FinTechs and the City:
Agglomeration Economies of Financial Services Firms in Midtown Manhattan

by

Michael Aaron Pearce

Bachelor of Arts, Yale University (2009)

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

June, 2019

Author …….
Michael Aaron Pearce
Department of Urban Studies and Planning
May 23, 2019

Certified by ……
Dennis Frenchman
Class of 1922 Professor of Urban Design and Planning
Director, Center for Real Estate
Thesis Supervisor

Accepted by …….
Ceasar McDowell
Professor of the Practice
Co-Chair, MCP Committee
Department of Urban Studies and Planning

FinTechs and the City:
Agglomeration Economies of Financial Services Firms in Midtown Manhattan

by

Michael Aaron Pearce

Submitted to the Department of Urban Studies and Planning

on May 23, 2019, in partial fulfillment of the

requirements for the degree of

Master in City Planning

Abstract

Although agglomeration is widely studied through wages, industrial output, capital, and innovation, research using real estate rates remains sparse; this is the case even though commercial real estate rents may serve as the most effective measure of agglomeration. To remedy this gap in the literature, this thesis seeks to understand the rental premium associated with the agglomeration of financial services firms and startups in Midtown Manhattan. The research relies upon hedonic regression to calculate the marginal impact of additional financial services firms, employment, annual sales, startup funding, and job postings on the rent paid by financial services firms in Midtown Manhattan. To understand whether the agglomeration effects diminish rapidly over space, I conduct these analyses at three radii: 100 meters, 250 meters, and 500 meters. Ultimately, my research confirms the statistically and substantially significant presence of agglomeration among financial services in Midtown. Furthermore, this thesis contributes to existing agglomeration economics research by specifying, in the same study, agglomeration impacts deriving from a given industry's established businesses as well as its startups.

Thesis Supervisor: Dennis Frenchman

Title:
FinTechs and the City:
Agglomeration Economies of Financial Services Firms in Midtown Manhattan

Keywords: Econometrics, Agglomeration, Financial Services, Financial Technology, FinTech, Startups, New York City, Midtown, Manhattan, Commercial Real Estate

Acknowledgments

If it takes a village to raise a child, then I have been the very lucky beneficiary of several diverse cities whose residents and visitors have instilled in me the skills needed to get through MIT, including writing and editing this thesis. Family first, I'm told. Thus, I thank you, S.J., for always inspiring me to think faster and more pointedly as well as to fight for my positions. S.S. taught me how to see multiple perspectives and mediate between competing voices. Sima, you have brought joy to my life these past two years and I could not have gone through this two-year process without you holding my hand the entire way.

Several professors at DUSP made my two years here into the learning experience I craved. A deep thank you to Amy Glasmeier who introduced me to the vastness of agglomeration research, my research advisor Sarah Williams, my academic advisor Larry Vale, and the DesignX team and mentors. Many thanks also to my thesis advisor, Dennis Frenchman, and readers Andrea Chegut and Garnette Cadogan.

I thank the people who taught me how to work hard and relentlessly: my fencing coaches and professional mentors. Without you, everything would seem difficult, insurmountable, and unachievable: Cole Harkness, Peter Burchard, Norman Mattox, George Nonomura, John Charles, Henry Harutunian, Steven Fingerhood, Dan Still, Michael Feldman, Peter Sullivan, Erik Foraker, Katie Darling, Stu Shiff, and Ted Knetzger.

I could not have written this thesis without serious help learning the dovetailed crafts of researching, writing, and editing. Key mentors in these endeavors include David Hathwell, Ed Kamens, Maurice Samuels, Carter Wiseman, Elihu Rubin, and Garnette Cadogan. Combined, you have taught me how to think and how to express my thoughts. I cannot thank you enough.

Finally, there is one person who has invested more blood, sweat and tears than these individuals combined: L.E. For doing everything listed above and, most importantly, for never letting me give up on myself, thank you from the bottom of my heart. I dedicate this work to you.

Contents

Chapter 1

Introduction

Driving across the country in the middle of the night, you look at your dashboard and remember your gas light has been blinking for miles. You keep your eyes peeled for a canopy of neon lights but, for the next ten minutes, you see nothing but your own headlights bouncing off the asphalt. Then, all of a sudden, a green dinosaur emerges on the horizon, a Hess station. Next to it, a bright orange ball hovers 20 feet above the ground of a 76 station across the street. As you step out of your car to fill up the tank, you remark to yourself that, like gas stations, many firms and individuals tend to locate near similar entities. Why do they do this and how do they benefit from this counterintuitive pattern of clustering?

Agglomeration economics is the strand of economic research that concerns itself with the ways in which geographic concentrations of people, jobs, and firms yield outsized returns to scale. This thesis pushes forward agglomeration economics research by determining how varying concentrations of financial services firms and startups in Midtown Manhattan impact commercial real estate rents. Said another way, I will address the question of whether financial services companies are more productive (and thus, able to pay a higher rent for office space) when surrounded by other financial services companies and startups.

Though many early economists and researchers pointed to the importance of spatial considerations in economic models (Von Thünen, 1826, among others), Alfred Marshall (1920) was the first to identify spatial factors as an area of important economic research. Over the past century, researchers have developed several ways of studying and measuring the impacts of agglomeration. These include industrial productivity, wages, patents, and commercial real estate prices. Researchers

suggest that, of all data used to quantify agglomeration economies, commercial real estate rents may be the best proxy because these figures represent actual monetary transactions undertaken in exchange for access to the best workers, superior service providers, and access to a city's knowledge networks (Koster, van Ommeren, and Rietveld, 2012; Koster, 2013). In addition to the economic rationale that real estate data improves agglomeration research, there are many qualitative aspects related to real estate data that may also help explain the mechanisms underlying agglomeration, even though traditional analytic methods "rarely measure the appearance of the outside neighborhood" (Glaeser et al, 2018). These include being able to model the vertical dimension of buildings, to analyze the variation of types and ages of buildings within a city, and to adjust estimates based on the presence or absence of active street life.

Despite the recognition of real estate data's importance, conducting agglomeration research using this type of information has been limited for a number of factors. First, the data are difficult to access since they are fiercely guarded by those who generate and utilize the data. Companies such as CoStar derive millions of dollars from selling data and investors such as Blackstone make billion-dollar decisions based on the proprietary data they gather and assemble from market interactions. It is understandable why they would not want to release their information to the broader world. Second, data that are accessible may not be of sufficient size to utilize in a statistical model. Because real estate transactions are infrequent, any potential experiment's sample is necessarily limited in size. Third, data that are accessible and of sufficiently large size are generally difficult to utilize, as there is no industry consensus on formulas used to calculate crucial parameters such as effective rent.

Despite these roadblocks, economists recognize the value of developing a better understanding of the real estate dynamics in cities. The urbanist Richard Florida emphasizes the need to "encourage research to focus on the competition for space that stems from the concentration of innovation, entrepreneurship, and creativity in a relatively small number of

superstar cities and knowledge hubs" (Florida, Adler, and Mellander, 2017). Thus, in order to begin addressing this lack of research in agglomeration economics grounded in commercial real estate data, my thesis builds a hedonic regression model to quantify the impact of nearby financial services firms and startups on commercial real estate prices.

I am able to contribute original research in this realm thanks to three data sources: Dun & Bradstreet, CB Insights, and CompStak. Dun & Bradstreet provides information about businesses of all sizes across the world including their address, year founded, annual sales, and number of employees. CB Insights aggregates, analyzes, and disseminates data and information about startups. CompStak is a commercial real estate startup that gathers information on lease transactions. Combined, I created a data set of over 7,000 leases that contains lease-specific factors such as effective rent, lease term, and tenant improvement dollars, along with business-environment factors such as the number of nearby financial services firms, the total funding raised by startups, and the number of nearby employees. In order to account for spatial variation in the business environment, I also create three distance variables (100 meters, 250 meters, and 500 meters) that aggregate the business-level factors within each of those buffers.

In order to consolidate and clean the data, I scrubbed location information, categorized them by industry, and created several new variables to parse insights within the data but obscured by the raw structure. After cleaning and consolidating the data, I specified a hedonic regression that incorporates the impact of these factors on the rent paid by financial services companies in Midtown Manhattan. Hedonic regression is the sensible tool for this study because it is easily implemented, easily understood, and makes "the most sense for valuing amenities that are spatially delineated," such as lease transactions (Glaeser et al, 2018).

Ultimately, I found that the presence of financial services firms has a strong, statistically significant relationship with commercial real estate prices. One additional financial services firm

within a 100-meter radius yields a 0.1% increase in the annual per square foot rent paid by financial services companies. When accounting for startups and financial technology startups, however, it is not the number of firms that has a positive impact on the rent, but rather the cumulative funding those startups have accumulated. At a 100-meter radius, an additional $1.0 million in funding raised by startups is associated with a 0.04% increase in the annual per square foot rent. For financial technology startups, an additional $1.0 million in funding raised by startups is associated with a 0.1% increase in the annual per square foot rent. This points to one of the main contributions of my research to the field of agglomeration economics: it is not merely the presence of firms that leads to agglomeration but the underlying business relationships and networks that have significant impact.

The second chapter presents a review of the literature related to agglomeration economies, focusing on its geographic components and methods of measurement. The key method in this thesis is using real estate rents to quantify agglomeration, despite the variety of issues facing researchers who pursue this method. The chapter concludes with an overview of emerging and future trends in real estate data that will facilitate future use of real estate rents to measure agglomeration.

The third chapter describes my data and methods, including how I cleaned, summarized, and mapped each data set. This sets the stage for several hypotheses that connect the underlying business dynamics that cause firms to agglomerate with the economic manifestation of that clustering. The chapter concludes with a description of how a hedonic model parses the marginal impacts on commercial rent and a general specification of what variables will be included in the final regressions.

Finally, I provide and interpret the results of my regressions, surface shortcomings of the research, and propose future areas of work. Ultimately, this thesis reaffirms prior research regarding the existence and distance of attenuation for agglomeration. Yet it also moves forward agglomeration research in two key ways. First, previous research has not considered the impact of

multiple types of business on agglomeration; in contrast, my research asserts that both established businesses and startups within a given industry play a significant role in the impacts of agglomeration. Second, my research suggests that economic and statistical work should be augmented by qualitative research into the built environment and industry dynamics to more effectively assess the underlying sources and causes of agglomeration.

Chapter 2

Literature Review

When more people gather together, they are increasingly productive. As dense accumulations of people, jobs, markets, and transportation networks, cities play a pivotal role in local, regional, national, and international economies. In his 1920 textbook, "The Economics of Industry," Alfred Marshall termed this phenomenon "agglomeration economics." Although researchers as far back as Von Thünen (1826) had previously described the economic value of cities, Marshall's book was the "first careful economic analysis of" the agglomeration phenomenon (Rosenthal and Strange, 2003). In it, Marshall identifies three sources of the benefits that derive from increased proximity: input sharing, labor pooling, and knowledge spillovers (Marshall, 1920).

Since Marshall, research into agglomeration economies has dramatically increased and become more robust, with the greatest expansion in related inquiries emerging over the past forty years (Dunse and Jones, 1998; Rosenthal and Strange, 2004). Most notably, in 1991, Paul Krugman coined the term "New Economic Geography" and used it to promote greater awareness of and research into the impacts of geography on economics (Krugman, 1991). Krugman asserts that "the study of spatial economics – of the location of production – has a long if somewhat thin history" (Krugman, 1998). He traces this history back to Von Thünen and suggests that geography was not discussed in economic circles because "geography turns out to be perhaps the most naturally 'non-linear' area of economics," meaning that it is difficult to incorporate the heterogeneous impacts of distance and proximity into quantitative economic models (Krugman, 1998). For proposing how to marry the scientific approaches of economics with the messy constraints and confounding factors

introduced by geography, Paul Krugman won the 2008 Nobel Prize in Economics (Krugman, 2009). Recognizing, as the Nobel selection committee did, that Krugman's work integrates "trade patterns and [the] location of economic activity," other researchers have spent significant time and effort in the flourishing field of economic geography (Nobel Committee, 2008).

This literature review discusses advances in agglomeration economics ranging from improving the geographic specificity with which researchers can understand agglomeration to tracing elements of Marshall's forces that were previously considered outside the scope of potential research. Because this chapter emphasizes literature that utilizes commercial office rents as a proxy for the presence of agglomeration, I also provide an overview of cutting edge real estate data enabling this type of research. The chapter concludes with an introduction to Midtown Manhattan, as well as to the sector I study in this thesis, financial technology.

2.1 Geographic Scope of Agglomeration

One of the most crucial components of agglomeration economics is the question of the geography and scale at which the forces of agglomeration operate. That is because "the concept of location is now defined as a geographic unit over which interaction and communication is facilitated…[and] economic activity is enhanced" (Feldman, 1999). For example, the Bay Area derives significant benefit from the agglomeration economies of the software industry located there; yet, within the region, Marshall's three forces are not uniformly distributed. Venture capital, for example, is an input that is highly concentrated on Sand Hill Road, the famed five-mile-long road in Palo Alto, CA. Thus, while interaction and communication may be occurring within a very compact area, economic activity across the entire region is enhanced as a result of this dense, highly-localized communication. Feldman's research provides valuable insight into the importance of geography: whereas early research in agglomeration focused on larger geographic areas, Liusman and his

colleagues (2017) emphasize "the importance of the propinquity of support services for office users." Conducting research at small geographic scales, such as within a city or even a neighborhood, is important because "different spatial arrangements [at different scales] can lead to distinct results, [even] using the same data and analysis methods" (Melanda, Hunter, and Barry, 2016). This issue is termed the Modifiable Areal Unit Problem (MAUP). The reason the MAUP is particularly important to economic geography is because "agglomeration economies…attenuate rapidly across geographic space" (Rosenthal and Strange, 2004). This means that research into agglomeration at larger geographies could miss some of the micro-agglomerations that occur within very small geographic boundaries. Arzaghi and Henderson (2008) argue pointedly in their study of Madison Avenue advertising executives that most of the agglomeration economic benefits occur within a short distance of individuals' offices. Rosenthal and Strange (2003) demonstrate the dramatic impact distance can have on the attenuation of agglomeration economies, finding that "the effect of own-industry employment in the first mile [is]…ten to 1,000 times larger than the effect two to five miles away." Without fine-grained data and analysis, these findings would be obscured as a result of the MAUP. One way to deal with the MAUP is to conduct analysis "on non-aggregate data, when that's possible" (Melanda, Hunter, and Barry, 2016). This allows the researcher the flexibility to conduct analysis at an aggregate level, if so desired, but also allows her to dig into the underlying trends that may be revealed at a more granular scale.

Although researchers recognize the importance of studying agglomeration over short distances, they have not yet reached consensus on the most appropriate geographic scale to study these forces and, in fact, "know little about distances between individual firms within regions" (Wallsten, 2001). Some early agglomeration research could only conduct research at the scale of tens of miles, while more contemporary research can pinpoint agglomeration within a quarter of a mile (Feldman, 1999; Wallsten, 2001; Drennan, 2018). This uncertainty about the appropriate distances to

study agglomeration results from three sources: the tools used to study agglomeration, the impact of different methods of geographical aggregation, and the difference between administrative and economic geographies. From a tools perspective, agglomeration research suffers from the fact that "empirical practice has employed relatively 'spaceless' statistical tools, despite frequent mentions in the literature of observed violations of the assumptions underlying the optimality of such tools" (Pace, Barry, and Sirmans, 1998). For example, the standard regression methodology, Ordinary Least Squares (or OLS), produces inaccurate results when analyzing data that violates its underlying assumption of heteroscedasticity; spatial and real estate data often violate this assumption.

The second issue associated with agglomeration economics research emerges because many data sets are available at only high levels of aggregation representing large geographic regions, such as MSAs, the Metropolitan Statistical Areas that aim to reflect a given region's economic extents. These areas can cover many counties across state lines; for example, New York City's MSA covers parts of New Jersey, Westchester County, and southwestern Connecticut. Previous agglomeration economics research primarily used aggregate data because that was the only level of aggregation available. In the 1990s and 2000s, "data typically force[d] researchers to aggregate up to large geographic units," forcing them to study agglomeration economies uniformly across larger geographic extents such as metropolitan areas (MSAs), states, or countries (Jaffe, Trajtenberg, and Henderson, 1993; Wallsten, 2001; Rosenthal and Strange, 2003). These data sets were not ideal and researchers such as Rosenthal and Strange (2003) recognized that "agglomeration should ideally be studied at a much more refined geographic level than has been the norm." As mentioned before, the MAUP "introduces bias into any estimation by discarding firm-specific variation" (Wallsten, 2001). Furthermore, "conclusions reached when the underlying data are aggregated to a particular set of boundaries (say, counties) may differ markedly from conclusions reached when the same underlying data are aggregated to a different set of boundaries (say, MSAs)" (Buzard et al, 2015). The example

Buzard et al (2015) provide compares their findings at hyper-local spatial scales with the earlier research of Jaffe, Trajtenberg, and Henderson (1993). Buzard et al (2015) assert that "at the smallest spatial scales, our localization statistics are on average much larger than those Jaffe, Trajtenberg, and Henderson report for the metropolitan areas included in their tests." This suggests that using more fine-grained data may surface previously unquantifiable findings.

The final concern that emerges when studying agglomeration is the difference between administrative geographies such as zip codes and geographic areas that better reflect the underlying economic realities of a given city. In research that pertains to commercial real estate, some data, such as US Census statistics, may be available only at the block group level; the commercial real estate market, however, maps industry fundamentals such as rent and vacancy to submarkets, such as Midtown Manhattan. These industry-specific geographies often do not conform neatly to administrative boundaries such as zip codes or census tracts. Melo et al (2008) write that "since economic data are often more available for administrative divisions than for geographic units with more economic meaning, it is common to find estimates of agglomeration based on administrative spatial units. Administrative boundaries, however, can be less appropriate to capture the spatial scale of agglomeration effects." The mismatch between administrative and economically-relevant boundaries diminishes the explanatory power of analyses mixing these types of geographies.

Although there are some roadblocks to conducting agglomeration economics research, several recent developments have made more granular analysis possible. Today, there are more widely-available "microrecord data sets for studying economic and social interactions at detailed levels" that have significantly helped research in the agglomeration economics field (Carlino and Kerr, 2015). These data "better represent firm organizing behavior,…provide greater data variability, and reduce multicollinearity and aggregation bias resulting from unobserved heterogeneity" (Melo et al, 2008). This better data has yielded research at a significantly more refined geographic scale than

was previously unimaginable. For example, whereas previous research could identify agglomeration only at the state or MSA level, Koster (2013) found that the geographic scope of agglomeration economies was approximately 15 kilometers, a distance much smaller than states and even many MSAs. 15 kilometers is one of the broadest geographic bandwidths described in research from the past twenty years. For example, Buzard et al (2015) found "the strongest evidence for the spatial concentration of R&D labs occur at very small spatial scales (such as within a two- to three-block area)," or a quarter of a mile. The smallest distance, however, was reported by Wallsten (2001), who found agglomeration at the scale of one-tenth of a mile. Clearly, we have come a long way from studying agglomeration at the MSA-level to being able to peer into the micro-scale foundations of these forces.

2.2 Measuring Agglomeration Economies

After understanding the geographic concerns inherent in researching agglomeration economies, we can proceed to the practical ways that researchers gather data about the presence and magnitude of agglomeration economies. One main concern for early agglomeration research was that the forces of agglomeration "cannot be directly measured" (Drennan, 2018). Over time, however, researchers have developed a number of ways to quantify the impact of clustering on a region's economy (Drennan, 2018). The literature refers to three primary ways that researchers can measure the effects of agglomeration: through increased wages or productivity, through elevated innovation activity, and through real estate premiums.

2.2.1 Wages & Productivity

Using wages and productivity as a proxy for agglomeration makes sense, as intuition suggests that, "in more productive locations, wages will…be higher" (Rosenthal and Strange, 2004). Wages

are a simple and accessible way to measure agglomeration economies because they are documented in widely-available public data sets, including the United States Census and the Consumer Population Survey (Rosenthal and Strange, 2004). These publicly-available data sets include information on average wages, average hours of work, as well as productivity measures such as industrial output. This makes measuring manufacturing and industrial jobs and, in turn, relative wages, a straightforward process. By extension, this makes agglomeration relatively straightforward to measure for those industries. For professional services, however, wage data are available, but the level of output or the corresponding number of hours worked are not. This is because, in contrast to industrial work which is based on the 40-hour workweek, professional services salaries are not tied to a specific number of hours worked or to a specific piece-rate, meaning "productivity cannot be [directly] estimated" (Drennan and Kelly, 2011). As a result, "the urban economic literature analyzing agglomeration economies in office activities such as producer services [was, at one point,] non-existent" (Drennan and Kelly 2011). Yet, as the economy has become increasingly intertwined with technological innovation, researchers have determined other, perhaps more effective, ways to measure agglomeration of professional services. Much has changed and, as a result of the developments in the following section, researchers are able to study the agglomeration of professional services; they have found that the effects of urban agglomeration are much stronger for service industries than for manufacturing and geographic concentration is much more pronounced in certain industries such as software. (Rosenthal and Strange, 2003; Melo et al, 2008).

2.2.2 Innovation

One of Marshalls' three sources of agglomeration is knowledge spillovers, which represents how concentrations of workers share information, methods, and insights. In a given business environment, the more people there are in a given area, the more likely they are to engage with one

another and share ideas, thus generating new knowledge. Innovation is closely tied to the context of knowledge spillovers because as workers exchange ideas and knowledge from their respective jobs, they may spark previously-unseen connections for their colleagues, thus increasing their innovative capacity and, in turn, productivity.

