Towards an Epidemiology of Gentrification: Modeling Urban Change as a Probabilistic Process using k-Means Clustering and Markov Models

By

Emily Binet Royall

B.S. Neuroscience, B.A. Plan II
University of Texas at Austin
Austin, Texas (2011)

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

Master in City Planning

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2016

Signature redacted

Author_____
Department of Urban Studies and Planning
(January 26th, 2015)

Signature redacted

Certified by _____
Dr. Andrea Chegut
PhD., Center for Real Estate
Thesis Supervisor

Signature redacted

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

Towards an Epidemiology of Gentrification: Modeling Urban Change as a Probabilistic Process using k-Means Clustering and Markov Models

By

Emily Binet Royall

B.S. Neuroscience, B.A. Plan II
University of Texas at Austin
Austin, Texas (2011)

Submitted to the Department of Urban Studies and Planning
on January 14th, 2016 in Partial Fulfillment of the
Requirements for the Degree of
Master in City Planning

ABSTRACT

Gentrification is viewed as both as a tool and a force--as a systematized vehicle for class-based oppression and racism, and an empirical force of change based on social, environmental and economic interactions. This complexity makes it challenging for researchers to study the impact of gentrification, for planners to anticipate the effects of gentrification with planning policy, and for developers to foresee investment outcomes. Current planning policy addresses the symptoms of gentrification, without defining the underlying construct of the process. This thesis examines latent constructs of gentrification through a data-driven process that identifies emergent states of change and assigns them to a Markov process, i.e. a process that assigns probabilities to potential "state" changes over time. For census block groups in four boroughs of New York City, this model takes three steps: 1) cluster census block groups into latent states defined by ACS socioeconomic and demographic data, 2) derive a Markov model by tracking transitions between states over time, and 3) validate the model by testing predictions against historic data and qualitative documentation. Using this process I was able to find emergent typologies of urban change, locate gentrifying neighborhoods without any spatial input, and uncover sequences of patterns that reliably predict socioeconomic outcomes at the census block group level. Through the design of a machine learning framework for gentrification I reflect on the importance of using algorithms that learn rather than reproduce assumptions, value of distilling large and complex data relationships into nuanced intuitions, and challenges of embedding computational modeling in political frameworks.

Thesis Supervisor: Andrea Chegut
Title: PhD, Research Scientist Center for Real Estate

Biographical Note

Emily was raised in San Antonio, Texas. She received a B.S. in Neuroscience and a B.A. in Plan II from the University of Texas at Austin. In her transition to city planning, Emily is interested in transforming and re-imagining our relationships to cities in the age of Information. In her graduate work at MIT, Emily has studied digitally mediated urban development, and its impact on human behavior and urban culture. She hopes to carry forward her knowledge and experience to promoting healthy, inclusive cities in the 21st century.

Acknowledgement

This body of work incorporates the talents, generosity, care and genius of a network of incredible people. I came up with the thesis concept during my employment at City Form Lab under the guidance of Andres Sevstuk, also a reader of this thesis. Andres helped me hone my quantitative thinking and encouraged me to keep the needs and perspectives of communities at the forefront of analysis. Thomas Wortmann, at Singapore University of Technology & Design and an author on the original paper presented at CUPUM was a significant contributor to this work, and amazing intellectual collaborator, staying up late nights despite time zone differences to discuss ideas. Neil Gershenfeld introduced me to tools in machine learning. Brent Ryan at MIT guided me in the decision to turn a crazy idea into a thesis. Sarah Williams graciously donated her expertise to help me solidify my data processing and mapping methods, and gave me wonderful insight into the personalities of the New York boroughs. I would like to extend my deepest gratitude to Andrea Chegut my advisor, who listened and cared. Her ongoing support, encouragement, and insistence on rigor is the core of this thesis, and helped me figure out "what to do with an idea." I must also thank Jana McCann, my mentor and role model and Zach Linge, my greatest friend, whose wisdom and love shaped my constitution. Finally, thanks to my parents Don and Caroline, who gave me the opportunity to explore.

