lees, L. (2003). Super-gentrification: The case of Brooklyn heights, New York City. *Urban Studies*, *40*(12), 2487–2509. https://doi.org/10.1080/0042098032000136174

Lees, L., Slater, T., & Wyly, E. K. (2008). *Gentrification*. New York: Routledge/Taylor & Francis Group. Retrieved from http://www.worldcat.org/title/gentrification/oclc/77574645

Ley, D. (1986). Alternative Explanations for Inner-City Gentrification: A Canadian Assessment. *Annals of the Association of American Geographers*, *76*(4), 521–535. https://doi.org/10.1111/j.1467-8306.1986.tb00134.x

Ley, D. (1994). Gentrification and the politics of the new middle class. *Environment Planning D Society Space*, *12*(1), 53. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=9501302900&site=ehost-live

Ley, D. (2001). *The new middle class and the remaking of the central city*. Oxford [u.a.]: Oxford

University Press.

Ley, D. (2003). Artists, aestheticisation and the field of gentrification. *Urban Studies*, *40*(12), 2527–2544. https://doi.org/10.1080/0042098032000136192

Ley, D., & Dobson, C. (2008). Are There Limits to Gentrification? The Contexts of Impeded Gentrification in Vancouver. *Urban Studies*, *45*(12), 2471–2498. https://doi.org/10.1177/0042098008097103

Lipton, S. G. (1977). Evidence of Central City Revival. *Journal of the American Institute of Planners*, *43*(2), 136–147. https://doi.org/10.1080/01944367708977771

Liu, C., & O'Sullivan, D. (2016). An abstract model of gentrification as a spatially contagious succession process. *Computers, Environment and Urban Systems*, *59*, 1–10. https://doi.org/10.1016/j.compenvurbsys.2016.04.004

Living in Salt Lake City. (2012). Retrieved May 1, 2017, from http://www.summitsothebysrealty.com/eng/article/living-in-salt-lake-city

Longley, P. A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, *30*(2), 369–389. https://doi.org/10.1080/13658816.2015.1089441

Lu, Y., Zhang, P., Liu, J., Li, J., & Deng, S. (2013). Health-Related Hot Topic Detection in Online Communities Using Text Clustering. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0056221

Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. Retrieved May 1, 2017, from http://manovich.net/content/04-projects/067-trending-the-promises-and-the-challenges-of-big-social-data/64-article-2011.pdf

Markosian, R. (2007). Artspace in Salt Lake City. Retrieved May 1, 2017, from http://utahstories.com/2007/01/artspace-in-salt-lake-city/

McKenzie, G., Janowicz, K., Gao, S., Yang, J.-A., & Hu, Y. (2015). POI Pulse: A Multi-granular, Semantic Signature–Based Information Observatory for the Interactive Visualization of Big Geosocial Data. *Cartographica*, *50*(2), 71–85. https://doi.org/10.3138/cart.50.2.2662

McKinnish, T., Walsh, R., & Kirk White, T. (2010). Who gentrifies low-income neighborhoods? *Journal of Urban Economics*, *67*(2), 180–193. https://doi.org/10.1016/j.jue.2009.08.003

Mills, C. A. (1988). "Life on the upslope": the postmodern landscape of gentrification. *Environment and Planning D: Society and Space*, *6*(2), 169 – 189. https://doi.org/10.1068/d060169

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064417

Nagel, A. C., Tsou, M. H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., … Sawyer, M. H. (2013). The complex relationship of realspace events and messages in cyberspace: Case study of influenza and pertussis using tweets. *Journal of Medical Internet Research*, *15*(10). https://doi.org/10.2196/jmir.2705

Nathalie P. Voorhees Center. (2014). The Socioeconomic Change of Chicago's Community Areas. Retrieved May 2, 2016, from

https://docs.wixstatic.com/ugd/992726_a60305a8ecc34951a0f48e55f5366c5b.pdf

National Association of Neighborhoods. (1980). *NAN Bulletin*. Washington, DC.

National Urban Coalition. (1978). *City Neiborhoods in Transition*. Washington, DC.