Although it is difficult to measure knowledge spillovers, the tight connection between innovation and knowledge spillover allows us to use patents, an output of the innovation process, as a proxy for knowledge spillovers. Until the 1990's, people considered knowledge spillovers very difficult to quantify. Krugman (1991), for example, asserted that "knowledge flows…are invisible; they leave no paper trail by which they may be measured and tracked." Even until the 2000's, researchers had "no doubt…that knowledge spillovers are difficult to identify empirically" (Rosenthal and Strange, 2004). These researchers saw the numerous potential benefits of being able measure the impact of innovation on agglomeration economies, but did not have a method. Two years after Krugman's assertion, however, Jaffe, Trajtenberg, and Henderson proposed a novel way of quantifying innovation activity by studying the "knowledge flows [that] do sometimes leave a paper trail, in the form of citations in patents" (Jaffe, Trajtenberg, and Henderson, 1993). Their paper, the first to quantitatively measure knowledge spillovers, found "evidence that these trails…are geographically localized" (Jaffe, Trajtenberg, and Henderson, 1993). Patents are useful to study agglomeration for several reasons: in addition to the fact that patents are a direct outcome of innovation, they are available publicly without confidentiality restrictions, they are released at a relatively granular level, and they are available over an extended time horizon (Carlino and Kerr, 2015). Since 1993, researchers have used patent data to quantify agglomeration economies, finding, in some cases, that "differences in growth rates may result from increasing returns to knowledge" (Feldman, 1999). More specifically, Carlino et al (2007) found that per capita patents increased by 20% in metropolitan statistical areas with twice the number of jobs per square mile of other

metropolitan areas. This suggests a strong relationship between concentrations of firms and innovative productivity.

That is not to say, however, that patent data do a perfect job of quantifying the agglomeration economies present in a region. Jaffe, Trajtenberg, and Henderson (1993) even admit that the geographic limits of "patent and citation data make it difficult to go much farther with such questions" regarding the differing rates of knowledge spillover at different geographic aggregation levels. Thus, in addition to patents, researchers have developed other ways to measure and quantify knowledge spillovers and innovation. Carlino and Kerr (2015) proposed an input-output methodology to measure innovation. Inputs are monetary factors such as research and development spending or venture capital, while outputs are intellectual developments such as patents and citations or physical developments such as new product announcements. According to these researchers, "one advantage of venture capital-based measures is that they are now available at the microlevel through services such as VentureXpert" (Carlino and Kerr, 2015). Following the use of venture funding as a proxy for agglomeration, Florida and King (2018) found "significant clustering of venture-capital-backed start-ups within metros at the ZIP [*sic*] code level." Their research extended beyond geographic clustering and found that these limited number of zip codes also accumulate the greatest share of startup funding. Across the United States, "less than one percent…of all zip codes…attracted more than $100 million in venture capital investment. These neighborhoods account for 60.7% of all venture capital investment" (Florida and King, 2016). Because these zip codes are not evenly distributed, but are instead concentrated "in distinct neighborhood micro-clusters across the United States," Florida's research allows us to understand that venture capital funding tends to concentrate in small areas and that these concentrations produce disproportionately more innovation than the average zip code (Florida and King, 2016; Florida, Adler, and Mellander, 2017).

Although using venture capital data to study agglomeration "affords researchers extensive flexibility in metric design," there are also challenges associated with using venture capital data, primarily the fact "that these investments are concentrated in specific technological areas…and types of firms…thus making them quite incomplete for describing innovation broadly" (Carlino and Kerr, 2015). Additionally, "the data on venture-capital-backed start-ups are in the hands of several private providers who charge relatively high fees to access it," making it difficult for a general research audience to analyze (Florida and King, 2018).

Another aspect of the relationship between venture capital, patents, and agglomeration economies is the fact that venture capital is used to start and fund companies that, in turn, generate patents. To quantify the relationship between agglomeration and company formation, Acs and Varga (2005) developed a framework that linked entrepreneurial activity and knowledge spillovers, ultimately determining that the agglomeration of entrepreneurial activity, as measured by new firm creation, positively impacts the spillover of new knowledge. Furthermore, companies benefit from being in such a dynamic environment, as Feldman (1999) describes Lerner (1996)'s research finding that "small start-up firms benefit from being in a location that is attracting venture capital investment." This is because "employment and sales growth were significantly higher if the award was made to a firm located in a zip code that received private venture capital activity" (Feldman, 1999). We thus see that high concentrations of certain types of economic activity will create positive momentum for that area's economic engine. Rosenthal and Strange (2003) support this finding, suggesting that "a more competitive and entrepreneurial environment enhances growth." In that research, Rosenthal and Strange measure agglomeration not through innovation but through firm births, suggesting that "if agglomeration economies are present, then births will occur near concentrations of existing employment, all else equal. If agglomeration economies are absent, then births will tend to disperse" (Rosenthal and Strange, 2003). Thus, we see that higher innovation

leads to the birth of more firms, especially of startups. Rosenthal and Strange (2003) also found that, in turn, small firms, especially startups, had a greater impact on the births of other, new, small firms than a comparable level of employment at larger companies. DeSilva and McComb (2012) support Rosenthal and Strange's findings and augment them further by adding a geographic dimension to their findings: similar to other research on the geographic reach of agglomeration, DeSilva and McComb found that the positive effects of agglomeration on company formation were "confined to a radius of only one mile or less." All of this research suggests that there is a tightly-clustered radius in which innovation begets further innovation, emphasizing the role and importance of dense centers of innovation in yielding greater productivity (Moretti, 2012).

2.2.3 Real Estate Values

The final way that researchers measure the impact of agglomeration economics is through local real estate prices. This is an especially important method, since some argue that rapidly rising "rents in urban areas are likely explained by agglomeration economies" (Koster, 2013). The intuition behind this method of studying agglomeration is that, if companies are more productive in concentrations, they will generate greater revenue and should be willing to pay more rent to locate there (Gyourko and Tracy, 1991; Blomquist, Berger, and Hoehn, 1998; Gabriel and Rosenthal, 2004; Drennan and Kelly, 2011). Said another way, in contrast to previous methods of measuring the value of agglomeration that required researchers to impute a dollar amount based on the intensity of the economic activity, real estate leases represent actual dollar-denominated financial transactions that companies sign, indicating their willingness to pay for access to labor pools, knowledge spillovers, and specialized service providers (Koster, van Ommeren, and Rietveld, 2012; Koster, 2013). Drennan and Kelly (2011) push this idea even further, hypothesizing that office rents best represent agglomeration economies for two reasons: first, rent represents the largest capital input component

of professional services firms. Because rent is such a large expenditure for businesses, they are only going to spend money where they think they can be most productive. Second, the fact that office rents vary within metropolitan areas and within cities allows researchers to understand the price an individual or firm is willing to pay to locate near amenities such as parks, restaurants, or transportation, as well as more agglomeration-related factors such as other firms, service providers, skilled labor, or knowledge networks.

Yet, there is another unique aspect of real estate and other spatial data that economists rarely consider: the qualitative aspects of the built environment (Salesses et al, 2013). For example, Glaeser et al (2018) found that "physical attribute[s] can add predictive power to models" but few studies incorporate elements of the built environment because of "a lack of data on the physical attributes of urban space" (Glaeser et al, 2018). A few recent papers such as Liu et al (2018) and Puri et al (2018) begin to integrate physical attributes and statistical, economic findings. In the former paper, researchers found that going up one floor in a building was associated with a 4.0% increase in rent. In the latter work, the researchers determined that less well-designed spaces (as measured by Gensler's Work Performance Index) secured lower-paying tenants. These papers provide examples of how future research will explicitly integrate factors of the built environment into economic models and research.

More generally, we must address the actual methods researchers use to quantify the impact of these spatial elements on real estate prices. Although "determining the value of commercial real estate remains elusively hard," one of the most prominent tools researchers employ to understand the underlying value of real estate is the hedonic method (Kok, Koponen, Martínez-Barbosa, 2017). A hedonic model uses linear regression as a tool to understand the marginal impact of environmental factors on real estate prices (Aurélio and González, 2016). According to Aurélio and González (2016), hedonic methods were proposed by Court in 1939, and further developed by

Griliches and Rosen in the 1970's. Since then, hedonic modeling has become "widespread in the urban economy" (Aurélio and González, 2016). This is because "relatively few attributes are necessary for estimation and the resulting models are generally easily comprehensible" (Melanda, Hunter, and Barry, 2016). Hedonic models have been utilized to price various factors that impact the value of real estate, from nearby air quality to environmentally friendly construction components (Kim, Phipps, and Anselin, 2001; Chegut, Eichholtz, and Kok, 2015).

That is not to say that hedonic regressions are a perfect modeling tool; there are a number of issues associated with using hedonic models to price real estate. First, real estate assets are relatively unique, meaning it is difficult to extrapolate consistent findings from heterogeneous, non-normally-distributed samples (Francke and Minne, 2018). Second, there are relatively few transactions, meaning the sample size is small (Francke and Minne, 2018). Third, the data describing the characteristics of these properties are not well gathered or distributed (Francke and Minne, 2018). One final drawback of using the hedonic method is that the spatial component is not always included, although it is a crucial component and including spatial factors "improves the overall goodness of fit" of the model (Pace, Barry, and Sirmans, 1998; Wilhelmsson, 2002; García, Gámez, and Alfaro, 2007; Francke and Minne, 2018). To address some of the weaknesses of hedonic analysis, researchers have developed other methodologies ranging from indices to neural networks and automated valuation models (AVMs) (García, Gámez, and Alfaro, 2007; Kok et al, 2017). An in-depth discussion of these alternative methods is out of the scope of this research, but it is important to note that researchers have developed other analytical methods besides hedonic regression to study real estate prices.

After understanding how researchers model real estate values, we can examine the underlying data issues mentioned above. The first key difficulty in using rents to measure agglomeration is finding high-quality data (Rosenthal and Strange, 2004). Though Rosenthal and

Strange pointed out this difficulty over 10 years ago, researchers still face many of the same issues. The primary issue with conducting agglomeration research using rents is the lack of widely-available and consistent real estate fundamental data at the building level. Additionally, elements such as operating expense reimbursement, tenant improvement amortization, and months of free rent make comparability between lease transactions difficult. Compounding this is the fact that there is no central clearinghouse for this data and most companies guard their lease details jealously. Despite a variety of data sources ranging from brokerage houses such as JLL, real estate data providers such as CoStar, and startups such as CompStak, real estate data are obscure, hard to source, and hard to compare.

A second issue with using real estate data is that, even when researchers do work with high-quality data, these data often reflect only "average rents in market areas which obscures the individual office building characteristics that influence rents such as age, size, height, building class, location, and high bandwidth capacity" (Drennan and Kelly, 2011). This issue is similar to the previously-mentioned MAUP, which suggests that the best spatial analysis is conducted with the most granular data possible because it provides the researcher the greatest flexibility in research design and, in the context of real estate research, the greatest reflection of the underlying assets or submarkets under study. Using average rents in a geographic area prevents researchers from pinpointing specific areas of firm concentration or how lease rates vary between individual buildings.

As a result of these issues with data, very few studies have utilized real estate prices to measure agglomeration economics (Puga, 2010; Drennan and Kelly, 2011; Koster, van Ommeren, and Rietveld, 2012; Liusman et al, 2017; Liu, Rosenthal, and Strange, 2018). However, this research could be transformative, as the "identification of relationships between movement, commercial transactions, social activities, and real estate value is a potential game changer that will increase the value and quality of the urban environment" (Donner, 2018). Despite the limited body of work, the

papers that have utilized micro-grained data reflecting individual building suites reveal compelling findings regarding the small scale at which agglomeration occurs. For example, Liu, Rosenthal, and Strange (2018) found that "within-building employment has a much stronger relationship with commercial rent than does zip code employment outside of the building." This suggests that the scale of agglomeration may be a single building; the benefits a firm derives by locating in a building with concentrations of firms and individuals may exceed the benefits of that same firm locating in a neighborhood with many firms. To understand the value of this kind of research, Figure 2.1 below summarizes papers that used building- or suite-level real estate data to quantify the impact and value of agglomeration economies.

Figure 2.1: Summary of Findings from Agglomeration Research Utilizing Real Estate Rents

Author(s)	Year	Finding
Jennen and Brounen	2009	Doubling the number of buildings in an office cluster is associated with a 4.5% increase in commercial rents
Drennan and Kelly	2011	Agglomeration economies are present in the CBDs of larger metropolitan office market, but not in suburban office markets
Koster, van Ommeren, and Rietveld	2012	Doubling job density is associated with a 3.5% increase in office rents
Koster	2013	One standard deviation increase in employment density is associated with an approximately 10% increase in rents
Liusman et al	2017	Mixed-use buildings in Hong Kong have higher rents than single-use buildings
Drennan	2018	Service industries in New York City experience agglomeration economies at small distances, even in the face of external shocks such as 9/11
Liu, Rosenthal, and Strange	2018	Doubling zip code employment is associated with a 10.7% increase in office rent The rent increase associated with moving up one floor in a building has the same effect as adding roughly 3,500 employees to a zip code

2.3 Current State of Real Estate Data

If the previous section has not yet driven home the point that real estate data are crucial to studying agglomeration, another aspect of the data's value is that, as with other financial industries,

"distributed information improves market efficiency" (Florance et al, 2005). This quote from CoStar's 2005 patent for a "system and method for collection, distribution, and use of information in connection with commercial real estate" reflects how increased data availability has positively impacted industries marked by asymmetrical information such as used cars, labor, and insurance (Garmaise and Moskowitz, 2004; Florance et al, 2005). At the same time, it is important to emphasize the notable differences between the real estate market and other financial industries. These include the heterogeneity of assets, the large costs associated with purchasing, maintaining, and improving assets, unique tax treatments, the illiquidity of transactions, the decentralized nature of sales that transact without a centralized clearinghouse, and the inability to short assets (Quan and Quigley, 1991; Garmaise and Moskowitz, 2004; Ghysels et al, 2012).

Despite these differences, there is evidence that more available and accessible data positively impact the real estate market. For example, the existence and use of the online marketplace Craigslist has led to improved market efficiency in residential real estate, including a 10% decrease in average vacancy rates and a three-week reduction in the average time for lease-up of vacant units (Kroft and Pope, 2014). This research proves the significant impact that data and information technology are having on the real estate sector including the development of "large databases…that track sales transactions of properties, both residential and commercial" (Barkham, Bokhari, and Saiz, 2018; Kok, Koponen, and Martínez-Barbosa, 2017). These databases have become so popular that "the collection and aggregation of real estate data have become an industry unto itself with companies such as LoopNet, Real Capital Analytics, and CoStar" storing and disseminating information about individual properties such as rental rate, occupancy, and operating expenses (Winson-Geiderman, 2018). Other services provide information about the performance of real estate funds over time, including the index compiled by NCREIF, the National Council of Real Estate Investment

Fiduciaries (Peng, 2018). Some of these organizations have become huge, publicly-traded companies such as CoStar whose market capitalization, as of May, 2019, is $18 billion.

Despite increased availability of data, there is still a strong need for data standardization and consolidation as well as the development and refinement of data-driven analytical techniques (Barkham, Bokhari, and Saiz, 2018). With new generations of data compilers and analyzers, startups and companies have emerged to bolster real estate research and valuation processes with a "combination of proprietary and public information with real estate implications" (Winson-Geiderman, 2018). Before this, "our collective understanding of how people interact with cities and the built environment has historically been limited to surveys with low spatial and temporal resolution" (Kontokosta and Johnson, 2017). More specifically, Sobolevsky et al (2014) reveal how public data such as "census survey data only capture[s] resident population based on place of primary residence. This obscures the true population density in a neighborhood at any given time, as it ignores temporal fluctuations in worker and visitor populations" (Sobolevsky et al, 2014). To counter this, Ratti et al (2006) used mobile phone data records to better understand the ways in which humans move through and use space. Hashemian et al (2015) also used bank card transactions to understand similar dynamics in Spain. Other cutting-edge data sources include "social media data, police records, and amenities related to economic vibrancy, [all of which] add significant value to pricing models" (Kok et al, 2017). A number of startups and companies that have started using these alternative data sources in their real estate decision-making include House Canary and Zillow (Kok et al, 2017).[1]

With the dramatic increase in availability and usability of both traditional and alternative real estate data, we must also keep an eye on how it will be useful. Donner et al (2018) write that we can

[1] Jennifer Conway's 2018 MIT thesis provides a detailed survey of startups that are attempting to increase the availability, comprehensibility, and usefulness of real estate data. Please see the references for a full citation.

now use "predictive analytics to generate new insights and financial models across a wide range of vertical urban services." They also provide more specifics, including "better forecasts for building utilization, more accurate assessment of the purchasing power of users of real estate, and better risk assessment of real estate users" (Donner et al, 2018). Despite the lauded benefits of using big data, we must also remember that "big data will enable us to run hedonic price regressions with more explanatory variables but it will not on its own enable us to run better-identified regressions" (Glaeser et al, 2018). Data and analytics will help us make better-informed decisions, but these cannot, on their own, make smarter decisions. That will be the work of research such as this thesis.

2.4 Introduction to Midtown Manhattan

Information was, at one time, expensive and time-consuming to transmit. News of the price of grain, international affairs, and new technological developments could be disseminated only by travelling on foot, on a horse, on a boat, or on a train throughout the country. The invention of the telegram and subsequent developments of the telephone and the world wide web eliminated these costs of transmitting information to zero. These developments also made clear that the cost of transmitting knowledge, or analysis based on facts and information, was not the same as the cost of transmitting raw information. Though Frances Cairncross predicted "the death of distance" in 1997 as a result of the rise of the internet, the pace of agglomeration has accelerated in the intervening 25 years, suggesting that distance is not dead at all. In fact, the opposite is true: proximity is power.

So how did New York City become and remain the major information hub that it is today? Before the invention of electronic modes of communication, water was the most cost- and, in many cases, time-efficient way of transmitting information. According to Ed Glaeser (2005), New York rose to prominence because of its waterways; this includes its access to a deep-water harbor, and, eventually, its access to the interior of the country through the Hudson River and the Erie Canal.

New York's location and access to bodies of water that connected it with other parts of the country and the world enabled it to serve as the informational hub for international, as well as regional, news. The rise of the financial sector in New York City emerged directly from this dynamic; Glaeser points out that "the success of the New York financial sector owes a great deal to the ability of New York to be a place where the latest news can be picked up quickly" (Glaeser, 2005). This is because "in no other industry are the returns to knowing the latest fact greater, this meant that once New York had a slight edge, this slight edge turned into a complete preponderance as the financial community came to the city to get access to the latest information" (Glaeser, 2005). Thus, New York City emerged as a financial hub and has remained so to this day as a result of the high value of instant access to the most up-to-date knowledge.

With this brief history of the rise of New York City as an information and knowledge hub, what does Midtown look like today? The submarket is home to major train stations such as Grand Central Terminal and Pennsylvania Station, as well as subway stations that saw well over 100 million riders in 2017 (the most recently-available data); that year, Grand Central's 4-5-6-S station alone saw nearly 45 million passengers (MTA, 2017). In addition to various forms of transportation, the submarket is home to many of the world's largest financial institutions including JP Morgan, Morgan Stanley, and CitiGroup. Iconic buildings in the submarket include the Chrysler Building and the Empire State Building. Yet the area is not merely an office park, but houses attractions for visitors and residents alike including museums, hotels, sports arenas, Times Square, Broadway and the southern edge of Central Park.

As a result of the multitude of ways to experience Midtown Manhattan, I must describe how it will be used and mapped in this thesis. As discussed earlier in the chapter, part of the challenge in using administrative data in conjunction with industry-specific geographies is that they do not always match up. This is the case in real estate economic research, where administrative data are maintained

at the zip code or census geography level, while the real estate industry uses submarkets that do not map neatly to those units; submarkets encompass many zip codes but cannot be mapped exactly to a specific subset because their borders do not perfectly match up. Further complicating the issue is the fact that brokerage houses utilize different geographic delineations for their submarket definitions. In order to determine which zip codes to include as part of the analysis, I relied upon JLL's geographic definition of the Midtown submarket, which stretches from the East River to the Hudson, between 30th Street and 65th Street. I then selected the zip codes that corresponded to, but sometimes extended beyond the boundaries of Midtown Manhattan, as described by JLL. Please see Figure 2.2 for an overview of the Midtown zip codes' locations within Manhattan, a more detailed map zooming in on Midtown, and the image from JLL's market report outlining the boundaries of Midtown.

Figure 2.2: Overview & Detail of Manhattan and Midtown Zip Codes

Source: JLL

2.5 Financial Technology Overview

To conclude this chapter, I delve into the financial technology, or "FinTech," industry, a broadly-defined segment of the startup world that touches every aspect of the economy involving

monetary transactions. FinTech includes insurance, payments, regulatory software, real estate equity and lending, as well as wealth management, personal finance, and capital markets infrastructure (Lindsay, 2019). The largest FinTech startups, as of 2019, include Stripe, a payment-processing firm headquartered in San Francisco and valued at $22.5 billion, Coinbase, a crypto-currency company valued at $8.0 billion, and Robinhood, a free brokerage app that is valued at $5.6 billion and headquartered in Menlo Park, CA (Kauflin, 2019). This preponderance of California-based startups is not an illusion: eight of the ten largest FinTech startups, valued at a combined $59.2 billion, are headquartered in the San Francisco Bay Area. One of the ten largest is located in Boston (Circle), and one is based in Chicago (Avant). Despite their geographic concentration, these companies span the breadth of the FinTech world, ranging in focus from payroll and human resources that impact the lives of individual workers (Gusto) to Capital Markets Infrastructure that helps banks and other financial institutions secure and optimize their software (Plaid).

There are a number of trends shaping the direction of this sector including the rise of open banking, broadening product offerings, new and updated asset classes, and the dissemination of technology throughout the real estate value chain. Open banking means that, instead of each bank providing its information over proprietary frameworks and networks, consumers, companies, and governments can access a given bank's data through open APIs, rather than the proprietary networks that banks currently use to deliver customers information. With increased standardization and openness of the API frameworks, entrepreneurs will develop services that consolidate data streams from a variety of financial services and serve this data to consumer and business customers alike. This means an application will have data from a variety of banks, financial institutions, and other services. The idea has gained traction in Europe, South America, and Southeast Asia. Yet many banks and financial institutions, especially in the United States, remain wary of implementing open standards for two reasons: security concerns and business model concerns. Security is always a

concern in the financial services industry and opening networks to outside app developers understandably makes many organizations nervous. From a business model perspective, banks do not want to lose customers to outside parties so they have incentives to keep consumers using their services through proprietary frameworks and networks. (Lindsay, 2019).