Foreword: An Ethics of Computation for Cities

The historic city, as artifact[1], is an interface between human perception and environmental constraint. This dynamic tension has produced a genealogy of urbanism spanning from slum settlements in Denpasar to the grand boulevards of Paris. These places exhibit complex and networked patterns of behavior (Jacobs, 1961). As Christopher Alexander observed, such patterns form a language that one can learn to speak through design (Alexander, 1977). Ultimately, he outlined a theory of design focused on learning the latent principles that generate healthy cities by careful observation of patterns of human interaction in the built environment over time. Michel Cantal-Dupart applies this in practice in what he calls an "Aesthetic of Equity," where attentive design strategies embed projects in local culture and facilitate social dynamics.

A distinctly different urban condition is present today. Contemporary urbanism reflects a societal perception that is progressively informed by digital culture. This can be seen in emerging developments that harmonize more with digitally traded capital, personalized media, and Internet memes. Such properties trade well internationally as REITs[2] in retirement portfolios but are often experientially static and detached from the human condition. Today, exponential population growth,

---

[1] Simon, Herbert. *The Sciences of the Artificial,* 1969. "An artifact can be thought of as an "interface"...between an 'inner' environment, the substance and organization of the artifact itself, and an 'outer' environment, the surroundings in which it operates (pg. 6)."

[2] Real Estate Investment Trusts; a type of security that invests in real estate through property or mortgages and often trades on major exchanges like a stock.

globalized labor and capital, and ubiquitous reliance on digitally mediated infrastructure make the rapid development of cities through algorithmic means more common (Sassen, 1994). Algorithms are ubiquitous in determining the form, function, and value of the built environment. Is it possible to speak the language of Alexander's *Quality*[3] or build an Aesthetic of Equity using algorithms? Perhaps this is no longer a question but rather an imperative: to design urban systems that remain compassionate in an overpopulated, overheated, digitally mediated future we need to shape a common language of computational ethics for urban development.

Algorithms are embedded in many systems that shape human experience, and are likewise shaped by our cultural perceptions. In *Economic Complexity,* Brian Arthur demonstrates how making choices based on modeled predictions actually brings forecasted expectations into reality (Arthur, 2014). In *Software Sorted Geographies,* Stephen Graham applies this theory to urbanism by demonstrating how computational culture affects our experience of cities through geographic information systems software (GIS), public space surveillance, and transportation systems (Graham, 2005). Despite these important vignettes, both professional practice and academic discourse lack a critical theory of algorithmic urbanism.

What would a language of computational ethics for urbanism look like? Answering this question is beyond the scope of this thesis, yet I believe what is

[3] The "Quality" as described by Christopher Alexander in A Timeless Way of Building: "There is a central Quality which is the root criterion of life and spirit in a man, a town, a building or a wilderness. This quality is objective and precise, but it cannot be named."

presented here is a first step. Computational ethics for urbanism would leverage the opportunities of big data to learn a pattern language of urbanism in a variety of contexts; both formal and informal, rural and urban, home and abroad. Using Algorithms that *learn* rather than reproduce assumptions, presents an opportunity for researchers to see the unseen qualities in their data, distilling large and complex data relationships into nuanced intuitions. The conclusions of such an analysis may reproduce a human or professional understanding of a neighborhood trend, but by arriving at such an intuition using algorithmic design, we become aware of the causes, mechanisms or correlations that result in the intuited pattern. Marrying data driven insight with community engagement can help us direct policy that intervenes in mechanisms, rather than treating the side effects of urban phenomena like gentrification.

The field of machine learning is currently undergoing a renaissance, and the suite of new techniques emerging from this fast-evolving field pose a horizon of opportunity for data scientists and urbanists alike. However I learned in the process of building the machine learning framework outlined in this thesis, that designing a machine learning algorithm (or any algorithm at all), is a highly subjective process requiring the artful arrangement of technical elements each capable of telling a different story about data. Therefore, an ethics of computation for urbanism would also recognize the role that data sources and political institutions play in the validity of the data-generating process.

Through writing this thesis, I've come to appreciate what Pablo Picasso meant by

"computers are useless, they only give you answers." Designing a quantitative tool is in fact an art, and the subjectivity therein is an opportunity to embed participation and local knowledge. Earlier attempts to apply machine learning to the analysis of segregation in inner city L.A. neighborhoods[4] dismissed critical elements of perception, culture and community participation. I hope to move past the political confines of cybernetics[5], by combining these elements into analytical tools embedded in equitable participatory frameworks.

