Evaluating the Impact of Online Influencers on Retail Property Rent - A Case Study in New York City

By

Xudong Sun

Master of Architecture, Tsinghua University (2016)
Bachelor of Architecture, Tsinghua University (2015)

Submitted to the Department of Urban Studies and Planning
in partial fulfillment of the requirements for the degree of

Master in City Planning

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2019

Author_____
Department of Urban Studies and Planning
Jan 17, 2019

Certified by _____
Dr.Andrea Marie Chegut
MIT Center of Real Estate, Thesis Supervisor

Accepted by_____
Associate Professor P. Christopher Zegras
Chair, MCP Committee
Department of Urban Studies and Planning

Evaluating the Impact of Online Influencers on Retail Property Rent - A Case Study in New York City

By

Xudong Sun

Submitted to the Department of Urban Studies and Planning

on Jan.18, 2019, in partial fulfillment of the requirements

for the degree of

Master in City Planning

Abstract

This study proposes a framework of analyzing online influencer behavior and evaluating its impact on retail rent using spatial econometric methods, in which we also examined the spatial autocorrelation and heterogeneity in New York's retail rent market. We use social media data mining and network analysis techniques to examine influencers and information diffusion in Instagram and develop metrics to quantify the impact. Using spatial econometric models, we construct models of retail rents that include the effect of online influencers and traditional hedonic features. The result suggests that online influencer behavior have a significant correlation with effective rents of retail real estate in the case study area of New York. We also examine the spatial spillover effect and spatial heterogeneity of the influencer effect. Our results provide the first analysis to link online behavior to retail real estate, it also proposes a framework to study the real estate by linking online and offline world, which is meaningful for retail real estate challenged by e-commerce and other forms of new economy.

Keywords: Retail Rent; Online Influencer; Spatial Econometrics; Social Media Data Mining

Thesis Advisor: Andrea Marie Chegut, MIT Center of Real Estate

Thesis Reader: David Geltner, Associate Director of research, Center for Real Estate

Professor of Real Estate Finance, DUSP

Acknowledgements

This project would not have been possible without the support from Compstak and REMeter company, who are in partnership with MIT Real Estate Innovation Lab to provide access to the dataset used in this research.

I would like to extend my sincere gratitude to everyone who helped me during the research:

In particular, I would like to thank my thesis advisor Dr. Andrea Marie Chegut who provided me with this precious opportunity to work on the project during my academic life at MIT. Thanks to Dr. Andrea's advice, guidance, and support, I'm able to explore my focus of interest to better contribute my knowledge to the field of data analytics, spatial econometrics, and real estate.

I would also like to thank Prof. David Geltner who offered me valuable comments throughout the research and directed me to ideas and resources to enrich my knowledge and thinking.

Besides, I would like to thank MIT Real Estate Innovation Lab for continuous support and inspiration, especially its forgiving culture that promotes countless innovating ideas. It was fantastic to have the opportunity to work with the lab team. Also, thanks Prof. Joseph Ferreira for giving me advice to construct my thesis at the beginning. I am also grateful to Dr. Yi Zhu.

Finally, I must express my very profound gratitude to my parents and to my partner, Zixiao Yin, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. Without your love and caring, I wouldn't be able to gain this amazing experience at MIT in the past two years, which will keep me moving forward in the future of my life.

Content

Chapter 1: Introduction

The Challenge of Retail Real Estate in the Digital Economy

Compared to other property types, retail real estate seems to be more subject to "the unceasing forces of change" (Stephen Roulac, 1994). In recent years, new technologies, together with new activities and customer behaviors generated by them, are changing every aspect of the retail real estate. Among all these new changes, the e-commerce and online shopping have the greatest impact. Many studies have found that online retail sales growth has outpaced that of the traditional retail industry. According to the monthly retail and e-commerce report from the US Census, the percentage of e-commerce sales in total quarterly rent sales increased by 17% from 2017Q1 to 2018Q3, compared to a 16.0% increase from 2016 to 2017. Much of the growth in e-commerce comes from the consumer's "redirected" purchase from traditional types of retail spacc such as shopping centers or stores, which evokes a growing concern among both the landlords and tenants of retail real estate.

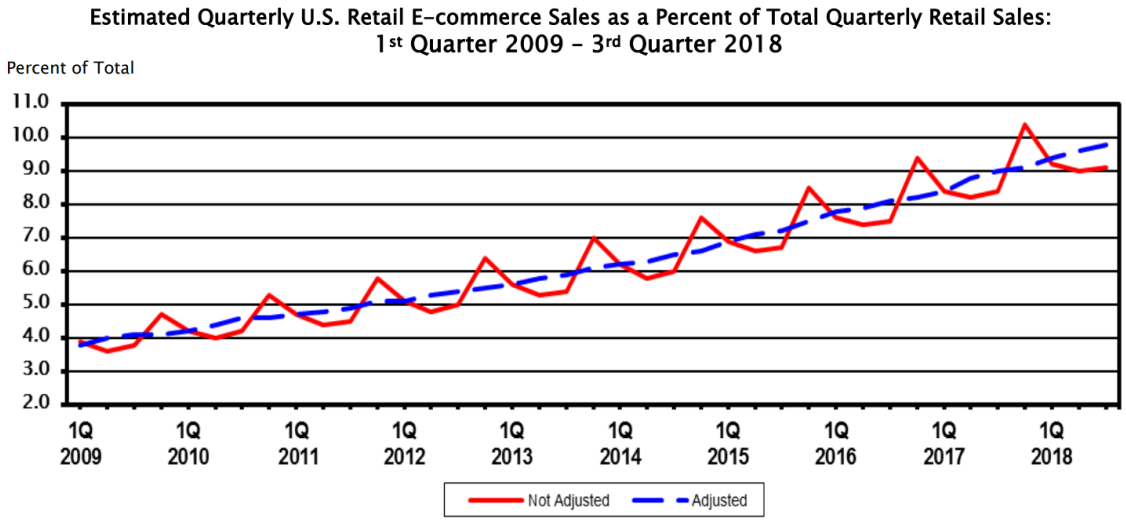

Figure 1-1: Estimated quarterly retail e-commerce sales as a percent of total quarterly retail sales. The increase proportion of e-commerce is a serious challenge to traditional retailers as much of the growth in e-commerce comes from redirected purchase (Source: US Census, Advance Monthly Sales for Retail And Food Services, November 2018)

Challenged by e-commerce and other forms of new economy, retailers are reconsidering their strategy to attract customers. For example, some retailers are transforming their physical stores into experiential destinations; some retailers combine their online and offline marketing to build customer loyalty; and there are also examples of new media-based (VR and AR, for example), highly personalized service in physical retail space to differentiate themselves. Nevertheless, all these strategies to enhance physical retail spaces are based on an in-depth understanding of new customer behavior pattern using cutting-edge technologies such as data science and machine learning, internet of things (IoT), artificial intelligence, and so forth.

For the retail real estate industry, it is equally important to understand the "new retail" and equip themselves with new technologies. In the digital economy, the productivity, or rent-generating capability, of retail space is no longer the traditional function of sales, services,

and building features. Instead, the true productivity of retail space should be a function of how the site enables the retailer to serve their customer successfully in both physical and virtual space (Norman G Miller, 2000). To attract tenants and maintain high productivity of the retail space, it is a good idea for the landlord and property manager to use cutting-edge technologies, study the retail customers, and grasp the new trends throughout the rent lease.

1.2 Online Influencer Marketing and Retail Real Estate

One of these new behaviors in the e-commerce era is online influencer marketing. The influential user (influencer) posts specific marketing-related content and affect the purchase decision over a large number of users in social media, who are potential customers. Compared to traditional marketing methods that directly focus on target customers, online influencer marketing focuses on influential users on social media websites.

In the early stage, influencers were mostly celebrities or experts who are also influential in the real world, but the demographics of influencers soon changed. The majority of influencers nowadays are “micro influencers” who are not necessarily celebrities but have the power to affect a particular group of audience. For example, a “micro” expert who has knowledge and authority in a particular niche which they actively engage with, or a popular figure among a particular group of people, as some researchers conclude that “everyone is an influencer (Bakshy, Eytan, et al., 2011).” In this study, we use this generalized definition of influencer: the social media user that post specific marketing-

related content and potentially affect the purchase decision over a large number of users in social media, regardless they are sponsored or not.

Online influencer marketing is proved to be very effective. A Neilsen marketing survey (Tapinfluence, 2017) suggests that influencer marketing outperforms most marketing methods, including digital marketing and celebrity endorsement, regarding return on investment (ROI). As a result, online influencer marketing is increasingly welcomed by brands. Recent influencer marketing reports by Forbes estimates that in 2017, nearly 50% of the brands have established specific funds in hiring social media influencers to promote their brands (Forbes 2017), which is a significant trend in retail industry.

Online influencer marketing is one of the new behaviors generated by new technologies and applications. Over the last decade, we have seen the rapid growth and increasing importance of social media. According to a report on social media use in the U.S. by Pew Research Center, roughly two-thirds of adults (68%) in U.S. report that they are Facebook users, and almost three-quarters of them are daily users. YouTube's user group is approximately three-quarters of adults and 94% of young adults between 18 and 24 in the U.S. Instagram, who has the most influencer group until 2018, is used by 35% of adults and 71% 18-to-24-year-olds. The large user group and ubiquitous influence of social media provides a great potential for online influencer marketing and its impact on retail real estate.

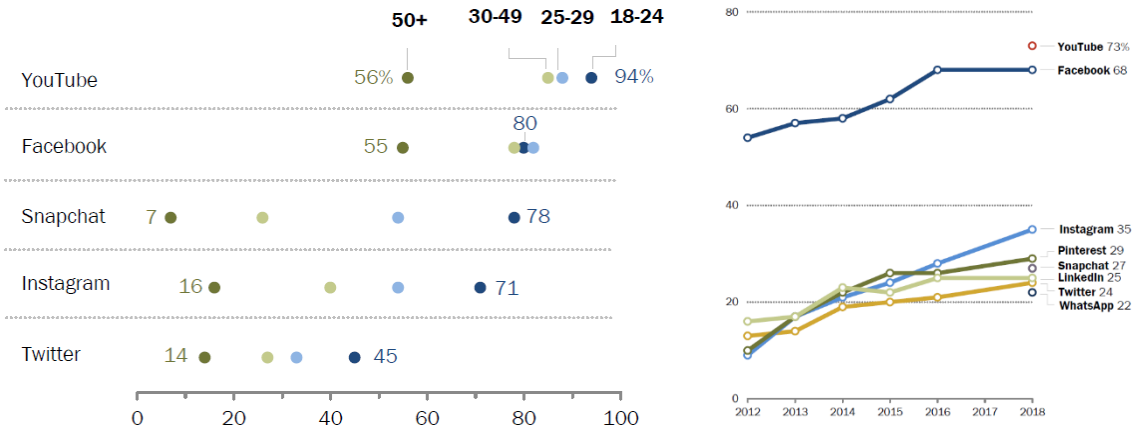

Figure 1-2: Survey of social media user group. Left: percentage of U.S. adults who report themselves as social media users in each age group; Right: percentage of U.S. adults who report themselves as users of common social media websites or mobile. The survey shows the rapid growth and increasing importance of social media. Specifically, Instagram is especially popular among young adults. (Source: Social Media Use in 2018, Pew Research Center)

More importantly, although it has some controversial side-effects, social media has become a part of the online lifestyle and a platform for opinions and online behaviors. As we discussed earlier, retailers nowadays are trying multiple strategies to redirect the customers back to physical stores. Online influencer marketing plays a critical role in this process for many reasons. It directly links the online and offline spaces and combines the advantages of both sides: the speed and efficiency of information diffusion in online space, and the richness of experiences in physical space. We can say that online influencer is not only an agent through which retailers and retail real estate broadcast their new experiences but also a part of new retail "process" that combines online with offline and constitutes new lifestyles.

Therefore, it is meaningful for real estate researchers to study influencers and other similar online behaviors and evaluate their impact on the real estate market. Although many have

studied online influencer marketing from multiple perspectives, few studies connect the topic of online influencers with retail real estate.

1.3 Research Goal

In this study, we try to propose a framework of analyzing online influencer behavior and evaluating its impact on retail rent using spatial econometric methods, in which we also examined the spatial autocorrelation and heterogeneity in New York's retail rent market.

Specifically, this research starts with the questions that what is the online influencer and influencing marketing, how does it work, and how can we quantitatively evaluate its impact? Then we investigate how influencers affect the retail rent in both theoretical and quantitative ways.

Given access to granular retail rent data (Compstak dataset, discussed in Chapter 4), the research intends to develop a quantitative method for evaluating the impact of online influencer behaviors on retail rents. Using network analysis and spatial econometrics, the method is designed to be replicated and applied to other kinds of online behaviors in social media. Furthermore, based on statistics, this research explores a new method to model and predict rental value based on online behaviors. Through such coupling, the study aims to suggest the correlations between the online world and real estate, in the hope of providing guidance in retail marketing, and real estate research, management, and investment: Is the online influencer marketing effective? How can we quantitatively evaluate the effect? How does it impact the real estate market?

By leveraging the richness of online data and spatial econometrics methodology, the research aims to create a dialogue between online influencers' behavior and property value in the physical environment and provide a new perspective to study the real estate by linking online and offline world.

1.4 Research Framework

This study is structured as follows. In Chapter 2, we first reviewed the studies on how scholars model the retail rent to evaluate a series of related features and predicting the profit-generating performance of the retail property. For online influencer and influencing marketing, we went through the theoretical research and quantitative analysis on several central questions: who influencers are, how they influence retail customers, and how we can study the influencing process and quantitatively measure their impact. Additionally, as we mainly use spatial econometrics tools in this study, we also reviewed the studies on these tools.

Chapter 3 briefly proposes theories on how influencers affect the retail rents. Based on the theories on the factors that affect retail rents, especially the theories on customers' behavior pattern, we proposed two possible theories: influencing people and influencing the place. We also made assumptions for the modeling.

Chapter 4 introduces the study area and three study datasets: Compstak, and Instagram dataset. In the last section of this chapter, we briefly introduced the method of using Instagram API to query influencer posts and corresponding follower networks.

In Chapter 5 to 7, we focus on the model framework, development, and findings, including 1) Investigate the magnitude of influencing effect, which we call “influencing score”, of each identified influencer in Instagram dataset; 2) Model the rent price using a spatial model that incorporate influencers’ location and magnitude; 3) Compare models and estimation results. Finally, we briefly summarize the limitations of this study and discuss the proposal for future studies.

To sum up, this study brings a new field of research on social media and information diffusion into real estate. We develop a framework to quantitatively evaluate the impact of online influencers based on the network model and associate the influencers’ online behavior on retail rents. Using spatial econometric methods, we find a significant and quantifiable effect.

Chapter 2: Literature Review

2.1 Online Influencers & Influencer Marketing

2.1.1 Theoretical Research on Influencers and Influencing Marketing

Many have addressed the topic of online influencers as an effective marketing method, and most studies attribute the effectiveness of influencers to social media. Freberg et al. (2011) suggest that online influencers "have emerged as a dynamic third-party endorser" given the huge user group and the ubiquitous influence of social media.

People have developed many theoretical models to describe online influencer marketing. One of the mostly applied theories is social learning theory by Bandura (1963) that describes several steps, including observational learning, mediational process, output behavior, and evaluation, that can describe how social interactions change an individual's behavior. Many studies apply the social learning theory on the topic of online influencer marketing and explain the impact of influencers on consumption behaviors (Makgosa et al. 2010). However, social learning theory cannot describe information diffusion through social media.

Fisherman's Influence Model is a simple description of the influencing process through social media, that the marketing message starts from an influencer, spreads throughout the influencer's social (follower) graph. This process is called amplification. Through several

amplifications the marketing message will be received by a potential customer (Brown et al. 2014).

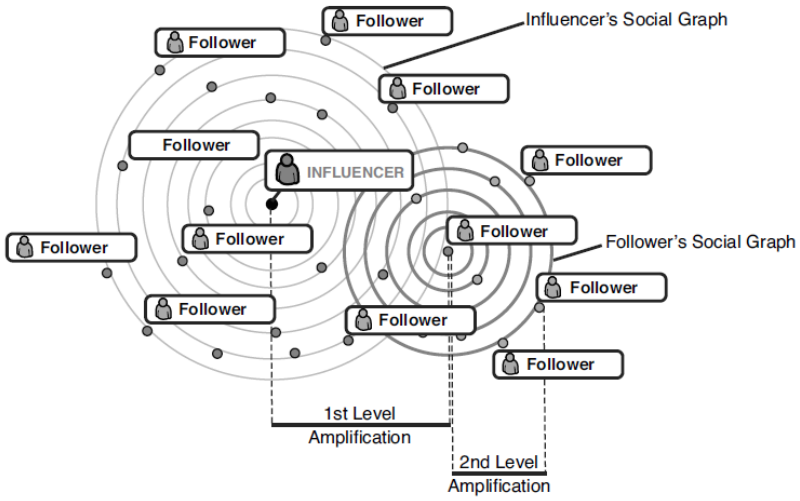

Figure 2-1: Fisherman's influence model. In this model, the marketing message starts from an influencer, spreads throughout the influencer's social (follower) graph and finally arrives customer. (Source: *Influencer marketing: Who really influences your customers?* Brown et al. 2014)

Guided by this model, the effect of an influencer is largely decided by the size of the follower graph. An influencer with more followers can drive greater brand awareness and will be more likely to trigger purchases.

The Fisherman's influence model has an assumption that each potential customer that receives the marketing message has the same probability of purchasing. In some cases, this assumption is violated. For example, we discuss a special kind of influencer in Chapter 1.2 called "micro influencers" that have authority only in a niche. A micro influencer can only affect a particular group of people on buying a particular kind of product.

Compared to the Fisherman's influence model that is centered on the influencer, Brown et al (2017) proposes the customer-centric influencer model that positions the customer at the center, and one customer might get "influenced" by multiple influencers whose social graphs intersect the customers'. Guided by this model, the best influencers for a brand should be those that are close to the potential customers in their social graph, and a customer's purchase decision is more likely to be triggered by multiple influencers.

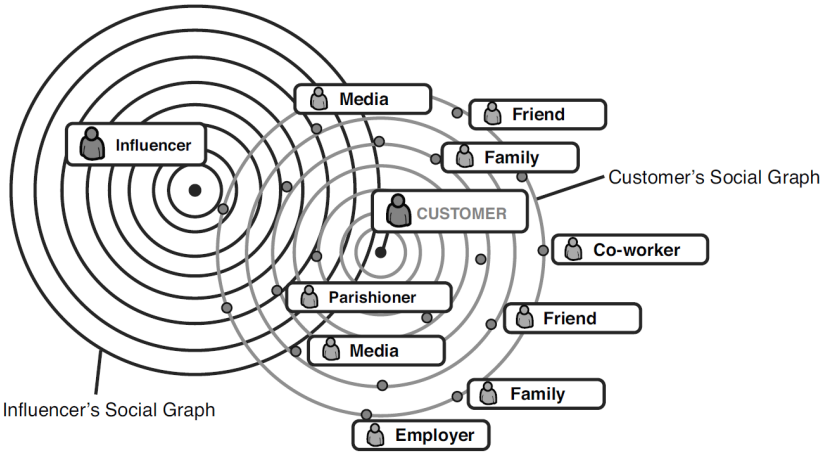

Figure 2-2: Customer-centric influencer model. In this model, the customer is the center of the model who gets marketing message from potentially multiple influencers through overlapping social graph. (Source: *Influencer marketing: Who really influences your customers?* Brown et al. 2014)

2.1.2 Measuring Influencing Effect

Researchers are also studying on how to measure the influencers' effect, especially in quantitative ways. The quantification of influencing effect is the basic step for multiple commercial applications (Lagrée, Paul, et al. 2017).