O'Sullivan, D. (2002). Toward micro-scale spatial modeling of gentrification. *Journal of Geographical Systems*, *4*(3), 251–274. https://doi.org/10.1007/s101090200086

Pang, B., & Lee, L. (2006). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in InformatioPang, B., & Lee, L. (2006). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 1(2), 91–231. doi:10.1561/1500000001n Retrieval*, *1*(2), 91–231. https://doi.org/10.1561/1500000001

Papachristos, A. V., Smith, C. M., Scherer, M. L., & Fugiero, M. A. (2011). More coffee, less crime? The relationship between gentrification and neighborhood crime rates in Chicago, 1991 to 2005. *City and Community*, *10*(3), 215–240. https://doi.org/10.1111/j.1540-6040.2011.01371.x

Parsons, D. J. (1980). *Rural gentrification : the influence of rural settlement planning policies*. Brighton, UK: University of Sussex.

Pattison, T. J. (1983). The stages of gentrification: the case of Bay Village. In P. L. Clay & R. M. Hollister (Eds.), *Neighborhood policy and planning* (pp. 77–92). Lexington, MA: LexingtonBooks.

Pereira, L., Rijo, R., Silva, C., & Martinho, R. (2015). Text Mining Applied to Electronic Medical Records: *International Journal of E-Health and Medical Communications*, *6*(3), 1–18. https://doi.org/10.4018/IJEHMC.2015070101

Popowich, F., Vx, C., & Va, C. (2005). Using Text Mining and Natural Language Processing for Health Care Claims Processing. *ACM SIGKDD Explorations Newsletter*, *7*(1), 59–66. https://doi.org/10.1145/1089815.1089824

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Porter, M. (2001). Snowball: A language for stemming algorithms. *Snowball*. Retrieved from http://snowball.tartarus.org/texts/introduction.html

Powell, J., & Spencer, M. (2002). Giving Them the Old "One-Two": Gentrification and the K.O. of Impoverished Urban Dwellers of Color. *Howard Law Journal*, *46*(3), 433. https://doi.org/10.3868/s050-004-015-0003-8

Rose, D. (1984). Rethinking gentrification: beyond the uneven development of marxist urban theory. *Environment and Planning D: Society and Space*, *2*(1), 47–74. https://doi.org/10.1068/d020047

Rothenberg, T. (1995). "And she told two friends": Lesbian creating urban social space. In D. Bell, G. Valentine, & J. Silk (Eds.), *Mapping Desire: Geographies of Sexualities* (pp. 165–181). London: routledge.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sampson, R. J. (2013). *Great American city : Chicago and the enduring neighborhood effect*.

Chicago: The University of Chicago press.

Schill, M. H., & Nathan, R. P. (1983). *Revitalizing America's cities : neighborhood reinvestment and displacement*. Albany: State University of New York Press.

Short, J. R. (1989). Yuppies, Yuffies and the New Urban Order. *Transactions of the Institute of British Geographers*, *14*(2), 173–188. https://doi.org/10.2307/622811

Silva, T. H., de Melo, P. O. S., Almeida, J. M., Salles, J., & Loureiro, A. A. F. (2013). A Comparison of Foursquare and Instagram to the Study of City Dynamics and Urban Social Behavior. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing* (p. 4:1--4:8). New York, NY, USA: ACM. https://doi.org/10.1145/2505821.2505836

Slater, T. (2006). The Eviction of Criticial Perspectives from Gentrification Research. *International Journal of Urban and Regional Research*, *30*(4), 737–757. https://doi.org/10.1111/j.1468-2427.2008.00772.x

Smith, D. (2005). "Studentification": the gentrification factory? *In R. Atkinson & G. Bridge (Eds.). Thevnew Urban Colonialism: Gentrification in a Global Context*, (Lees 1999), (pp. 72–89). https://doi.org/10.4324/9780203392089

Smith, D., & Holt, L. (2007). Studentification and "apprentice" gentrifiers within Britain's provincial towns and cities: Extending the meaning of gentrification. *Environment and Planning A*, *39*(1), 142–161. https://doi.org/10.1068/a38476