As the urban condition evolves in tandem with digital culture, it's necessary that we consider what a computational ethics for urban development implies. How does digital culture shape our perception of human needs in urban environments? How does perception in turn shape computational modes of design for places? How and where has algorithmic design failed or succeeded in history? Can we meet the growing needs of urban metropolises using technology that responds to the complex nuances of human behavior and culture? These are some of the questions I hope to shed light on in the work that follows.

---

[4] Light, Jennifer S. *From Warfare to Welfare: Defense Intellectuals and Urban Problems in Cold War America*. Baltimore: Johns Hopkins UP, 2003.

[5] Goodspeed, Robert. "Smart Cities: Moving beyond Urban Cybernetics to Tackle Wicked Problems." *Cambridge Journal of Regions, Economy and Society* 1.13 (2011).

Contents

1 Introduction

This thesis uses machine learning to uncover a syntactic structure in socioeconomic data that defines gentrification. By applying this method to ACS data in five New York City boroughs, I was able to find emergent typologies of urban change, locate gentrifying neighborhoods without any spatial input, and uncover sequences of patterns that reliably predict socioeconomic outcomes at the census block group level. I believe this is the start of uncovering a data-driven "pattern language," following Alexander and Dupart, as the algorithm learns from observed behavior to find relationships that contribute to outcomes like gentrification. Finding these latent relationships may help planners design policy that targets mechanisms of gentrification, rather than treat its symptoms.

Gentrification is an urban phenomenon described by complex interactions among social, economic, political and environmental forces and whose definition is disputed in academic literature. The very nature of gentrification is subject to hot debate centered on whether or not it is a product of nature or artifice. In Chapter 1, I examine the gentrification debate and propose that a conceptual reframing of gentrification as the form of artifice described by Simon, lends it to computational study using machine learning--- closing the gap between natural and artificial interpretations of the phenomenon. Furthermore, I examine the history of computational analysis of urban change and gentrification, situating it in the political construct of Cybernetics as it emerged in Los Angeles in the 1970s, which ultimately led to the demise of quantitative practices in urban planning and

design (Light, 2011). I suggest that the conceptual framework provided here, can help shift practitioners and theorists away from the historic pitfalls of Cybernetics and towards a more inclusive, genuine practice of computational analysis in urbanism. Furthermore, I argue that this perspective is drastically needed to ameliorate the "cultural split" (Portugali, 2006) between the natural and social sciences, and "space" vs. "place" perspectives of urban geography known to stifle innovation and progress in urban studies. More importantly, gentrification is a condition of modern urban culture that is unsustainable for many communities around the world. Current policy frameworks combat gentrification by treating its symptoms: rent control programs insulate neighborhoods from urban change, stabilization vouchers try to assuage displacement, property tax freezes protect longtime residents from the effects of rising property values. Notably, many of these programs emerged under strenuous bureaucratic and political climates and are more or less successful given these powerful constraints. However, I hope that this work hints at the possibility of uncovering *mechanisms* of gentrification, where policy and design intervention may be more successful. At the very least, this research provides insight into the phenomenon itself, by identifying typologies of gentrification and revealing interdependencies that may be responsible for its emergence. *Chapter 1* reviews the gentrification debate, the qualitative and quantitative techniques used to model gentrification, and situates computational perspectives in the Cybernetics discourse. *Chapter 2* is an overview of Pattern Classification and Machine Learning techniques, and highlights the selected methods for this analysis. *Chapter 3* reviews the

advantages and disadvantages of American Community Survey data employed in this project, and provides descriptive statistics for the results of a feature extraction process. *Chapter 4* is a complete journal-ready paper (pending publication) that describes the machine learning techniques used to analyze urban change and gentrification, and reports the results. *Chapter 5* is a broader discussion of results, implications for planning and future research opportunities.

Why does it matter? Mechanisms can be understood and redesigned, whereas symptoms can only be treated. Interventions should occur as a design processes at the mechanism level, not the surface level. Jane Jacobs (1961) said as much in terms of the direction future urban design should take:

> My idea, however, is not that we should try to reproduce, routinely and in a surface way, the streets and districts that do display strength and success as fragments of city life...if we understand the principles behind the behavior of cities, we can build on potential assets and strengths, instead of acting at cross-purposes to them (140).

Our job as planners and designers is to uncover the mechanisms that generate the emergent outcomes we observe. In the case of gentrification, where the symptoms are so controversial and contribute keenly to the commodification and decomposition of cultures and communities, further study and action is desperately needed.