The first step to measure influencing effect is to identify the influencers among a great number of social media users. Some studies suggest that most influencer behaviors are non-sponsored and largely self-motivated, and these non-sponsored influencers are mostly not distinguishable from the sponsored only from their post content (Bakshy, Eytan, et al., 2011). Lahuerta-Otero, et al. proposed a method to identify influencers on Twitter using graph theory and social influence theory (Lahuerta-Otero, et al 2016). The result of this study suggests that influencers have some features that can be identified using graph theory (specifically, network centrality metrics).

The graph theory approach is also used in tracking the influencing process and measuring influencers' impact. Kwak et al. (2010) used two network centrality measures: degree centrality (number of followers) and page-rank, together with the number of retweets, as measures of influence. Bakshy at al. (2011) used the size of diffusion tree in Twitter as the measure of influence. However, the diffusion tree method only applies to the social media websites that have "repost" or similar functions so that we can track the information diffusion by reposts. Manikonda et al. (2014) examined multiple network metrics of Influencer users as a whole, including homophily, reciprocity, and clustering coefficient. But there are few studies addressing the identification and impact measurement of Instagram influencers.

For measuring influencers' impact, Miller, Rohan, and Natalie Lammas (2010) suggested that volume metrics, such as network degree centrality, number of hits, likes, or website traffic, cannot fully measure the impact. Instead, Angel and Sexsmith (2009) proposed metrics that also incorporate qualitative features such as tones and the quality of

interaction. Fisher (2009) incorporated more features like content freshness and relevance, relevant actions, conversation size, author credibility, and so forth.

2.2 Spatial Econometric Models in Real Estate Research

2.2.1 Spatial Econometric Models

As a subfield of econometrics, spatial econometrics addresses spatial effects in regression model data (Paelinck and Klaassen, 1979; Anselin, 1988a). Spatial econometrics was initially applied to some specialized fields that deal with spatial or geographical data, for example, geography, urban and regional economy (Pace et al., 1998). But recently, spatial econometrics has increasingly been applied in other fields of economics. Spatial econometrics has become an essential part of mainstream applied econometrics (Anselin, Luc, 2010).

To address various spatial problems, researchers have developed a series of spatial models over the last few decades. We can divide these models into two groups: those addressing spatial autocorrelation, and those deal with spatial heterogeneity.

Spatial Autocorrelation

Spatial dependence, or spatial autocorrelation, means that the observation at one location depends on other observations at other locations. It usually results from (a) the existence of spillover effects, in our example of retail rents, the impact of rent changes in one retail property on the rent prices of its neighbors; (b) spatially correlated omitted variables; or

(c) measurement error or misspecification of the functional form. Cliff and Ord (1973) proposed spatial autoregressive regression (SAR) model, one of the mostly used spatial models to address spatial autocorrelation. And Anselin's (1988) and Le Sage's (1998) books on spatial econometrics summarizes most spatial modeling techniques.

For SAR model with time fixed effects, Lung_fei Lee et al. (2009) suggested that SAR panel data models using a limited number of time fixed variables might have different asymptotic properties with non-panel-data SAR models. Nicolas Debarsy et al. (2010) proposes a method to assess spatial autocorrelation in a fixed effects panel data model. Using LM and LR tests, the method can distinguish two types of spatial autocorrelations: spatial lag and spatially autocorrelated errors.

Another approach to address spatial autocorrelation is to use the spatial fixed effect in the multiple regression model. Ciccone (2002) suggests that "the introduction of increasingly detailed spatial fixed effects allows to control for spatially correlated omitted variables." McMillen DP (2003) proposes a Monte Carlo experiment and finds that incorrect functional form can lead to spurious spatial autocorrelations, which can be corrected with fixed effect variables. However, some researchers such as Luc Anslin and Daniel Arribas-Bel suggest that spatial fixed effect cannot address true spatial dependence but is just a form of spatial heterogeneity (Anselin, Luc, et al. 2013). When the data generating process contains spatial autocorrelation or spatial error dependence, the method of spatial fixed effect becomes more spurious. The spatial fixed effect can only successfully remove the spatial autocorrelation when it only exists in each "spatial subset" of samples, such as a state, a town, a or a block.

Spatial Heterogeneity

Spatial heterogeneity shows up regarding spatial heteroscedasticity or spatially varying parameters. For example, the effect of an independent variable on the dependent variable in a regression model can change at different locations.

Using local models is one easily-implemented method to deal with spatial heterogeneity. Similar to the data processing method of spatial fixed effect model, we divide the study area into distinct geographic subsets and estimate a local model for each subset (Schnare and Struyk 1976; Goodman 1981, 1998; Michaels and Smith 1990; Bourassa et al. 2003). But this method has its limitations. For example, to fulfill the assumptions of OLS regression model, there should be no spatial heterogeneity inside each spatial subset. But it is usually difficult to define subsets that can accurately grasp the pattern of spatial heterogeneity and eliminate spatial heterogeneity in each subset (Helbich, Marco, et al. 2016).

The spatial expansion methods are a series of modeling methods addressing spatial heterogeneity pioneered by Cassetti (1972). By replacing the independent variable whose effect has spatial heterogeneity with a function of some location-specific features, the spatial expansion methods "expand" the parameters and allow varying parameters in the OLS model framework.

There are several critical modified versions of the spatial expansion model. The Tucson, Arizona, Fik, et al. (2003) proposed a fully interactive expansion model. The study uses property coordinates and submarket dummy variable in a second-order polynomial

function of housing attributes. Many other studies also use spatial expansion function on coordinates (Clapp 2001; Pavlov 2000). However, due to the complexity of the expansion functional form, the Tucson (2003) model only includes three housing attribute variables, which might cause biased results due to omitted variables.

The geographically weighted regression (GWR) is another approach addressing spatial heterogeneity(Brunsdon et al. 1996; Fotheringham and Brunsdon 1999; Fotheringham et al. 2002). GWR is similar to both the local model method and spatial expansion methods. It estimates local models and allows varying parameter estimates over space, which is similar to the local model method. However, GWR does not rely on spatial subsets. Similar to the Tucson (2003) spatial expansion model, GWR takes the coordinates as the basic spatial unit. For each study point, GWR estimates a local model using observations whose values are weighted by (a function of) their distance to the study point (Fotheringham et al. 2002). Some studies suggest that GWR has better explanatory power and prediction accuracy than spatial expansion model (Bitter, Christopher, et al. 2007). Another method called moving window regression (MWR) is a form of GWR when the local model does not use all weighted observations but only the unweighted values of N nearest observations (Brunsdon et al. 1996). Some studies suggest that MWR has slightly less prediction accuracy compared to GWR, and GWR results are more robust for a wider range of window size selection (Páez, Antonio, et al. 2008).

GWR/MWR model also has limitations. Many studies address the local multi-collinearity problem of GWR/MWR (Wheeler and Tiefelsdorf, 2005; Griffith et al. 2008). Since GWR/MWR selects a subset of all observations and uses weighted values of these

observations as input, the multicollinearity can be introduced to the local input in this process (Páez et al. 2011).

In recent years, researchers have been modifying the GWR/MWR model and propose new spatial models that address spatial heterogeneity. Marco Helbich et al. (2015) evaluate several spatial models including the spatial expansion model, MWR, GWR and compare their spatial patterns of local parameter estimates, and propose a new model addressing spatial heterogeneity called eigenvector spatial filtering (ESF). ESF model outperforms GWR and MWR in prediction accuracy, and more importantly, ESF does not have the local multi-collinearity problem like GWR/MWR. But it is less intuitive then GWR and harder to interpret.

2.2.2 Applications to real estate

The importance of location in determine real estate values is axiomatic (Mats, 2002). Researchers in both theoretical and applied econometrics, including real estate, have acknowledged spatial autoregression and spatial heterogeneity (LeSage and Pace 2009). But it is challenging to incorporate space effect into traditional models. The primary motivation of applying spatial techniques in real estate is to increase the precision in estimation the property value (Dubin, Robin, et al., 1999).

The multiple regression method in which we use multiple property features to predict the property value was introduced into real estate research by Eisenlauer (Dubin, Robin, et al., 1968) and Blettner (Blettner, Robert A. 1969). However, the locational features are usually hard to observe and quantify and omitted in non-spatial model specifications. Many studies

suggest omitting spatial effect, both spatial autoregression and spatial heterogeneity, can cause geographic errors. For example, Thrall (1988) categorized common errors caused by misuses of geographic data, including spatial autocorrelation that all regression models using geographic measurement are potentially subject to.

For spatial autocorrelation, Dubin, et al proposed a model specification that combine a spatial autoregression with traditional multiple regression specification(Dubin, Robin, et al, 1999), in which spatial effect is modeled by a spatial weight matrix. The study also suggested its application in retail site selection.

There is an extensive literature that addresses spatial autoregression or uses spatial autoregressive model for real estate research. For example, Can and Megbolugbe (1997) examines the spatial spillover effect in house price; Brasington (1999) models the effect of school quality on property values using both traditional hedonic model and spatial autoregressive model; Angel Ibeas et al (2012) investigated the house price variation that is affected by changing transportation conditions using multiple linear regression and SAR model.

For spatial heterogeneity, expansion methods are widely used for real estate research. Ayse Can (1992) uses both the spatial expansion model and SAR model. The study addresses spatial autocorrelation in dependent variable using SAR model and spatial heterogeneity using expansion function by market segmentation (census tract). Thériault et al. (2003) use spatial expansion model that has two sets of expansion function to transform housing attributes: accessibility and neighborhood attributes. The main limitation of these study is the granularity of the basic spatial unit.

Although GWR and MWR model has not been used in real estate context, there is an increasing interest in using GWR and MWR to examine spatial heterogeneity. Compared to spatial expansion models that use dummy variables of spatial subsets, GWR and MWR have the advantage that the marginal prices and other parameter estimations can vary continuously across space. Another appealing feature of GWR/MWR is that it "partly mimics appraiser's sales comparisons and price adjustment processes" (Bitter, Christopher, et al. 2007). In Bitter (2007), GWR is used to model the spatial variation in housing attribute prices and outperforms spatial expansion models.

Additionally, as the local model of GWR has the same framework with traditional multiple regression model used in real estate, it is relatively easy to use GWR to modify traditional models and increase performance (Helbich, Marco et al. 2016).

2.3 Limitations

We have reviewed the studies on online influencers, spatial econometric models, and the modeling methods for real estate research. Although there are some studies addressing the effect of influencer marketing on customers or urban physical environment, and people have developed multiple methods to quantify an influencer's impact, there are few studies relate this topic to retail real estate. Meanwhile, real estate researchers have started to incorporate features that are associated with the new economy into their models, but few studies address the topic of online influencer marketing or other forms of online behavior that can be connected to real estate.

Chapter 3: How Influencers Affect the Retail Rent: A Theoretical Approach

3.1 The Factors That Affect Retail Rents

To evaluate the impact of the influencers on retail rents, we first discuss the factors that affect retail rents from a theoretical perspective. We hope to answer the following questions: how effective rents are decided and how could influencers change some of these factors.

Theoretically, the equilibrium rent of retail space should represent a proportion of the "excess of income over expenditure of a trade carried on in the premises" (Emeny et al. 1984). In reality, we can hardly observe the equilibrium rent in the form of effective rent of a lease. Part of the reason is that most retail rent leases have relative long lease terms except for the pop-up retail space (Kim, Hyejeong, et al. 2010; Ryu, Jay Sang, 2011), which is still a new form of retail and not widely adopted.

Appraisers' perspective on retail rents can help us understand the question of how effective rents are decided. Appraisers use multiple methods to determine a proper price for a retail space, in which the concerns of both landlords and tenants are considered. The basic method is the comparison, which is based on the simple idea that the rent of two identical properties should be the same (Fisher, Martin, 1994). When comparing two retail properties, appraisers consider the features of the space when deciding the rent price, for example, the frontage, depth, size of the

store, the service, and equipment quality of the building, and so on Crosby et al., 1992; Adair et al., 1996a). These factors are usually used in hedonic models by real estate researchers (MacFarlane & Fibbens, 1990).

Other than features of the retail space, appraisers should take the location into account, as an old cliché of real estate says that "location, location, and location" are three most important factors for real estate. The location of a property represents a series of features: population, transportation, and accessibility affect the number of potential customers; the demographics, local industries decide the behavior pattern of local customers; the quality of public services change the management cost, and so forth.

Appraisers also make adjustments on rent price based on details of the rent lease (O'Roarty, Brenna, et al. 1997)[1]. For example, transaction size, lease term, rent bump, and more importantly, the price of previous or nearby rent leases. The effect of previous rent price is captured by time series models such as the AR model, and the impact of nearby rent price can be modeled by spatial autoregressive models.

To sum up, other than the features of the retail space and rent lease details, the rent value of a retail property is primarily decided by a series of location-related factors including the number of potential customers, transportation and accessibility, the quality of urban environment and public services, the quality and popularity of a retail district, and so forth. We can roughly categorize these factors into two groups: those related to customers (for example, demographics, customer behavior) and those related to place (for example, urban environment and retail district).

[1].

3.2 Influencing People & Influencing Place

Influencer marketing is targeted at potential customers. Therefore, we think it will affect retail rents by influencing the behavior of potential customers. We call this process "influencing people." The theory of influencing people is based on many studies on marketing.

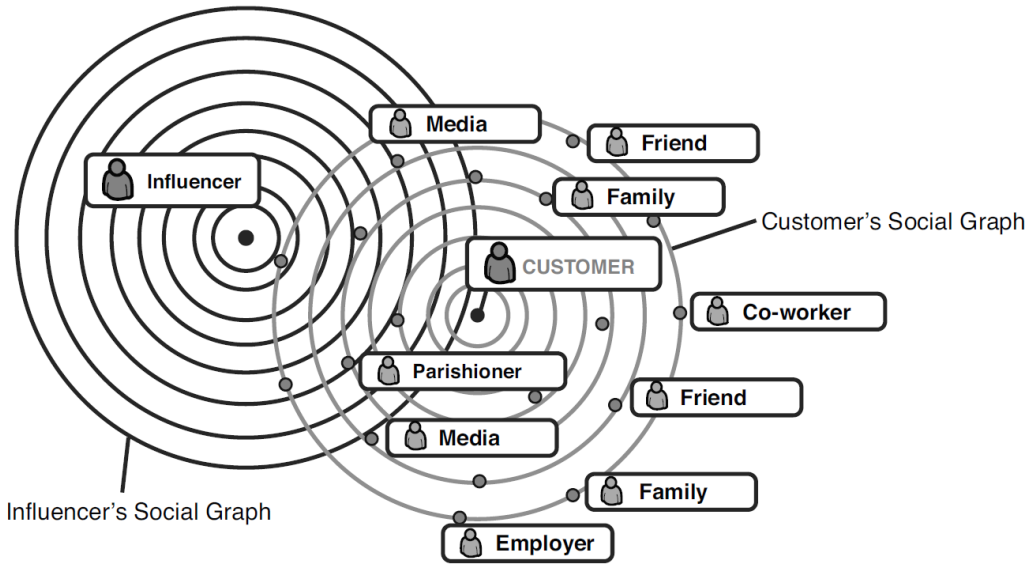

Figure 3-1: The implication of the customer-centric influencer model. A customer's purchase decision is more likely to be triggered by multiple influencers who share overlapping social graph with the customer. (Source: *Influencer marketing: Who really influences your customers?* Brown et al. 2014)

This theory seems to be very intuitive: if more customers are influenced, a store can generate more revenue, and we expect higher rent prices. However, this process only applies to certain lease types that part of the rent is proportional to sales (percentage lease). For other lease types, this kind of effect can only be evaluated when the current lease term ends, and a new one starts. Also, this theory means the rent price is somehow affected by the previous tenant, which becomes counter-

intuitive given the fact that in the real world two consecutive tenants are seldom similar. For example, consider a retail space that accommodated a clothing store with a lot of influencers. This fact is less likely to increase the rent price of the next tenant if the next tenant is a book store.

Another theory is "influencing place." Although each influencing post has hashtags that denoting a particular target, for example, a store or restaurant, we think the post has wide-area effect, and other retail spaces nearby can also benefit from it. People are not only attracted to the targeted location, but they also perceive the nearby area. For instance, when people see an influencer post targeted on a store in Hudson Yards, the post not only provides the information about the store, it also refreshes the memories or stimulates the curiosity for Hudson Yards: its history, its redevelopment, and its new spaces. Also, if people are attracted to the store, they are very likely to get into some other stores nearby when they are shopping.

In the context of retail rents, we think the "influencing place" theory makes more sense than the "influencing people" thaory. For example, when the rent price for retail space is decided, the landlord, broker, and tenant will consider the influencer behaviors in surrounding areas. If there are a great number of influencer posts nearby, the tenant will be more likely to accept higher rent prices for a retail space because it is a sign of popularity or online visibility.

3.3 Basic Assumptions

Based on the theoretical discussions, we made several basic assumptions for this study as follows.

First, all influencer posts have wide-area effects on proximate retail rents, and the wide-area effect is the predominant effect of the influencer behavior. As we discussed in Chapter 3.2, we are more

interested in the wide-area effect of the influencer behavior as a predictor of retail rents, which is a more intuitive way to explain how retail rents are affected.

Second, we use the "averaged" influencer behavior of the year to proximate the overall temporal effect. The influencer behavior could affect the rent price before and after the commencement date of the rent lease, and the process can be complicated. As preliminary research, we consider using the yearly averaged influencer behavior. For example, a rent lease on Jan 8th, 2018 is affected by the influencer behavior of the previous year, from Jan 8th, 2017 to the commencement date.

Third, although the effect of influencer posts will diminish with both time and distance, we weight their effect only by the Euclidean distance from the post to the retail rent. The temporal diminishing process is closely linked with the information diffusion pattern in the social network, which differs for every influencer (Bakshy, Eytan, et al., 2011). We will further investigate the temporal effects in future studies.

Chapter 4: Study Area and Research Dataset

4.1 Compstak Retail Dataset

The table 4-1 lists all used variables in Compstak Dataset

	Variable	Format	Description
Transaction Detail	Sublease	binary	if the transaction is a sublease
	Free_rent	float	free rent period in years
	bump_rate	float	rent bump rate of the lease
	bump_year	float	rent bump year
	Lease_term	int	rent lease term in years
	Lease Type	categorical	lease types: gross, net, etc.
	Transaction Size	float	the size of the retail space
	Commencement Date	date	commencement date of the lease
Tenant Information	Tenant Industry	categorical	the industry of the tenant
Building Information	Building Age	date	Age of the building
	Building Renovation Time	date	the nearest building renovation time
	Floor Occupied	int	the level of the lease
	Building Class	categorical	The class of the building, A, B, C or unknown
	Property Type	categorical	The type of the building where the retail space is in

4.3 Instagram Data

To identify multiple influencers among all social media posts and evaluate the impact, we need two aspects of information. First, we want to know the content of posts and its location within the study area. Second, for those posts with related content, we are interested in a “network” or "graph" for following relations among users, from which we can analyze the diffusion pattern of information and identify influencers.

4.3.1 Content of Post

Using public API, we scraped down all public Instagram posts that meet our criteria from September 20, 2015, to November 20, 2018) within our study area. We focused on the posts that have certain typical characteristics.