Smith, N. (1979). Toward a theory of gentrification: A back to the city movement by capital, not people. *Journal of the American Planning Association*, *45*(4), 538–548. https://doi.org/10.1080/01944367908977002

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2011.604636

Tan, A. H. (1999). Text mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 65–70. Retrieved from http://www3.ntu.edu.sg/sce/labs/erlab/publications/papers/asahtan/tm_pakdd99.pdf%5Cnpapers2://publication/uuid/43DC91E7-1D1A-44AE-8BB3-163AD208411D

Torrens, P. M., & Nara, A. (2007). Modeling gentrification dynamics: A hybrid approach. *Computers, Environment and Urban Systems*, *31*(3), 337–361. https://doi.org/10.1016/j.compenvurbsys.2006.07.004

Tsou, M.-H. (2011). Mapping Cyberspace: Tracking the Spread of Ideas on the Internet.". In *In Proceeding of the 25th International Cartographic Conference*. Paris, France. Retrieved from http://icaci.org/files/documents/%0AICC_proceedings/ICC2011/Oral Presentations PDF/%0AD3-Internet, web services and web map%0Aping/CO-354.pdf

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, *42*(sup1), 70–74. https://doi.org/10.1080/15230406.2015.1059251

Tsou, M.-H., Yang, J.-A., Lusher, D., Han, S., Spitzberg, B., Gawron, J. M., … An, L. (2013). Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and*

Geographic Information Science, *40*(4), 337–348. https://doi.org/10.1080/15230406.2013.799738

Tsou, M. H., & Leitner, M. (2013). Visualization of social media: Seeing a mirage or a message? *Cartography and Geographic Information Science*. https://doi.org/10.1080/15230406.2013.776754

Van Criekingen, M. (2009). Moving In/Out of Brussels' Historical Core in the Early 2000s: Migration and the Effects of Gentrification. *Urban Studies*, *46*(4), 825–848. https://doi.org/10.1177/0042098009102131

Vermeulen, P. (2016). Gentrify Everything: New Forms of Critical Artistic Agency. *MaHKUscript. Journal of Fine Art Research*, *1(2)*, 18. https://doi.org/http://doi.org/10.5334/mjfar.11

Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7227 LNCS, pp. 231–238). https://doi.org/10.1007/978-3-642-29047-3_28

Wyly, E. K., & Hammel, D. J. (1998). Modeling the Context and Contingency of Gentrification. *Journal of Urban Affairs*, *20*(3), 303–326. https://doi.org/10.1111/j.1467-9906.1998.tb00424.x

Wyly, E. K., & Hammel, D. J. (2004). Gentrification, segregation, and discrimination in the American urban system. *Environment and Planning A*, *36*(7), 1215–1241. https://doi.org/10.1068/a3610

Xu, C., Wong, D. W., & Yang, C. (2013). Evaluating the "geographical awareness" of individuals: An exploratory analysis of twitter data. *Cartography and Geographic Information Science*, *40*(2), 103–115. https://doi.org/10.1080/15230406.2013.776212

Yelp. (2014). Cheers To 10 Years of Yelp! Retrieved May 1, 2017, from https://www.yelpblog.com/2014/08/cheers-to-10-years-of-yelp-

Zukin, S. (1989). *Loft living : culture and capital in urban change*. New Brunswick: Rutgers University Press.

Zukin, S. (2000). *Landscapes of power : from Detroit to Disney World*. Berkeley, Calif.: University of California Press.

Zukin, S. (2011). *Naked city : the death and life of authentic urban places*. New York: Oxford University Press.

Zukin, S. (2014). *Loft living : Culture and Capital in Urban Change*. New Brunswick New Jersey: Rutgers University Press.

Zukin, S., Lindeman, S., & Hurson, L. (2015). The omnivore's neighborhood? Online restaurant reviews, race, and gentrification. *Journal of Consumer Culture*, 146954051561120. https://doi.org/10.1177/1469540515611203