2 Gentrification

2.1 The Gentrification Debate

In 2014, Tom Slater, Reader of Urban Geography at the University of Edinburg published a seething critique of an article written by Philip Ball, the editor of the American scientific journal, *Nature*. "Gentrification is a Natural Evolution," described gentrification as a natural force underpinning the evolution of cities. The piece reviewed a recent paper "The Form of Gentrification" (Veneradi, 2014) published earlier that year in *Physics and Society,* which took a Complexity Theory of Cities (CTC) approach to identifying emergent properties of neighborhood characteristics and statistically correlated them to gentrification. As Ball put it, the work suggested that "cities obey laws beyond the reach of planning," likening the study of cities to that of the evolution of biological organisms[6]. Slater wrote an incensed piece condemning Veneradi's work and Ball's support of it, arguing that Complexity theorists failed to consult the broad body of urban studies literature on the subject. Slater rejected the reinvention of the wheel on the part of so called "urban scientists" who naively drew conclusions about a phenomenon that had been heavily researched by a historic parade of architects, geographers and urban planners (Slater, 2014). Slater (2014) contends:

---

[6] See Gentrification is a Natural Evolution: http://www.theguardian.com/commentisfree/2014/nov/19/gentrification-evolution-cities-brixton-battersea

> The authors think that the role of 'urban form' in processes of gentrification has not been subject to enough scrutiny, yet as far back as 1986 scholars were writing about the production and consumption of a 'gentrification aesthetic', and have continued to do so with theoretical guidance from Pierre Bourdieu's immense body of work on class, habitus, field, and taste (para. 6).

Slater's abrupt outbreak warns of a commonly perceived danger associated with the CTC analysis of gentrification: that if an urban process resulting in the destruction of communities is understood to be "natural," it will be justified by the political elite to further marginalize and exploit minorities and low income communities (Slater, 2014). Slater's retort is echoed by public commentary on the article, as well as in his own work published immediately after the outrage: *The Eviction of Critical Perspectives from Gentrification Research*, which laments the lack of research engagement with the very communities that suffer displacement and class-based oppression as a result of gentrification.

In his blog, "Homunculus" Ball responds to Slater's attack acknowledging the lack of awareness of existing gentrification research in the scientific community, but charging that the urban theorist had confused urban science with Social Darwinism. As Ball (2014) writes in response to Slater;

> I can understand this point of view to some degree, for certainly a 'naturalism' based on a misappropriation of Darwinian ideas has been used in the past to excuse the rapaciousness of capitalist economies. But to imagine that this is what a modern "complexity" approach to social phenomena is all about seems to me to reflect a deep and possibly even dangerous confusion. The aim of such work is, in general, to understand how certain consequences emerge from the social and institutional structures we create. These consequences might sometimes be highly non-intuitive in ways that simple cause-and-effect narratives can't hope to capture (The Science and Politics of Gentrification, para. 6).

Ball's response is echoes the broader position of CTC theorists: that Complexity offers methodologies designed to capture 'non-intuitive' or latent socioeconomic structures of urban processes that are assumed to behave as complex systems. The attraction of these modeling capacities according to CTC theorists is the possibility that they may uncover hidden insights into a living system. CTC theorists, to the chagrin of their poststructuralist counterparts, take a weak positional stance regarding the political frameworks surrounding their work.

This public debate between a scientist and humanist about cities reflects a critical rupture in today's urban development community. At the core of this rift is gentrification, viewed both as a tool and a force---as a systematized vehicle for class-based oppression and racism, and an empirical force of change based on

social, environmental and economic interactions[7]. Is gentrification a political or natural process? Are political and natural processes necessarily different? This question has long motivated researchers of urban change and gentrification studies, and continues to stimulate heated contemporary debate. The two sides of the gentrification debate reflect a broader, historic divide between quantitative, positivists and qualitative, poststructuralists (Portugali, 2009). The long and turbulent history of gentrification studies emerging from these two opposing perspectives has left us with few answers; while gentrification continues disrupt the composition of our communities, transforming cultures into commodities.