According to truth-in-advertising laws and standards of the FTC (Federal Trade Commission), a sponsored influencer should hashtag both the name of the brand and the word “sponsored.” However, in our discussion of retail real estate, we think the influencer group is not limited to the sponsored professional influencers. Every Instagram user that acts like an influencer should be considered. Also, the “influenced” should include not only brands but also the physical environment, such as the building, the street, or the neighborhood. Therefore, our filtering rules for influencer posts are as follows:

Figure 4-1: An example of Instagram influencer post. (Source: Instagram)

1. The post has hashtags that is related to the retail property in our retail rent dataset, for example, every post that hashtags the name of the brand, store, or building are recognized as influencer posts regardless whether they are sponsored.

2. The post has the geolocation information. Instagram influencers does not necessarily post their locations, especially when they are posting only for a brand. In our study, the assumption is that only posts with geolocation information are recognized as valid influencing posts.

3. The user has more than a certain number of followers.

The scraping process starts with building the targeted hashtag list. We applied data fusion techniques to find the "full list" of unique related hashtags. The data fusion process consists of 3 parts: first, in Compstak dataset we have the tenant store name information, which can be directly used for hashtags; second, the building name; third, as some retail property have multiple units, we supplement the hashtag list with POI (point of interest) dataset from NYC OpenData. We find the retail stores in POI dataset that are in the same building.

The Instagram API has limitations on the queuing method and we can only use the username or hashtags to extract data. We built a web scraper to scrape all posts that containing targeted hashtags.

Table 4-1: The variable list of Instagram post dataset

Variable	Type	Description	Variable	Type	Description
username	String	A user's name shown in Instagram	**comment_count**	Float	Number of comments of this post
user_id	String	A user's unique ID in Instagram database	**lng/lat**	Float	The geographic coordinates
follower_count	Integer	Number of followers at the post time	**date**	Date/time	The timestamp of the post
like_count	Integer	Number of likes of this post	**tags**	String	The hashtags of the post

After the filtering and data cleaning, we got the dataset of Instagram posts containing the information about the user, post content, the list of hashtags, geographical coordinates, number of

followers, number of likes and comments of this post. The table 4-1 shows all variables of final Instagram posts dataset.

4.3.2 Follower Network

The impact of an influencer post also depends on how "influential" he or she is. For social networking sites (SNS) like Instagram, Twitter, or Facebook, we can use a network graph to model the users' relations, measure the importance of each user, and study the information diffusing pattern.

We have a list of influencers from the Instagram posts dataset. Using Instagram API, we can queue the list of followers of each influencer, then we repeated the process and find the followers of followers. In theory, we can repeat this process until we have enumerated all users, but in this study, we only repeated this process twice. Finally, we obtained a follower graph in which one user is connected with other users by a directed edge.

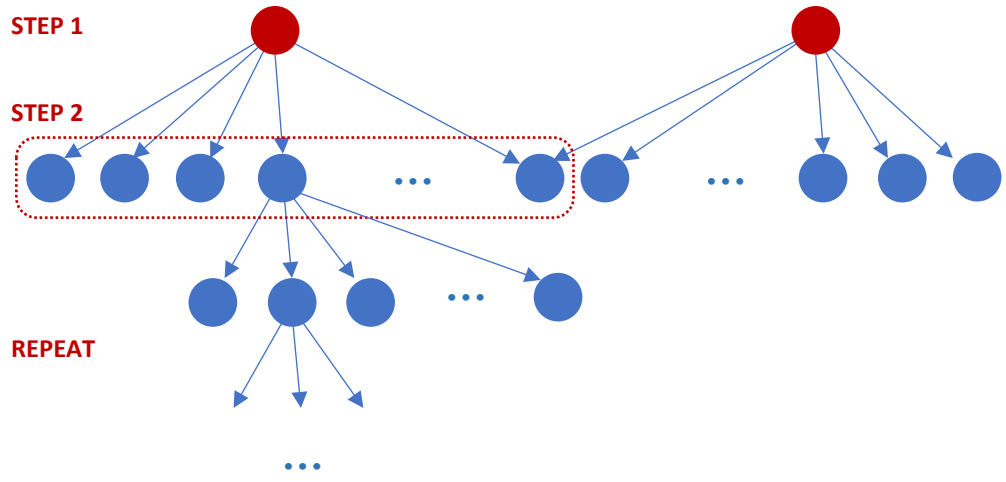

Figure 4-2: The process of mining the follower network from Instagram API.

Chapter 5: Model Development

5.1 Overview of Approach

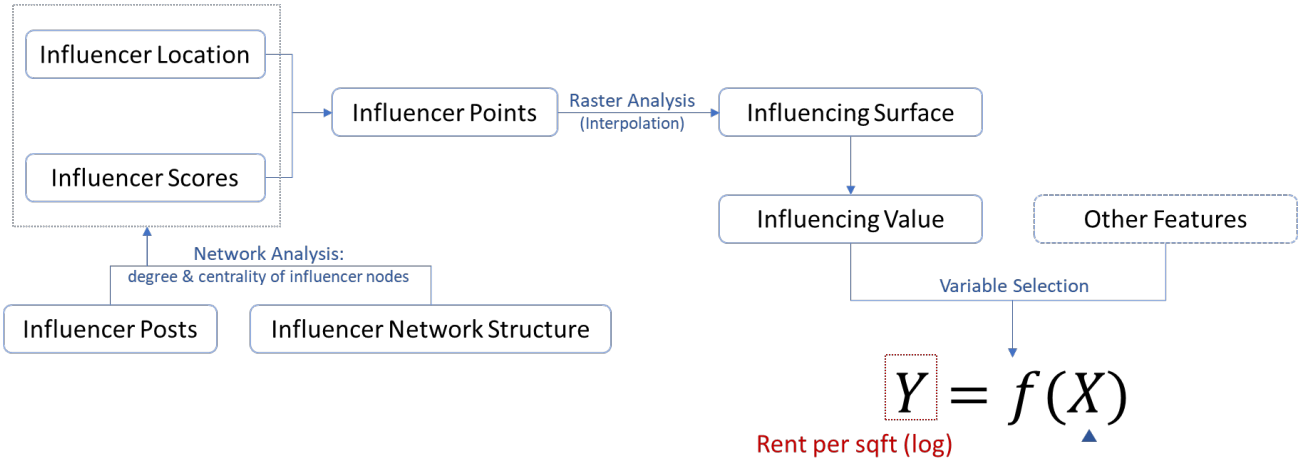

Figure 5-1: The modeling framework of this research.

The goal of this study is to develop a model that can quantify the Instagram influencer behavior and measure its impact on retail rents, which can enhance our understanding of the connection between online behaviors and urban placemaking. Using the influencing score that measures the joint effect of both direct impact of the influencer post and the potential impact, we proximate the information diffusion process. Combining the influencer score and the location of the post, we get point-observations, which we use to generate a "surface" of influencing value for every point in our study area. The surface is generated using spatial interpolation methods.

Finally, the influencing value, together with other features related to building features, transaction features, time and location fixed variables, will be used as explanatory variables for our models.

The dependent variable is the effective rent per square feet. In the rest of this chapter, we will discuss the calculation and descriptive statistics of these variables in detail.

5.2 Influencers Analysis

The best way to decide the general impact of one influencer post is to track the information diffusion process after it is posted and investigate how the information passes from one user to another, how far does it go, and how it evolves through the diffusion process. The information diffusion process is dependent on the structure of the follower network and the user's position.

In this research, we proximate the information diffusion with influencer score. The influence score is the measurement of the joint effect of the direct impact of a post, which can be measured by the number of likes or comments, and the indirect or potential effect, which is dependent on the influencer's position in the follower network.

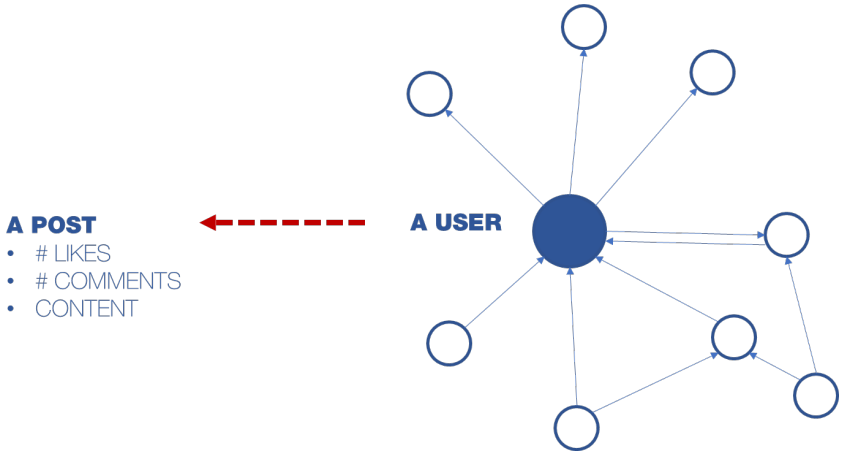

Figure 5-2: The joint effect of the direct impact of a post: the direct impact of a post, which can be measured by the number of likes or comments, and the indirect or potential effect, which is dependent on the influencer's position in the follower network

For example, we compare the influencing score of two posts by different influencers. We think that the post with more likes or comments (until the day) has a greater direct impact. If we have the same immediate impact, but one user has more followers or followed by other users who have many followers, the information from this influencer is more likely to be conveyed to more users and become more influential. Therefore, we need to further investigate the influencer's network features and the information diffusion process.

5.2.1 Network Analysis on Influencers

To decide the relative importance of the influencer, we calculate the centrality measurements in the follower network. Three centrality measures are related to our topic: degree centrality, eigenvector centrality, and betweenness centrality. As the follower network is directed[2], and we are more interested in the "downstream" information diffusion, all centrality measures we talk about are out-centralities.

The degree centrality is a simple centrality measure that counts how many neighbors a node has. In a directed network, the out-degree means the number of outgoing links. In the context of SNS, it is the number of followers. In most cases, the degree centrality is a good measurement of how influential a user is. But in some cases, the degree centrality is not a complete measurement. For example, if a piece of information is passed through multiple users, the degree centrality can only measure the first step.

The eigenvector and betweenness centrality are better choices for complex follower networks. In graph theory, eigenvector centrality (also called eigen-centrality) is a measure of the influence of

[2] A directed edge from node A to node B in the follower network means A is followed by B, or A's post passes to B.

a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. In our context, high eigenvector centrality means the users are followed by other influential users.

Betweenness centrality is based on the path that goes through the node. It represents the degree of which nodes stand between each other. A node with higher betweenness centrality would have more control over the network, because more information will pass through that node.

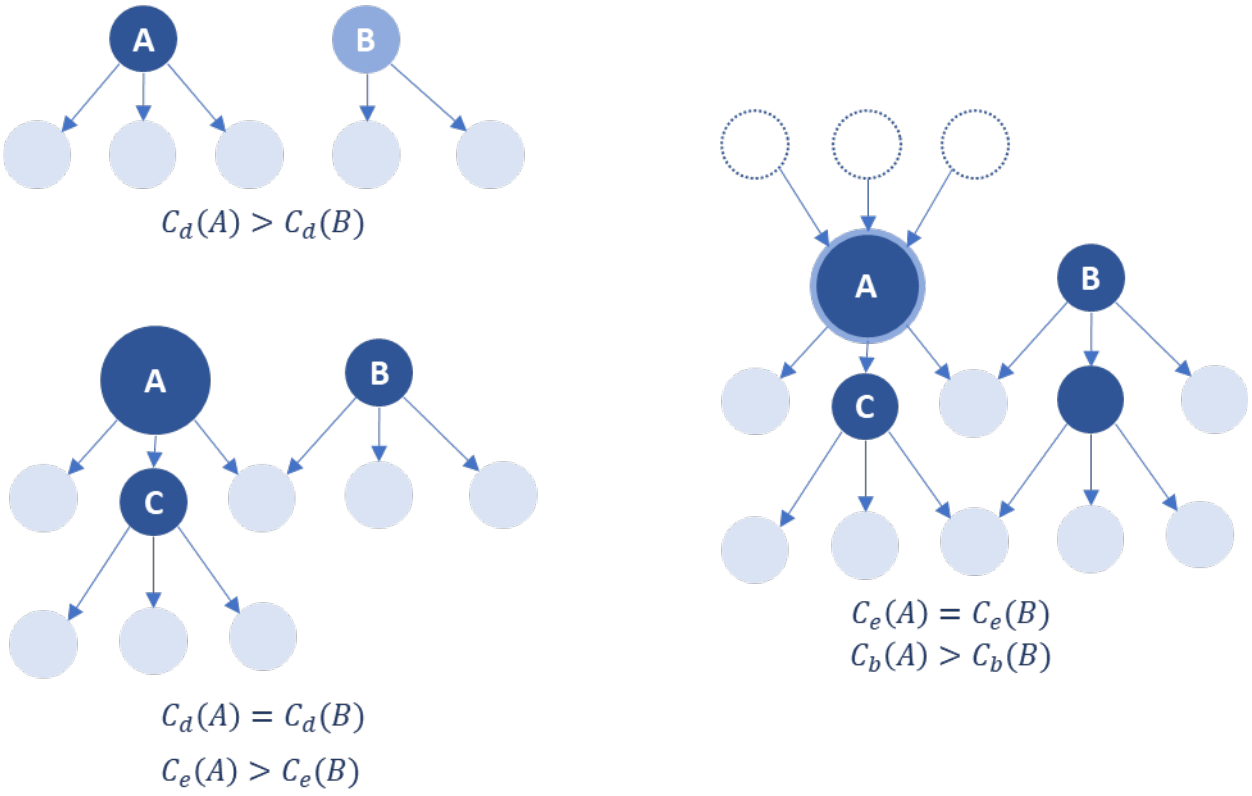

Figure 5-3: 3 centrality measures: degree, eigenvector centrality, and betweenness centrality

The figure 5-3 illustrates the difference among these centrality measures. In the first section, the influencer A and B have different number of followers, thus A has higher degree centrality than B; in the second case, A and B have the same number of followers, but one of A's followers, C, is more influential than the rest of followers (assume they don't have other followers). Therefore, A has higher eigenvector centrality than B; in the last case, A and B has the same number of followers with same relative importance, but A is on the path of some other influencers, but B is not. Thus, A has higher betweenness centrality than B.

The choice of centrality measure also depends on the network structure we want to investigate. For example, betweenness centrality is an ideal measure for Twitter influencers, because Twitter influencers often repost instead of post new messages. The eigenvector centrality is a good way to remove the effect of inactive followers or "bots" but require high computational power. In this research, we will calculate the eigenvector centrality on the simplified network that has max depth 2, which means only the influencer's followers and their followers. The distribution of centrality is shown in the following figure.

5.3.2 Calculation of Influencing Score

The influence score is the measurement of the joint effect of the direct and potential impact of a post. However, things can be more complicated if we consider the whole process of information diffusion.

Suppose the influencer makes a post at time 0, then we get the data, including the number of likes, comments, and followers at time t. First, the number of likes and comments are good indicators for the direct impact of the post (Bakshy, Eytan, et al, 2011), but there often exist a larger group

of "invisible audience" that get the information but leave no likes or comments (Bernstein, Michael S., et al, 2013). Second, after the post is made, the number of likes and comments will grow as more followers will see the post. If *t* is large enough, the number of likes and comments will be stable (Lu, et al, 2014), but if *t* is small, we are more likely to underestimate the direct impact of the post. Third, the number of followers can also change.

Each SNS website has a slightly different information diffusion pattern, which is affected by the function of the website. For example, the influencing posts on Twitter are usually spread by the user's repost, and we can study the information diffusion process by tracking the repost of a post (Bakshy, Eytan, et al, 2011). However, the Instagram has no repost function, and the only indicators of information diffusion are the number of likes, comments, and the follower network. It also has a "flatter" follower network compared with Twitter (Goel, Sharad, et al, 2012), which means the first one or two "steps" of information diffusion (an influencer's followers and their followers) contribute most to the overall impact.

In this research, we use the following influencing score calculation to proximate this complex process. The influencing score of a post, p, by the user, u, is defined as:

$$Infl(p,u) = N_l(p)C_e(u)^{\frac{N_c(p)}{C_d(u)}}$$

The $N_l(p)$ and $N_c(p)$ are the number of likes and comments of the post. The $C_e(u)$ and $C_d(u)$ are the eigenvector centrality and degree centrality of the user. Intuitively, the eigenvector centrality indicates the potential of the influencer, but each post has different response from the same group of followers, which we model with $N_c(p)/C_d(u)$. Generally, given the same group of followers, the more comments a post has, it is more likely to have great impact; if the proportion of followers

that response to the post is low, the impact could be low even with great number of followers. After calculating the influencer score for each post, we map the influencer points as shown in the following map, in which the dark color means higher influencing value.

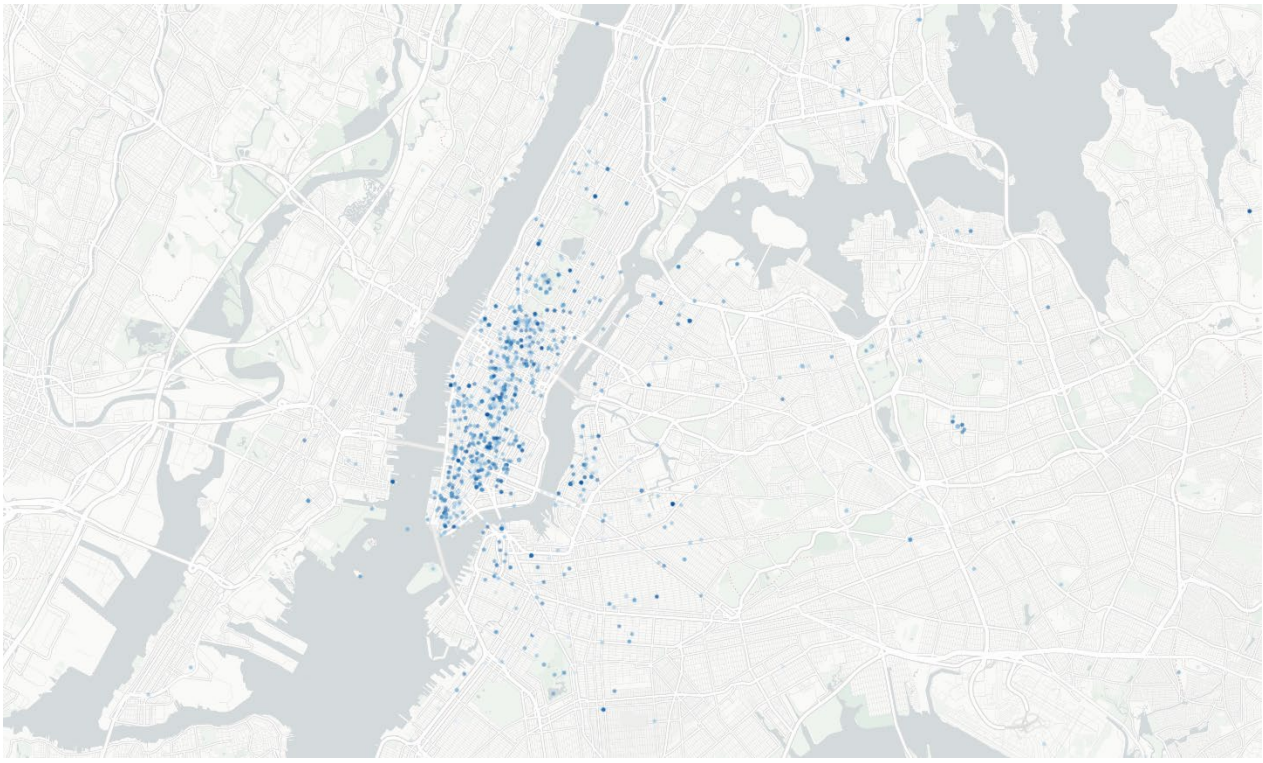

Figure 5-4: The map of influencer posts in New York City from 2014/1/1 to 2018/12/20. The color of the dots shows the score of the posts that are calculated by $Infl(p,u) = N_l(p)C_e(u)^{\frac{N_c(p)}{C_d(u)}}$, where p is the post and u is the user.