The second major trend in FinTech is that companies are broadening the products that they provide customers and clients. In the consumer FinTech space, this has manifested in two major trends: companies trying to integrate as closely as possible with payroll services and companies starting to offer debit cards. They want to integrate with paychecks for an obvious reason: this creates recurring cash flow through their system from which they derive benefit. By investing the money until the end user needs it, these companies can earn return on the money housed temporarily within their ecosystems. By offering debit cards, companies are also engraining themselves in consumers' daily lives. From a business perspective, FinTech startups can also charge higher fees on debit cards than banks, thanks to the Dodd-Frank Act's Durbin Amendment (2010), which limits the fees that banks, but not other institutions, can charge for debit cards.

In addition to new products, FinTech companies are opening access to asset classes that were previously not investable or available only to the extremely wealthy. These include luxury cars, boats, music, municipal bonds, and even art. One existing asset class that has seen a dramatic influx of FinTech startups is real estate. There are a number of reasons for this including the magnitude of dollars involved in the industry. For example, US household mortgage debt in 2018 exceeded $9 trillion, providing a huge addressable market for startups. For context, the next largest major household debt category by total dollar size is student loans, at $1.4 trillion. Auto loans total $1.25 trillion, and credit card debt is the next largest category at $850 billion (Lindsay, 2019). Another reason startups are drawn to real estate is the opportunity to transform the way the industry conducts business. There has been limited adoption of technology into real estate companies'

systems, so the majority of leases and other operational documents are maintained in paper files. As startups realize the opportunity to transform a huge industry, they are building tools and services that integrate technology into real estate processes, lending, auctions, and analytics. But emerging financial innovation is not solely aimed at real estate companies. There are also new financial products emerging to help consumers with housing, ranging from new home equity lines of credit to rent-to-own services.

Because of the vast opportunities in the industry, the FinTech sector is experiencing a significant boom in the number and dollar volume of venture capital deals. In 2018, there were just over 1,700 FinTech deals, representing a 15% increase over 2017 and a 93% increase over 2014 (Lindsay, 2019). These deals represented a collective $39.6 billion, a 120% increase in dollar volume over 2017 and a 374% increase in dollar volume over 2014 (Lindsay, 2019). This dramatic increase in the dollar volume of venture funding masks one enormous funding round of $14 billion, raised by Ant Financial, a Chinese FinTech company (Lindsay, 2019). Excluding that outlier, FinTech companies raised $25.6 billion, a 42% year-over-year increase, more in line with the 36% average annual growth rate between 2014 and 2017 (Lindsay, 2019). Along with secular growth in funding, the number of $100 million or larger venture capital rounds has accelerated, increasing by 44% between 2017 and 2018 (Lindsay, 2019). This is especially pronounced among FinTech unicorns, companies valued at $1 billion or more. As of 2017, there were only 23 FinTech unicorns but by the end of 2018, that number almost doubled to 39, valued at a collective $147.4 billion (Lindsay, 2019).

Despite the broad growth opportunities that are driving the expansion of FinTech, there are a number of industry headwinds including data security and limited paths to exit. Data security is a key concern for any web service, but because FinTech companies involve individuals', companies', and governments' money, the standard of scrutiny is much higher. I will not detail the issues around data security, since they are common to other sectors, if somewhat more pronounced in FinTech.

The second industry hurdle is the fact that the two traditional paths for firms to exit are currently staunched. The first usual way for startups to exit is for a large, existing firm to acquire it. This is not happening to FinTech startups, as existing financial institutions are instead creating low-stakes partnerships with FinTech startups. Established firms do this to test the value of a startup's service while still maintaining the flexibility of testing other startups and not committing significant capital to a single acquisition. This provides FinTech startups valuable partnerships but often stymies their ability to provide their investors the large returns venture capitalists seek.

In addition to the lack of exits through acquisition, there have been fewer IPOs of FinTech startups than startups in other industries. In 2018, only three FinTech unicorns (private startups with a valuation in excess of one billion dollars) went public (Lindsay, 2019). This dynamic results directly from the fact that there are an increasing number of significant venture capital funding rounds, meaning FinTech startups do not have to raise money through the public market. In 2018, there were 52 rounds of fundraising that provided FinTech companies $100 million or more (Lindsay, 2019). With increased cash on their balance sheets, FinTech startups are able to stay private longer. Because both traditional paths to exit are currently underutilized, there is a risk that venture capitalists who would like to realize their investments are unable to do so. Without making their desired return, they may not be able to continue investing in early-stage deals, and thus strangle the next generation of FinTech startups currently emerging.

In conclusion, FinTech companies are seeing a dramatic increase in the number and size of funding deals, as well as more partnerships with existing financial institutions. The standardization of financial API frameworks will facilitate increased innovation and adoption, as will startups' focus on providing customers with better user experiences. Weaknesses in the industry, however, include data security concerns paired with difficult paths to exit. Financial technology is a fascinating, rapidly growing part of the startup ecosystem. By studying it through the lens of agglomeration in

conjunction with existing firms, we can begin to understand how New York City's traditional role at the center of the financial services industry is evolving in the face of 21st century technological change.

Chapter 3

Data & Methodology

This thesis utilizes three main data sets in order to understand the agglomeration economies of financial services firms in Midtown Manhattan. This chapter describes the data sources, outlines the cleaning process, and relates the results of exploratory data analysis. I also detail how I consolidated the disparate, cleaned data sets into a single file that will serve as the basis for my regression models. In addition, I introduce the way I will use regressions to determine the presence and magnitude of agglomeration economies. The final section of the chapter documents additional data sources that would be necessary to uncover the underlying reasons that individuals and firms agglomerate, an aspect of agglomeration economics that economists have not yet been able to uncover.

3.1 Data Sources

3.1.1 CompStak

CompStak is an eight-year-old startup based in New York City with an innovative business model. Their crowdsourced marketplace for real estate information enables platform users to contribute information to CompStak's system in exchange for credits used to purchase other information that the users wouldn't otherwise know. To address the potential issue of data quality, CompStak's users "up-vote" and "down-vote" information based on its accuracy, ensuring the most reliable information possible for all users. This has resulted in one of the most comprehensive, well-structured databases of commercial real estate that contains a consistently calculated effective rent. This last attribute is the crucial factor that enables me to model the presence and impact of

agglomeration, since it is the dependent variable in my regression. The actual methodology used to determine agglomeration will be discussed in-depth later in the chapter.

The raw CompStak data included 7,178 observations, consisting exclusively of new lease transactions within Midtown Manhattan executed between 2009 and 2019. This data set was relatively clean and only required a few minor tweaks. First, I converted columns that contained the lease term and months of free rent from text columns to numeric ones. The term column contained the words "years" and "months" to describe the lease term, rather than a real number such as 5.5. The free months column contained the word "months" at the end of each number, rather than an integer value.

The CompStak data also specified whether a lease exists only on one floor or comprises space on multiple floors. Recognizing that higher floors cost more (Koster, van Ommeren, and Rietveld, 2014), my second feature engineering operation determined how to incorporate the presence of multiple floors. I solved this in three ways: first, I created one column that took the value of the highest floor on the lease (for a lease that occupies floors 2 through 10, this method would assign the record a value of 10). The second column I created took on the average value of all floors contained within the lease (using the same theoretical lease floor range, this method assigns the record a value of 6). Because the data did not provide the number of square feet on each floor, it was impossible to create a weighted average of floors, so a standard average had to suffice. Third, I created a column that took on a binary value to signify whether a lease was contained on a single floor or had space on multiple floors. Because of concerns with multicollinearity (strong linear relationships between independent variables in a regression), I ultimately relied solely on the third column when specifying the actual regressions.

The column in the original data that specified which floors a lease contained also specified whether a lease included ground floor space, part of a basement, or space on a mezzanine or

concourse. Recognizing that the effective rent accounts for all floors in a given lease and that these types of spaces can increase rents (ground floor often contains retail, which commands higher rent) or decrease them (subterranean basements are often used for storage and are often much cheaper), I created a binary column for each of the four 'special' floor types: ground, mezzanine, basement, and concourse. These variables would allow me to parse out the marginal impact of these spaces on the effective rent.

In a final cleaning step, I created a binary variable to categorize the tenant as financial services or non-financial services. The CompStak data contained an industry classification column, but it was sparsely populated; 80.7% of the entries contained blank values. Thus, I created several text-based filters that I applied to the tenant name to categorize that company as financial services or not. I labeled tenants as financial services if their tenant category included "Banks," "Real Estate," "Financial Services," or "Insurance." If the tenant name included "Mortgage," "Securities," "Capital," "Bank," "Wealth," "Broker," "Realty," "Real Est," "Invest," "Insur," or "Financ," I also categorized them as a financial services firm.

After cleaning the data, it is important to dig into the trends at a building level as well as at the lease level. In order to do the former, I eliminated duplicates in the address field, ensuring that each building was represented only once in the data set. This resulted in a set of 1,084 unique buildings representing over 300 million square feet. An important factor in the calculation of rents is the class of the building, and, although there are no hard rules defining a building as Class A, B, or C, the categorization reflects qualitative aspects of the building such as construction year, renovation year, ownership and management, the status of its building systems including HVAC, and the presentation of its lobbies and elevators. As we see below in Figure 3.1, there are the greatest number of Class B buildings, but Class A buildings contain the greatest number of square feet.

Figure 3.1: Number and Square Feet of Buildings by Class

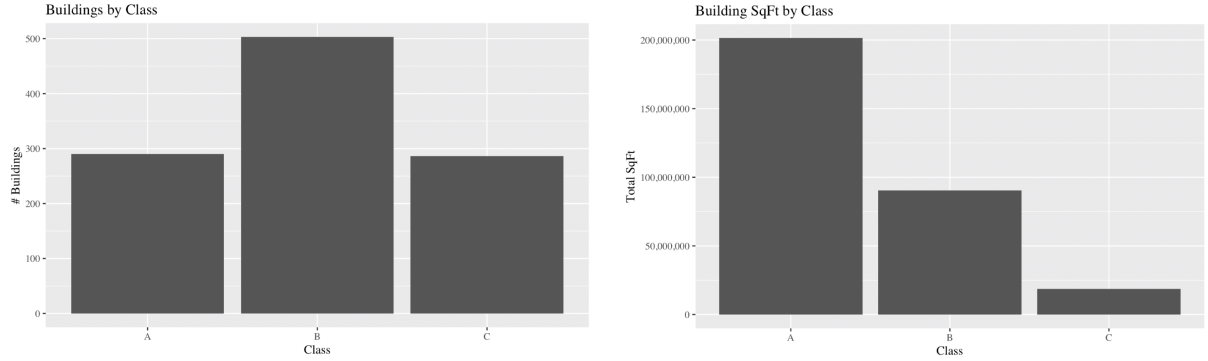
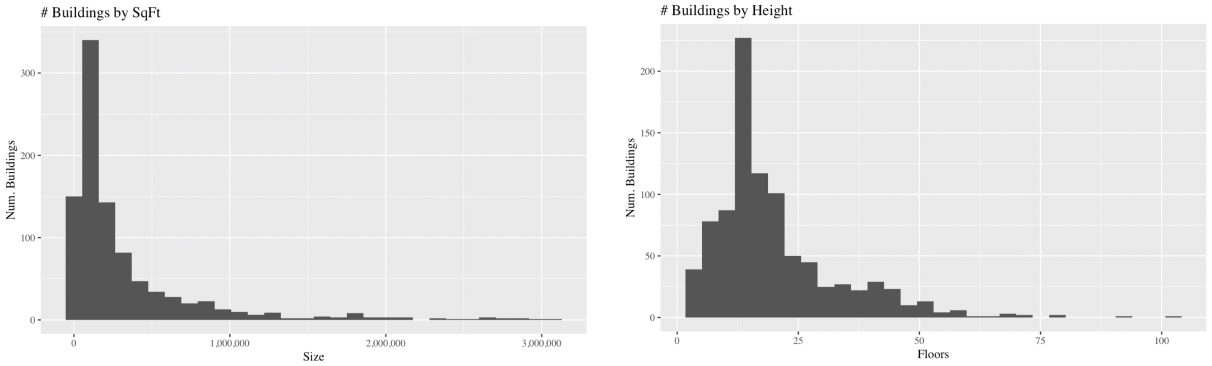


Figure 3.2 shows the frequency distribution of buildings by their number of square feet and their height, both of which show very skinny and right-skewed distributions. The summary of these two distributions is included in the table as part of Figure 3.2

Figure 3.2: Distribution of Buildings by Square Feet and Number of Floors

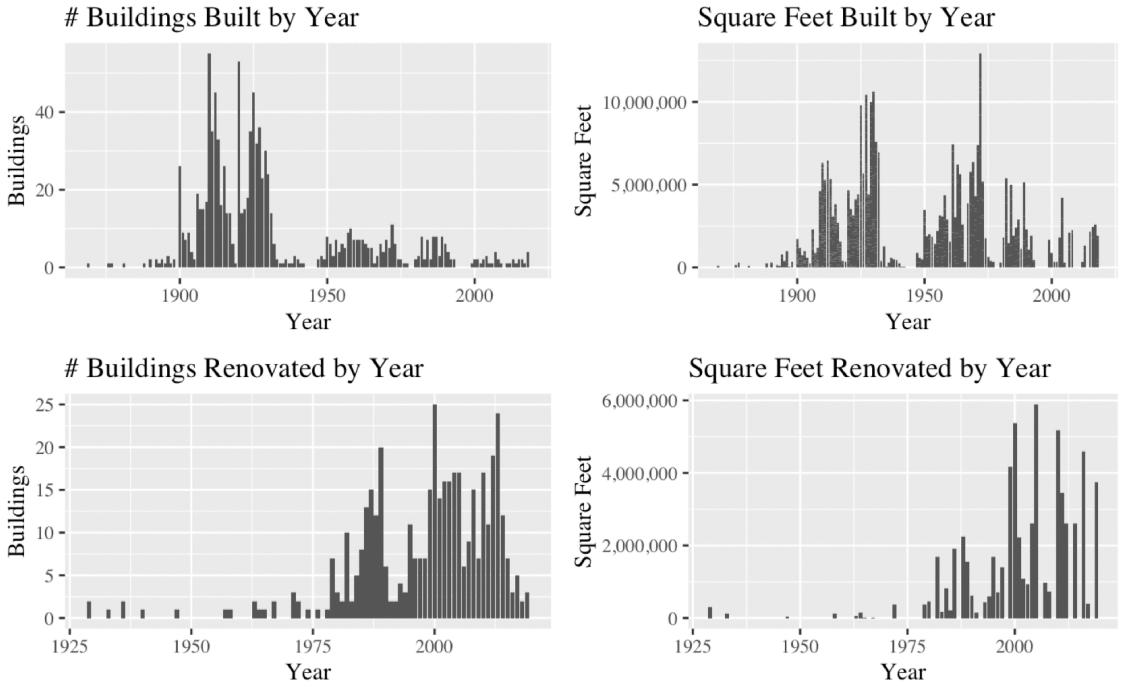


Distribution	Min	Median	Mean	Max	Std. Dev.
Square Feet	2,376	150,418	328,430	3,078,513	463,899
Floors	3	16	20	102	13

Figure 3.3 reveals the number of buildings and the sum of their total square feet arranged by year of construction and year of renovation. This figure accurately reflects the large inter-war building boom, followed by a post-WWII boom that contained fewer buildings but more square feet, as technological advances allowed developers to build taller buildings with larger floorplates. Regarding renovations, we see that most have occurred since 2000, with large peaks every couple of years,

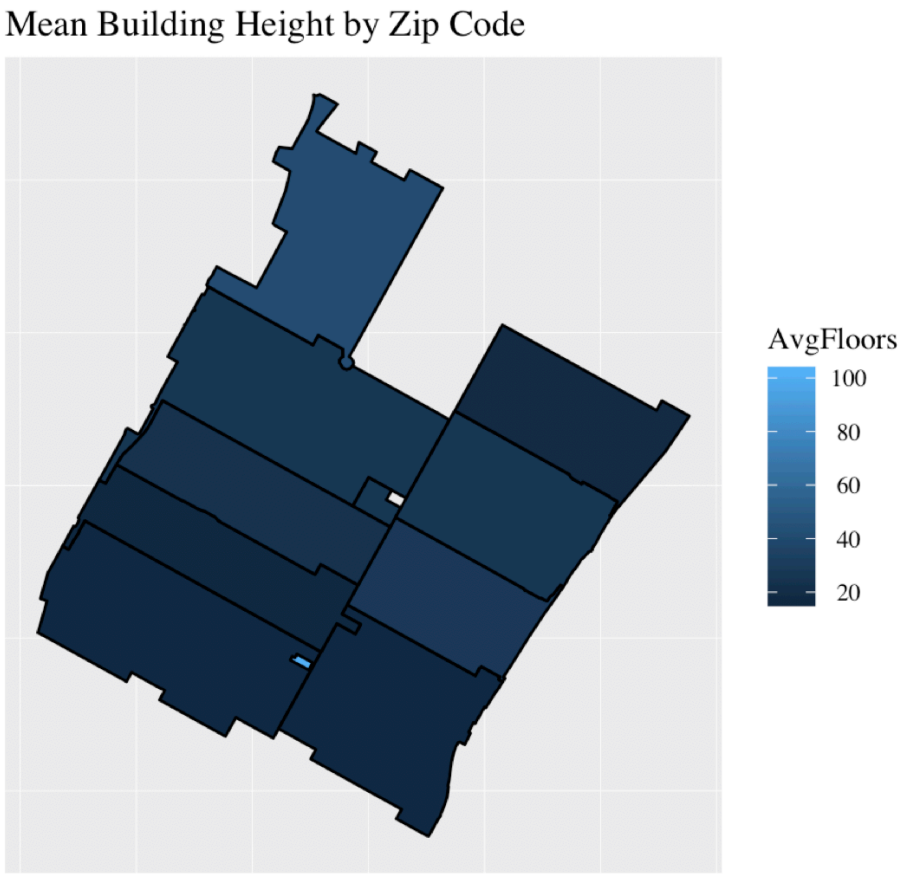
reflecting the seven- to ten-year hold periods of private equity firms that purchase and renovate these buildings.

Figure 3.3: Number and Square Feet of Buildings Constructed and Renovated by Year



Although these high-level trends are informative, it is even more beneficial to examine these trends geographically so we can understand how the vary over space. Figure 3.4 shows the mean building height by zip code. At first glance, mean height seems to be relatively consistent across the submarket. This is because there is a small zip code towards the southern edge of Midtown that has the highest average height and skews the sample. This zip code consists wholly of the Empire State Building, which dwarfs all the other office buildings in the submarket.

Figure 3.4: Mean Building Height, Construction Year and Renovation Year by Zip Code



As mentioned before, zip codes are useful regions of analysis because they are consistently delineated but become less useful when they do not map to underlying economic reality, such as in real estate research. Thus, CompStak's submarket column also provides an additional filter through which I can consider the height of buildings in Midtown. Comparing the previous map with Figure 3.5 which shows the building heights of individual buildings colored by CompStak's submarket attribute, we see how aggregation at the zip code level hides two important trends. First, we see that some submarkets and some classes, in general, contain taller buildings or shorter buildings than other submarkets or other classes. For example, Hudson Yards has a preponderance of taller buildings, while Chelsea is dominated by shorter ones. This makes intuitive sense, as Hudson Yards is a recently-developed, master-planned neighborhood, while Chelsea has developed over New York

City's history. In the same figure, we see that Class A buildings have a much larger average height than Class C buildings, which also makes sense since Class A buildings tend to have been constructed more recently. The second finding from Figure 3.5 is that, within a given submarket or class, there is also significant variation in the height of buildings. For example, the Empire State Building, which skewed the zip code analysis in Figure 3.4, shows up again here as an outlier within its own submarket of Murray Hill. Similarly, Class A buildings vary significantly in height. Variability between and within submarkets and classes limits the rigorous performance of statistical methods and points to the value of qualitative analysis and considering the built environment factors in future economic geography studies.

Figure 3.5: Building Heights by Submarket and Class

The construction and renovation maps in Figure 3.6 show trends that are the inverse of one another, with the most recent construction located in the northwestern corner but more recent renovation everywhere else in the submarket. It is important to note that these mean construction and renovation year maps do not reflect a building-square-feet-weighted average, so these maps may be obscuring some large-scale, recent construction such as Hudson Yards. This again, points to the value of considering built environment factors when preparing data for econometric analysis in the real estate sphere.

Figure 3.6: Mean Construction Year and Renovation Year by Zip Code

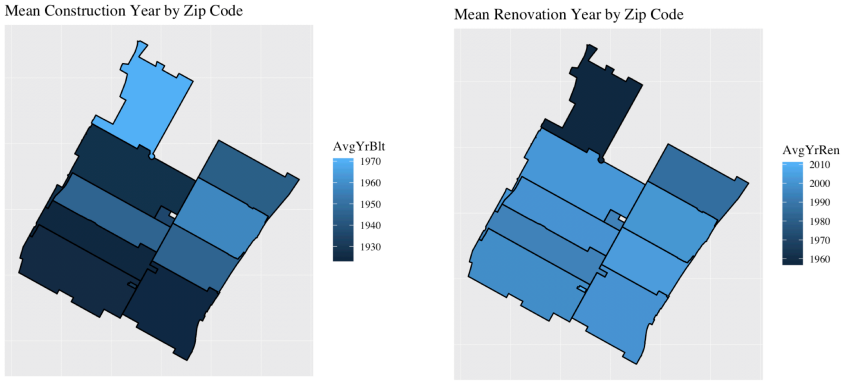


Figure 3.7, the penultimate map describing building-level statistics, shows the total number of square feet in each zip code as well as the total number of square feet of each building class within each zip code. We see Class A buildings distributed throughout the submarket, though the greatest number of square feet are in the zip codes containing and nearby Grand Central Terminal. Class B and Class C buildings, by contrast, are concentrated in the southwest corner, closer to Midtown South.