Given how detrimental the experience of gentrification can be for many communities, the question as to whether or not gentrification is "natural' is moot. At its best, gentrification is inefficiency. Because development capital fails to reinforce cultural capital, real estate investment will always be behind the curve: it will always rely on cultural actors to drive investment choices resulting in the inevitable displacement, and the subsequent "death" of authentic communities. In the remainder of this chapter, I argue that there is a cultural split between academic groups that focus on studying communities as spaces and those that study their subjective qualities as places. The split between the quantitative, positivist perspectives of communities (the "space" perspective) and qualitative, poststructuralist perspectives (the "place" perspective) resulted in a body of

[7] This is a Newtonian reading of "force" whose third law describes a force as an interaction between different bodies. The analogy of "force" as applied to gentrification is intended to reflect the recent contributions of physicists to the debate on gentrification studies and urban change.

gentrification literature that can't agree on virtually any aspect of gentrification, much to the detriment of bringing positive change to residents in gentrifying neighborhoods. The development community has embraced the former view resulting in the "colonization" of cultural capital, which reproduces the gentrification phenomena. The CTC approach to cities, which to date has been vastly misunderstood by the urban theorist community, is proposed as a framework for bridging this gap between "space" and "place" perspectives, and shapes a sustainable development practice that can transform the nature and impact of gentrification on our communities.

2.2 Gentrification: Embodying the Cultural Spilt

Key Elements

- *The Gentrification Debate illustrates broader cultural split between natural and social sciences on urban issues.*
- *The two opposing views of gentrification represent "space" and "place" based scientific cultures.*
- *Outdated modeling practices have led the development community to embrace a "colonization culture," that contributes to gentrification.*

The two sides of the gentrification debate reflect the broader, historic split between quantitative, positivist (QP) and qualitative, poststructuralist humanist (QPH) views on the development of cities (Portugali, 2009). This split is what Juval Portugali refers to in *Complexity Theory as a Link Between Space and Place,* as a cultural split between the two "cultures of science" (Portugali, 2006),

the humanities and natural sciences (Snow, 1964). The split between these two cultures is well documented in geography and endlessly demonstrated in urban studies, particularly surrounding issues of urban change and gentrification. The bifurcation emerged in the 1970s when leading figures of the Cybernetics movement representing the quantitative, positivistic perspective of urban geography became self-critical of the optimization movement, the use of computational optimization tools to generate "ideal" urban scenarios (Harvey, 1969). Specifically, Civil Rights activists in America after the Vietnam War criticized the outcomes of quantitative approaches to urban planning, calling out the discrimination these approaches justified in the name of optimization through practices such as redlining (Light, 2011)[8]. This troubling history equipped David Harvey's *Social Justice and the City*, a seminal critique from the Marxist-structuralist perspective of quantitative approaches to understanding and designing communities (Harvey, 1973). Humanistic geography emphasized "place" through examination of the intimate experiences and connections between people and their homes, neighborhoods and cities, over "space"--- the simplistic quantitative analysis of topography detached from human cognition.

The bifurcation lives on in the gentrification debate. On one hand, urban theorists backed by decades of scholarly work and qualitative studies on urban change advocate that gentrification is not a "natural phenomena" but instead an

[8] See documentation of the Los Angeles Community Analysis Bureau for a specific history of how machine learning applications were used to target low income communities for redlining and crime prevention tactics.

institutionalized vehicle of oppression dictated by the uneven flow of capital (Slater, 2006; Slater, 2014; Lees, 1994). Because gentrification is the result of capital flows, scholars in this camp argue for an upheaval of political cultures as the best means to combat gentrification. This position is defined by the seminal Rent Gap model (Smith, 1979; Smith, 1987), which illustrates the intentional disinvestment in neighborhoods for the purpose of exploiting marginalized communities for capital gain. The model captures the position of its advocates by suggesting that the gap produced between capitalized ground rent, and potential ground rent is intentionally designed (Slater 2015). Therefore, the Rent Gap model initiated a reading of gentrification that was built primarily on the flow of "capital not people" (Smith, 1979). According to Slater, the uneven distribution of capital has produced a landscape of tension between gentrification and disinvestment. Calling gentrification a "systemic and structural problem" Slater builds from this view, proposing to change the culture around gentrification in order to shape it in the interest of local communities. However, some tension exists within the urban theorist camp regarding economic versus cultural explanations for gentrification, and more recently an effort has emerged amongst scholars to connect opposing views within the camp. In *Rethinking Gentrification: Beyond the Position of Economics or Culture*, Loretta Lees argues that spatially, economic capital mirrors cultural capital (Lees, 1994).