5.3.3 Spatial Interpolation and influencing Value

The next step of our study is to estimate influencer's impact on the properties in Compstak dataset. Some of the posts are direct influencer post for some properties in Compstak dataset, but most posts are for other properties nearby. We think an influencer post has not only direct impact on its targeted property, but also indirect impact on its nearby properties. The influencing value of each

property in Compstak dataset should be the overlapping impact of all its nearby influencer posts. Therefore, we need to predict the influencing value at unmeasured locations (properties in Compstak dataset) using measured locations (Influencer posts).

We use spatial interpolation to predict the value. The process can be defined as follows: given the N values of a studied phenomenon $z_j, j = 1, \dots, N$, at discrete points p_j within a certain study region, find a d-variate function $F(r)$ which fulfills the condition that $F(r_j) = z_j$. There exist infinite number of $F(r)$ and corresponding interpolation method (Mitas, Lubos, et al, 1999), including local neighborhood, smoothness, and spatial statistical approaches.

In this study, we use the inverse distance weighted interpolation (IDW), which is one of the most frequently used interpolation methods (Lu, George Y et al, 2008). The general idea of inverse distance weighted interpolation (IDW) is that things that are close to one another are more alike than those that are farther apart. Therefore, we can predict the studied value of any unmeasured location using the weighted average of "nearby" values, which can be defined as:

$$F(r) = \sum_{i=1}^{m} w_i z(r_i) = \frac{\sum_{i=1}^{m} \frac{z(r_i)}{|r - r_i|^p}}{\sum_{i=1}^{m} \frac{1}{|r - r_i|^p}}$$

Where m is the number of nearest neighbors and p is the power at which the weight decreases with the distance. When $p = 0$, there is no decrease with distance, and the prediction will be the mean of all the data values in the search neighborhood. When p is high, the weight will decrease rapidly with distance, which means only the surrounding points will influence the prediction.

To find the values of m, p that yield best prediction, we use the optimization process as follows: First, split the measured locations (influencer posts) set, P , into two groups: the training set P_{train} (80% of the samples in P) and P_{test} (20% of the samples in P), and we take the points in training set as "measured" and test set as "unmeasured". Second, we set the initial value of m, p and run the IDW and "predict" the value of test set points using training set values. Third, we compare the predicted value and the true value of test set points and calculate the root mean squared error (RMSE), update the value of m and p using gradient descent method. Finally, we iterate these steps until converge. The following maps show the optimization process with different value of m (Nearest Neighbor, NN) and p (Inverse Distance Power, IDP) using 2014-2018 influencer posts.

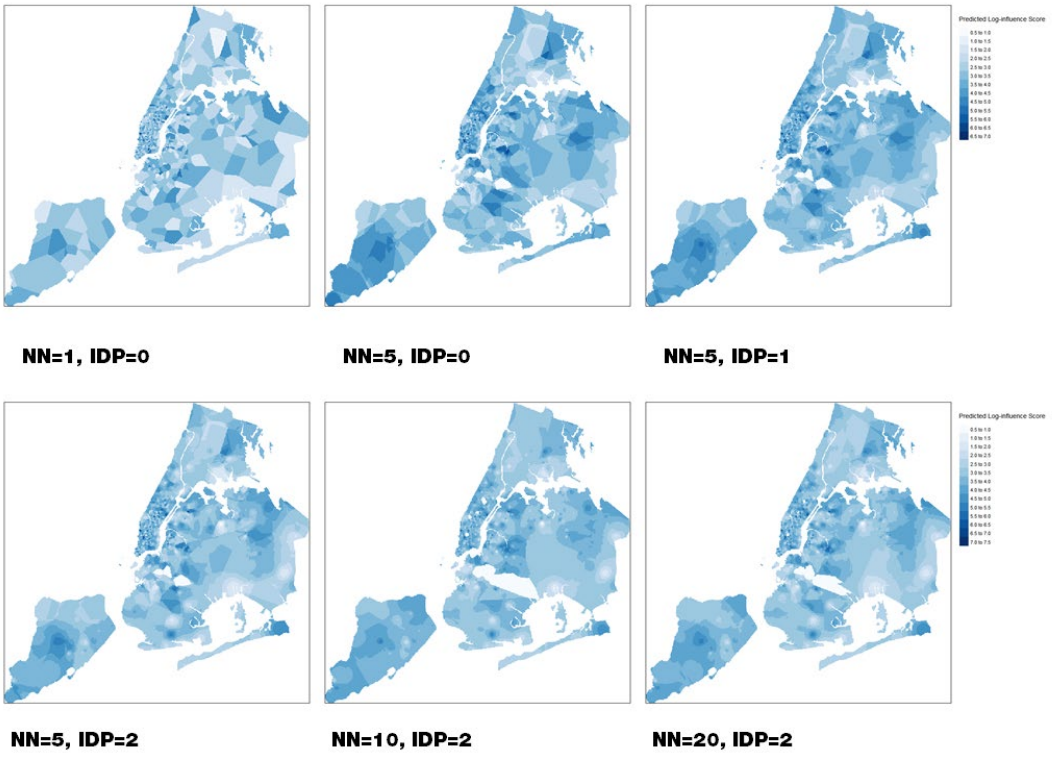

Figure 5-5: Six snapshots of the optimization process. Different number of nearest points (NN, m) and power of the inverse distance (IDP, p) lead to different interpolation results.

When m = 1, p = 0, the prediction at a location is solely dependent on its nearest measured location. In this case, we divide the study area with Voronoi polygons. As the value of m increases, we take more points into consideration and the distribution pattern becomes more complex. Considering that many influencer posts are concentrated in Manhattan, the p value is less likely to be 0, which means an influencer far away has the same impact with another influencer nearby. But p also cannot be too large, which might omit the indirect impact of nearby posts.

We tried this optimization method on each year's posts and mapped the result of the log influencer value as follows:

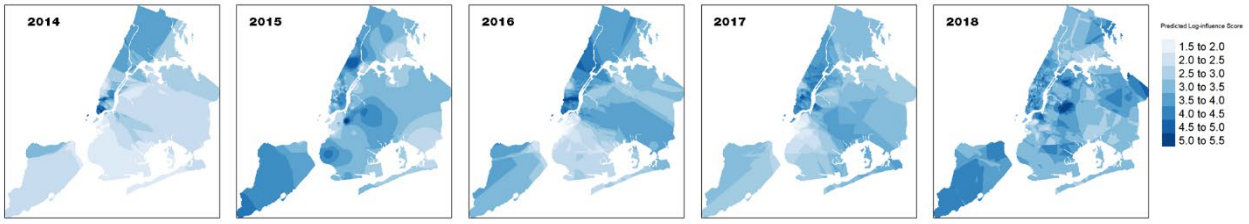

Figure 5-6: Interpolated influencer value surface using yearly data from 2014-2018

Using the calculated surface, we can predict the influencing value for each point in our study area. The next step is to predict the influencing value at locations of retail properties in Compstak dataset. For each retail rent lease in Compstak dataset, we filter out influencer posts of the previous year and apply the interpolation on selected posts, then predict the influencing value at the location of the rent lease.

5.3 Variable Preparation and Exploratory Analysis

5.3.1 Variable Preparation

To evaluate the impact of influencers on retail rent, we need to include explanatory variables other than the influencing value into our model to avoid biased result from omitted variables. The Compstak dataset provides detailed information on rent leases.

The following table is the final set of candidate features used for modeling, including the influencing value, features related to transaction detail (lease term, sub-lease, free rent, lease type, transaction size), tenant industry, building information (building age, renovation time, building type and class, etc.)

Table 5-1: Selected variables including the influencing value, features related to transaction detail (lease term, sub-lease, free rent, lease type, transaction size), tenant industry, building information (building age, renovation time, building type and class, etc.)

	Variable	Format	Description
Influencer			
Behavior	inff	float	influencing score calculated by IDW
Transaction			
Detail	Sublease	binary	if the transaction is a sublease
	Free_rent	float	free rent period in years
	bump_rate	float	rent bump rate of the lease
	bump_year	int	rent bump year

Lease_term		rent lease term in years
yr1	binary	1 year or less lease term
yr5	binary	1 to 5-year lease term
yr10	binary	5 to 10-year lease term
yr15	binary	10 to 15-year lease term
yr20	binary	15 to 20-year lease term
yrM20	binary	More than 20-year lease term
Lease Type		a dummy variable for each lease type (including unknown)
type.Full_Service	binary	all-inclusive rent
type.Gross	binary	all-inclusive rent gross lease
type.Modified_Gross	binary	modified gross lease with negotiable nets
type.Net	binary	Single Net Lease, base rent plus a pro-rata share of the building's property tax
type.Net_of_Electric	binary	Single Net Lease, base rent plus electricity fee
type.NN	binary	Double Net Lease, base rent plus a pro-rata share of property taxes and property insurance
type.NNN	binary	Triple Net Lease, property taxes, insurance, and CAMS--on top of a monthly base rent
type.Other	binary	other lease options
Transaction Size		dummy variable for transaction scale in >500, 500-1000, 1000-2000,2000-5000,and >5000 sqft

	area0	binary	under 500sqft
	area500	binary	500sqft-1000sqft
	area1000	binary	1000sqft-2000sqft
	area2000	binary	2000sqft-5000sqft
	area5000	binary	More than 5000sqft
Tenant			a dummy variable for each tenant industry
Information	**Tenant Industry**		(including unknown)
	ind.Apparel	binary	
	ind.Banks	binary	
	ind.Capital_goods	binary	
	ind.prof_service	binary	
	ind.Consumer_Durables	binary	
	ind.Education	binary	
	ind.Financial_Services	binary	
	ind.Leisure&Restaurants	binary	
	ind.Healthcare	binary	
	ind.Food&Beverage	binary	
	ind.Automobile&Components	binary	
	ind.Warehousing	binary	
	ind.Energy	binary	
	ind.Leisure & Restaurants	binary	
	ind.Retail	binary	
	ind.Non-Profit	binary	
	ind.Media	binary	
	ind.Telecommunication	binary	
	ind.Public	binary	
	ind.Legal	binary	

	ind.Other	binary	
Building Information	**Building Age**		dummy variable for building age in <10, 25, 50,75, 100 years
	age10	binary	less than 10 years
	age25	binary	10 yers-25 years
	age50	binary	25 years-50 years
	age75	binary	50 years-75 years
	age100	binary	75 years-100 years
	ageplus100	binary	more than 100 years
	unknown	binary	unknown building age
	Building Rennovation Time		dummy variable for building renovation time
	renov_5yr	binary	renovated after 2013
	renov_10yr	binary	renovated between 2008-2013
	renov_15yr	binary	renovated between 2003-2008
	renov_plus15yr	binary	renovated earlier than 2003
	unknown or never renovated	binary	unknown renovation time
	Floor Occupied		the floor of the lease
	floor.basement	binary	the lease is on the basement
	floor.ground	binary	ground floor
	floor.lower_level	binary	floor 2-5
	floor.high	binary	more than 5
	floor.multiple	binary	Occupy multiple floors
	Building Class		The class of the building
	ClassA	binary	

	ClassB	binary	
	ClassC	binary	
	Unknown	binary	
	Property Type		The type of the building where the retail space is in
	ptype.Hotel	binary	The retail space is in a hotel building
	ptype.Industrial	binary	The retail space is in an industrial building
	ptype.Mixed_Use	binary	The retail space is in a mixed-use building
	ptype.Multi_Family	binary	The retail space is in a multi-family building
	ptype.Office	binary	The retail space is in an office building
	ptype.Retail	binary	The retail space is in a retail building
	ptype.Other	binary	Other or unknown building types
Time Fixed Variables		binary	The time fixed variables for each quarter from 2014 to 2018
Location Fixed Variables	Asset turnover rate	float	The asset turnover rate of the retail property in each submarket in NYC

5.3.2 Exploratory Analysis

It is not feasible to include all related variables into the model. Therefore, we need exploratory analysis to explore distributions of the independent explanatory variables, and their correlation with the dependent variable.

Influencing Value

Using spatial interpolation, we calculated the influencing value surface for each year from 2014 to 2018 and matched the influencing value to retail rent leases by their location and transaction time. Considering the limited number of pre-2018 influencer posts we can get, we made a simplified assumption that a rent lease is affected by the influencer behavior in the same year that is calculated using 2-dimensional IDW. The ideal method is using spatial-temporal interpolation to predict the influencing value.

Another interesting finding is that from 2014 to 2018, the log-rent of retail properties in Compstak dataset is positively correlated with the influencing value at the location. The mean influencing value increased from 2015 to 2018, which are shown in the following table 5-2. Also, the magnitude of influencing value increases from 2014 to 2017, which roughly corresponds to the development of influencer marketing in recent years[3].

Table 5-2: Descriptive statistics of influencing value from 2014 to 2018.

	2014	**2015**	**2016**	**2017**	**2018**
Mean	2.75	0.16	0.23	0.29	0.35
Standard Deviation	1.49	0.12	0.16	0.45	0.36
# Observations	126	514	1101	1179	903
# Influencer Posts	71	145	256	399	2735

[3] Due to Instagram's recent "depreciation" policy that the data of old postings will be not available to public API, this observation might be biased.

The figure 5-7 shows the distribution of log influencing value and its correlation with the log effective rent value from 2014 to 2018.

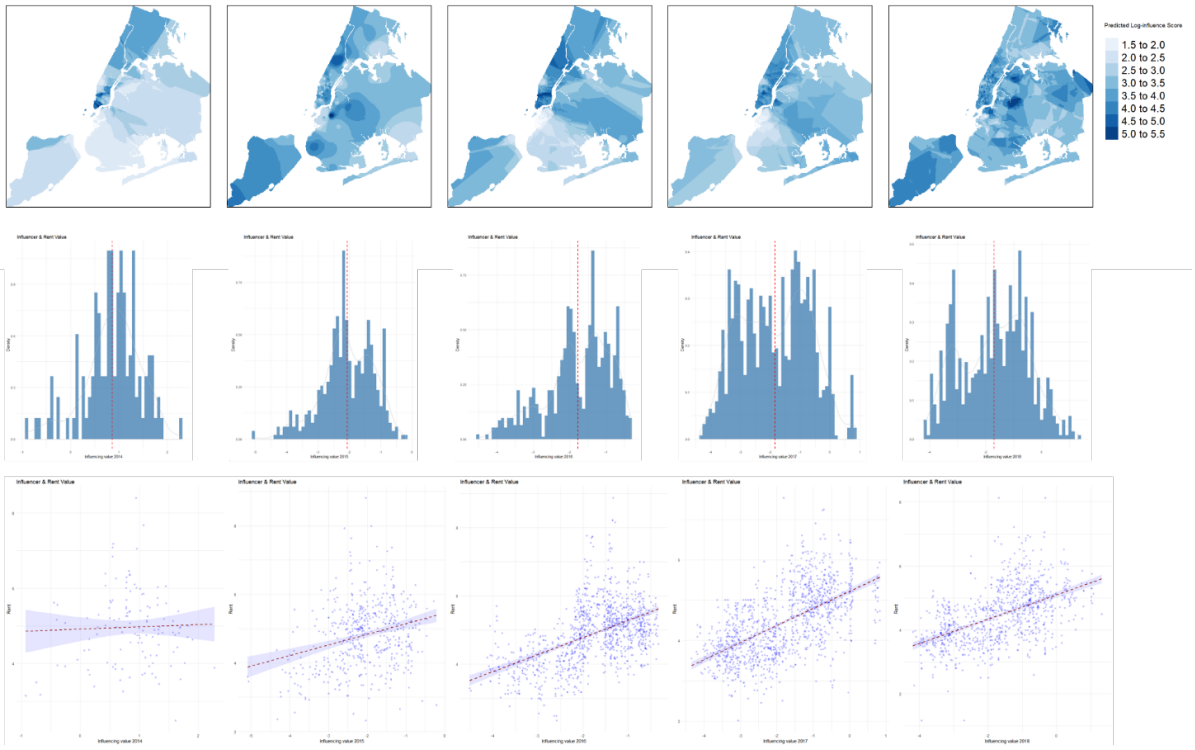

Figure 5-7: Interpolated influencer value surface and descriptive statistics using yearly data from 2014-2018 The mean influencing value increased from 2015 to 2018, which are shown in the following table 5-2. Also, the magnitude of influencing value increases from 2014 to 2017, which roughly corresponds to the development of influencer marketing in recent years

Asset Turnover

The asset turnover measures the efficiency of a company's assets to generate revenue (Fairfield, Patricia M, et al., 2001). The ratio is the percentage of net sales in total assets. From the REMeter dataset, we get the average asset turnover of retail-related industries for each zip-code in New York City. We take the asset turnover as a location fixed variable. High average asset turnover means the companies in the region are more efficient in generating sales. It could be positively

correlated with log rent because regions with high asset turnover are usually more attractive to retail companies, which will increase the demand. A negative correlation is also possible where the property or land is generally cheap. The distribution of asset turnover is shown in the figure.

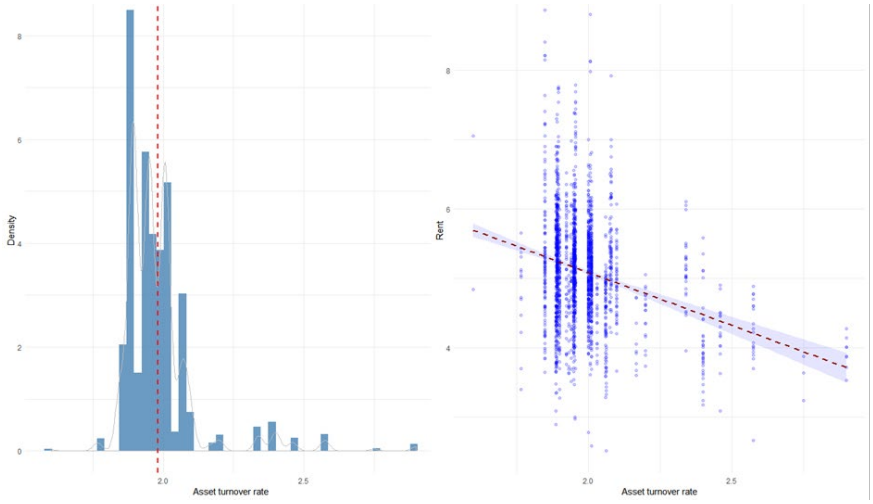

Figure 5-8: Distribution and correlation with effective rents of local asset turnover

Transaction Detail

Transaction Size: The transaction size is generally considered an important factor affecting the rental value of retail space (Brooks, Chris, et al., 2000). Intuitively, the retail space in different scale has different pricing. Therefore, we categorized the total transaction area into four groups: under 500sqft, between 500 and 1000sqft, between 1000 and 2000sqft, between 2000 and 5000sqft, and over 5000sqft. In our 3823 samples, the 1000-2000 group has most samples (32.61%), followed by 2000-5000 (26.70%), 500-100 (22.75%), over 5000 (11.16%), and below 500 (6.74%).