Figure 3.7: Square Feet per Zip Code by Building Class

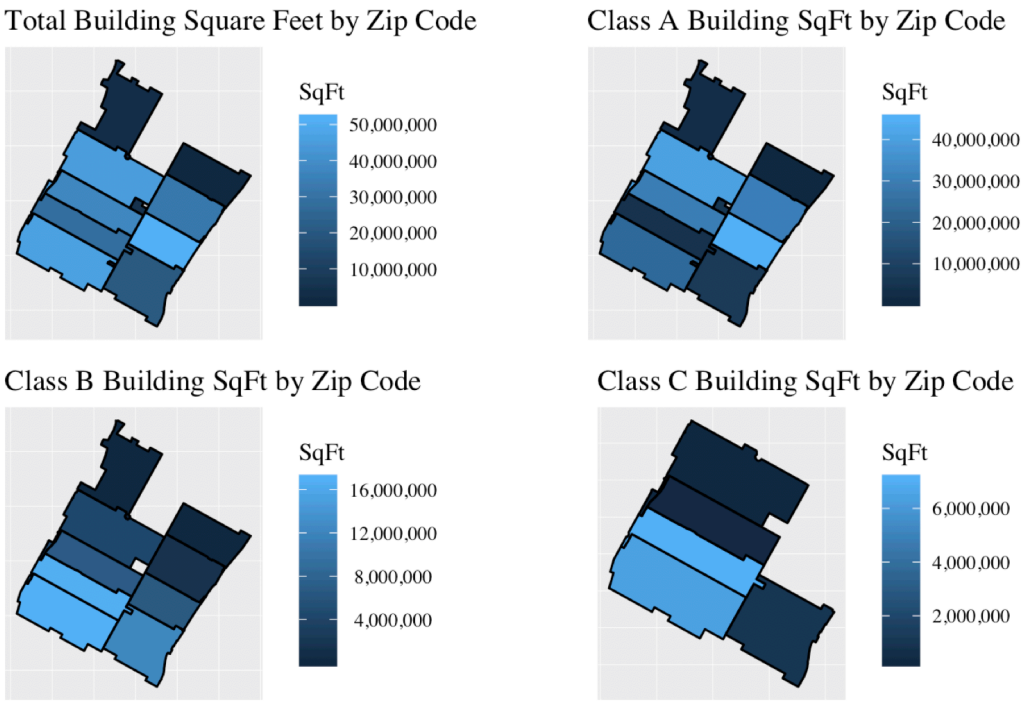


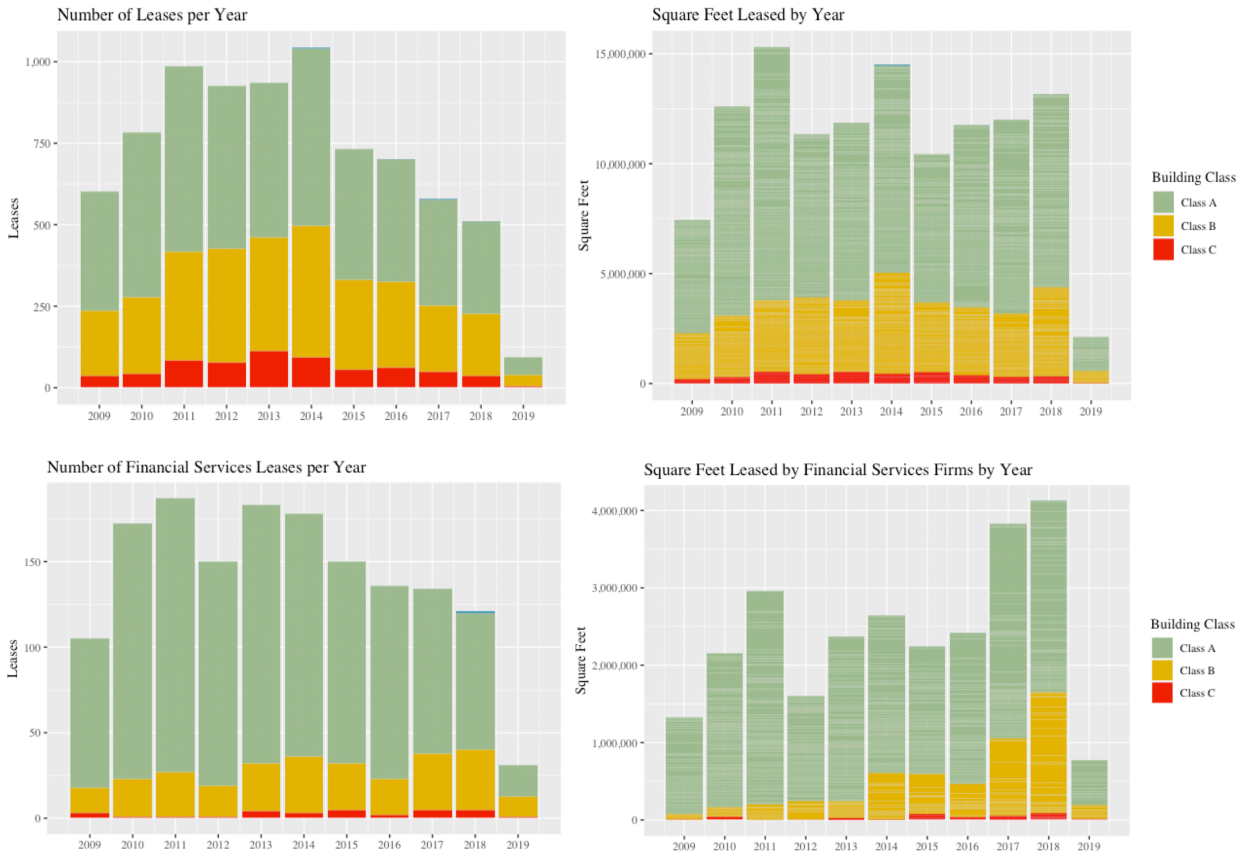
Figure 3.8 provides additional, granular insight into these trends, showing the variability of building sizes between submarkets and classes as well as the variability of building square feet within a given submarket or class. Similar to the differences between Figures 3.4 and 3.5, we see significant variation both between and within submarkets. For example, Hudson Yards is dominated by much larger buildings, while, on average, the buildings in Times Square South are much smaller. For within-submarket variation, look again at Murray Hill, where the Empire State Building is significantly larger than any other asset in the submarket. We also see significant variation within and between classes in Figure 3.8. These trends reflect the age of the buildings and certainly have an impact on where financial services firms and startups choose to locate. As such, future economic geography should emphasize aspects of the built environment as they specify studies in order to account for this endogenous variability.

Figure 3.8: Size of Building by Submarket and Class

After looking at building-level trends, we can return to the full CompStak data set to examine financial services leasing activity over the 2009 – 2019 study period. Figure 3.9 aggregates the total square feet leased by financial services firms in each year, broken up by building class,

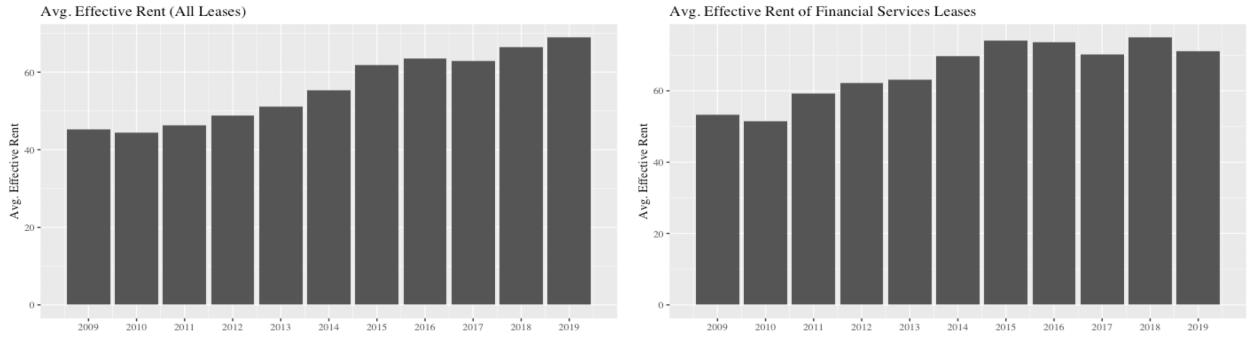
where we see the preponderance of activity across all years took place in Class A buildings. This is especially pronounced among financial services leases, which occupy the lower row of Figure 3.9.

Figure 3.9: Number and Square Feet of Leases by Year and Class



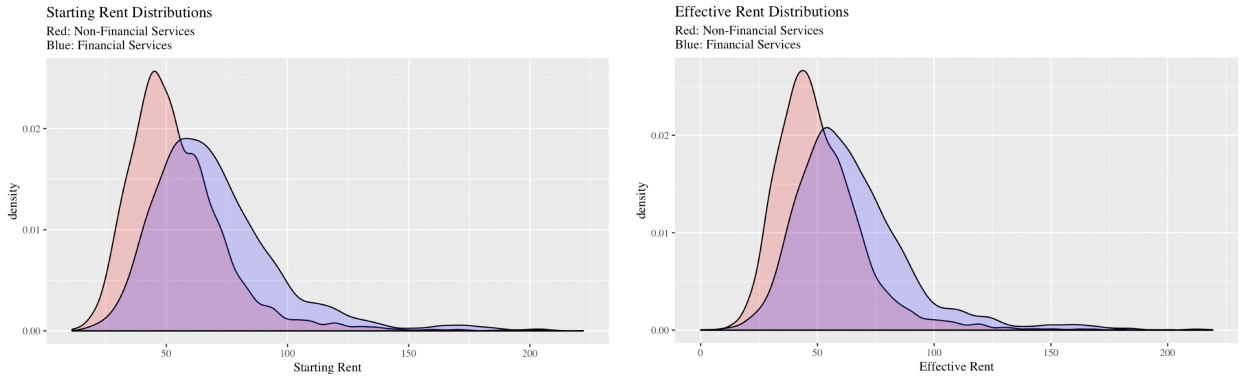
With Figure 3.10, we see the growth in rent over the study period. Average effective rents dropped from $45.32 in 2009 to $44.48 in 2010, a function of the Great Financial Crisis, then rose uninterrupted through 2016 to $63.46. 2017 experienced a slight decline in average prices to $62.95, but then rent picked up again through 2019 to reach $68.97. For financial services firms, we see similar declines into 2010 paired with subsequent growth. From a 2010 nadir of $51.43, rents grew to their zenith in 2015 of $74.14. Since 2015, rent has vacillated but stands at $71.08 as of March, 2019, when this data set was downloaded from CompStak.

Figure 3.10: Average Effective Rent 2009 – 2019



Though we are primarily interested in financial services leases, it is important to compare their distributions with non-financial services firms, as we have been doing with the leasing trends up to this point. Figure 3.11 reveals the differences in effective rent between leases to non-financial services companies and financial services firms. Financial services firms have a higher mean effective rent ($65.05 versus $51.12 for non-financial services companies) and fatter distribution than non-financial services (standard deviation of $25.03 versus $19.69 for non-financial services firms).

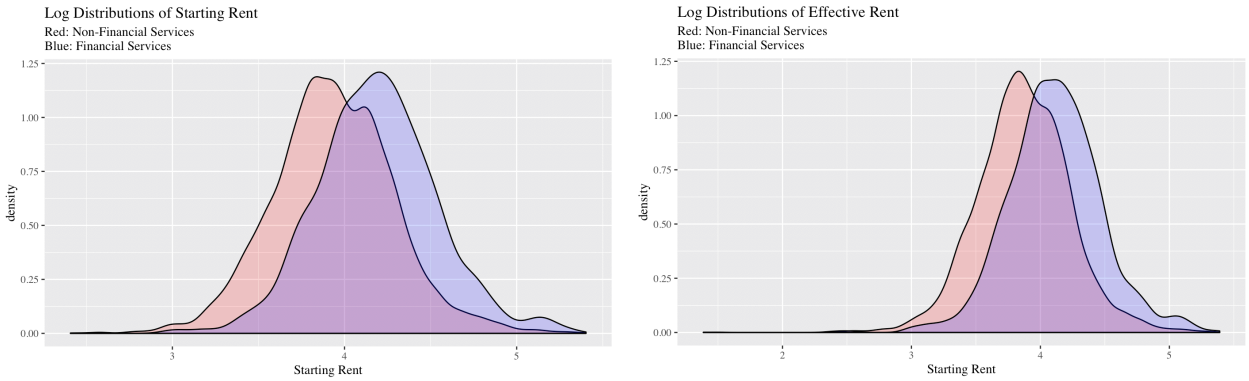
Figure 3.11: Distribution of Starting and Effective Rent



But because these distributions are skewed, we must log transform them. We do this to ensure our dependent variable is normally (or nearly normally) distributed, which ensures the unbiasedness of the coefficient estimates in our regression. Figure 3.12 below shows the log

transformed distributions of effective rent, revealing that they are now much closer to a normal distribution.

Figure 3.12: Distribution of Lease Sizes and Terms by Class



We can also categorize the leases in terms of their size and duration. Figure 3.13 reveals the distributions of leases by size and term, split up by class of building.

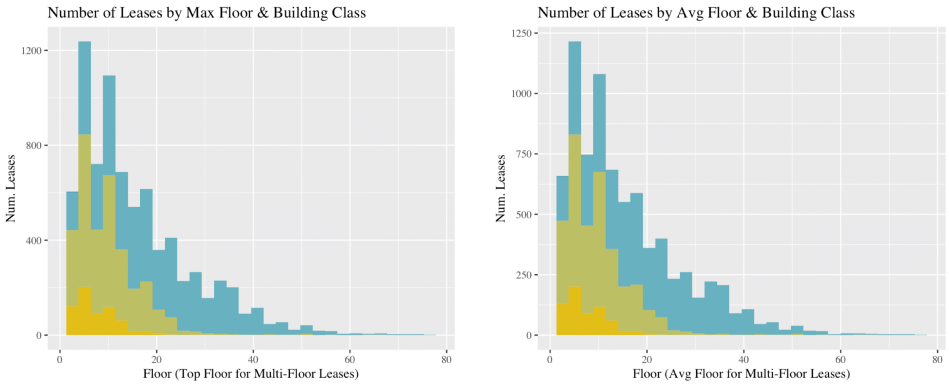
Figure 3.13: Distribution of Lease Sizes and Terms by Class

Distribution	Min	Median	Mean	Max	Std. Dev.
Square Feet	100	6,630	15,526	902,000	35,691
Term	1	7	7.3	51.8	3.9

One of the tricky aspects of working this data set was, as described before, how to account for the fact that many of these leases contained multiple floors. Figure 3.14 compares the two

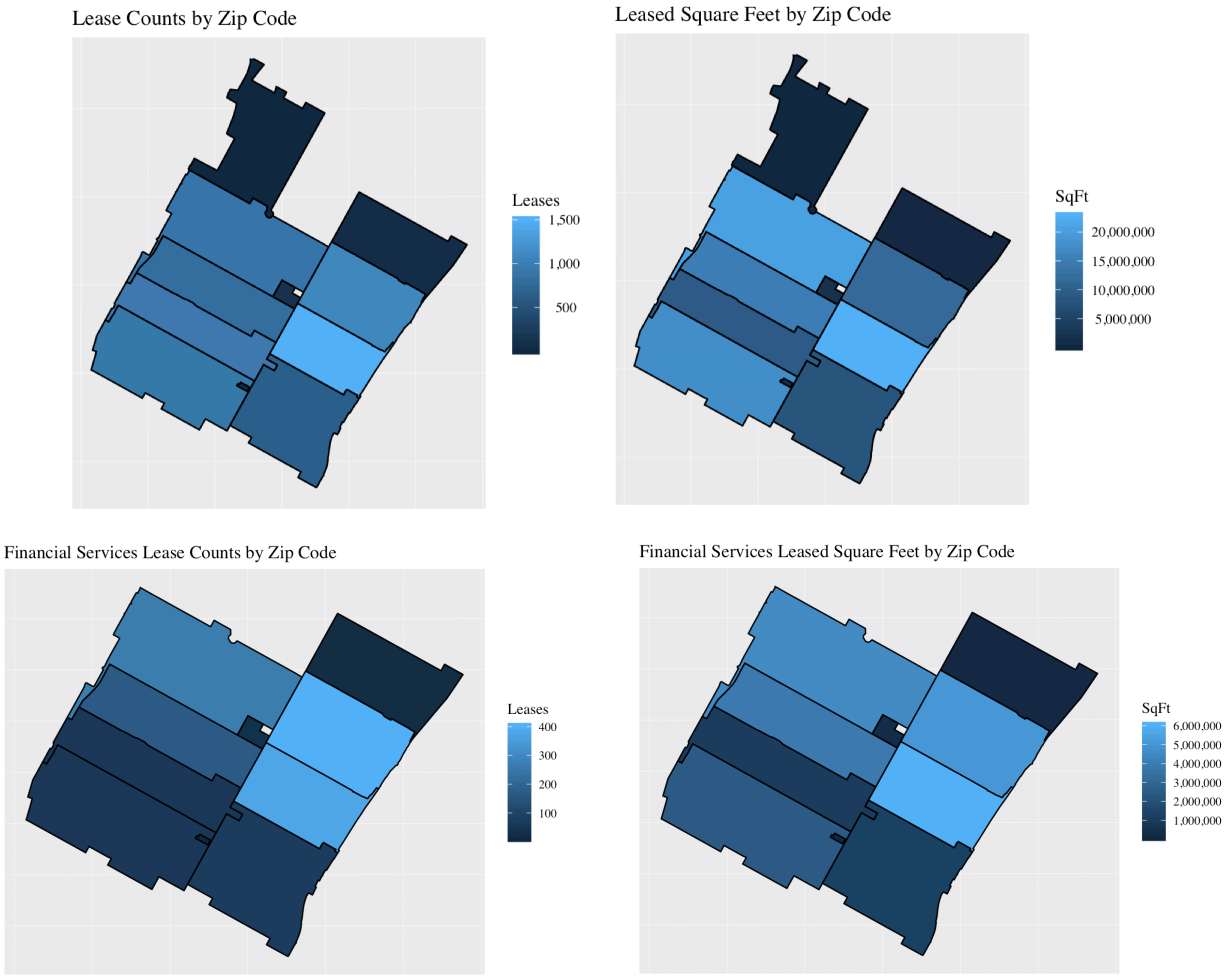
methodologies: assigning the floor to the highest floor in the range and assigning it to the average floor of the range.

Figure 3.14: Distribution of Multi-Floor Metrics



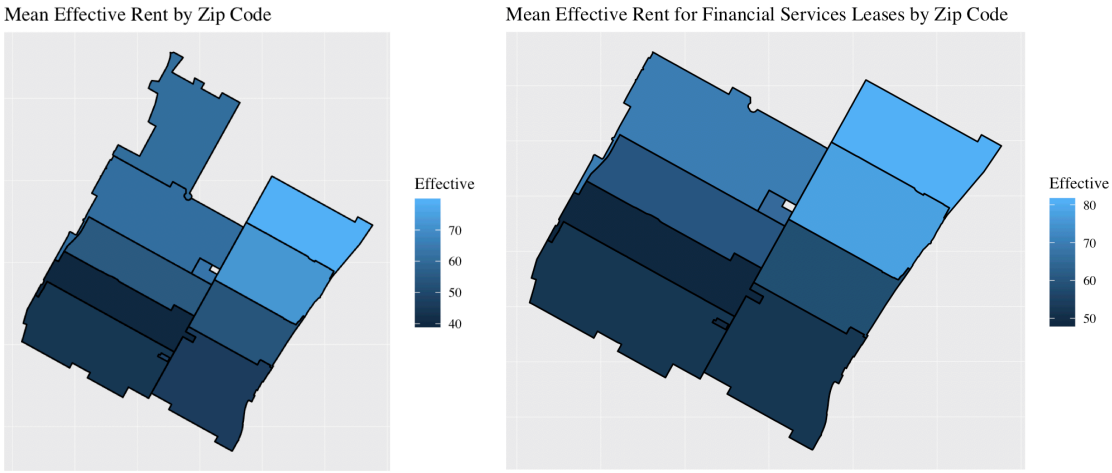
Similar to the other data sets, we can map these data to understand their distributions and dispersal throughout Midtown. Figure 3.15 shows the number and square feet of CompStak leases in the submarket. Please note that there are no financial services leases in zip code 10023, in the northwestern corner of Midtown, hence the maps display slightly differently.

Figure 3.15: Lease Counts and Square Feet by Zip Code



Finally, Figure 3.16 shows us that the highest mean effective rent is in zip code 10065, while the lowest rent is found in the southwestern corner of the city, where we see the majority of Class C buildings. The fact that rents rise nearly perfectly monotonically in a northeastern direction suggests spatial autocorrelation, which will negatively impact the validity of our regression estimates and emphasizes the need for more in-depth, urban-focused research that could explain and quantitatively adjust for this phenomenon.

Figure 3.16: Mean Starting and Effective Rent by Zip Code



3.1.2 Dun & Bradstreet

Dun & Bradstreet is a large, well-known business information database subscription service that MIT provides to its students. The company researches and provides information about businesses of all sizes, from mom and pop stores in rural America to global Fortune 500 companies in major cities. To narrow down these potentially broad searches, users can filter their queries to make requests more specific with geographic filters such as, city, state, zip code, and metropolitan area, as well as business-specific filters such as NAICS code or whether a company is publicly listed on a stock exchange. For this analysis, the crucial information I extracted includes the company location, its annual sales, and its number of employees. These figures become independent variables in the regression that determines the presence of agglomeration. If, for example, an increase in the number of companies or any of these other factors is associated with an increase in effective rent, we will be able to conclude that agglomeration is present. To ensure I gathered information only about financial services firms in Midtown, I narrowed the analysis by limiting the searches to firms located in New York, NY that are categorized as three-digit NAICS codes 522 and 523. These

NAICS codes correspond to Credit Intermediation and Related Activities and Securities, Commodity Contracts, and Other Financial Investments and Related Activities, respectively. To understand the more detailed and descriptive business functions associated with each company listed within my data, I detail in the table below all of the four-digit NAICS codes under 522 and 533, their corresponding category names, and summary statistics.