On the other hand, Complexity scientists, gathering momentum as a movement

in the 1990s, advocate for a "Complex Systems Theory of Cities (CTC)" perspective on gentrification, which understands gentrification as a "Complex System," one characterized by measurable, but generalized properties that are shared with other living, social and economic entities. These properties describe stochastic systems, for example the property of "emergence" where local interactions between agents produce emergent effects that in turn shape the behavior of individual agents (Batty, 2005; Alexander,1977; Bettencourt, 2013). CTC theorists liken cities to living organisms, taking Jane Jacobs as their patron saint who argued "Cities happen to be problems in organized complexity, like the life sciences (Jacobs, 1961)." CTC theorists argue that because urban processes are dynamic and complex, they are difficult to model and predict. This calls for the use of a suite of simulation tools borrowed from the natural sciences (Batty, 2005). As a result, CTC is associated with the broader "City Science" movement, which emphasizes the use of simulation, modeling and quantitative methods for the study of cities. City Science is challenged as being the next generation of the flawed Cybernetics movement of the 1960s, which produced analyses that supported the destruction and deterioration of urban communities during the urban renewal movement (Light, 2011; Light, 2003). Cybernetics in the United States was crippled by lack of computational power, poor data availability, and positioned in an institutional framework that emphasized using military strategy for urban development.

Cybernetics, and several models of urban change and gentrification of this era

were informed by a scientific revolution in the ecological sciences, resulting in spatial modeling practices that mimicked spatial modeling problems in living ecosystems such as migration and succession. This type of simulation lent itself well to issues of race and space facing inner city American neighborhoods, most notably in Los Angeles (Light, 2011)[9]. Because ecological models enabled the study of migration patterns and patterns of segregation, they were instrumental in justifying a race and class-based planning agenda in many cities around the turn of the century. Perhaps the most notorious example occurred in the 1930s when the scientific revolution in ecology inspired planners who adapted ecological models to analyze migration patterns in America's segregated neighborhoods. The outcome of this cross-pollination was a suite of tools used to standardize risk-assessments in national real-estate industries following the Great Depression (Light, 2011). In some municipalities this directly led to the notorious redlining practice. The suspicion of the City Science movement from the perspective of urban theorists and planners, and by default of CTC methods, is derived from this important history. However, CTC proposes a distinctly different modeling framework, rooted instead in emerging concepts in computer science, neurobiology and bioinformatics (Batty, 2005).

[9] See "How LA used Big Data to build a Smart City in the 1970s." http://gizmodo.com/uncovering-the-early-history-of-big-data-in-1974-los-an-1712551686

2.3 Conflicting Definitions Gentrification: Evidence of the Split

Key Elements

- *Lack of consensus regarding the definition of gentrification is a result of the split between "space" and "place" based cultures.*
- *The definition of gentrification has changed alongside scientific revolutions, new technologies and political agendas.*
- *A consensus about gentrification's definition is needed for productive research to ensue.*

QP and QPH scholarly camps have produced a large body of work of gentrification studies, but also a lack of consensus regarding both causes and outcomes of the process. Likewise, the term's meaning has evolved in harmony with scientific revolutions and political agendas, as well as different computational models. Emerging cultural and scientific zeitgeists always seem to offer new ways of thinking about the problem, but once tools are developed they are often implemented as a means to the ends of political actors with specific agendas.

Urban change has a long history of empirical study (Du Bois, 1899; Weber, 1899). Gentrification, a specific typology of urban change was documented relatively recently however, with the language to describe it formally coined by Ruth Glass in 1964. The emergence of gentrification in both idea and form during the mid 1960s posed a challenge to traditional ideas about how urban change worked. Observations of gentrification described by inner-city reinvestment, socioeconomic and demographic compositional changes, and rising property

values at the urban core challenged the conventional wisdom that urban change was a process restricted to the periphery. Historically, neighborhood change was viewed as a process of perpetual expansion on the periphery, driven by class and capital in which the upper classes were thought never to return to older neighborhoods (Hoyt, 1939). Likewise, Burgess's traditional concentric model predicted growth and change on the urban fringe (Burgess, 1923). These early perspectives understood urban change as a symptom of suburban growth in a post-war housing market. Scholars of urban change in the 1960s and early 1970s began documenting how instances of gentrification deviated from convention and these cases demonstrate some of the early frameworks shaping gentrification studies of this period (Freeman, 2005).