Rent Bump and Free Rent: In commercial real estate, the rent bump means the periodic adjustments on the rental rates. For example, if the bump rate is a %, the bump year is b, and a lease is initially \$c/sqft, the rent will increase by $c * a\%$ every year in the first b years. The free rent period is a portion of the lease term in which the tenant rent the space for free. Both methods can be positively correlated with effective rent because landlord usually use rent bump and free rent to attract tenants to accept relthe atively high price. But they could also mean concessions granted by the landlord when the retail space is in over-supply.

In our exploratory analysis, we find that the distribution of all three variables are left-skewed, and they are all positively correlated with the effective rent. The distribution and correlation are shown in figure.

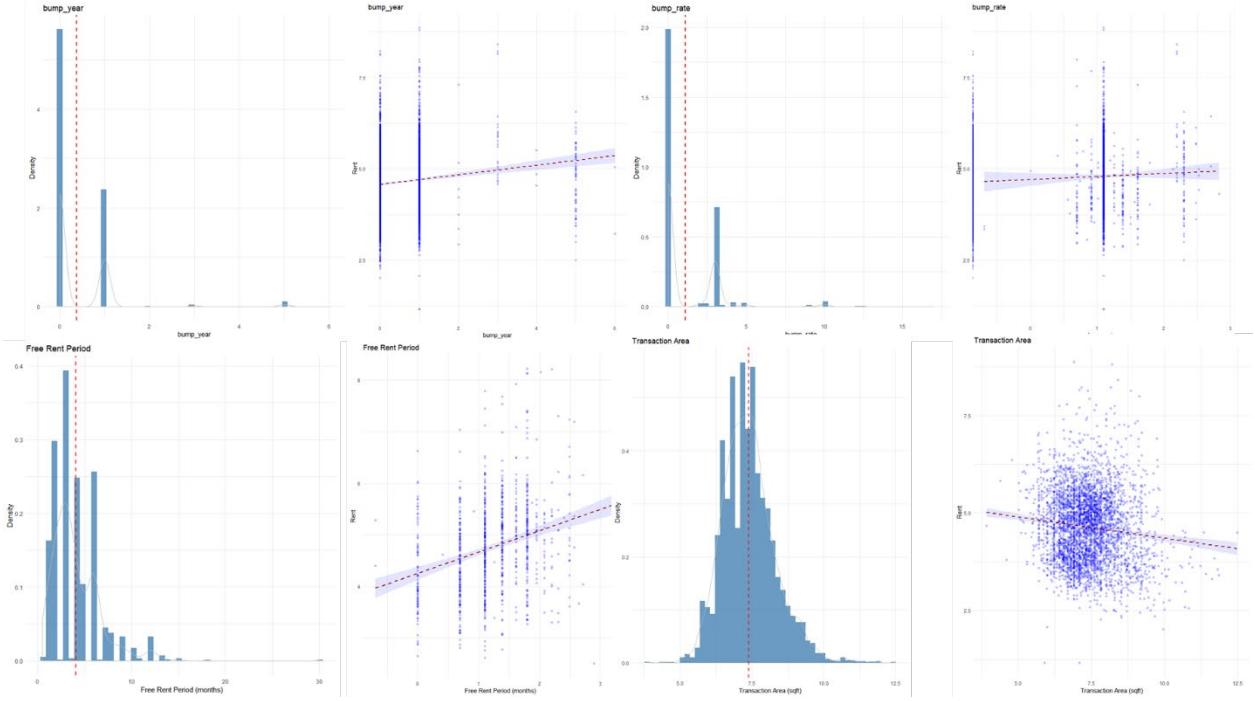

Figure 5-9: Distribution and correlation with effective rents of rent bump year, rent bump rate, free rent period, and transaction area.

Lease Term: The average lease term in 3823 observations is 9.01 years, with 3.81 standard deviations. We categorized the lease term into 5 groups: less than 1 year (1.75%), between 1 to 5 years (24.48%), between 5 to 10 years (53.36%), between 10 to 15 years (11.32%), between 15 to 20 years (1.80%), and over 20 years (0.36%). More than half of all observations are in the 5-10 years group. Although the maximum lease term is 50 year, there are only 14 observations in the "over 20 years" group.

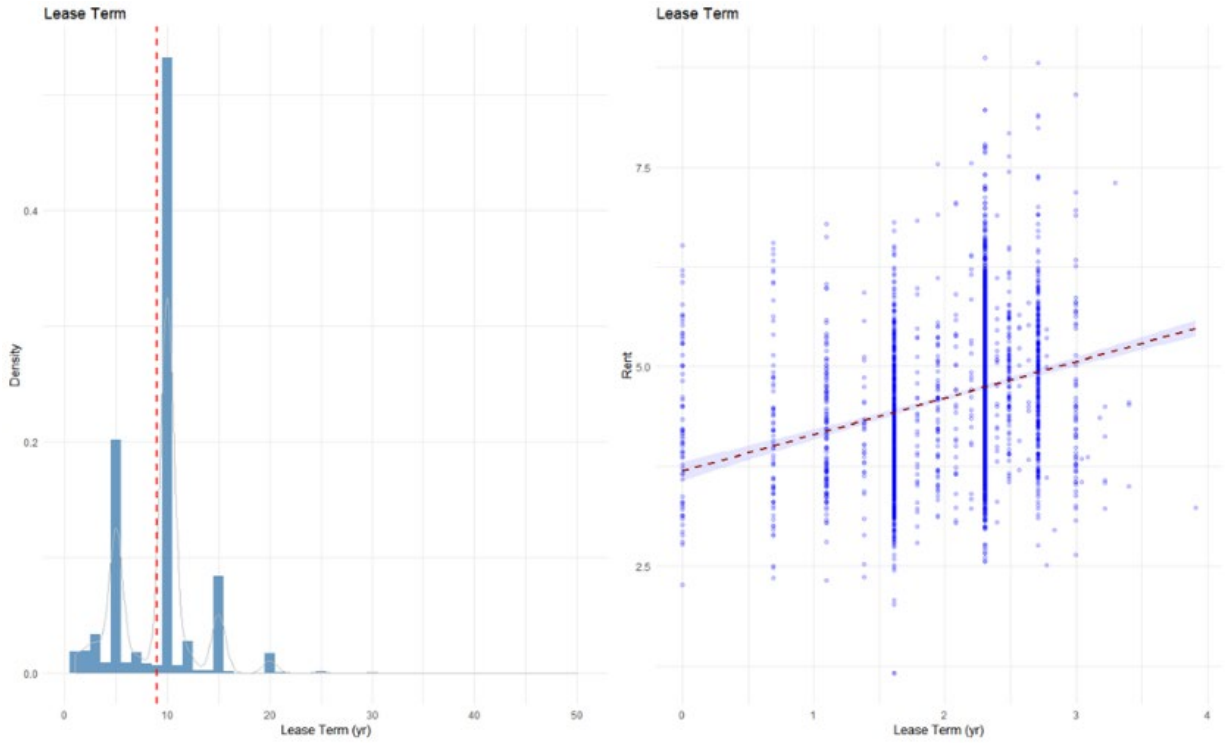

Figure 5-10: Distribution and correlation with effective rents of lease term

Building and Tenant Features

Building Time & Renovation Time: We consider the building features that might affect the rent value, including the building age, renovation time, and property type (figure). Intuitively, newly built or renovated buildings could have higher rent. But the opposite could also be true. For example, historic buildings are more expensive than average new buildings. From our exploratory analysis, we can only find that the building age and renovation year are both negatively correlated with the effective rent, which aligns with common sense that new buildings have higher rent. But the magnitude is close to zero.

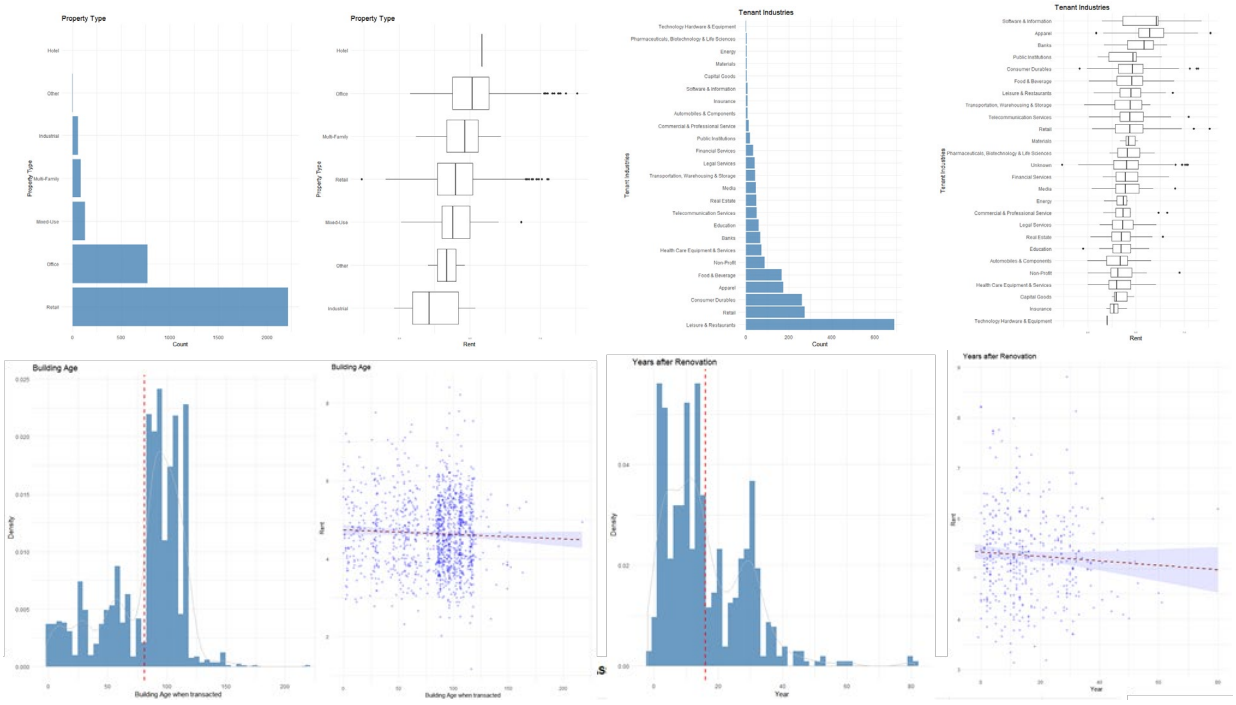

Figure 5-12: Distribution of property type and tenant industry; distribution and correlation with effective rents of building age and renovation time.

In analyzing the effect of building age, we created seven categories: from less than ten years up to 100 years at 25-year intervals, more than 100 years and unknown building age. Similarly, we used five categories for the renovation year, which can be found in the table.

Building Floors Occupied: The study categorized the building floors occupied into five groups: basement, ground floor, lower levels (between 2 and five floors), more than five floors, and multiple floors. Among 3823 samples, most are on the ground floor (81.87%), followed by multiple floors (15.02%), lower levels (2.22%), more than five floors (0.49%), and basement (0.36%).

Property Type: The property type can also affect the effective rent. Most retail spaces are located in retail buildings, others in the office, mixed use, multi-family, and industrial buildings (figure). Our exploratory analysis shows that retail spaces that are located in hotel or office buildings have higher average effective rent, while those in industrial buildings have the lowest average effective rent.

Tenant Industry: We also consider the effect of tenant industry. Retail space can accommodate various retail sub-industries (Guy, Clifford, et al. 1998), or other similar industries (Onkvisit, Sak, et al., 1981) such as restaurants, apparel, consumer durables, etc. We categorized the total of 3,823 tenants in the Compstak dataset according to their industry classification. From our exploratory analysis, we find that the average rent of industries like apparel or food & beverage is higher than traditionally defined “retail” industry (figure).

Time Fixed Effects

We use time-fixed variables for each quarter between 2014 to 2018 to grasp the overall market trend in the different period. From our explanatory analysis, we find that the average effective rent in 2018 decreases compared to 2017.

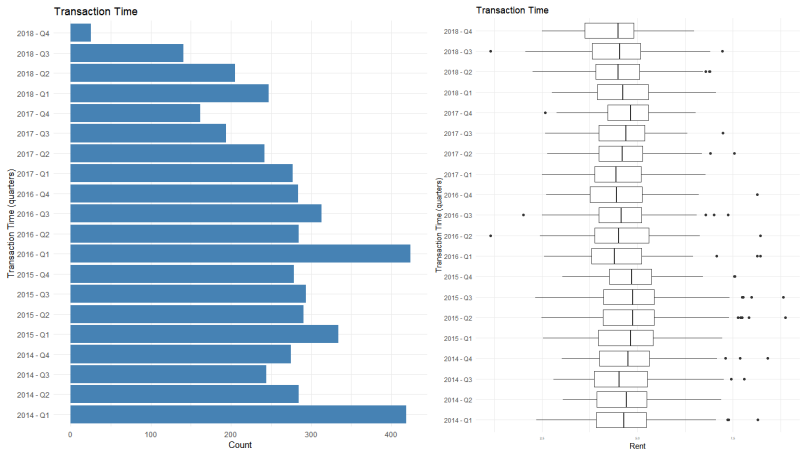

Figure 5-13: Sample count and average effective rent for each quarter from 2014 to 2018.

This finding aligns with a recent report from CBRE[4] that the average asking rents in New York fell in 2018 by a little more than 12%, following years of sky-high rents after the Great Recession that forced many businesses to halt expansion or shutter their shops.

[4] L Thomas, Rents keep dropping in New York as a new wave of retail moves in, July 17, 2018. https://www.cnbc.com/2018/07/16/rents-keep-dropping-in-new-york-as-a-new-wave-of-retail-moves-in.html

5.3 Modeling & Statistical Analysis

The effective rent of retail space could be affected by various attributes. Using the influencing value calculated with spatial interpolation as a proxy for online influencer behaviors and hedonic model framework, we seek to identify the correlation between the influencing value and effective rents paid by tenants. In this study, we employ a hedonic regression framework and use three types of model: non-spatial OLS regression model, spatial autocorrelation regression (SAR) model, and geographically weighted regression (GWR) model.

5.3.1 OLS Model

We estimate a semi-log equation relating the effective retail rent per square foot to the influencing value and hedonic features of retail space as represented by,

$$logY_i = \alpha + \beta X_i + \gamma Inf_i + \varepsilon_i \qquad (5.1)$$

In the equation above, the dependent variable is the logarithm of the effective rent per net square foot Y in retail space i; the Inf is the influencing value at the location of i, a continuous numeric variable; X is a vector of building features (building age, renovation time, building class, floor occupied, building type), tenant information (tenant industry), transaction details (rent bum ,free rent, lease term, lease type, transaction size), locational and time fixed variables (local average asset turnover, submarket, transaction time in year and quarter); α is the intercept and β, γ are coefficients of independent variables; ε is the error term.

The OLS regression requires our independent variables, and the error term fulfill several assumptions, including strict exogeneity ($E[\varepsilon|X^*] = 0, X^* = [X\ Inf]$), spherical errors ($Var[\varepsilon|X^*] = \sigma^2 I_n$), normality ($\varepsilon|X^* \sim N(0,\ \sigma^2 I_n)$) , ε independent and identically distributed, and no linear dependence in independent (iid) variables.

5.3.2 Spatial Autocorrelation and Spatial Weight Matrix

Tobler's "first law of geography," said, "Everything is related to everything else, but close things are more related than things that are far apart."(Tobler, 1979) This rule is also true in real estate. We can imagine that a landlord is very likely to increase the rent if the neighbor's rent price is high. The opposite could also happen that a landlord might decrease the rent to attract tenants and compete against neighbor landlords who have similar retail space. In both cases, the non-spatial model could leave out the effect of "proximity." Although the non-spatial OLS model includes location-fixed variables, omitting the spatial dependency of the dependent variable can still lead to biased or inconsistent results (Anselin and Bera, 1998). Additionally, the iid assumption of OLS regression will be violated if the spatial dependency or spatial autocorrelation[5] exists.

Spatial autocorrelation has some definitions that are used in different contexts (M.Sawada, Mike, 2001). The intuitive definition is that the mapped data has some organized pattern. (Upton and Fingleton, 1995) For example, the spatial autocorrelation exists if we can find clear "clusters" of high rent price. Cliff and Ord (1973) defined the spatial correlation as a quality's presence makes its presence in its neighbor more or less likely. More specifically, it is the correlation that is only caused by spatial proximity (Griffith, 2003). According to its definition, the spatial autocorrelation

[5] In this study, we use spatial dependency and spatial autocorrelation inter-changeably. Although in other applications they are slightly different. (Anselin and Bera, 1998)

can be found by mapping the OLS residuals and eyeballing the spatial distribution pattern. We can also use statistical tests. In this study, we use Moran's I test.

Moran's I

The Moran's I test is one of the most used statistics to test spatial autocorrelation. We use the global Moran's I test,

$$I = \frac{N\left(\sum_i \sum_j w_{ij}(x_i - \bar{x})\left(x_j - \bar{x}\right)\right)}{\sum_i \sum_j w_{ij} \sum_i (x_i - \bar{x})^2}$$

where N is the number of spatial units (retail space) indexed by i and j; x is the tested variable (effective rent); and w_{ij} is the item (i, j) of the spatial weight matrix, W.

Spatial Weight Matrix

The spatial weight matrix, W, is usually defined as:

$$w_{ij} = \begin{cases} 1, & if\ j\ \in N(i) \\ 0, & otherwise \end{cases}$$

$N(i)$ is the set of neighbors of location j. The neighbor set are usually defined based on adjacency of geographical zones such as towns, neighborhoods, or states. In this study, we define the "neighbor" based on the Euclidean distance because our retail rent transactions are "points". We also need to modify the values in the spatial weight matrix from binary to float. The spatial matrix we use in this study is defined as:

$$w_{ij} = \begin{cases} rd_{ij}^{-p}, & if\ j\ \in N(i)\ and\ d_{ij} < d_{threshold} \\ 0, & otherwise \end{cases}$$

Figure x shows the distance distribution of our observations. We set a threshold distance, $d_{threshold} = 1\ km$. If the distance between two retail spaces are greater than the threshold distance, we set the spatial weight value to 0, which means they are not proximate enough to affect each other. If the distance is less than the threshold distance, the weight value is proportional to d_{ij}^{-p}, where p is the Inverse Distance Power (IDP). We add a constant r to avoid weight value explode for very close points.