Figure 3.17: Summary Table of Four-Digit NAICS Codes

Four-Digit NAICS Code	**Business Function**								
5221	Depository Credit Intermediation								
5222	Non-Depository Credit Intermediation								
5223	Activities Related to Credit Intermediation								
5231	Securities and Commodity Contracts Intermediation and Brokerage								
5232	Securities and Commodity Exchanges								
5239	Other Financial Investment Activities								
Four-Digit NAICS Code	**#**	**Sales ($MMs)**				**Employees**			
		Min	**Median**	**Mean**	**Max**	**Min**	**Median**	**Mean**	**Max**
5221	1,020	0.00	0	208.95	99,624.00	0	12	562	252,539
5222	873	0.00	0.24	75.33	43,281.00	0	4	129	55,000
5223	318	0.00	0.16	3.75	441.10	1	3	20	592
5231	1,243	0.00	0.38	84.60	43,642.00	1	5	233	57,633
5232	48	0.00	0.63	46.97	1,110.22	1	10	178	2,902
5239	3,419	0.00	0.14	21.30	27,499.40	1	2	69	24,353
Complete Data Set	6,921	0.00	0.16	66.51	99,624.00	0	3	177	252,539

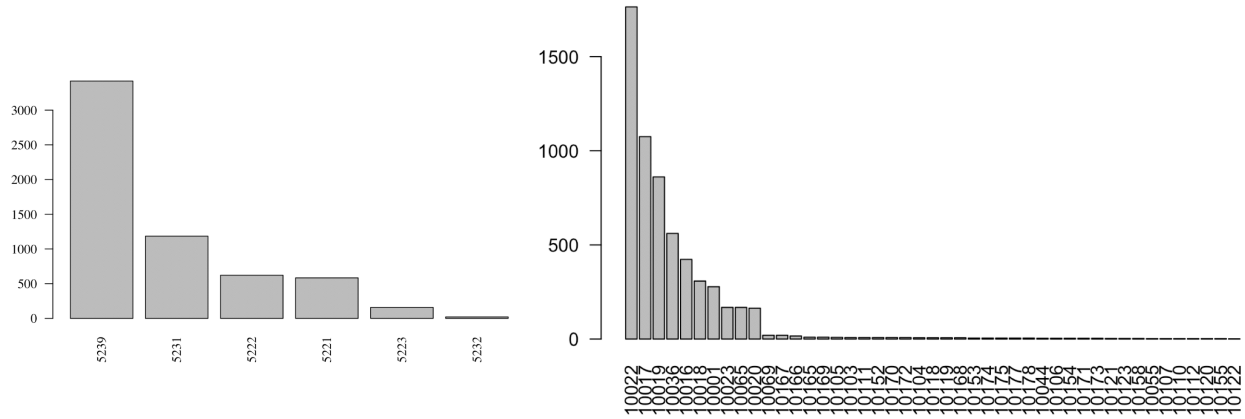
Because Dun & Bradstreet's output includes the address of each business, geocoding these points with the help of ArcGIS's built-in geocoder was simple and straightforward. The geocoder assigns each address a score between 0 and 100, corresponding to the confidence ArcGIS has in the accuracy of its assignment. When a geocoder is unable to locate an address, it gives that entry a score of 0. In the case of the Dun & Bradstreet data, the raw data contained 7,140 total rows of data, 219 of which were assigned a score of 0 during the geocoding process. Upon further examination, the reason for these scores of zero was Dun & Bradstreet did not include those companies' addresses in

its data. Removing these left 6,921 companies with addresses to begin cleaning at the zip code level, which is how I will ensure they are in Midtown.

There was one company with NA as its zip code and 11 companies with 0 as their zip codes. The one with NA as its zip code did have an address that was not picked up by the geocoder, so I added that manually. The companies with 0 as their zip codes also had addresses the geocoder mislabeled as belonging to zip code 00000, so I was able to manually adjust those to reflect their actual zip codes. After ensuring that all of the entries could be mapped to a zip code, I narrowed down the data to those with zip codes actually mapping to zip codes within Midtown Manhattan.

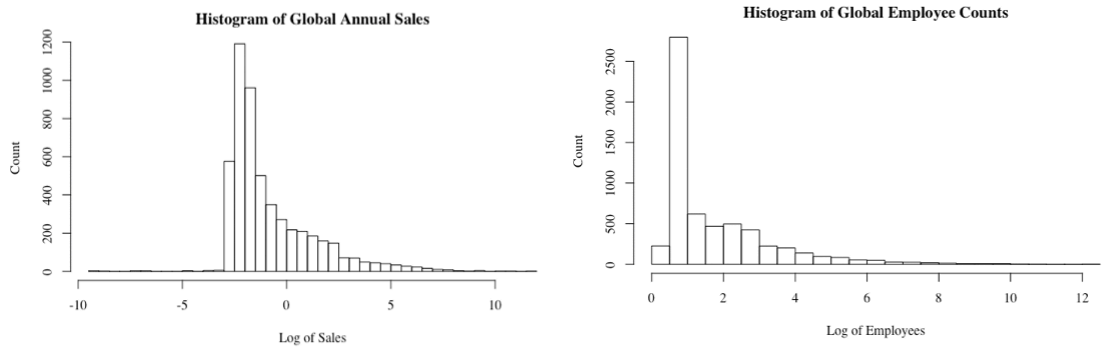
With the Midtown entries isolated, I could conduct preliminary analysis on the types of financial services companies located in Midtown. First, I wanted to understand which NAICS codes were most common. As shown in Figure 3.18 below, 5239 was by far the most common with 3,419 companies, followed by 5231 (1,185), and 5222 (622). This is unsurprising because 5239 represents "Other Financial Investment Activities" – a catch-all category – so any company that does not fit neatly within a specific NAICS code would be aggregated into this one. At the same time, I wanted to understand how many companies are located in each zip code. Also shown in Figure 3.18, the preponderance of financial services firms is in 10022 (1,764 companies), followed by 10017 (1,075), and 10019 (861). I dig into the geographic location of these zip codes later in the section.

Figure 3.18: Frequency of NAICS Codes & Location of Financial Services Firms



Next, I wanted to determine the distribution of all financial firms' employment and annual sales, both of which are shown below in Figure 3.19. As mentioned before, these factors will serve as the independent variables in the regression to determine the presence of agglomeration.

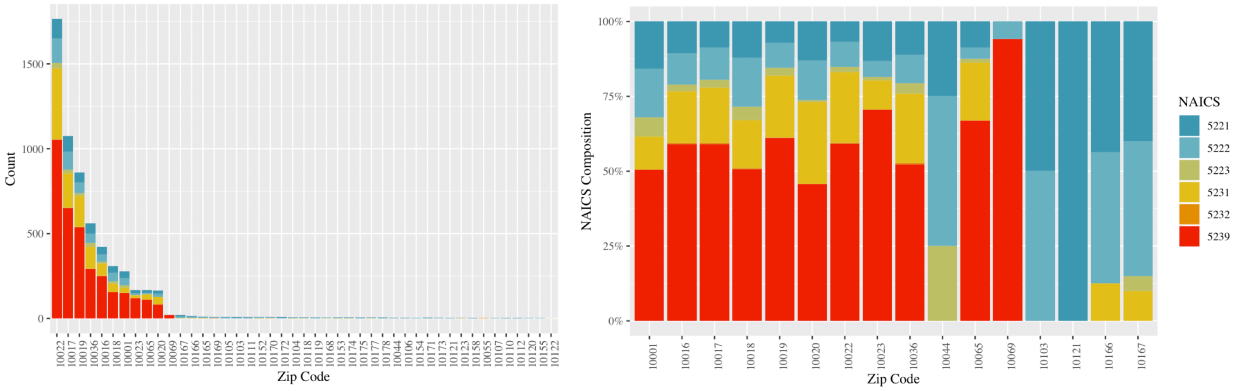
Figure 3.19: Log Frequency of Employees and Annual Sales



After understanding the three key factors of firm location, annual sales, and employees in isolation, we achieve a deeper understanding by studying the relationships between them. Figure 3.20 builds on Figure 3.18 by breaking out the number of companies in each zip code by their four-digit NAICS code. In 10022, the zip code with the greatest number of companies, for example, NAICS code 5239 is the most numerous, consistent with many of the Midtown zip codes. This is

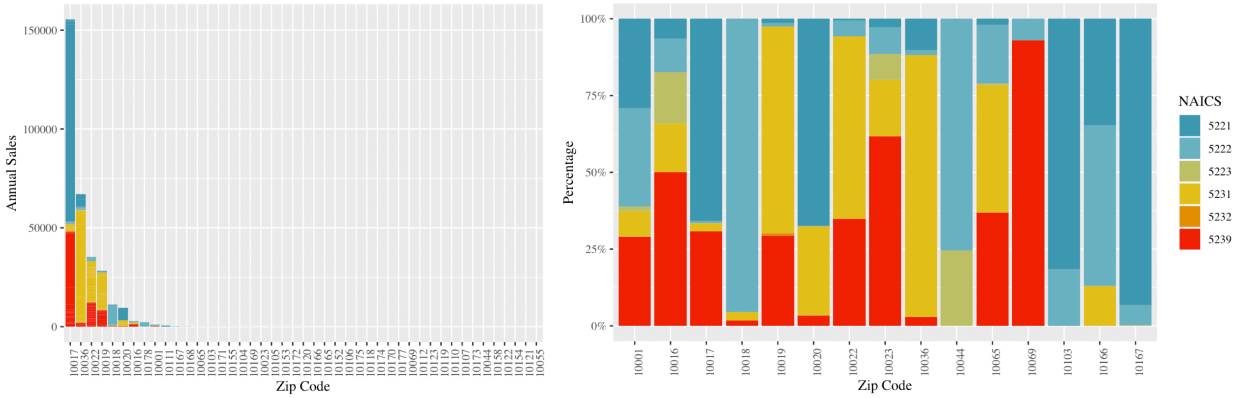
unsurprising since 5239 represents a general, catch-all category. Yet there are some zip codes with no companies in 5239, including 10103 and 10121.

Figure 3.20: NAICS Composition of Zip Codes



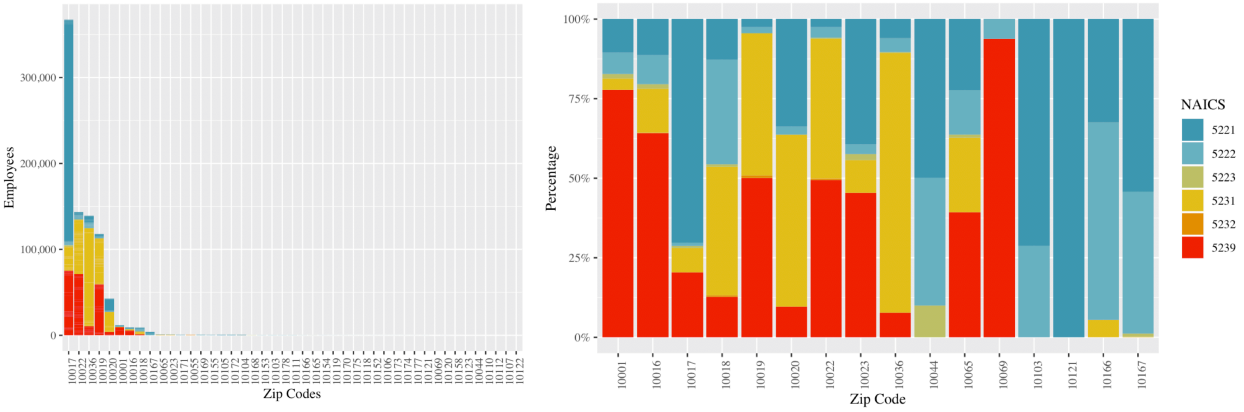
Sales by zip code is similarly a mix of industry composition, with some still dominated by 5239 (10009, for example), though many fewer zip codes contain above 50% of their annual sales attributable to companies in 5239.

Figure 3.21: Annual Sales per Zip Code and NAICS Code



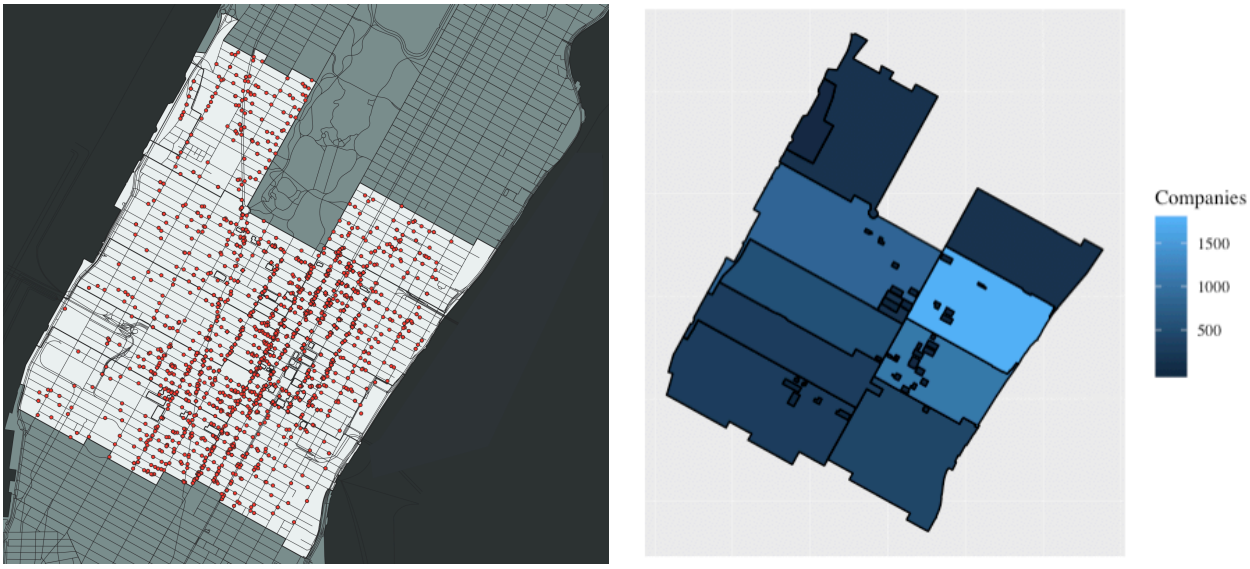
The bar charts in Figure 3.22 show that, similar to the sales figures, there are a few zip codes whose employment figures are dominated by companies in the 5239 NAICS category, although there are many that reveal a healthy mix of NAICS codes.

Figure 3.22: Employees per Zip Code and NAICS Code



As a final step in understanding the data, I built maps to show these trends spatially. The first map returns to our first analysis, showing which zip codes have the greatest number of financial services firms. We see, as expected, a high concentration on the east side of Midtown, primarily around and just to the north of Grand Central Terminal.

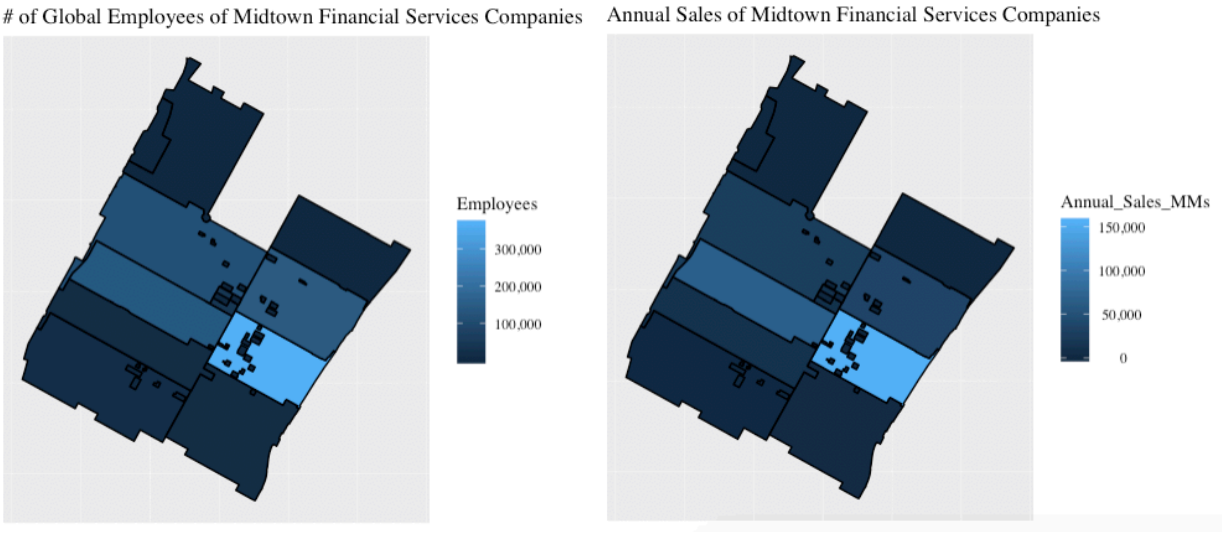
Figure 3.23: Midtown Financial Services Companies: Locations and Counts by Zip Code



Although firms are concentrated in zip code 10022, the greatest number of employees and largest total annual sales are located in the zip code immediately to the south, which actually contains Grand Central Terminal. This may reflect the fact discovered in the previous section that there are

more leasable square feet in that zip code, though there may be other factors at play, such as firms wanting to locate where they will minimize the commutes of their workers or executives.

Figure 3.24: Global Employees & Annual Sales by Zip Code



3.1.3 CB Insights

CB Insights is an information platform that provides information and analysis about the startup universe; the company also happens to be a startup itself. Their information includes overviews of sectors and individual companies, as well as in-depth analysis of industry-wide technology and funding trends. One offering CB Insights provides includes downloadable lists of companies that, similar to Dun & Bradstreet, include company location and number of employees, among other data points. The main difference between this data set and the Dun & Bradstreet data is that CB Insights focuses on startups, rather than Dun & Bradstreet's broader information about businesses of all stages.[2] As a result of this focus on startups, CB Insights also provides information

[2] Through the cleaning process, I found that this is not completely true, as CB Insights' data unexpectedly includes established companies, in addition to startups.

on a given company's technology stack (what programming frameworks they use for their application), who their investors are, and the total amount of funding they have raised. I received the CB Insights data used in this thesis through the MIT Real Estate Innovation Lab's subscription. As with the Dun & Bradstreet data, information from the CB Insights companies will be included as independent variables in the final regressions. Including independent variables that describe the business environment both in terms of established firms and startup companies makes this research unique; no other studies identified have attempted this methodology that incorporates businesses from two life cycle stages to identify and quantify the financial impact of agglomeration.

In order to process this data, I first geocoded the entries via the same method as the Dun & Bradstreet companies. I then dropped variables such as the technology each company uses, lists of past investors, and the website URL because I believed they would have limited use in this specific study, though future research might make use of them. The columns that remained included address, company name, number of 'contacts' (a proxy for employees), total funding received, sector, industry, sub-industry, whether the company is venture-backed, its exit round, its exit date, the date of last funding, the round of last funding, the number of current job postings, and the number of job postings in the past six months.

Next, similar to the Dun & Bradstreet data, I eliminated companies that had a geocoding score of 0 (meaning there was no address in the company entry), which left 6,581 companies that I could filter based on their zip codes. In this data set, there were three companies with NA for the zip code. These were "Antiques at the Armory Show," "Columbia University," and "Pier Antiques Show," so it was possible to drop these without impacting the analysis. This was also the first indication that the data set may not contain startups exclusively. Despite the small number of zip codes with NA values, there were 98 entries with zip codes equal to 0. All of them had addresses and, on further examination, I noticed that they represented a common set of buildings. After

identifying the common buildings, I only had to look up only 34 addresses. Using Google maps, I manually matched each address to its corresponding zip code.

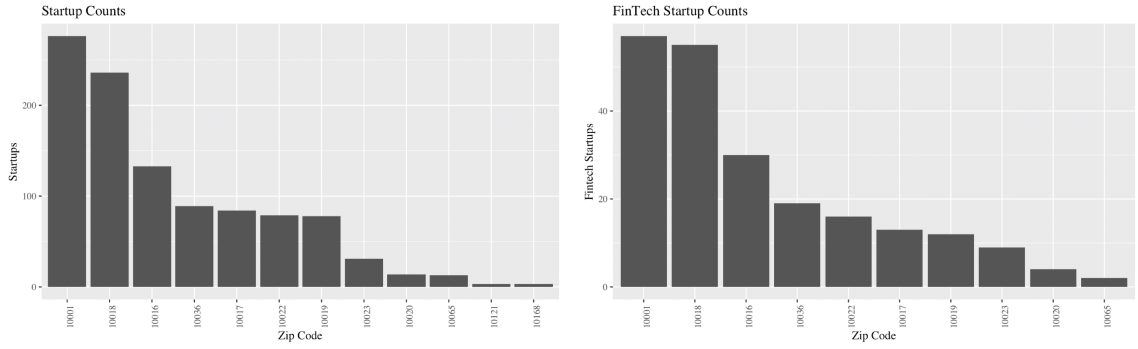
After assigning each company an address and zip code, I could then limit my search to the companies within the Midtown Manhattan zip codes according to the same process used to filter the addresses in the Dun & Bradstreet data set. The final cleaning step was to verify that the data set contained only startups. Although I was able to filter the Dun & Bradstreet data in order to ensure I was retrieving only financial services firms, the CB Insights data set actually contained companies of all sizes and stages of development, contrary to their claim that their data represent only startups. Thus, the first step of this cleaning procedure was determining whether a given company was a startup. At first, I relied on two columns, whether the company was venture backed, and whether the exit date was blank. A company with a non-blank exit date indicates that it either went public or was acquired, so would no longer be considered a startup. After applying these two filters, 1,039 companies remained, but on closer examination, this filtering left several companies that I knew were not startups such as the consulting firm AT Kearny and the financial services giant Alliance Bernstein. Thus, I utilized a third column, the date of last funding, to make the final filter. The final startup filter included only companies with a non-blank entry for last funding and resulted in a list of 667 startups.

Given the research's focus on financial services firms, it is also important to determine which companies were FinTech startups; the list of startups still contained companies from a variety of industries so the next step was limiting the list of startups to only financial services startups. This proved difficult because of the lack of uniquely identifying information that could indicate whether a given startup was a FinTech. Although CB Insights provides three columns that classify each company – Sector (of which there are 21 unique values), Industry (165 values), and Sub-Industry (207 values) – there were a variety of combinations of sectors, industries, and sub-industries

associated with known FinTech startups. Thus, I had to carefully cut the data to ensure only financial services companies were included and, in the end, filtered the data to a list of 217 FinTech startups.