Late 20th century gentrification studies confirmed that a process of urban change was occurring in many inner-city communities across the US during the 1970s (Clay 1979; Sumka 1979; National Urban Coalition, 1977). Seminal views emerging from these observations have characterized gentrification as urban re-investment resulting in displacement of the poor (Lees et al., 2008; Clay 1979), a result of the "rent-gap" hypothesis (Smith, 1979) and a symptom of regional economic or demographic change (Clay, 1989). While these scholars agreed that a unique pattern of urban change was emerging from America's inner city neighborhoods---they were unable to settle on causes or definitions of the process. As an interdisciplinary problem gentrification was difficult to define and both the social and natural sciences had their own views. The widening cultural

divide between natural and social scientists further aggravated this lack of consensus in the late 20th century (Portugali, 2006).

A robust definition of gentrification does not exist today. The U.S. Department of Housing and Urban Development currently defines gentrification as "the process by which a neighborhood occupied by lower-income households undergoes revitalization or reinvestment through the arrival of upper-income households (U.S. Department of Housing and Urban Development, 1979.)." This definition does not include displacement as a necessary outcome of gentrification. Alternatively, the Brookings Institution cites gentrification as "the process of neighborhood change that results in the replacement of lower income residents with higher income ones (Brookings Institution, 2001. pg 11)." The Brookings Institution draws a clear connection between gentrification and displacement caused by the arrival of a specific actor: a high-income demographic. However, a measure of gentrification relying on solely income could dismiss gentrifying neighborhoods that experience a high influx of an educated, but not necessarily high-income class (Clay, 1979; Freeman, 2005). Alternatively, the *Encyclopedia of Housing* (Smith 1998) defines gentrification as "the process by which central urban neighborhoods that have undergone disinvestments and economic decline experience a reversal, reinvestment, and in-migration of a relatively well-off, middle and upper middle class population." Unlike the Brookings Institution, the Encyclopedia of Housing does not ascribe agency to any one actor as responsible for gentrification, and dismisses the typology of gentrification that is

documented in well-invested, but relatively low-income neighborhoods. While these authorities agree that gentrification is fundamentally a "process," the causes and effects remain open for interpretation, and are often in conflict with each other.

Quantitative approaches to defining urban change also reflect the QP and QPH cultural split. As a result, models of urban change and gentrification often fall into binary categories of causality (government-assisted vs. market-driven, spatial vs. social), outcomes (displacement vs. succession) and modes (political agents vs. natural forces). Two important perspectives emerge from the division: the "space" perspective and "place" perspective (Portugali, 2006), where models emphasize data-driven spatial outcomes, or humanistic experiences and political outcomes. Overall, researchers over the decades have drawn little consensus about the causes and outcomes of gentrification, and models have disappeared and re-emerged in the popular discourse.

2.4 Features Emerging from QP and QPH Modeling Studies

Despite the fact that QP and QPH modeling studies have not reached conclusions regarding the outcomes and causes of gentrification, this thesis proposes to assesses whether a data-driven, CTC-based method would be more successful in identifying patterns of urban change related to gentrification than previous methods. Pattern classification, or "machine learning" methods allow

data to "speak" by recovering models responsible for generating observed patterns (Duda, 2011). This thesis proposes to compare pattern classification approaches where elements of gentrification patterns are either assumed or unknown. In the case where features related to gentrification are assumed, we rely on the diversity of quantitative analyses discussed previously to target features that may be related to gentrification. These features are income, (Brookings Institution, 2011; Smith 1979), education level (Clay, 1979; Freeman, 2005; Ley, 1986), density and percentage family households (Hoover & Vernon, 1959; Birch, 1979, Marcuse 1985, Kolko, 2010), structure age (Hidalgo, 2015; Birch, 1979), land use (Smith, 1979; Burgess, 1923), employment status (Ley, 1986) and race (Beauregard 1986; Hamnett, 1991; Slater, 2006; Schelling, 1971). The selected features span across both physical and social dimensions of urban development, and accordingly the results may provide insight into the recent discussion about whether or not physical changes in neighborhoods anticipate social change (Hidalgo & Glaeser, 2015). It is hypothesized that patterns of gentrification must contain some or all of these elements, as they represent the findings of the greatest portion of quantitative literature on the subject over the last 60 years.