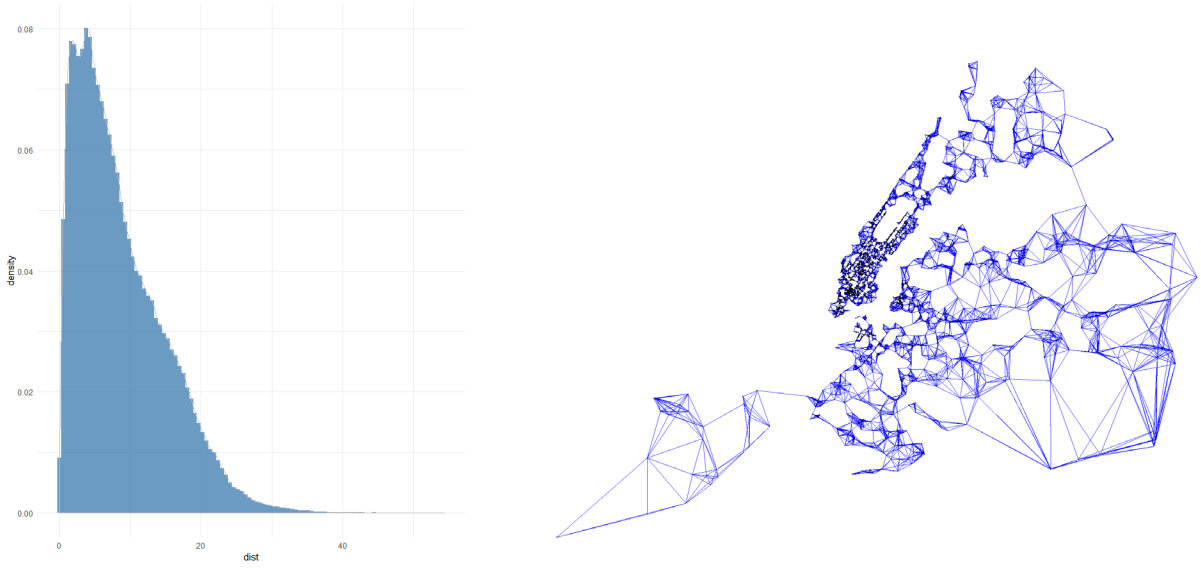

Figure 5-14: Left: distribution of distance between each pair of retail transactions in our dataset from 2014 to 2018; right: the neighboring diagram using nearest points.

Another way to define the spatial weight matrix is using nearest points. The $N(i)$ is defined as N nearest points of location i. The spatial weight value w_{ij} is set to 1 when j is one of the nearest neighbors of i. This approach does not require too much computational resources in calculating

the spatial weight matrix and estimation of the SAR model. However, for some outlier points, their nearest neighbors include long-distance points (Figure).

5.3.3 SAR Model

To quantify the spatial autocorrelation in effective rent in the dependent variable, we employ the spatial autoregression model (SAR):

$$Y = \rho WY + X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$$

where the parameter ρ quantifies the spatial dependency (or more intuitively, spillover effect) in Y. If $\rho = 0$, there is no spatial dependency in Y, which is a vector of cross-sectional observations; β is the vector of the explanatory variable's coefficient. The SAR model can be written as:

$$Y = (\mathrm{I_n} - \rho \mathrm{W})^{-1} X\beta + (\mathrm{I_n} - \rho \mathrm{W})^{-1} \varepsilon$$

The maximum likelihood estimator of β is:

$$\hat{\beta} = (X'X)^{-1}X'Y - \hat{\rho}(X'X)^{-1}X'WY$$

$$= (X'X)^{-1}X'Y - \hat{\rho}(X'X)^{-1}X'WY$$

$$= \hat{\beta}_0 - \hat{\rho}(X'X)^{-1}X'WY$$

where $\hat{\beta}_0$ is the maximum likelihood estimator of the OLS model. The coefficient $\hat{\beta}$ can be interpreted as the sum of direct and indirect impact.

We use the SAR framework to modify our OLS model specification, a semi-log equation relating the effective retail rent per square foot to the influencing value and hedonic features of retail space as represented by,

$$logY_i = \boldsymbol{\rho W log Y_i} + \alpha + \beta X_i + \gamma Inf_i + \varepsilon_i \qquad (5.2)$$

The equation 5.2 can be written as:

$$logY_i = (I_n - \rho W)^{-1}(\alpha + \beta X_i + \gamma Inf_i) + (I_n - \rho W)^{-1}\varepsilon_i \qquad (5.3)$$

where ρ is the parameter of spatial spillover effect; W is the spatial weight matrix we calculated in 5.3.2; and ε is the iid and standard Gaussian distributed error term

5.3.4 Spatial Heterogeneity and GWR Model

With the SAR model, we can quantify the spillover effect in rent prices and specify the direct and indirect effect of explanatory variables. However, both the OLS and SAR model we used thus far are based on a key assumption that the parameters remain constant over the study area, which means there is no local variation in the parameter value. For example, an influencer post in Manhattan has the same effect on its nearby retail spaces with another post in Brooklyn or suburb Queens. Although we use location fixed variables to grasp different rent price level in different submarket, many studies suggest that the model performance and explanatory power can be improved using more complicated methods such GWR/MWR (Marco Helbich et al. 2015). As we are interested in accounting for potential spatial heterogeneity in parameters, especially the coefficient of influencer

behaviors, we use a geographically weighted regression (GWR) to investigate the spatial heterogeneity.

The GWR model permits the parameters to be estimated locally (Fotheringham et al. 2003). In GWR, the linear model is rewritten in "local" form:

$$Y_i = X\beta_i + \varepsilon \quad (5.3)$$

where i is the location where the local parameters β_i are estimated using local observations Y_i, using a weighted scheme:

$$\hat{\beta}_i = (X'W_iX)^{-1}X'W_iY$$

The X observations are weighted by kernel function, which can take many forms. The kernel function takes the distance between observations and the study point as input and assigns greater weight to closer observations. The two mostly used kernel functions are the Gaussian and bi-square kernel (Bidanset, Paul E, et al. 2014).

In this study, we use the Gaussian kernel function, and the weight matrix is defined in a Gaussian form:

$$w_{ij} = e^{-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2}$$

where d_{ij} is the Euclidean distance between the location of observation i and j; and h is the "bandwidth" of the local model. For the Gaussian kernel function, the curve of weight decay identical to the Gaussian curve. If the distance is greater than the bandwidth, the weight will rapidly decrease and close to 0 when the distance increases to $2 * h$. For the bi-square kernel function, the weight of observations out of the bandwidth is equal to 0.

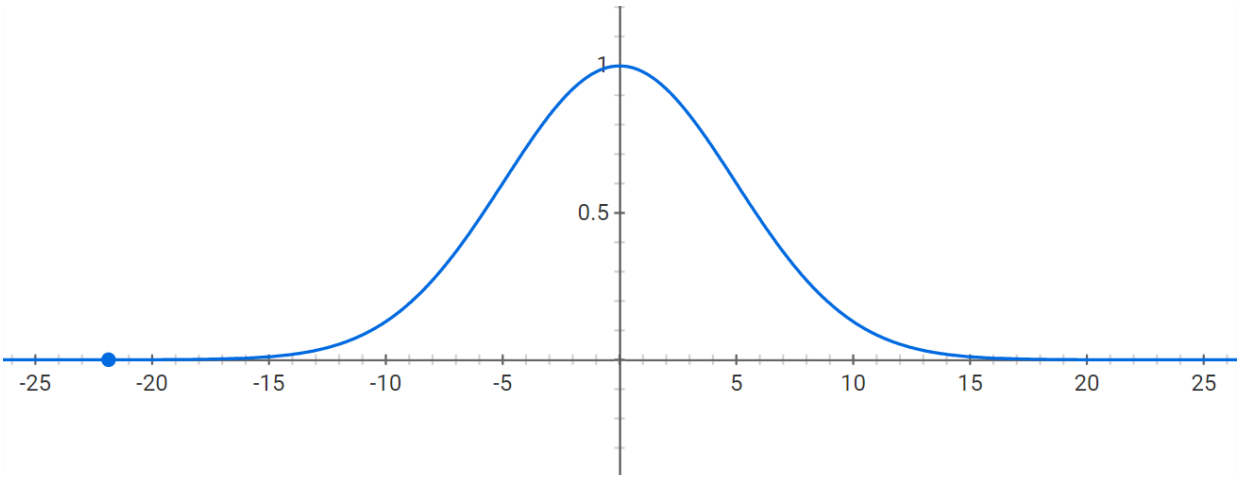

Figure 5-15: Gaussian kernel function using bandwidth = 7. If the distance is greater than the bandwidth, the weight will rapidly decrease and close to 0 when the distance increases to 2*h.

Intuitively, the GWR model runs a multivariate regression model at each location using the weighted observations within the bandwidth. The bandwidth can be manually defined, but we use an optimization process to minimize the MRSE of equation 5.3 (Fotheringham et al. 2002).

We employ a simplified model specification in the GWR part:

$$logY_i = \alpha + \beta X_i^* + \gamma Inf_i + \varepsilon_i \qquad (5.2)$$

where X^* only includes the variables that do not have local multicollinearity.

The problem of local multicollinearity is addressed by a number of studies as the main limitation of GWR/MWR model (Wheeler and Tiefelsdorf, 2005; Griffith et al. 2008; Páez et al. 2011). Since GWR/MWR selects a subset of all observations and uses weighted values of these observations as input, the multicollinearity can be introduced to the local input in this process. Suppose we are looking at the location i, and estimate a local model using observations within the bandwidth. It is possible that all these observations have the same value in multiple variables. For example, if i is in a newly built downtown commercial district, all observations have 0 value for the dummy

variable “property type: industrial,” meanwhile all values for the dummy variable “building age: over 100 years” are also 0. Although there is no global multicollinearity, the model cannot be estimated locally.

There are several ways to select the explanatory variables for GWR meanwhile avoid the local multicollinearity, including using expert opinion, stepwise variable selection, selecting from alternatives based on the Akaike information criterion (AIC) value(Lu, B, Charlton, et al. 2014). In this study, we only select the influencing value, asset turnover, lease term, transaction size, lease type, building class (whether it’s class A), and building type (whether it’s mixed-use) for GWR.

Chapter 6: Findings and Discussion

6.1 Initial Findings

6.1.1 OLS Regression Results

The table 6-1 shows the estimation results of the OLS model.

Table 6-1 OLS Regression Result

Variables	ESTIMATE	STD.ERROR	STATISTIC	P.VALUE
Intercept	-1.33836	0.106708	-12.5423	2.22E-35
Influencer Behavior				
Log Influencing Value	0.399865	0.014291	27.98051	1.17E-156
Lease Term (Base Case: Over 20-year lease term)				
1 year or less lease term	-0.14902	0.100188	-1.48738	0.136999
1 to 5-year lease term	-0.22107	0.051086	-4.3274	1.55E-05
5 to 10-year lease term	0.238311	0.047167	5.052474	4.57E-07
10 to 15-year lease term	0.337947	0.057101	5.918447	3.54E-09
15 to 20-year lease term	0.363655	0.100229	3.628237	0.000289
Rent Bump & Free Rent				
free rent period in years	0.025742	0.012658	2.033727	0.042049
rent bump rate	0.10184	0.037207	2.737153	0.006227
rent bump year	-0.05977	0.046965	-1.27266	0.203219
Lease Type (Base Case: Full Service Lease)				

Gross lease	-0.10239	0.028025	-3.65362	0.000262
Single Net Lease	-0.23483	0.048447	-4.84708	1.30E-06
Double Net Lease	-0.1994	0.141385	-1.41034	0.158523
Triple Net Lease	-0.19853	0.067983	-2.92026	0.003518
Transaction Size (Base Case: Over 5000sqft)				
under 500sqft	0.833912	0.0613	13.60377	3.54E-41
500sqft-1000sqft	0.651948	0.047333	13.77376	3.81E-42
1000sqft-2000sqft	0.474555	0.044062	10.77019	1.16E-26
2000sqft-5000sqft	0.375832	0.04372	8.596399	1.19E-17
Tenant Industry (Base Case: Other)				
Apparel	1.063606	0.059596	17.84705	1.89E-68
Banks	0.581653	0.092071	6.317443	2.97E-10
Capital Goods	0.252202	0.328235	0.768359	0.442322
Commercial & Professional Service	-0.05387	0.195823	-0.27509	0.783259
Consumer Durables	0.202743	0.048831	4.151966	3.37E-05
Education	-0.04138	0.097093	-0.42621	0.669977
Finance	0.199288	0.12695	1.569816	0.116542
Food & Beverage	0.022609	0.059477	0.380131	0.70387
Health Care Equipment & Service	-0.05032	0.088265	-0.57007	0.568664
Insurance	-0.23463	0.276574	-0.84833	0.396308
Legal Services	-0.16373	0.115576	-1.4166	0.156683
Leisure & Restaurant	0.047341	0.033603	1.408836	0.158966
Media	-0.15043	0.10782	-1.39519	0.163041
Non-profit Organization	-0.14419	0.080244	-1.7969	0.072431
Pharmaceutical, Biotech & Life Sciences	-0.12406	0.364788	-0.34009	0.733809
Public Institutions	0.103132	0.172632	0.597409	0.550271
Real Estate	0.075679	0.106791	0.708658	0.478581
Retail	0.111507	0.048561	2.29621	0.021719
Software & Information	0.832755	0.277058	3.005709	0.002667
Hardware & Equipment	-0.62014	0.728241	-0.85156	0.39451
Telecommunication	0.401808	0.10493	3.82928	0.000131
Building Age (Base Case: Over 100 years)				

75 years-100 years	-0.0152	0.038385	-0.396	0.692127
50 years-75 years	-0.04214	0.036009	-1.17031	0.24195
25 years-50 years	-0.01331	0.059915	-0.22223	0.824148
10 yers-25 years	0.007578	0.066522	0.11392	0.909307
less than 10 years	0.009047	0.327315	0.027639	0.977952
Building Renovation Time (Base Case: No renovation)				
renovated after 2013	0.272585	0.077973	3.495868	0.000478
renovated between 2008-2013	0.202851	0.097398	2.082714	0.037345
renovated between 2003-2008	0.249342	0.070577	3.532929	0.000416
renovated earlier than 2003	0.27104	0.071343	3.799111	0.000148
Floor Occupied (Base Case: Multiple or unknown)				
basement	-0.00093	0.034772	-0.0268	0.978623
ground floor	-0.66163	0.198847	-3.32731	0.000885
floor 2-5	-0.45661	0.085935	-5.31341	1.14E-07
more than 5	-0.88142	0.173273	-5.0869	3.82E-07
Building Class (Base Case: Class C or unknown)				
Class A	0.26909	0.066501	4.046397	5.31E-05
Class B	-0.10255	0.051514	-1.9908	0.046575
Property Type (Base Case: Other Types)				
Hotel	0.164435	0.732365	0.224526	0.82236
Industrial	-0.60172	0.102987	-5.84275	5.57E-09
Mixed-use	-0.23076	0.073116	-3.15604	0.001612
Multi-family	-0.07974	0.086565	-0.9212	0.357008
Office	0.208022	0.050693	4.103528	4.16E-05
Retail	-0.03539	0.035259	-1.00384	0.31552
Time Fixed Effect (Base Case: 2014)				
T.2015	0.990514	0.081495	12.15425	2.28E-33
T.2016	0.895468	0.076518	11.70273	4.25E-31
T.2017	0.665866	0.076709	8.680416	5.78E-18
T.2018	0.648442	0.076545	8.471413	3.43E-17

Location Fixed Effect				
Asset Turnover	-0.05113	0.012215	-4.18588	2.91E-05
Multiple R-squared	**0.4818**			
Adjusted R-squared	**0.4729**			
F-statistic	**54.6 ON 64 AND 3758 DF**			**< 2.2E-16**

In the estimation result of the OLS model, the coefficient of the influencing value is significant, which means the influencer behavior has a significant impact on the effective rents. If the influencing value increases by 1%, we'd expect effective rents to increase by 0.3998%.

Our model relates effective rents to location-fixed effect in the form of local average asset turnover. In theory, companies with relatively high sales and low asset costs tend to have high asset turnover. If the local asset turnover is high, it could be due to low rent price. Therefore, it might be negatively correlated with effective rents. In our estimation results, if the local average asset turnover changes by 1%, we'd expect effective rents to change by -0.0511%.

The transaction size also affects effective rents. Our estimation result shows that smaller sized retail spaces tend to have higher effective rents (per square feet). The retail space under 500sqft, 500-1000 sqft, 1000-2000sqft, and 2000-5000sqft have 83.39%, 65.19%, 47.45%, and 37.58% higher effective rents per square feet compared to the base case of over 5000sqft.

There are other features related to transaction details. For lease term, we find that the leases with less than 5-year term tend to have lower effective rents than the base case (over 20-year lease term). The leases with 5 to 10 year, 10 to 15-year, and 15 to 20-year term have 23.83%, 33.79%, and

36.36% higher effective rents than the base case (over 20-year lease term). For the lease type, we find that the full-service lease (base case) has the highest effective rents per square feet, and other types, the gross, single net, double net, and triple net lease, have 10.23%, 23.48%,19.94%, 19.58% lower effective rents. This result aligns with common sense that in a full-service lease, the landlord pays for all operating expenses such as maintenance, utilities, insurances, and taxes, so the landlord is more likely to increase the base rent to compensate these costs. For the tenant industry, we find that the apparel has the highest effective rent (106.36% higher than the base case), and the hardware & equipment has the lowest effective rent (62.01% lower than the base case).

For the hedonic features, we find that the effect of building age is not significant, while the renovation time has a significant effect on effective rents. Compared to the base case of no renovation, we'd expect 27.25%, 20.28%, 24.93%, and 27.10% increase in effective rents for properties renovated after 2013, between 2008-2013, between 2003-2008, and earlier than 2003. For the floor occupied, ground floor, basement, lower and high floors have lower effective rents. We also find that properties with high class (class A) have 27% higher effective rents compared to the base case (class C or less). Additionally, we investigate the effect of building types to effective rents of the retail space. We find that retail spaces in office or hotel buildings have higher effective rents (20.80% and 16.44%), while in industrial or mixed-use buildings the rent will be lower by 60.17% and 23.07%.

We also include time-fixed effects in our OLS model. The estimation result shows that compared to the base case of 2014, the effective rents are 99.05%, 89.54%, 66.58%, and 64.84% higher in 2015-2018. The decreasing trend from 2015 to 2018 aligns with a recent report on New York retail rent market by CBRE, in which the researcher found that average rent prices in a dozen of 16 main

retail corridors in New York fell in the past 12 months. We think that the landlords and property managers are more likely to lower the rents to stimulate the activity when the vacancy rate is high but the landlords "are more optimistic about the future in this market (L. Thomas, 2018)."

6.1.2 Spatial Autocorrelation and SAR Regression Results

As we discussed in Chapter 2, although some studies suggest that the spatial autocorrelation can be fixed using spatial fixed effect. However, the spatial fixed effect can only successfully remove the spatial autocorrelation when it only exists in each "spatial subset" of samples, in our case, the zip-code zones. In our study area, the spatial autocorrelation pattern is less likely to exist only in zip-code zones. If there exists spatial autocorrelation that is not corrected by spatial fixed effect, the assumption of OLS model is violated. To test if spatial autocorrelation exists, we employ Moran's I test on residuals of our OLS model. The table 6-2 shows the test statistics.

Table 6-2: Moran's I Test Statistic

Moran I statistic standard deviate = 36.889, p-value < 2.2e-16

Observed Moran I	Expectation	Variance
3.484382E-01	-1.731134e-03	9.010565e-05

The Moran's I value is significantly positive, which means there exists a significant positive spatial autocorrelation in OLS residuals[6]. We include spatial autoregressive terms and run the SAR model and try to minimize the effect of spatial autocorrelation.

We map the residuals of OLS model and SAR model and find clear spatial patterns (Figure 6.1). Compared two maps, the OLS residual map clearly has spatial patterns. For example, the model tends to underestimate the rent price in upper east and upper west New York, while in the SAR residual map the residuals are more randomly distributed.