Because of the small number of FinTech startups and the imperfect filtering process to identify startups and FinTech startups, I utilize both the startup data set and its subset, FinTechs, throughout the analysis. As a result, the following exploratory data analysis will calculate figures and display charts for both startups and FinTech startups. As seen in Figure 3.25 below, zip code 10001 has the greatest number of both startups and FinTech startups, followed closely by 10018. This is in contrast to the Dun & Bradstreet companies, which were located primarily in 10022.

Figure 3.25: Number of Startups and FinTech Startups per Zip Code



We can further segment these figures by sector: Figures 3.26 and 3.27 below tells us that the plurality or majority of startups and FinTech startups are categorized with an industry of "Internet," which is unsurprising, given they are startups.

Figure 3.26: Summary of Startups by Zip Code and Sector

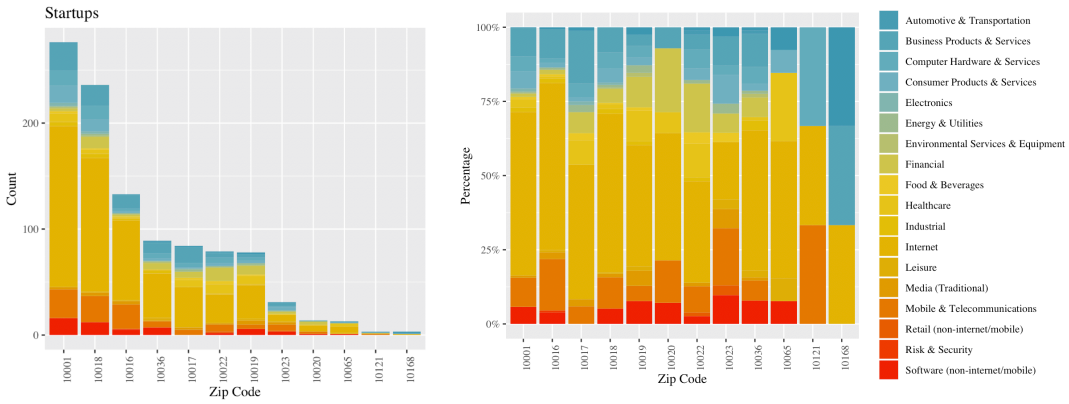
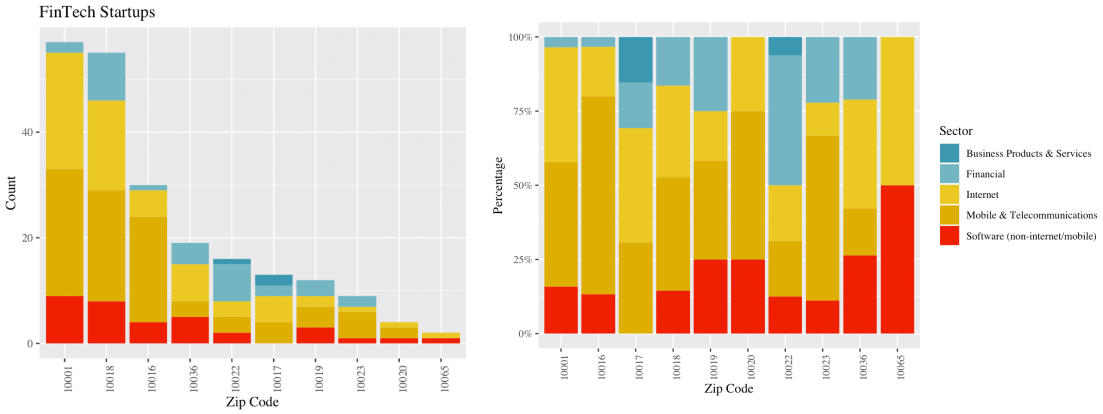


Figure 3.27: Summary of FinTech Startups by Zip Code and Sector



We can also describe the distribution of the total funding of startups and FinTech startups, displayed below in Figure 3.28. The distribution of total startup funding has a mean of $17.8 million, a median of $1.2 million, and a standard deviation of $66.4 million. The distribution of total FinTech funding has a mean of $32.3 million, a median of $5.6 million, and a standard deviation of $99.9 million.

Figure 3.28: Distribution of Total Funding (MMs USD)

Data Set	Min	Median	Mean	Max	Std. Dev.
Startup	$1.0	$1.2	$17.8	$1048.0	$66.4
FinTech	$0.0	$5.6	$32.3	$1048.0	$99.9

Because the CB Insights data set contains the category of last round of funding, in addition to analyzing the total amount raised, we can dig into the most recent round these startups have raised. The original CB Insights data set contained 48 categories of funding, many of which were duplicative. For example, the data set included "Series A" as well as "Series A – II." After consolidating all of the duplicative elements using regular expressions, 19 categories of funding remained, the most common of which for startups was "Unavailable" (409 companies), while the most common for FinTech startups was Series A (39). By dollar amount, the largest categories for startups were Unattributed ($2,442.0 million) and Series B ($2,356.8 million), while the largest categories for FinTech startups were Debt ($1,540.6 million) and Series A ($854.6 million). Because of the preponderance of Series A deals and dollars raised associated with FinTech companies, we can deduce that, in this data set, the FinTech startups are likely earlier stage than the overall startup data set. For more detail, please see Figures 3.29 and 3.30 below.

Figure 3.29: Count and Sum of Startup Funding by Funding Type

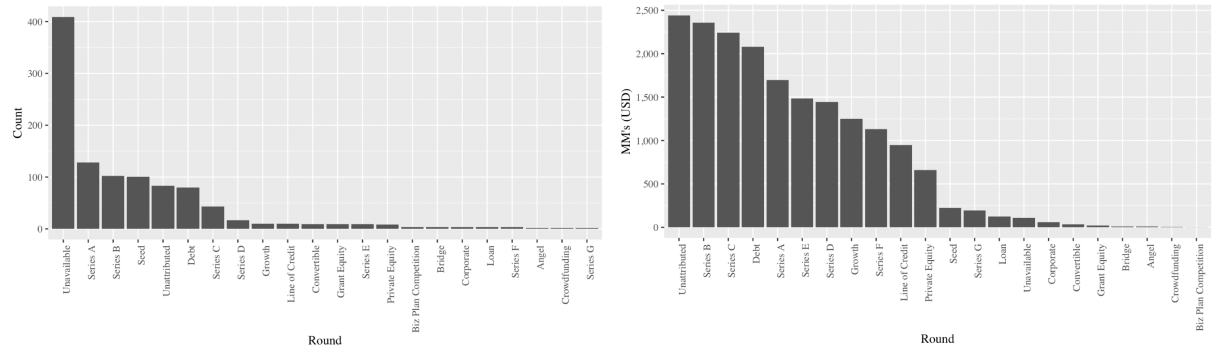
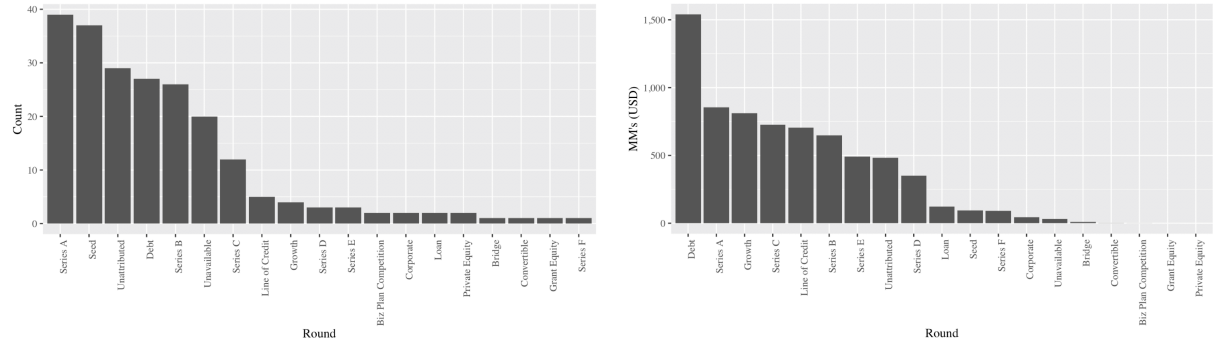


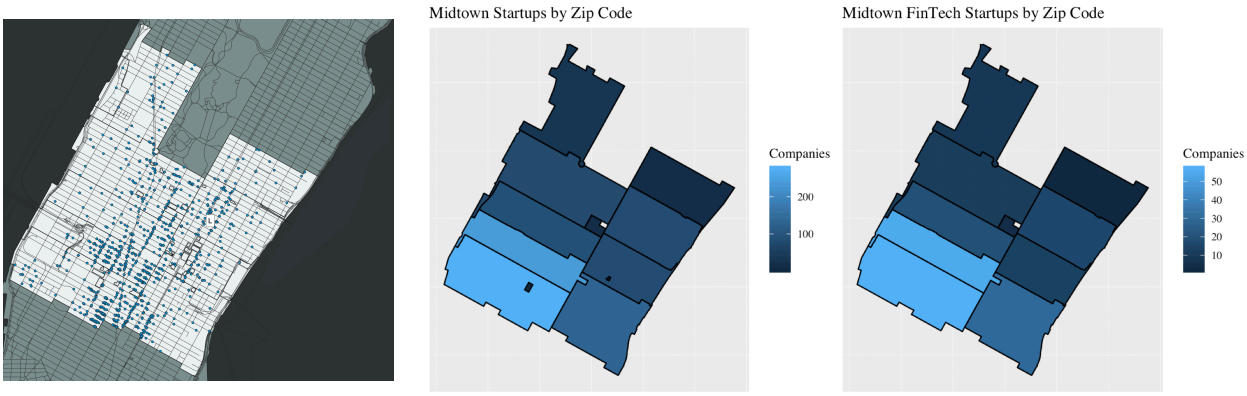
Figure 3.30: Count and Sum of FinTech Funding by Funding Type



Finally, although they are an important factor to study, I do not display the distributions of number of employees or number of jobs posted in the past six months because there are so many blank or zero values. For example, there are only 318 positive values for the number of contacts at startups (versus 721 NA values or zeroes), and the median of that distribution is zero (mean is 18) with a standard deviation of 177. Similarly, among startups there are only 274 positive values for the number of job postings in the past six months (765 values of NA or zero), and that distribution has a median of zero, a mean of 11, and a standard deviation of 93. This skewed and incomplete data set is a significant limitation of this aspect of the CB Insights data sets; I will further consider this limitation in the following chapter.

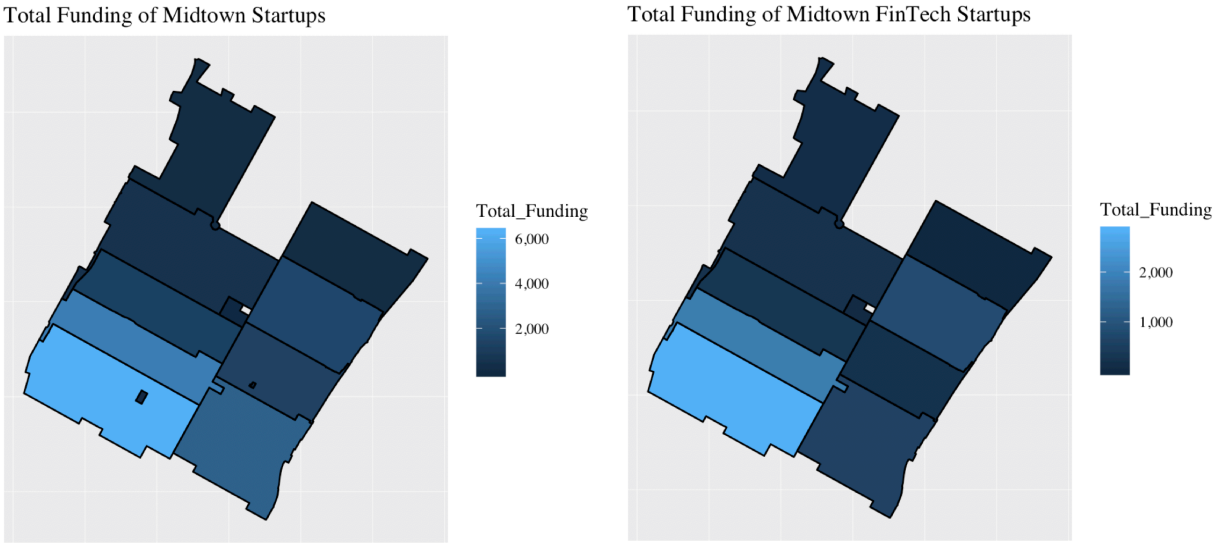
As before, the most useful exploratory data analysis for this thesis consists of understanding these factors spatially and Figure 3.31 below shows us the number of startups and FinTech startups in each of the Midtown zip codes. As shown before in Figure 3.26, the zip codes with the greatest number of startups (both overall and FinTech) are 10001, 10016, and 10017. This reveals that, in contrast to the Dun & Bradstreet data, the concentration of startups is located in the southwest corner of Midtown, suggesting startups in Midtown tend to locate closer to Midtown South, a well-known startup hub.

Figure 3.31: Startup & FinTech Locations in Midtown



This pattern of startups and FinTech startups concentrating in the southwestern corner of Midtown continues when considering the sum of total funding garnered by startups. In Figure 3.32, we see the concentration of funding raised by companies housed in Midtown is similarly located near Midtown South.

Figure 3.32: Total Funding of Midtown Startups and FinTech Startups by Zip Code



Yet this pattern breaks down when considering the number of employees. For startups, Figure 3.33 reveals that the greatest number of employees are located in 10022, which more closely parallels the employee concentration of financial services firms from the Dun & Bradstreet data. The

corresponding concentration of FinTech employees follows the trend of firm locations and total funding. Reflecting the prior maps, 10001 has the largest number of FinTech employees. The discrepancy between locations of startup and FinTech employees suggests that some startups may choose to locate near the traditional centers of financial services, perhaps when they reach a certain headcount.

Figure 3.33: Startup and FinTech Startup Employees by Zip Code

3.2 Data Consolidation

With all of the data cleaned and explored, the final task is to combine them into one consolidated data set. This consolidated data set will have the effective rent, lease attribute data, and data describing the business environment. Because effective rent is the dependent variable, we must measure the concentration of business environment factors around each of the CompStak leases. In order to test the extent of the forces of agglomeration, we can construct concentric buffers of varying radii with each CompStak point at the center. In this study, I used radii of 100 meters, 250 meters, and 500 meters, based on a review of the literature. Figure 3.34 below shows two maps that

encapsulate this process: one with all the data points from CompStak, Dun & Bradstreet, and CB Insights, and the other showing the CompStak points at the center of a series of buffers.

Figure 3.34: Map of All Data Points and All CompStak Buffers

After creating these buffers, I joined the information from the Dun & Bradstreet and CB Insights data sets with each CompStak buffer. This enabled me both to count the number of service companies, startups, and FinTech startups within each buffer, as well as sum up the attributes such as funding, sales, or employees.

After calculating the values for each buffer, I merged them into one large data set. This resulted in a table with 7,184 records and 65 columns. The columns represented the lease-level factors from each CompStak record, as well as each business-level factor at each radius distance. From this I created four subsets: the first, representing a base model, contained only lease-level factors. Then for each of the three business data sets (Dun & Bradstreet, CB Insights startups, and CB Insights FinTechs), I constructed subsets that contained the business environment factors for that given data set at the three buffer distances. With these data sets in hand, I was prepared to construct the agglomeration regressions.

3.3 Methods: Agglomeration & Regressions

Regressions provide researchers one of the most reliable and interpretable quantitative tools to study agglomeration. This is because it enables researchers to understand the marginal impact of a variety of individual factors on a single variable of interest. As described in Chapter 2, the variable of interest when studying agglomeration could be an area's productivity, employment, innovation, or individual firms' real estate prices.

A standard OLS regression formula includes the dependent variable (Y), an intercept term (β_0), a matrix of the data (X), a matrix of coefficients for the data (β_1), and an error term (ε). For this thesis, the dependent variable is the log of each lease's annual per square foot effective rent. The

typical OLS regression equation is written below and all regressions I model will build off this foundation.

$$Y = \beta_0 + X\beta_1 + \varepsilon$$

As a result of joining business information from Dun & Bradstreet and CB Insights with the lease-level transaction data from CompStak, each lease transaction now has two categories of data associated with it: lease-specific data and business environment data. Before modeling the impact of the business environment on effective rents, it is important first to establish a baseline model that includes only the lease-level factors. In the equation below, β_1 represents the matrix that contains coefficients for all of the lease-level factors.

$$Y = \beta_0 + \beta_1(Lease\ Factors) + \varepsilon$$

These lease-level factors include transaction year, floor type ("Entire" or "Partial"), whether the lease is on a single floor or has multiple floors, whether the lease contains the ground floor, whether the lease contains a mezzanine, whether the lease contains basement space[3], whether the lease is a sublease, the lease term in years, the months of free rent, the tenant improvement dollars spent, whether the building is Class A, Class B, or Class C, the year the building was built, the year the building was renovated, and the submarket.

After modeling the baseline regression, we can add the business environment factors that I engineered during the cleaning and modeling process. These consist of two components: specific business attributes derived from the consolidated data set and a distance factor. The business attributes are the information from the Dun & Bradstreet and CB Insights columns in the consolidated data set. The business environment variables available from the Dun & Bradstreet columns are number of financial services firms, annual sales of financial firms, and number of

[3] Although I initially ran the regressions with the basement variable, I omitted it from the final regression specification due to the amount of statistical noise it introduced into the rest of the coefficients. Further research should include a more in-depth analysis of the structure of this basement variable and its impacts on effective rents.

employees at financial firms. The information derived from the CB Insights data include number of startups / FinTech startups, as well as those companies' total funding, headcount, and number of jobs posted in the past six months.

The distance factor represents the three buffers I created at 100 meters, 250 meters, and 500 meters. This means that for each business data set, the regression is modeled three times, once for each buffer distance. I do this in order to account for the fact that agglomeration varies over space and to determine whether my analysis is sufficiently sensitive to pick up this variation. The three equations below represent the regressions that I will run for each business data set with β_2, β_3, and β_4 representing the business factors at each buffer distance.

$$Y = \beta_0 + \beta_1(Lease\ Factors) + \beta_2(Business\ Environment)_{D\&B100} + \beta_3\ (Business\ Environment)_{D\&B250} + \beta_4(Business\ Environment)_{D\&B500} + \varepsilon$$

$$Y = \beta_0 + \beta_1(Lease\ Factors) + \beta_2(Business\ Environment)_{CBIstartups100} + \beta_3\ (Business\ Environment)_{CBIstartups250} + \beta_4(Business\ Environment)_{CBIstartups500} + \varepsilon$$

$$Y = \beta_0 + \beta_1(Lease\ Factors) + \beta_2(Business\ Environment)_{CBIfintechs100} + \beta_3\ (Business\ Environment)_{CBIfintechs250} + \beta_4(Business\ Environment)_{CBIfintechs500} + \varepsilon$$

3.4 Missing Methods: Dynamics Driving Agglomeration

Although we can gather data, process it, and specify regressions aimed at identifying the quantitative impacts of agglomeration, economists have not yet been able to identify the underlying driving factors that cause agglomeration. For example, Rosenthal and Strange report that evidence of the sources of agglomeration is "suggestive rather than conclusive" (2006). According to them, this is because "it is difficult to be certain about causality. Agglomeration causes workers to be more productive. But skilled workers may also be drawn to urban areas, both because of high urban wages and also because of consumption…This complicates efforts to identify the impact of agglomeration on productivity" (Rosenthal and Strange, 2006). As an example of this uncertainty, Ellison et al

(2010) point out that "industrial relationships may be the result of collocation instead of the cause of collocation." Said another way, the benefits of agglomeration may result from proximity, rather than being a driving force for firms to locate near one another. Mariotti et al (2010) also point out that "although spatial proximity is important for generating knowledge spillovers, it is not sufficient as (i) proximity does not necessarily imply interaction and (ii) interaction does not necessarily mean positive spillovers." This quote brings into question the value of looking at agglomeration through knowledge spillovers since proximity does not necessarily mean firms will interact or that interaction will necessarily promote innovation or increased productivity. Saxenian (1994) provides a great example of the differences between types of interactions, comparing the success of Silicon Valley with the decline of Boston's Route 128 corridor as an example of how differently knowledge spillovers can occur, even within the same industry. Thus, many economists, including those who conceived of New Economic Geography, believe that "advancing the micro-foundations of knowledge diffusion and informational externalities is a future research direction of major importance" (Fujita and Krugman, 2004).

Among financial services firms and startups in Midtown Manhattan, I hypothesize that the presence of anchor institutions, the thick labor market, and the power of social networks all play key roles in the reasons that agglomeration exists among financial services firms in Midtown Manhattan.

As mentioned in the introduction to FinTech startups, these companies rely heavily on partnerships with large, established financial services firms. More broadly, the financial services industry also relies on taking their cues from large investment firms and banks. The desire for proximity to the key companies and individuals in the industry may incentivize firms of all sizes to cluster near one another.

An additional potential reason for agglomeration may be the area's robust labor market. When a region contains a large number of employees in the same or adjacent industries, also called

'labor pooling,' the companies and employees in that region face lower risks of business disruption because of lower search costs of matching a job with a suitable employee (Krugman, 1991). One major example of this is Silicon Valley, where, "although the cost of labor and land in the valley is very high, firms continue to do business there. This is entirely consistent with the idea of localization economies." In the case of Midtown Manhattan, firms may not care what price they have to pay for real estate in order to access the best talent in the financial services industry.

Furthermore, because the financial services industry is so human-capital intensive, there is great value in sharing ideas and knowledge with others. Although the cost of disseminating information such as stock prices "has been rendered invariant by the telecommunications revolution, the marginal cost of transmitting knowledge," or applied, useful analysis of that information, actually increases with distance (Audretsch, 1998). As mentioned before, New York City rose to prominence because it was the location that people could exchange information most efficiently. It retains that role today and firms want to capitalize on the ability to access the most current information in their industry.

Despite the concordance between my hypotheses and contemporary economic research, little quantitative work has been conducted to measure the underlying sources of agglomeration. To bring the sources of agglomeration into better focus, I turn to research that has explored the value of knowledge spillovers in another industry centered in New York City, the arts and culture sector. From 2004 to 2007, Elizabeth Currid interviewed over 300 tastemakers in New York City and used her qualitative findings to understand how agglomeration impacts individuals in the industry and undergirds the structure of the sector as a whole. One of her key findings was that the "social milieu appeared to be the most important mechanism by which the cultural economy operates," suggesting that access to strong networks for sharing information is a significant component of why these firms agglomerate in New York City (Currid, 2007). According to Currid, "social life enables interaction

across subgroups" suggesting proximity serves as a cross-pollination function, enabling greater innovation and productivity (Currid, 2007). Furthermore, the social life enabled people to meet others "who were important to their careers" suggesting support for the value of thick labor markets – in such a market, job searchers have an easier time finding employers seeking their skillsets and vice versa (Currid, 2007). An additional discovery was that, for fashion designers, "access to the 7th Ave garment district, along with pattern and fabric stores were important to their ability to produce clothing" (Currid, 2007). This suggests the presence of inputs and anchor institutions also promote agglomeration. Finally, New York's cultural agglomeration "established the city as a global tastemaker and cultural producers want their products associated with it" (Currid, 2007). This means that there is a perceptual benefit of being a firm that New York City is known for and locating in New York City. That would certainly be the case for financial services firms, as well.

Currid's findings emphasizes how difficult it is to use traditional econometric and statistical tools to determine *why* firms agglomerate. By using qualitative, rather than quantitative analysis, she succeeded in uncovering the impetus behind agglomeration. In contrast, economists and statisticians have struggled to design a data set, sampling mechanism, or experimental design that can comparably prove causality. This suggests that qualitative analysis could add significant explanatory power to the existing econometric and statistical agglomeration literature.

Chapter 4

Results, Limitations, & Future Work

Running a regression as I described in Chapter 3 requires only one command in R, the programming language I used throughout this thesis. The more difficult task is interpreting the results of these regressions, understanding the limitations of the findings, and predicting what future work could improve the results and push similar research forward. This chapter aims to fulfill these three tasks.

4.1 Results

The results indicating agglomeration result from regressions that reveal an association between environmental factors (data from Dun & Bradstreet and CB Insights) and the dependent variable, annual per square foot effective rent (from CompStak). But before running regressions to determine the presence of agglomeration, I built a baseline model with only lease-specific factors to test the reasonableness of the regression. This table and all other regression tables are included at the end of Section 4.1, starting on page 85, in order not to disrupt the analysis narrative.

4.1.1 Model One: Base Model

Figure 4.1 displays the baseline regression with only lease-specific attributes and no additional business information. There are two models in Figure 4.1, one that includes CompStak's default "Submarket" variable and one that does not. At first, the model that includes submarket

appears to be the superior specification.[4] Of the two models, it has the higher Adjusted R^2, meaning its independent variables explain a greater amount of the variation in the dependent variable. Both models suggest that the factors with the greatest magnitude impact on a lease's effective rent include the transaction year, being a multi-floor lease, whether the lease is a sublease, the length of the lease, the number of free months of rent, the tenant improvement allowance, whether it is a Class A, B, or C building, the year renovated, and the year built. Interestingly, although the model that incorporates submarkets has a higher R^2, only two of the submarket coefficients are statistically significant.

Because the dependent variable is a log-transformed variable, the interpretation of the coefficients requires a bit of math to make them comprehensible. To translate a given regression coefficient into that covariate's percentage impact on the dependent variable, one must exponentiate it, subtract one, then multiply that value by 100. For example, the coefficient on transaction year in the model without submarkets is 0.039; this means that each additional year in the study period is associated with a $100 * (e^{0.039} - 1)$, or 4.0% increase in rent. This is consistent with our earlier data analysis that found a compound annualized growth rate for the effective rent of 4.1%. The coefficient for transaction year in the model with submarkets is comparable, at 0.041.

The coefficients associated with having a multi-floor lease is -0.075 without submarkets and -0.077 with submarkets, meaning tenants achieve overall rent reductions as a result of taking a larger portion of the building. This scale discount is associated with an approximately 7.2% reduction in rent. The linear nature of this coefficient is perhaps misleading; a firm would likely receive a greater discount for leasing ten floors in a building than the discount associated with leasing only two, but parsing out the form of this factor would require further research.

[4] We will see later in the chapter that the inclusion of the submarket variable is not ideal for determining the presence of agglomeration based on the initial research specification, though it is certainly a variable that future research should include and utilize.

With a coefficient of -0.207 and -0.216, signing a lease as a sublessor is associated with an 18 to 20% discount (depending on whether you use the model with submarkets or without), which makes sense because subleases are typically cheaper than signing a new lease as the primary tenant.

One of the more interesting dynamics is between the coefficients on lease term, free rent, and tenant improvement allowance. Each additional year in lease term is associated with an increase in effective rent of approximately 2%. Increases in free rent and tenant improvement allowance, however, have negative impacts, reducing effective rent by 0.9% for each additional free month of rent and by 0.1% for each additional tenant improvement dollar spent. In some respects, these coefficients make sense because the longer you are paying rent, the more rent you pay. Similarly, the more months of rent you do not have to pay, the less rent you are paying over the course of your lease. From a practical perspective, however, these findings are less intuitive. Landlords prefer longer leases and often charge premiums for shorter-term leases. Similarly, if a landlord provides tenant improvement dollars to a tenant, that expenditure is often recaptured through an increase in their rent, which would translate into a higher effective rent. As a result, to completely understand the dynamics at play would require further examination of these coefficients and the underlying data.

Going from Class A to Class B or to Class C also has a significant impact on rent. Based on the regression without submarkets, Class B rents are expected to be approximately 14% cheaper than comparable Class A transactions. Class C are expected to be approximately 24% cheaper than comparable Class A transactions. These values seem a bit high, so would warrant further investigation. Another factor related to the class coefficients that requires further examination is that introducing the submarket variable into the regression dramatically changes these coefficients. For example, when the submarkets are included, the Class B coefficient changes to represent a 6.5% reduction in price, as compared to a Class A lease. The Class C coefficient is no longer statistically significant with the presence of the submarket variables. The fact that these coefficients vary

dramatically with the inclusion of additional variables suggests there is strong linear relationship between the class variable and the submarket variable, which is unsurprising, given the relationships we saw between class and submarket in Chapter 3 (see Figures 3.5 and 3.8, specifically).

The year a building was built and renovated also has statistically significant impact, with a 0.4% rent increase for each additional year in which the building was renovated and a 0.5% rent increase for each additional year in which the building was constructed. Older buildings may be less valuable than newer ones because they may contain smaller floorplates that are hard to update to contemporary, open-office standards; their building infrastructure may not be up to current, higher building loads; their lack of HVAC; or they may have slower elevators or smaller windows.

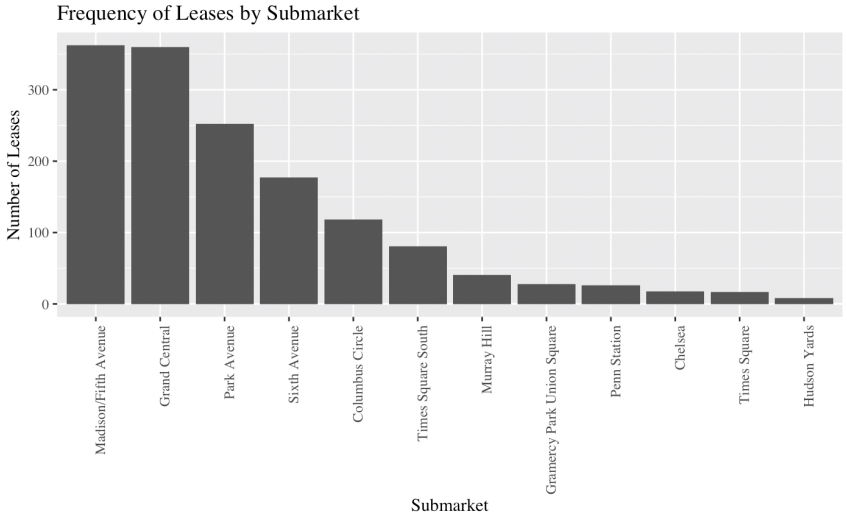
Taken together, these coefficients suggest that if a tenant's only consideration were cost, that firm should sublease multiple floors in an old, un-renovated, Class C building. Yet so few tenants consider solely costs when searching for office space. The divergence between an optimal solution from a cost perspective and the actual behavior of firms and office workers points to the fact that location, design, and the social life of neighborhoods matter significantly in location decisions. Perhaps models with business environment factors will begin to shed light on this divergence.

4.1.2 Model Two: Dun & Bradstreet

Because the baseline model with CompStak's submarket variable seemed to explain more of the variation in the dependent variable, I began analyzing the Dun & Bradstreet using the model that incorporated that categorical variable. As seen in Figure 4.2, however, none of the variables were statistically significant at the 100-meter buffer. Furthermore, none of the types of variables (number of firms, total annual sales, and number of employees) had statistically significant coefficients at all distance bands. As I have already suggested, there appears to be a conflict between the CompStak submarket variable and the rest of the research design. Because the business environment factors

were gathered at buffers around each CompStak point without regard to submarket, further splitting the analysis by submarket interrupts the signals that would otherwise surface in the regression based on the original research design. Additionally, as demonstrated in Chapter 3, there is significant variation of many factors between and within submarkets. In Figure 4.3, we see another example of that variability; the sample size varies significantly between each submarket; some have over 300 observations, while others have fewer than 10.

Figure 4.3: Relative Frequency of CompStak's Submarket Variable



The fact that the submarket variable introduces significant variability causes the regression to underperform. This is not to say that specifying the regressions with submarkets is incorrect. Due to the original research design, however, the additional factor of submarket actually detracts from the findings. In future work, researchers should create a research design that considers agglomeration in the context of submarket geography. I discuss this further in the concluding section of the chapter.

Ultimately, using the models that did not include CompStak's submarket variable provided the evidence of agglomeration I expected. As seen in Figure 4.4, the statistically significant coefficients on number of Dun & Bradstreet firms within 100 meters, 250 meters, and 500 meters were .001, .0002, and .0002, respectively, meaning an additional financial services firm within 100

meters is associated with a 0.1% increase in the expected rent. This may not seem like a lot, but that means an additional 100 firms within a 100-meter area would raise the expected annual effective rent by 10%. In Midtown Manhattan where financial services are prevalent, this is not out of the realm of possibility. More specifically, within this data set, there are 352 CompStak leases (or 23.7% of the total) with 100 or more financial services firms within 100 meters.

Similar to the 100-meter radius, an additional firm within 250 and 500 meters is associated with an expected increase in effective rent of 0.02%. In order to demonstrate that agglomeration dissipates over distance (in accordance with Rosenthal and Strange, 2003), I would have to conduct a statistical test that determines whether these coefficients are statistically different.

The other two variables derived from the Dun & Bradstreet data, employee count and sum of annual sales, are significant but their magnitudes are so small that they do not appear to suggest the presence of agglomeration. Because I found the statistically significant presence of agglomeration at all three distances, this model lends support to the idea that firms do not optimize their office location decisions only on costs, but also take into account agglomeration-related factors such as the presence of other, similar firms.

4.1.3 Model Three: CB Insights Startups

Using the CB Insights data, there are several interesting findings, as demonstrated in Figure 4.5. First, as opposed to the Dun & Bradstreet models that proved an association between more firms and higher effective rent, the model using CB Insights startup data suggests that an additional startup within a given radius is actually associated with a *decrease* in rent. This decrease is 1.8% for each additional startup within a 100-meter radius of a lease, 0.9% for one more startup within a 250-meter radius and 0.5% for one more startup within a 500-meter radius. While initially counterintuitive, there could be a reasonable explanation for this. Because real estate is one of the

largest components of a startup's fixed costs, they may be choosing to locate in areas that have cheaper rents. This would align with the findings in Chapter 3 that a large number of startups are located in the southern part of Midtown, where there is also the highest number of Class C buildings. Similarly, the number of startup employees has a negative impact on effective rent. The explanation here is similar to the count of startups – more startup employees are currently located in the southern part of the submarket, which is home to more Class C buildings, thus resulting in an association, but not necessarily a causal relationship, between the number of startup employees and effective rent.

The variable that suggests the existence of positive agglomeration, however, is the total amount of funding raised by startups. At a 100-meter distance, an additional $1.0 million of total funding translates into a 0.04% expected increase in rent. Similar to the Dun & Bradstreet firm coefficient, this seems like a small number. But given the scale at which companies are raising funds, an additional $100 million of total funding by startups within a given 100-meter radius would translate into an expected rental rate increase of 4.0%. In the startup data set, the mean amount of funding is $109.3 and 535 of the 100-meter buffers (or 35.6% of the total) contain more than $100 million in aggregate venture capital funding. The significance and magnitude of the funding variable remains even at 250 and 500 meters but, similar to the Dun & Bradstreet model, I would need to conduct statistical tests to determine whether the agglomeration diminishes over space. Attenuation notwithstanding, the fact that increased funding yields increase rents (and, by proxy, productivity), we must, in future research, examine what underlying force this reflects. Do firms that raise money in the same round want to share information? Do they share venture capitalists? These questions should be addressed in future research, especially with qualitative interviews of these firms and venture capitalists, in order to better describe the underlying agglomerative instinct.

4.1.4 Model Four: CB Insights FinTech Startups

The final model also uses CB Insights data but uses the CB Insights FinTech subset. As mentioned in Chapter 3, I created two data sets from the CB Insights data, one containing 667 startups and the other containing 217 FinTech startups. Here, as with the previous CB Insights model, the number of FinTech startups is associated with a decrease in expected rent. Figure 4.6 shows that, at 100 meters, the expected rent reduction with each additional $1.0 million dollars in funding is 3.6%, at 250 meters, the expected reduction is 1.5%, and at 500 meters, the expected reduction is 0.7%. Yet, despite the seeming contradiction between the presence of startups and the direction of rent, the justification is similar to the previous model, namely that because of the large amount of money going to real estate and startups' constrained budgets, startup companies may be clustering in areas that have, on average, cheaper rent. This was analyzed in our visual inspection of the location of startups and the average effective rent in those neighborhoods in Chapter 3.

Finally, similar to the CB Insights startup model, it is actually the total amount of funding raised by startups within a radius that has a positive impact on the expected rental rate. The expected increases for an additional $1.0 million of startup funding at 100 meters, 250 meters, and 500 meters are 0.1%, 0.02%, and 0.03%, respectively. This suggests that, similar to the startups, financial technology startups may want to share insight with other firms that have raised similar amounts of money and benefit from mutual knowledge spillovers.

Figure 4.1: Baseline Model with and without Submarket

	Log of Annual Effective Rent Per SqFt (USD)	
	(1)	(2)
Transaction Year	0.039***	0.041***
Entire Floor	−0.017	0.008
Partial Floor	−0.037	0.017
Multi-Floor	−0.075**	−0.077**
Has Ground	−0.072	−0.015
Has Mezzanine	−0.059	−0.125
Is Sublease	−0.207***	−0.216***
Lease Term (Yrs)	0.018***	0.021***
Free Rent (Mos)	−0.009**	−0.007**
TI Allowance (USD)	−0.001***	−0.001***
Class B	−0.151***	−0.067**
Class C	−0.274***	−0.162
Year Renovated	0.003***	0.004***
Year Built	0.005***	0.004***
Columbus Circle		0.143
Gramercy Park / Union Sq.		0.050
Grand Central		0.004
Hudson Yards		0.030
Madison / 5th		0.283***
Murray Hill		−0.009
Park Ave		0.242**
Penn Station		−0.138
Sixth Ave		0.145
Times Sq.		−0.048
Times Sq. South		−0.086
Constant	−89.642***	−93.078***
N	858	858
R^2	0.383	0.515
Adjusted R^2	0.373	0.500

*p < .1; **p < .05; ***p < .01

Figure 4.2: Dun & Bradstreet Model with Submarkets

	Log of Annual Effective Rent Per SqFt (USD)			
	(1)	(2)	(3)	(4)
Transaction Year	0.039***	0.041***	0.041***	0.041***
Entire Floor	−0.017	0.005	−0.014	0.003
Partial Floor	−0.037	0.017	0.008	0.017
Multi-Floor	−0.075**	−0.082***	−0.066**	−0.063**
Has Ground	−0.072	−0.015	−0.028	−0.018
Has Mezzanine	−0.059	−0.132	−0.158	−0.151
Is Sublease	−0.207***	−0.213***	−0.203***	−0.206***
Lease Term (Yrs)	0.018***	0.021***	0.022***	0.024***
Free Rent (Mos)	−0.009**	−0.007**	−0.007**	−0.008***
TI Allowance (USD)	−0.001***	−0.001***	−0.001***	−0.001***
Class B	−0.151***	−0.065**	−0.082***	−0.106***
Class C	−0.274***	−0.158	−0.180*	−0.193*
Year Renovated	0.003***	0.004***	0.004***	0.004***
Year Built	0.005***	0.004***	0.003***	0.003***
Columbus Circle		0.137	0.137	0.130
Gramercy Park / Union Sq.		0.046	0.051	0.040
Grand Central		0.003	0.032	0.026
Hudson Yards		0.025	0.034	0.043
Madison / 5th		0.269**	0.257**	0.226*
Murray Hill		−0.011	−0.014	−0.049
Park Ave		0.237**	0.238**	0.221*
Penn Station		−0.140	−0.136	−0.153
Sixth Ave		0.132	0.133	0.119
Times Sq.		−0.055	−0.060	−0.013
Times Sq. South		−0.090	−0.053	−0.112
Financial Services Firms (100m)		0.0001		
Financial Services Annual Sales (100m)		−0.00000		
Financial Services Employees (100m)		0.00000		
Financial Services Firms (250m)			−0.00002	
Financial Services Annual Sales (250m)			−0.00001***	
Financial Services Employees (250m)			0.00000***	
Financial Services Firms (500m)				0.0001**
Financial Services Annual Sales (500m)				−0.00000
Financial Services Employees (500m)				−0.00000
Constant	−89.642***	−93.390***	−92.897***	−93.535***
N	858	858	858	858
R^2	0.383	0.517	0.530	0.539
Adjusted R^2	0.373	0.501	0.514	0.524

*p < .1; **p < .05; ***p < .01

Figure 4.4: Dun & Bradstreet without Submarkets

	Log of Annual Effective Rent Per SqFt (USD)			
	(1)	(2)	(3)	(4)
Transaction Year	0.039***	0.039***	0.041***	0.041***
Entire Floor	−0.017	−0.037	−0.040	−0.028
Partial Floor	−0.037	−0.044	−0.020	−0.011
Multi-Floor	−0.075**	−0.085**	−0.059*	−0.062**
Has Ground	−0.072	−0.049	−0.058	−0.006
Has Mezzanine	−0.059	−0.118	−0.166	−0.225
Is Sublease	−0.207***	−0.203***	−0.194***	−0.188***
Lease Term (Yrs)	0.018***	0.019***	0.021***	0.023***
Free Rent (Mos)	−0.009**	−0.008**	−0.008**	−0.008***
TI Allowance (USD)	−0.001***	−0.001***	−0.001***	−0.001***
Class B	−0.151***	−0.116***	−0.103***	−0.139***
Class C	−0.274***	−0.209**	−0.216**	−0.198**
Year Renovated	0.003***	0.003***	0.003***	0.004***
Year Built	0.005***	0.005***	0.004***	0.004***
Financial Services Firms (100m)		0.001***		
Financial Services Annual Sales (100m)		−0.00001***		
Financial Services Employees (100m)		0.00000***		
Financial Services Firms (250m)			0.0002***	
Financial Services Annual Sales (250m)			−0.00001***	
Financial Services Employees (250m)			0.00000***	
Financial Services Firms (500m)				0.0002***
Financial Services Annual Sales (500m)				−0.00000***
Financial Services Employees (500m)				0.00000
Constant	−89.642***	−88.874***	−91.400***	−93.549***
N	858	858	858	858
R^2	0.383	0.420	0.486	0.503
Adjusted R^2	0.373	0.408	0.475	0.492

*p < .1; **p < .05; ***p < .01

Figure 4.5: CB Insights Startups without Submarkets

	Log of Annual Effective Rent Per SqFt (USD)			
	(1)	(2)	(3)	(4)
Transaction Year	0.039***	0.040***	0.041***	0.040***
Entire Floor	−0.017	−0.036	−0.018	−0.030
Partial Floor	−0.037	−0.029	−0.021	−0.025
Multi-Floor	−0.075**	−0.075**	−0.088***	−0.092***
Has Ground	−0.072	−0.053	0.006	−0.083
Has Mezzanine	−0.059	−0.223	−0.077	−0.151
Is Sublease	−0.207***	−0.204***	−0.208***	−0.200***
Lease Term (Yrs)	0.018***	0.019***	0.020***	0.022***
Free Rent (Mos)	−0.009**	−0.009***	−0.007**	−0.007**
TI Allowance (USD)	−0.001***	−0.001***	−0.001***	−0.001***
Class B	−0.151***	−0.119***	−0.088***	−0.152***
Class C	−0.274***	−0.155	−0.179*	−0.291***
Year Renovated	0.003***	0.003***	0.003***	0.005***
Year Built	0.005***	0.004***	0.004***	0.003***
Startups (100m)		−0.018***		
Startup Total Funding (100m)		0.0004***		
Startup Employees (100m)		−0.001***		
Startup Job Postings (100m)		0.0003***		
Startups (250m)			−0.009***	
Startup Total Funding (250m)			0.0002***	
Startup Employees (250m)			−0.0002***	
Startup Job Postings (250m)			0.00000	
Startups (500m)				−0.005***
Startup Total Funding (500m)				0.0001***
Startup Employees (500m)				−0.0001***
Startup Job Postings (500m)				0.0001***
Constant	−89.642***	−90.917***	−91.870***	−92.007***
N	858	858	858	858
R^2	0.383	0.442	0.451	0.481
Adjusted R^2	0.373	0.430	0.439	0.470

*p < .1; **p < .05; ***p < .01

Figure 4.6: CB Insights FinTechs without Submarkets

	Log of Annual Effective Rent Per SqFt (USD)			
	(1)	(2)	(3)	(4)
Transaction Year	0.039***	0.040***	0.041***	0.041***
Entire Floor	−0.017	−0.019	−0.024	−0.031
Partial Floor	−0.037	−0.034	−0.035	−0.032
Multi-Floor	−0.075**	−0.075**	−0.075**	−0.091***
Has Ground	−0.072	−0.116	−0.070	−0.013
Has Mezzanine	−0.059	−0.114	−0.030	−0.191
Is Sublease	−0.207***	−0.197***	−0.205***	−0.193***
Lease Term (Yrs)	0.018***	0.020***	0.019***	0.021***
Free Rent (Mos)	−0.009**	−0.010***	−0.008**	−0.006**
TI Allowance (USD)	−0.001***	−0.001***	−0.001***	−0.001***
Class B	−0.151***	−0.116***	−0.100***	−0.126***
Class C	−0.274***	−0.195*	−0.222**	−0.218**
Year Renovated	0.003***	0.003***	0.003***	0.005***
Year Built	0.005***	0.005***	0.005***	0.004***
FinTechs (100m)		−0.037***		
FinTech Total Funding (100m)		0.001***		
FinTech Employees (100m)		−0.0002		
FinTech Job Postings (100m)		−0.002**		
FinTechs (250m)			−0.015***	
FinTech Total Funding (250m)			0.0002***	
FinTech Employees (250m)			0.0001	
FinTech Job Postings (250m)			−0.0004	
FinTechs (500m)				−0.007***
FinTech Total Funding (500m)				0.0003***
FinTech Employees (500m)				−0.001***
FinTech Job Postings (500m)				−0.0003**
Constant	−89.642***	−92.535***	−92.748***	−95.234***
N	858	858	858	858
R^2	0.383	0.421	0.413	0.461
Adjusted R^2	0.373	0.408	0.401	0.449

*p < .1; **p < .05; ***p < .01

4.2: Limitations

Although I uncovered compelling evidence of the existence and magnitude of agglomeration economies among financial services companies in Midtown Manhattan, there are many limitations to this study.