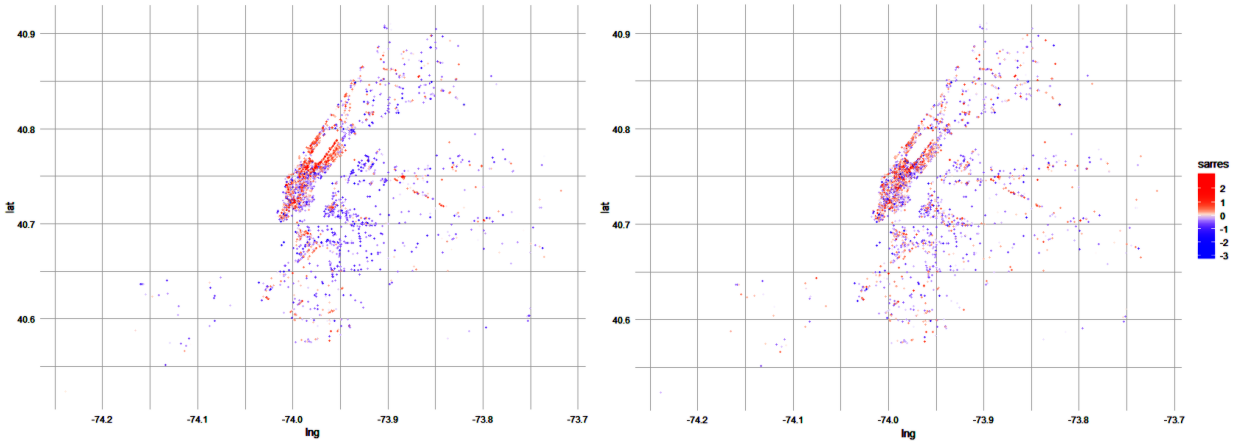

Figure 6-1: Residual maps of OLS model (left) and SAR model (right)

The table 6-2 shows the estimation results of the SAR model.

Table 6-2 SAR Regression Result

Variables	ESTIMATE	STD.ERROR	STATISTIC	P.VALUE
Intercept	-0.79125	0.083892	-9.4318	< 2.2e-16
Influencer Behavior				

[6] The significant Moran's I statistics is only one of multiple statistical tests that indicate spatial autocorrelation. A significant Moran's I only cannot fully justify the usage of SAR model.

Log Influencing Value	0.141019	0.012233	11.5274	< 2.2e-16
Lease Term (Base Case: Over 20-year lease term)				
1 year or less lease term	-0.20266	0.078023	-2.5974	0.009393
1 to 5-year lease term	-0.14237	0.039807	-3.5765	0.000348
5 to 10-year lease term	0.168151	0.036736	4.5773	4.71E-06
10 to 15-year lease term	0.243017	0.044476	5.464	4.65E-08
15 to 20-year lease term	0.266256	0.078054	3.4112	0.000647
Rent Bump & Free Rent				
free rent period in years	0.006401	0.009859	0.6493	0.516165
rent bump rate	0.100707	0.028974	3.4757	0.00051
rent bump year	-0.06185	0.036574	-1.6912	0.090799
Lease Type (Base Case: Full Service Lease)				
Gross lease	-0.01768	0.021835	-0.8095	0.418218
Single Net Lease	-0.10678	0.037747	-2.8288	0.004672
Double Net Lease	-0.00215	0.110124	-0.0195	0.984428
Triple Net Lease	-0.11903	0.052973	-2.247	0.024638
Transaction Size (Base Case: Over 5000sqft)				
under 500sqft	0.720178	0.047879	15.0417	< 2.2e-16
500sqft-1000sqft	0.549852	0.036999	14.8614	< 2.2e-16
1000sqft-2000sqft	0.405572	0.034379	11.7972	< 2.2e-16
2000sqft-5000sqft	0.294682	0.034094	8.6431	< 2.2e-16
Tenant Industry (Base Case: Other or unknown)				
Apparel	0.595696	0.046584	12.7875	< 2.2e-16
Banks	0.496989	0.071712	6.9303	4.20E-12
Capital Goods	-0.0312	0.25561	-0.1221	0.902841
Commercial & Professional Service	-0.14881	0.152496	-0.9759	0.329136
Consumer Durables	0.123858	0.038028	3.257	0.001126
Education	-0.04817	0.075611	-0.6371	0.524044
Finance	0.273692	0.098861	2.7685	0.005632
Food & Beverage	0.052992	0.046319	1.1441	0.252599

Health Care Equipment & Service	0.013326	0.068743	0.1938	0.846294
Insurance	0.019797	0.215392	0.0919	0.92677
Legal Services	-0.10997	0.090006	-1.2218	0.221772
Leisure & Restaurant	0.049786	0.026168	1.9025	0.057101
Media	-0.06492	0.083983	-0.773	0.439543
Non-profit Organization	-0.08822	0.062492	-1.4116	0.158062
Pharmaceutical, Biotech & Life Sciences	0.263376	0.284077	0.9271	0.35386
Public Institutions	-0.06558	0.134436	-0.4878	0.625669
Real Estate	0.074419	0.083165	0.8948	0.370874
Retail	0.080744	0.037818	2.1351	0.032754
Software & Information	0.587274	0.215791	2.7215	0.006499
Hardware & Equipment	-0.46776	0.567125	-0.8248	0.409492
Telecommunication	0.313014	0.081726	3.83	0.000128
Building Age (Base Case: Over 100 years)				
75 years-100 years	-0.02364	0.029892	-0.7908	0.42908
50 years-75 years	-0.03408	0.028042	-1.2154	0.224227
25 years-50 years	0.041584	0.046666	0.8911	0.372875
10 yers-25 years	-0.027	0.051809	-0.5211	0.602327
less than 10 years	0.263364	0.254895	1.0332	0.301499
Building Renovation Time (Base Case: No renovation)				
renovated after 2013	0.218098	0.060734	3.5911	0.000329
renovated between 2008-2013	0.162126	0.075848	2.1375	0.032556
renovated between 2003-2008	0.135241	0.054994	2.4592	0.013925
renovated earlier than 2003	0.114389	0.055599	2.0574	0.039649
Floor Occupied (Base Case: Multiple or unknown)				
basement	0.023012	0.027079	0.8498	0.395438
ground floor	-0.55013	0.154886	-3.5519	0.000383
floor 2-5	-0.45398	0.066949	-6.7809	1.19E-11
more than 5	-1.05285	0.134948	-7.8019	6.00E-15
Building Class (Base Case: Class C or unknown)				
Class A	0.167094	0.051878	3.2209	0.001278

Class B	-0.05006	0.040123	-1.2477	0.212137
Property Type (Base Case: Other Types)				
Hotel	-0.04401	0.570335	-0.0772	0.938498
Industrial	-0.32021	0.080342	-3.9856	6.73E-05
Mixed-use	-0.12634	0.056952	-2.2184	0.026527
Multi-family	-0.11493	0.067412	-1.7049	0.088205
Office	0.067598	0.039508	1.711	0.087083
Retail	-0.02099	0.027458	-0.7645	0.44459
Time Fixed Effect (Base Case: 2014)				
T.2015	0.364155	0.064513	5.6447	1.66E-08
T.2016	0.343107	0.060431	5.6776	1.37E-08
T.2017	0.261715	0.060275	4.342	1.41E-05
T.2018	0.251108	0.060039	4.1824	2.88E-05
Location Fixed Effect				
Asset Turnover	-0.01702	0.00955	-1.7817	0.074794
Rho	**0.6077**			**< 2.2E-16**
LR Test Value	**1541.1**			
Asymptotic standard error	**0.012217**			
Wald statistic	**2474.3**			**< 2.2E-16**
Log-likelihood	**-3396.816 FOR LAG MODEL**			
LM test for residual autocorrelation	**63.71**			**1.4433E-15**
AIC	**6927.6**			
AIC FOR LM	**8466.7**			

The spatial autoregressive parameter (0.6077) is significant, and the LR test shows the inclusion of the spatial autoregressive term does improve the model.

The interpretation of the SAR model is more complicated than the OLS model. In OLS, we have $\frac{\partial Y_i}{\partial X_{ik}} = \beta_k$ and $\frac{\partial Y_i}{\partial X_{jk}} = 0$, where Y_i, X_i is the dependent variable and independent variable vector for the ith observation. The change in X_i will only affect Y_i by the magnitude of β and there is no indirect effect. However, for SAR model, we have:

$$Y = (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}\varepsilon$$

Therefore, the impact of an independent variable is:

$$\frac{\partial Y}{\partial X} = (\mathrm{I_n} - \rho \mathrm{W})^{-1}\beta = S_r(W)$$

When $i = j$, we have the direct impact $\frac{\partial Y_i}{\partial X_{ik}} = S_r(W)_{ii}$; when $i \neq j$, we have the indirect impact $\frac{\partial Y_i}{\partial X_{jk}} = S_r(W)_{ij}$. We can use 3 metrics: average direct impact, which is the average of $S_r(W)_{ii}$, similar to the traditional interpretation; average indirect impact, which is the average impact of one observation's neighbors on its outcome; and average total impact, which is the total of direct and indirect impacts of an independent variable on the outcome of an observation.

We use Monte Carlo simulation to obtain simulated outcome of these impact values. Table 6-3 shows the direct and indirect effect of each independent variable, and table 6.4 shows the p-value of these impact values.

Table 6-3 SAR Regression Impact Analysis

Variables	DIRECT	INDIRECT	TOTAL
Influencer Behavior			
Log Influencing Value	0.155738	0.203727	0.359466

Lease Term (Base Case: Over 20-year lease term)			
1 year or less lease term	-0.22381	-0.29278	-0.51659
1 to 5-year lease term	-0.15723	-0.20568	-0.36291
5 to 10-year lease term	0.185702	0.242924	0.428625
10 to 15-year lease term	0.268382	0.351081	0.619464
15 to 20-year lease term	0.294047	0.384655	0.678702
Rent Bump & Free Rent			
free rent period in years	0.007069	0.009248	0.016317
rent bump rate	0.111219	0.14549	0.256708
rent bump year	-0.06831	-0.08936	-0.15767
Lease Type (Base Case: Full Service Lease)			
Gross lease	-0.01952	-0.02554	-0.04506
Single Net Lease	-0.11793	-0.15426	-0.27219
Double Net Lease	-0.00237	-0.00311	-0.00548
Triple Net Lease	-0.13146	-0.17196	-0.30342
Transaction Size (Base Case: Over 5000sqft)			
under 500sqft	0.795348	1.040426	1.835773
500sqft-1000sqft	0.607243	0.794359	1.401603
1000sqft-2000sqft	0.447905	0.585922	1.033826
2000sqft-5000sqft	0.32544	0.425721	0.751161
Tenant Industry (Base Case: Other or unknown)			
Apparel	0.657873	0.860589	1.518462
Banks	-0.03446	-0.04508	-0.07954
Capital Goods	-0.16435	-0.21499	-0.37933
Commercial & Professional Service	0.136786	0.178935	0.315721
Consumer Durables	-0.0532	-0.0696	-0.1228
Education	0.302259	0.395397	0.697656
Finance	0.014717	0.019251	0.033968
Food & Beverage	0.021863	0.0286	0.050463
Health Care Equipment & Service	-0.12145	-0.15887	-0.28032

Insurance	0.054983	0.071925	0.126908
Legal Services	-0.07169	-0.09378	-0.16547
Leisure & Restaurant	-0.09742	-0.12744	-0.22487
Media	0.290866	0.380494	0.67136
Non-profit Organization	-0.07243	-0.09474	-0.16717
Pharmaceutical, Biotech & Life Sciences	0.082187	0.107512	0.189699
Public Institutions	0.089172	0.11665	0.205822
Real Estate	0.648572	0.848423	1.496995
Retail	-0.51658	-0.67576	-1.19234
Software & Information	0.345686	0.452205	0.797891
Hardware & Equipment	0.657873	0.860589	1.518462
Telecommunication	0.548863	0.71799	1.266853
Building Age (Base Case: Over 100 years)			
75 years-100 years	-0.02755	-0.03616	-0.06371
50 years-75 years	-0.03673	-0.0482	-0.08493
25 years-50 years	0.045548	0.059771	0.105319
10 yers-25 years	-0.02954	-0.03877	-0.06831
less than 10 years	0.272118	0.357089	0.629207
Building Renovation Time (Base Case: No renovation)			
renovated after 2013	0.243617	0.319689	0.563306
renovated between 2008-2013	0.18372	0.241088	0.424809
renovated between 2003-2008	0.151287	0.198527	0.349814
renovated earlier than 2003	0.125779	0.165055	0.290834
Floor Occupied (Base Case: Multiple or unknown)			
basement	0.025221	0.033096	0.058317
ground floor	-0.6072	-0.7968	-1.40399
floor 2-5	-0.50236	-0.65922	-1.16158
more than 5	-1.16829	-1.53309	-2.70138
Building Class (Base Case: Class C or unknown)			
Class A	0.183581	0.240906	0.424487
Class B	-0.05498	-0.07215	-0.12713

Property Type (Base Case: Other Types)			
Hotel	-0.05055	-0.06634	-0.11689
Industrial	-0.34986	-0.45911	-0.80897
Mixed-use	-0.14183	-0.18612	-0.32795
Multi-family	-0.12638	-0.16584	-0.29222
Office	0.074538	0.097813	0.172351
Retail	-0.02338	-0.03068	-0.05406
Time Fixed Effect (Base Case: 2014)			
T.2015	0.399974	0.52487	0.924844
T.2016	0.376542	0.49412	0.870662
T.2017	0.286283	0.375677	0.66196
T.2018	0.274154	0.35976	0.633914
Location Fixed Effect			
Asset Turnover	-0.01879	-0.02458	-0.04338

Table 6.4 SAR Regression Impact P-value

Variables	**DIRECT**	**INDIRECT**	**TOTAL**
Influencer Behavior			
Log Influencing Value	0	0	0
Lease Term (Base Case: Over 20-year lease term)			
1 year or less lease term	0.008023	0.008659	0.008262
1 to 5-year lease term	0.000315	0.000399	0.000345
5 to 10-year lease term	2.91E-06	4.33E-06	3.26E-06
10 to 15-year lease term	9.09E-08	2.64E-07	1.40E-07
15 to 20-year lease term	0.000931	0.001119	0.001002
Rent Bump & Free Rent			
free rent period in years	0.514654	0.515363	0.514936

rent bump rate	0.000993	0.001246	0.001097
rent bump year	0.133131	0.135643	0.134269
Lease Type (Base Case: Full Service Lease)			
Gross lease	0.410336	0.412169	0.411229
Single Net Lease	0.00545	0.005401	0.00532
Double Net Lease	0.939554	0.940415	0.940027
Triple Net Lease	0.01793	0.018702	0.018172
Transaction Size (Base Case: Over 5000sqft)			
under 500sqft	0	0	0
500sqft-1000sqft	0	0	0
1000sqft-2000sqft	0	0	0
2000sqft-5000sqft	0	0	0
Tenant Industry (Base Case: Other or unknown)			
Apparel	1.92E-11	1.68E-10	4.35E-11
Banks	0.868199	0.866824	0.867382
Capital Goods	0.318153	0.318387	0.318058
Commercial & Professional Service	0.002321	0.002221	0.002206
Consumer Durables	0.475598	0.473443	0.474213
Education	0.00326	0.003711	0.003438
Finance	0.257568	0.258145	0.257622
Food & Beverage	0.929018	0.929051	0.92902
Health Care Equipment & Service	0.912786	0.915137	0.914099
Insurance	0.202648	0.204088	0.2032
Legal Services	0.062415	0.064695	0.063387
Leisure & Restaurant	0.420477	0.424171	0.422385
Media	0.13235	0.135695	0.133925
Non-profit Organization	0.339015	0.338754	0.338634
Pharmaceutical, Biotech & Life Sciences	0.640963	0.640663	0.640703
Public Institutions	0.384154	0.38475	0.384295
Real Estate	0.034902	0.036185	0.035391
Retail	0.007384	0.007706	0.007454
Software & Information	0.37276	0.375217	0.373951

Hardware & Equipment	9.18E-05	9.98E-05	9.06E-05
Telecommunication	1.92E-11	1.68E-10	4.35E-11
Building Age (Base Case: Over 100 years)			
75 years-100 years	0.431309	0.433516	0.432384
50 years-75 years	0.214801	0.217033	0.215832
25 years-50 years	0.379733	0.377982	0.378558
10 yers-25 years	0.612145	0.613082	0.612574
less than 10 years	0.283911	0.286198	0.284962
Building Renovation Time (Base Case: No renovation)			
renovated after 2013	0.000158	0.000225	0.000183
renovated between 2008-2013	0.03353	0.035495	0.034404
renovated between 2003-2008	0.01142	0.011996	0.011594
renovated earlier than 2003	0.039787	0.041268	0.04035
Floor Occupied (Base Case: Multiple or unknown)			
basement	0.388316	0.389529	0.38881
ground floor	0.000296	0.000352	0.000313
floor 2-5	8.25E-12	5.37E-11	1.50E-11
more than 5	0	1.11E-15	0
Building Class (Base Case: Class C or unknown)			
Class A	0.001139	0.00128	0.00118
Class B	0.242329	0.242877	0.242394
Property Type (Base Case: Other Types)			
Hotel	0.927093	0.928188	0.927697
Industrial	0.000122	0.000194	0.000151
Mixed-use	0.026256	0.028221	0.027141
Multi-family	0.080383	0.081569	0.080793
Office	0.090557	0.090058	0.089976
Retail	0.423324	0.423218	0.423063
Time Fixed Effect (Base Case: 2014)			

T.2015	5.67E-09	2.81E-09	2.85E-09
T.2016	7.18E-09	5.17E-09	4.50E-09
T.2017	9.37E-06	7.42E-06	7.44E-06
T.2018	2.70E-05	2.09E-05	2.16E-05
Location Fixed Effect			
Asset Turnover	0.067317	0.069978	0.068533

We can see that not all independent variables have significant indirect impacts. Most importantly, the influencing score has a significant direct and indirect impact. 15.57% / 35.95% = 43.32% of the total effect is due to a retail property's own influencing value, while 56.68% of the total effect comes from the influencing value of neighboring properties. This finding aligns with our assumption that the influencer behavior has a spatial spillover effect. When customers are attracted to some place by influencer's posts, the impact will not be limited to the targeted retail space but also benefit nearby ones. From the impact analysis, we find that the indirect impact even has a slightly higher magnitude than the direct impact[7].

Some other independent variables also have significant indirect impacts, including the transaction scale, tenant industry of apparel and banks, renovation time within five years, floor occupied, building class and building type. Some of them are easy to interpret. For example, the significant indirect impact of building features (building type, class, and floor occupied by retail spaces) can be explained by the clustering of similar buildings and retail spaces. The clustering of similar tenant industries such as the apparel or bank can explain the corresponding indirect impact. The

[7] This result needs further check using difference in differences (DID) analysis. We need more granular influencer data and split the observations into study group and control group. For example, compare the effective rents of two retail properties with the same wide-area influencer effect in their neighborhood, but different influencer effect on each particular store.

indirect impact of renovation time could be related to the redevelopment in some neighborhoods, for instance, the Hudson Yards.

6.1.3 Spatial Heterogeneity and GWR Regression Results

In this section, we estimated the GWR model to explore spatial varying coefficients of explanatory variables, especially the influencing value. To avoid the local multicollinearity, the independent variables for the local model are restricted to those significantly affect effective rents in our previous models. But this method still cannot guarantee there is no local multicollinearity. We apply stepwise variable selection and get the independent variables used in local models.