First, there is a mismatch in timing between the date of lease transactions and the locations of the businesses. The lease transactions range in year from 2009 to 2019, whereas the business data sets (both Dun & Bradstreet and CB Insights) reflect point-in-time estimates of the locations of businesses when I ran the queries (January and March, 2019, respectively). This means that my data set assumes all businesses from Dun & Bradstreet and CB Insights were near all leases at the time of lease execution, an unrealistic expectation. I might address this in the future by limiting the lease transactions only to those that have occurred in the past 18 months, when it could be more realistically assumed that the leases and businesses were in the same place at the same time.

A second limitation is the non-normality of many of the covariates. As described in previous chapters, a number of the variables are not normally distributed, even after log transformations. This means that a linear regression may produce biased estimates of the coefficients, thus limiting the usefulness of the research. Though researchers can often log transform non-normal variables to make them more normally distributed, this research contained a number of non-normally distributed variables that could not be log transformed because they contained meaningful values of 0 that cannot be excluded. In the case of months of free rent and tenant improvement dollars these values of 0 could accurately signify that a tenant received no free rent or no tenant improvement dollars or suggest that the person who entered the lease did not know those figures. Because we cannot determine which answer is correct, we cannot exclude these variables, nor can we log transform them because log transforming 0 returns a value of negative infinity.

A third limitation of the study relates to the calculation of effective rent. Although using effective rent allows us to account for differences in lease term, tenant improvement work, and months of free rent, the rent is nonetheless associated with one of a variety of reimbursement methods (full-service gross, modified gross, etc.). This means that varying amounts of building-level operating expenses such as electricity, gas, water, or janitorial will be included in that effective rent number, preventing a perfect apples-to-apples comparison between leases that contain different reimbursement methods. If all buildings had the same operating expenses, this reconciliation would be simple; that is not the case, however, as older buildings are significantly less energy efficient and it is difficult to estimate the operating expenses for a given building, just given its age. I hoped to be able to standardize the different reimbursement methods by incorporating the operating expenses column CompStak provides, but that column was so sparsely populated with data that it was impossible to reconcile the different reimbursement methods.

Fourth, from a methodological perspective, I treated all aspects of Midtown as uniformly distributed. This includes the distribution of transportation access, locations of office buildings, and other clearly unrealistic assumptions. On a related note, as the hedonic literature indicates, it is impossible to account for all of the factors that impact the price of real estate. In this research, numerous other considerations that could impact the price of real estate include proximity to Central Park, distance to subway stations, and the views from taller buildings or suites. This omitted variable bias may lead to less unbiased coefficients and undermine the overall validity of the findings.

On another geography-related note, there was significant tension between the geographic units of analysis: buffers and submarkets. Though submarkets serve as the primary geographic unit of analysis for the real estate industry, significant agglomeration research has been conducted using buffer analysis. Buffers provide more effective statistical measures precisely because they ignore

underlying variation within the urban landscape, such as submarket. But including both types of geographic delineation muddies the water of interpretability, as we saw in Figure 4.2. Though buffers may not be ideal for use in a real estate setting, they are nonetheless useful in this research and set the scene for future research that will utilize the submarket boundaries as the geographic basis of analysis.

Finally, this modeling demonstrates the associations between variables, but is not structured as an experiment that could prove causality. Without better research and sampling design, we cannot definitively state that the agglomeration forces *caused* the increases in effective rent. Much of this is attributable to the fact that, as discussed in Chapter 3, researchers have not yet been able to test what causal factors lead to agglomeration. Though I believe the presence of anchor institutions, strong labor markets, and the power of the financial services social network drive agglomeration in Midtown, no statistical research related to agglomeration has been able to uncover these causal relationships in a statistically rigorous way. That is where we might turn to qualitative research similar to Currid (2007) and speak with the actual participants in the market to learn the main considerations that drive their decisions when choosing an office location.

Section 4.3: Future Work

Incorporating additional data sources and qualitative information are the key areas of future research. The factors that are missing include input from the landlords, brokers, and financial services firms that were studied in the statistical analysis. Ground-truthing the findings unearthed in this thesis would greatly benefit future research to understand where I have correctly identified trends as well as where I may have mistakenly found relationships that are not correctly specified.

Additional areas of expansion include broadening the focus to other sectors besides financial services. Because New York City is home to large concentrations of other industries such as media,

telecommunications, or art, this research could be applied to companies in those sectors just as easily. Such an expansion to the arts and culture sector could supplement Elizabeth Currid's 2007 study, as well as the follow-on study conducted by Currid and Sarah Williams in 2009 (Currid and Williams, 2009) that mapped the tight-knit relationships in the fashion industry. This expansion to other sectors would demonstrate how agglomeration patterns differ between industries within the same geographic area and would aim to uncover whether similar rationales exist for agglomeration irrespective of industry.

Finally, there are also often-utilized regression models that incorporate aspects of spatial autocorrelation including the spatial Durbin, spatial error, and spatial lag models. These incorporate additional parameters that summarize the amount of influence surrounding points have on a given observation. Future research should incorporate these models to understand not only the influence of agglomeration on rents, but also account for the tendency of neighboring buildings to impact the price in a given building.

Chapter 5

Conclusion

In many ways, the econometrics of urban geography has not developed significantly since Alfred Marshall's specification of the three forces that underline agglomeration, or even since Paul Krugman's specification of New Economic Geography. Though researchers have access to more granular and more diverse data, we can only continue to prove, again and again, that agglomeration exists; we have not yet determined statistical sampling techniques or econometric experimental designs that account for the driving factors that cause individuals and firms to agglomerate in space.

This thesis supports and contributes to traditional urban econometrics in a variety of ways. First, I determine that agglomeration exists within small distances (100 to 500 meters) and is statistically associated with higher effective rents. This confirms the findings of generations of urban economists with the assistance of a rarely-utilized data set, so it serves as a valuable contribution to the literature in its own right.

Yet this thesis also pushes forward traditional urban econometrics by incorporating two data sources that speak to the differences faced by established firms in an industry versus those faced by startup entrants to the sector. My thesis shows that proximity to other established firms results in higher rent, as does proximity to additional startup funding. This dual finding may serve as an entry point into further inquiry pertaining to the underlying forces that cause firms to agglomerate. Established firms may have different informational, job-market, or access needs than startups within the same industry, thus causing them to choose different locations, even within the same market.

Despite this thesis' statistical findings, I also consider the weaknesses in traditional urban geography's exclusive focus on econometrics. I speculate that combining statistical analysis with qualitative data will yield a better understanding of the problem that has been plaguing econometricians and statisticians working on agglomeration: causation. By layering qualitative information into the research design and statistical wrangling process, we may be better equipped to prove that agglomeration exists, as well as to uncover the as-yet-unproven forces that drive the agglomerative impulse among firms. Combined, statistical and qualitative analysis can make our research more robust and provide findings that are significant and satisfying to economists as well as urbanists, unearthing findings that maintain mathematical rigor while promoting interpretability and applicability.

References

Acs, Zoltan J., and Attila Varga. "Entrepreneurship, agglomeration and technological change." *Small Business Economics* 24.3 (2005): 323-334.

Arzaghi, Mohammad, and J. Vernon Henderson. "Networking off madison avenue." *The Review of Economic Studies* 75.4 (2008): 1011-1038.

Audretsch, David. "Agglomeration and the location of innovative activity." *Oxford Review of Economic Policy* 14.2 (1998): 18-29.

Barkham, Richard, Sheharyar Bokhari, and Albert Saiz. "Urban big data: city management and real estate markets." GovLab Digest: New York, NY, USA (2018).

Blomquist, Glenn C., Mark C. Berger, and John P. Hoehn. "New estimates of quality of life in urban areas." *The American Economic Review* (1988): 89-107.

Brounen, Dirk, and Maarten Jennen. "Local office rent dynamics." *The Journal of Real Estate Finance and Economics* 39.4 (2009): 385.

Buzard, Kristy, et al. "The agglomeration of American R&D labs." *Journal of Urban Economics* 101 (2017): 14-26.

Cairncross, Frances. "The death of distance: How the communications revolution will change our lives." (1997).

Carlino, Gerald A., Satyajit Chatterjee, and Robert M. Hunt. "Urban density and the rate of invention." *Journal of Urban Economics* 61.3 (2007): 389-419.

Carlino, Gerald, and William R. Kerr. "Agglomeration and innovation." *Handbook of Regional and Urban Economics*. Vol. 5. Elsevier, 2015. 349-404.

Chegut, Andrea, Piet Eichholtz, and Nils Kok. "The price of innovation: An analysis of the marginal cost of green buildings." Center for Real Estate MIT Working Paper Series29.5 (2015): 34-67.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. "The identification of agglomeration economies." *Journal of Economic Geography* 11.2 (2010): 253-266.

Conway, Jennifer. *Artificial Intelligence and Machine Learning: Current Applications in Real Estate*. Thesis. Massachusetts Institute of Technology, 2018. Accessible on the web at: https://dspace.mit.edu/handle/1721.1/120609 (Accessed May 21, 2019)

Currid, Elizabeth. "How art and culture happen in New York: Implications for urban economic development." *Journal of the American Planning Association* 73.4 (2007): 454-467.

Currid, Elizabeth, and Sarah Williams. "The geography of buzz: art, culture and the social milieu in Los Angeles and New York." *Journal of Economic Geography* 10.3 (2009): 423-451.

Davis, Lindsay. "2019 Fintech Trends to Watch" CB Insights (2019).

De Silva, Dakshina G., and Robert P. McComb. "Geographic concentration and high tech firm survival." *Regional Science and Urban Economics* 42.4 (2012): 691-701.

Donner, Herman, Kent Eriksson, and Michael Steep. "Digital Cities: Real Estate Development Driven by Big Data."

Drennan, Matthew P., and Hugh F. Kelly. "Measuring urban agglomeration economies with office rents." *Journal of Economic Geography* 11.3 (2010): 481-507.

Drennan, Matthew P. "Do agglomeration economies decay over short distances? Are they stable in the face of shocks? Evidence from Manhattan." *International Journal of Urban Sciences* 22.1 (2018): 1-16.

Dunse, Neil, and Colin Jones. "A hedonic price model of office rents." Journal of property valuation and investment 16.3 (1998): 297-312.

Ellison, Glenn, Edward L. Glaeser, and William R. Kerr. "What causes industry agglomeration? Evidence from coagglomeration patterns." *American Economic Review* 100.3 (2010): 1195-1213.

Feldman, Maryann P. "The new economics of innovation, spillovers and agglomeration: A review of empirical studies." *Economics of innovation and new technology* 8.1-2 (1999): 5-25.

Florance, Andrew, et al. "System and method for collection, distribution, and use of information in connection with commercial real estate." U.S. Patent No. 6,871,140. 22 Mar. 2005.

Florida, Richard, and Karen King. "Rise of the Urban Startup Neighborhood." Martin Prosperity Institute Working Paper, 2016.

Florida, Richard, and Karen M. King. "Urban Start-up Districts: Mapping Venture Capital and Start-up Activity Across ZIP Codes." *Economic Development Quarterly* 32.2 (2018): 99-118.

Florida, Richard, Patrick Adler, and Charlotta Mellander. "The city as innovation machine." *Regional Studies* 51.1 (2017): 86-96.

Florida, Richard, and Charlotta Mellander. "Rise of the startup city: The changing geography of the venture capital financed innovation." *California Management Review* 59.1 (2016): 14-38.

Francke, Marc, and Alex Van de Minne. "Dealing with Unobserved Heterogeneity in Hedonic Price Models." *Available at SSRN 3249256* (2018).

Fujita, Masahisa, and Paul Krugman. "The new economic geography: Past, present and the future." *Papers in Regional Science* 83.1 (2004): 139-164.

Gabriel, Stuart A., and Stuart S. Rosenthal. "Quality of the business environment versus quality of life: do firms and households like the same cities?" *Review of Economics and Statistics* 86.1 (2004): 438-444.

García, Noelia, Matías Gámez, and Esteban Alfaro. "ANN+ GIS: An automated system for property valuation." *Neurocomputing* 71.4-6 (2008): 733-742.

Garmaise, Mark J., and Tobias J. Moskowitz. "Confronting information asymmetries: Evidence from real estate markets." *The Review of Financial Studies* 17.2 (2003): 405-437.

Ghysels, Eric, et al. "Forecasting real estate prices." *Handbook of Economic Forecasting*. Vol. 2. Elsevier, 2013. 509-580.

Glaeser, Edward L. *Urban colossus: why is New York America's largest city?* No. w11398. National Bureau of Economic Research, 2005.

Glaeser, Edward L., et al. "Big data and big cities: The promises and limitations of improved measures of urban life." *Economic Inquiry* 56.1 (2018): 114-137.

González, Marco Aurélio Stumpf. "Automated Valuation Methods in Real Estate Market—a Two-Level Fuzzy System." *Advances in Automated Valuation Modeling*. Springer, Cham, 2017. 265-278.

Gyourko, Joseph, and Joseph Tracy. "The structure of local public finance and the quality of life." *Journal of Political Economy* 99.4 (1991): 774-806.

Hashemian, Behrooz, et al. "Socioeconomic characterization of regions through the lens of individual financial transactions." *PloS One* 12.11 (2017): e0187031.

Jacobs, Jane. "The death and life of great American cities. 1961." New York: Vintage (1992).

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. "Geographic localization of knowledge spillovers as evidenced by patent citations." *The Quarterly Journal of Economics* 108.3 (1993): 577-598.

Kauflin, Jeff. "The 11 Biggest Fintech Companies in America 2019." Forbes, https://www.forbes.com/sites/jeffkauflin/2019/02/04/the-10-biggest-fintech-companies-in-america-2019/ (Accessed May 6, 2019).

Kim, Chong Won, Tim T. Phipps, and Luc Anselin. "Measuring the benefits of air quality improvement: a spatial hedonic approach." *Journal of Environmental Economics and Management* 45.1 (2003): 24-39.

Kok, Nils, Eija-Leena Koponen, and Carmen Adriana Martínez-Barbosa. "Big Data in Real Estate? From Manual Appraisal to Automated Valuation." *The Journal of Portfolio Management* 43.6 (2017): 202-211.

Kontokosta, Constantine E., and Nicholas Johnson. "Urban phenology: Toward a real-time census of the city using Wi-Fi data." *Computers, Environment and Urban Systems* 64 (2017): 144-153.

Koster, Hans RA. "Rocketing rents: The magnitude and attenuation of agglomeration economies in the commercial property market." (2013).

Koster, Hans RA, Jos van Ommeren, and Piet Rietveld. "Is the sky the limit? High-rise buildings and office rents." *Journal of Economic Geography* 14.1 (2013): 125-153.

Kroft, Kory, and Devin G. Pope. "Does online search crowd out traditional search and improve matching efficiency? Evidence from Craigslist." *Journal of Labor Economics* 32.2 (2014): 259-303.

Krugman, Paul. "Increasing returns and economic geography." *Journal of Political Economy* 99.3 (1991): 483-499.

Krugman, Paul "What's New About the New Economic Geography?" *Oxford Review of Economic Policy* 14.2 (1998): 7-17.

Krugman, P. (2009). The increasing returns revolution in trade and geography. *American Economic Review*, 99(3), 561-71.

Lerner, Josh. *The government as venture capitalist: The long-run effects of the SBIR program*. No. w5753. National Bureau of Economic Research, 1996.

Liu, Crocker H., Stuart S. Rosenthal, and William C. Strange. "The vertical city: Rent gradients, spatial structure, and agglomeration economies." *Journal of Urban Economics* 106 (2018): 101-122.

Liusman, Ervi, et al. "Office rents, mixed-use developments, and agglomeration economies: a panel data analysis." *Journal of Property Investment & Finance* 35.5 (2017): 455-471.

Loesch, August. The economics of location: Translated from the second rev. German ed. by William H. Woglom with the assistance of Wolfgang F. Stolper. Yale University Press, 1954.

Mariotti, Sergio, Lucia Piscitello, and Stefano Elia. "Spatial agglomeration of multinational enterprises: the role of information externalities and knowledge spillovers." *Journal of Economic Geography* 10.4 (2010): 519-538.

Marshall, Alfred, and Mary Paley Marshall. "The Economics of Industry." Macmillan and Company, 1920.

Melanda, Edson, Andrew Hunter, and Michael Barry. "Identification of locational influence on real property values using data mining methods." *Cybergeo: European Journal of Geography* (2016).

Melo, Patricia C., Daniel J. Graham, and Robert B. Noland. "A meta-analysis of estimates of urban agglomeration economies." *Regional Science and Urban Economics* 39.3 (2009): 332-342.

Metropolitan Transit Authority (MTA). Annual Subway Ridership. Accessible online at: http://web.mta.info/nyct/facts/ridership/ridership_sub_annual.htm. (Accessed May 21, 2019).

Moretti, Enrico. The New Geography of Jobs. Houghton Mifflin Harcourt, 2012.

Nobel Committee. *The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2008*. https://www.nobelprize.org/prizes/economic-sciences/2008/summary/(Accessed May 21, 2009).

Pace, R. Kelley, Ronald Barry, and Clemon F. Sirmans. "Spatial statistics and real estate." *The Journal of Real Estate Finance and Economics* 17.1 (1998): 5-13.

Peng, Liang. "Benchmarking Local Commercial Real Estate Returns: Statistics Meets Economics." *Real Estate Economics* (2018).

Puga, Diego. "The magnitude and causes of agglomeration economies." *Journal of Regional Science* 50.1 (2010): 203-219.

Puri, Zoya, Suneeth John, and Andrea Chegut. "Does Work Performance Design Impact Value? Linking Design Metrics to Financial Performance in Cities." Working Paper accessible at: http://reilab.wpengine.com/research_article/does-work-performance-design-impact-value-linking-design-metrics-to-financial-performance-in-cities/ (Accessed May 21, 2019).

Quan, Daniel C., and John M. Quigley. "Price formation and the appraisal function in real estate markets." *The Journal of Real Estate Finance and Economics* 4.2 (1991): 127-146.

Quigley, John M. "Urban diversity and economic growth." *Journal of Economic Perspectives* 12.2 (1998): 127-138.

Ratti, Carlo, et al. "Mobile landscapes: using location data from cell phones for urban analysis." *Environment and Planning B: Planning and Design* 33.5 (2006): 727-748.

Rosen, Sherwin. "Hedonic prices and implicit markets: product differentiation in pure competition." *Journal of Political Economy* 82.1 (1974): 34-55.

Rosenthal, Stuart S., and William C. Strange. "Geography, industrial organization, and agglomeration." *Review of Economics and Statistics* 85.2 (2003): 377-393.

Rosenthal, Stuart S., and William C. Strange. "Evidence on the nature and sources of agglomeration economies." *Handbook of Regional and Urban Economics*. Vol. 4. Elsevier, 2004. 2119-2171.

Rosenthal, Stuart S., and William C. Strange. "The micro-empirics of agglomeration economies." *A Companion to Urban Economics* (2006): 7-23.

Salesses, Philip, Katja Schechtner, and César A. Hidalgo. "The collaborative image of the city: mapping the inequality of urban perception." *PloS One* 8.7 (2013): e68400.

Saxenian, AnnaLee. "Regional networks: industrial adaptation in Silicon Valley and route 128." *Regional Networks and Industrial Adaptation*. (1994).

Sobolevsky, Stanislav, et al. "Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain." *2014 IEEE International Congress on Big Data*. IEEE, 2014.

Von Thünen, Johann Heinrich. *Isolated state: an English edition of Der isolierte Staat*. Pergamon Press, 1966.

Wallsten, Scott J. "An empirical test of geographic knowledge spillovers using geographic information systems and firm-level data." *Regional Science and Urban Economics* 31.5 (2001): 571-599.

Wilhelmsson, Mats. "Spatial models in real estate economics." *Housing, Theory and Society* 19.2 (2002): 92-101.

Winson-Geiderman, Kimberly. "The Office Property and Big Data Puzzle: Putting the Pieces Together." NAIOP Research Foundation, August 2018.