The table 6-5 shows the selected independent variables and the estimation result of the GWR model.

Table 6-5 GWR estimation result

Variables	MIN.	1ST QU.	MEDIAN	3RD QU.	MAX.	GLOBAL
Intercept	-1.5626	-0.8432	0.1516	0.6106	1.9516	0.3541
Influencer Behavior						
Log Influencing Value	-0.2302	-0.0654	0.0024	0.0556	0.4609	0.3180
Lease Term (Base Case: Over 15-year lease term)						
less than 5-year lease term	-1.9028	-0.4353	-0.1503	0.0207	0.6695	-0.2783
5 to 10-year lease term	-1.3095	0.0063	0.2031	0.3403	0.9675	0.2885
10 to 15-year lease term	-0.9215	-0.0286	0.1973	0.4004	1.2074	0.3301
Lease Type (Base Case: Full Service Lease)						
Gross lease	-0.5297	-0.1923	-0.0550	0.0670	0.4303	-0.2099
All Net Lease	-1.5661	-0.2829	-0.1548	-0.0197	0.7157	-0.3711

Transaction Size (Base Case: Over 2000sqft)						
under 500sqft	-0.5858	0.0699	0.2302	0.4201	1.2195	0.2769
500sqft-1000sqft	-0.6921	-0.0577	0.1029	0.2502	0.8937	0.1238
1000sqft-2000sqft	-0.8593	-0.1177	0.0314	0.2053	0.9001	0.1136
Building Features						
Building Class: Class A	-0.9963	0.0740	0.4309	0.8618	1.8928	0.6903
Property type: Mixed-use	-1.2979	-0.2915	-0.0613	0.1501	0.9936	-0.1525
Residual Sum of Squares	1154.507					
AIC	7049.144					
Quasi-Global R2	0.6806					

Compared to the result of the OLS model, the global coefficients of GWR have not changed a lot. The global coefficient of influencer impact is 0.32 compared to 0.40 in the OLS model and 0.36 (aggregation of direct and indirect effect) in the SAR model. The value of this coefficient ranges from -0.23 to 0.46 with a median that slightly over 0. This is a surprising finding that for nearly half of our observations, the influencing value is negatively correlated with effective rents. To further explore the spatial distribution of influencer effect, we map the estimations of coefficient:

From the map, we can find a distinct spatial pattern of influencer effect. The observations whose effective rents and influencing value are negatively correlated are mostly clustered at 2 locations: East Midtown and Lower Manhattan. Other such observations are located in a small cluster near the Wall Street, Astoria, Claremont Village.

This finding could be due to a number of reasons. As we discussed in Chapter 2 and 3, the pattern of influencer behavior and the effect on consumers vary across different retail sectors. For example,

influencer effect for food or clothing is, by intuition, greater than that on durable goods. This problematic result might be caused by removing the tenant industry. The effect of influencer marketing also depends on the feature of particular consumer groups. Also, it is possible that influencer marketing has not affected rents yet. We need a spatial=temporal model to further explore the process of influencers' impact.

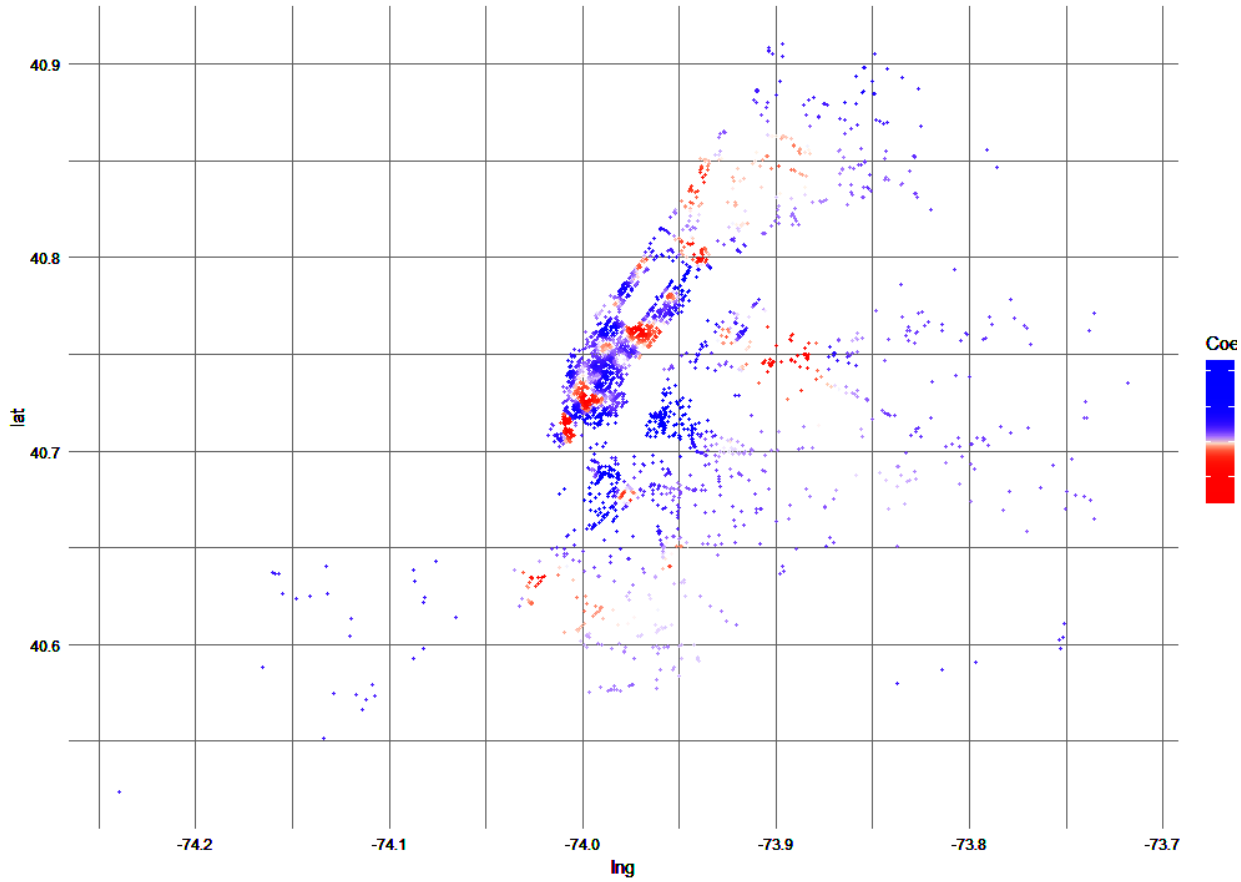

Figure 6-1: Spatial distribution of the coefficient estimation of influencing value. The value of the coefficient ranges from -0.23 (red) to 0.46 (blue). The white areas between the reds and blues are areas where the influencing value has no significant impact on effective rents. We can find two major clusters and one sub-cluster of negative influencer effect: East Midtown, Lower Manhattan, the Wall Street.

6.2 Discussions

6.2.1 Estimation Overview and Model Selection

From our estimation results, we find that the effect of influencing value is significant for both spatial and num-spatial models, which means influencers have an economically significant impact on effective rents of New York's retail rental market. Additionally, in the GWR model, we find the spatial pattern of influencers' impact.

The fitting of OLS model is not very impressive. The relatively low adjusted R-squared value can be partly attributed to spatial autocorrelation and omitted variables. It could also because we use logged effective rent per square feet as our dependent variable.

We can compare our models using the Akaike information criterion (AIC). As a model is never exact in representing the process that generates data observations, and there are some information losses. AIC estimates the relative information lost of a model in the following form:

$$AIC = 2k - 2\ln(\hat{L})$$

where k is the number of parameters and $\hat{L}$ is the maximum likelihood of the model. The models with less AIC value are preferred because AIC will reward models with higher log likelihood and penalize those with more parameters. We calculated the AIC values of our OLS, SAR, and GWR model as shown in Table 6-6. The GWR model has the lowest AIC value, which means GWR has the best tradeoff between model fitting and the number of parameters.

Table 6-6 Model AIC value comparison

	OLS	SAR	GWR
AIC	8466.7	6927.6	6885.7

The Bayesian information criterion (BIC) is also widely used for model selection. The only difference is the penalty on the number of parameters. The BIC is defined as:

$$BIC = \ln(n)k - 2\ln(\hat{L})$$

where n is the number of data points, and k is the number of parameters. In our study, the BIC analysis has similar results compared to AIC. However, some studies suggest that AIC is asymptotically optimal for selecting the regression models that minimize MRSE (Yang et al. 2005).

Our SAR model is only $e^{\frac{AIC_{min}-AIC_i}{2}} = e^{\frac{(6885.7-6927.6)}{2}} = 8 * 10^{-10}$ times as probable as the GWR model[8], which means the two models are very close. However, there is a significant difference between the OLS model and spatial models.

To sum up. For model selection, our spatial models perform significantly better than the non-spatial OLS model. Although GWR has a lower AIC value, the difference between GWR and SAR is very small. Additionally, since GWR only include a small part of our explanatory variables to avoid local multicollinearity, SAR model can explain more features that affect retail rents than GWR.

[8] The relative likelihood value of model i in the form of $e^{\frac{AIC_{min}-AIC_i}{2}}$ is vary similar to the likelihood ratio. However, the likelihood ratio test can only be used to compare nested models, but AIC has no such restriction.

6.2.2 Potential Application

This study provides a framework of analyzing online influencing behavior and evaluating its impact on retail rent using spatial econometric methods, in which we also examined the spatial autocorrelation and heterogeneity in New York's retail rent market. As discussed in previous findings section, in the case study of New York's retail rent market, the positive effect of influencing behavior is globally significant. The influencer behavior has both direct and indirect positive effect on retail rents. For different neighborhoods, the effect of influencer behavior may differ.

In general, this approach provides insight in two ways: modeling or predicting the retail rental value using metrics of online behavior; and designing an online marketing strategy to affect the retail rent price. Therefore, this evaluating and predictive model could be potentially used for investors and urban researchers.

6.2.4 Research Limitations

Nevertheless, the methodology still has several limitations.

First, the methodology is limited by our data source. Due to Instagram's "depreciation" policy, it's easier to scrape recent posts than old ones. If we use temporal analysis methods, it may cause biased results. Therefore, we have to use the yearly average value, which limits the granularity of the analysis. Also due to data source limitations, we did not analyze the content of influencer posts, which requires image processing and natural language processing (NLP) techniques. The data source also limits the explanatory power of our models and bring in possible endogeneity problem. Since we cannot track the information diffusion process, there is no guarantee that the causal

relationship between influencer behavior and effective rents exists. It is possible that landlords of high rent properties pay more on online branding.[9]

Second, we used IDW to predict the influencing value, which is a compromise considering the computation resources and the size of the dataset. However, there are other interpolation methods that may yield better predictions. If we can solve the data source problem, the ideal method is to use a spatial-temporal interpolation method to grasp the temporal effect of influencing behaviors. More importantly, to store and process a large volume of spatial-temporal data, we need a fundamental upgrade in database techniques.

Third, we used GWR to investigate spatial heterogeneity in parameter estimates, especially the coefficient of the influencing value. Our GWR model is based on Euclidean distance. However, in some recent research papers, using non-Euclidean distance (ND) instead of Euclidean distance (ED) can improve the performance of GWR regarding AIC value (Lu, B, Charlton, et al. 2014), especially for city-scale GWR modeling. This finding is not surprising because the ND is a better proxy for psychological distance than ED. Considering the importance of accessibility in retail space, using ND could improve the model performance.

Additionally, to avoid local multicollinearity, we only include a few independent variables in our GWR local model. This limits the explanatory power of our GWR.

[9] This effect has been minimized in our social media mining. As we do not distinguish sponsored and unsponsored influencers, and the number of non-sponsored posts far exceeds the number of sponsored posts.

Chapter 7: Future Work and Conclusions

7.1 Future Work

As mentioned in the limitations, in future studies, one of the most critical tasks is to improve the social media mining method. Firstly, with un-depreciated data, we can further explore the temporal effect of influencers and apply spatial-temporal models. Second, the metric of influencing effect can be further improved by introducing content analysis. We need natural language processing (NLP) and image recognition techniques to evaluate the content: how it is related to the targeted retail space, what its emotion or attitude is, and so forth. More importantly, we need an advanced database or information system techniques to store and process the high-volume data stream of spatial-temporal data. Furthermore, although Instagram is currently the main platform for influencers, several other websites or applications, such as Snapchat, are cultivating their influencer ecosystems. In future works, we need to incorporate more data sources to get a full view of influencer behavior.

We can also improve the study with new modeling methods. In our spatial heterogeneity analysis, we have to remove some independent variables from the GWR local model to avoid local multicollinearity, which limits the explanatory power of GWR. However, there are new modeling methods that can avoid multicollinearity without removing independent variables. For example, mixed geographically weighted regression (MGWR) allows not only varying estimates for features with spatially heterogeneous effects but also fixes estimates for those without spatial effect; Eigenvector spatial filtering (ESF) filters variables to avoid misspecification, and in some studies perform better then GWR in prediction accuracy (Griffith, Daniel A, 2013).

7.2 Conclusions

This research proposes a new way of evaluating and predicting retail rents through the lens of online behavior by correlating influencers with effective rents. We find that the effect of influencing value is significant for both spatial and num-spatial models, which means influencers have an economically significant impact on effective rents of New York's retail rental market. Additionally, we find the spatial pattern of influencers' impact using GWR model.

The research also develops a framework to quantify the impact of online influencer behaviors on retail rents. Using network analysis and spatial econometrics, the method can be replicated and applied to other kinds of online behaviors in social networks

Additionally, the research is not limited in influencer marketing, which is only one of the new activities generated by new technology, but rather inspire a further collaboration in the age of new economy among different stakeholders including landlords and tenants, social media, researchers, and all kinds of data providers for a better understanding of real estate market.

References

Anselin, Luc. "Thirty years of spatial econometrics." Papers in regional science 89.1 (2010): 3-25.

Anselin, Luc, and Daniel Arribas-Bel. “Spatial Fixed Effects and Spatial Dependence in a Single Cross-Section: Spatial Fixed Effects and Spatial Dependence.” Papers in Regional Science, vol. 92, no. 1, Mar. 2013, pp. 3–17. Crossref, doi:10.1111/j.1435-5957.2012.00480.x.

Bakshy, Eytan, et al. "Everyone's an influencer: quantifying influence on twitter." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

Bitter, Christopher, Gordon F. Mulligan, and Sandy Dall’erba. "Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method." Journal of Geographical Systems 9.1 (2007): 7-27.

Bidanset, Paul E., and John R. Lombard. "The effect of kernel and bandwidth specification in geographically weighted regression models on the accuracy and uniformity of mass real estate appraisal." Journal of Property Tax Assessment & Administration 10.3 (2014).

Blettner, Robert A. "Mass Appraisals Via Multiple Regression Analysis." The Appraisal Journal 37.4 (1969): 513-521.

Brooks, Chris, and Sotiris Tsolacos. "Forecasting models of retail rents." Environment and Planning A 32.10 (2000): 1825-1839.

Brown, Duncan, and Nick Hayes. Influencer marketing: Who really influences your customers?. Routledge, 2008.

Brown et al. “Influence Marketing: How to Create, Manage, and Measure Brand Influencers in Social Media Marketing.” Choice Reviews Online, vol. 51, no. 05, Jan. 2014, pp. 51-2752-51–2752. Crossref, doi:10.5860/CHOICE.51-2752.

Casetti, Emilio. "Generating models by the expansion method: applications to geographical research." Geographical analysis 4.1 (1972): 81-91.

Cliff, A.D. and Ord, J.K. 1973. Spatial autocorrelation. Pion, London

Eisenlauer, Jack F. "Mass versus individual appraisals." The Appraisal Journal 36.4 (1968): 532-40.

Dubin, Robin, et al. Spatial Autoregression Techniques for Real Estate Data. p. 17.

Fairfield, Patricia M., and Teri Lombardi Yohn. "Using asset turnover and profit margin to forecast changes in profitability." Review of Accounting Studies 6.4 (2001): 371-385.

Fotheringham, A.S., Brunsdon, C., and Charlton, M.E., 2002, Geographically Weighted Regression, Chichester: Wiley; Paez A, Farber S, Wheeler D, 2011, "A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships", Environment and Planning A 43(12) 2992-3010

Griffith, Daniel A., David WS Wong, and Thomas Whitfield. "Exploring relationships between the global and regional measures of spatial autocorrelation." Journal of Regional Science 43.4 (2003): 683-710.

Griffith, Daniel A. Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer Science & Business Media, 2013.

Guy, Clifford M. "Classifications of retail stores and shopping centres: some methodological issues." GeoJournal 45.4 (1998): 255-264.

Helbich, Marco, and Daniel A. Griffith. "Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches." Computers, Environment and Urban Systems 57 (2016): 1-11.

Kim, Hyejeong, et al. "Psychographic characteristics affecting behavioral intentions towards pop-up retail." International Journal of Retail & Distribution Management 38.2 (2010): 133-154.

Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." Proceedings of the 19th international conference on World wide web. AcM, 2010.

Lagrée, Paul, et al. “Algorithms for Online Influencer Marketing.” ArXiv:1702.05354 [Cs], Feb. 2017. arXiv.org, http://arxiv.org/abs/1702.05354.

Lahuerta-Otero, Eva, and Rebeca Cordero-Gutiérrez. “Looking for the Perfect Tweet. The Use of Data Mining Techniques to Find Influencers on Twitter.” Computers in Human Behavior, vol. 64, Nov. 2016, pp. 575–83. Crossref, doi:10.1016/j.chb.2016.07.035.

Lee, Lung-fei, and Jihai Yu. "Estimation of spatial autoregressive panel data models with fixed effects." Journal of Econometrics 154.2 (2010): 165-185.

Lu, B, Charlton, M, Harris, P, Fotheringham, AS (2014) Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. International Journal of Geographical Information Science 28(4): 660-681

Manikonda, Lydia, Yuheng Hu, and Subbarao Kambhampati. "Analyzing user activities, demographics, social network structure and user-generated content on Instagram." arXiv preprint arXiv:1410.8099 (2014).

McMillen DP (2003) Spatial autocorrelation or model misspecification? International Regional Science Review 26: 208–217

Miller, Norman. "Retail leasing in a web enabled world." Journal of Real Estate Portfolio Management 6.2 (2000): 167-184.

Onkvisit, Sak, and John J. Shaw. "Modifying the retail classification system for more timely marketing strategies." Journal of the Academy of Marketing Science 9.4 (1981): 436-453.

O'Roarty, Brenna, et al. "A Case-Based Reasoning Approach to the Selection of Comparable Evidence for Retail Rent Determination." Expert Systems with Applications, vol. 12, no. 4, May 1997, pp. 417–28. Crossref, doi:10.1016/S0957-4174(97)83769-4

Ryu, Jay Sang. "Consumer attitudes and shopping intentions toward pop-up fashion stores." Journal of Global Fashion Marketing 2.3 (2011): 139-147.

Sawada, Mike. "Global spatial autocorrelation indices-Moran's I, Geary's C and the general cross-product statistic." Laboratory of Paleoclimatology and Climatology, Dept. Geography, University of Ottawa,(Mimeo) (2001).

Thrall, Grant Ian. "Common geographic errors of real estate analysts." Journal of Real Estate Literature 6.1 (1998): 45-54.

US Census, Advance Monthly Sales for Retail And Food Services, November 2018

Yang, Yuhong. "Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation." Biometrika 92.4 (2005): 937-950.