2.5 QP Models of Urban Change: Space-based Perspective

Key Elements

- *Space-based models describe urban change in terms of spatial actors and outcomes.*
- *Space-based models use modeling tools to tie data to spatial boundaries, and are informed by advances in natural sciences.*
- *Stage models, Neighborhood Life Cycles and spatial simulations characterize the QP models of urban change.*

The "space"-based perspective sees urban change in terms of spatial actors and outcomes, and uses modeling tools that tie data to spatial boundaries for the purpose of organizing information in a computationally friendly way. This view was adapted from location theory, which was the foundation of QP approaches to urban geography (Portugali, 2006). From this perspective spatial interaction between bodies and settlements, central places and demand, is governed by spatial forces such as distance measured by transportation costs (Portugali, 2006). Largely informed by spatial modeling in ecology, these early models of urban form saw space as a landscape of physical forces and largely neglected dimensions of human experience and culture.

2.5.1 Neighborhood Life Cycles and Stage-models

Several scholars have framed urban change and gentrification in terms of lifecycles, stages of development, and evolution (Burgess, 1929; Hoyt, 1939; Smith, 1979; Birch, 1971; Hoover & Vernon, 1959; Vernon, 1959). For decades

Burgess's concentric model of urban growth was the conventional wisdom regarding urban change, and several urban change researchers built on the principles of his work. Burgess's model showed concentric rings of growth and disinvestment, suggesting that cities expanded on the periphery, as higher-income classes developed or purchased new property, and lower-income classes would move into the neighborhoods left behind. At the time the concentric ring model provided an adequate explanation for inner-city decline, a major concern of planners and statesmen of the mid 20th century. Furthermore, the model provided a framework for thinking about urban change as a succession of spatial states---it was the first time patterns of urban change were observable at the metropolitan scale.

Figure 1

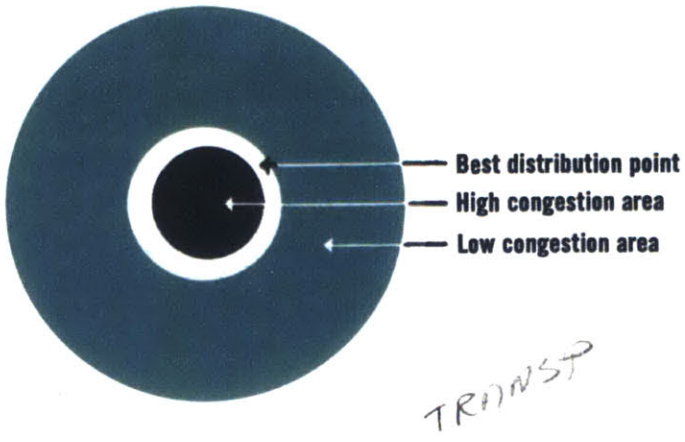

Figure 1 The concentric model as adapted by Vernon in "The Changing Economic Function of the Central City."

The concentric, stage model of urban development proved fruitful for quantitative

analysis and policymaking for several decades. In "The Changing Economic Function of the Central City," Vernon identifies an "early" and "late" state of gentrification, which would later inform the seminal "rent-gap" hypothesis (Vernon, 1959). Hoover and Vernon further elaborated on this model in *Anatomy of a Metropolis*, where they identified five distinct stages of development contributing to waves of concentric change: 1) residential development in single-family houses, 2) transitions to population growth resulting in high-density construction, 3) downgrading, where aging housing stock is adapted for higher density use, 4) shrinkage and decline as a result of changes in family household size, 5) renewal, in which obsolete housing is replaced by new multifamily housing (Hoover & Vernon, 1959).

In "Towards a Stage Theory of Urban Growth," David Birch used Hoover and Vernon's stages to measure change at the census block group level. Birch measured percentages for census block groups of five indicators obtained from factor analysis: new housing, multi unit structures, single family housing, high rent units, population decline and renter occupied units. These factors were then used to compute "scores" for each stage, and census blocks were categorized by several stages according to how they performed for each stage score. Scores were normalized and "stage intensity curves" were interpreted as probability density functions (PDF) of stage scores for individual census block groups. Birch acknowledged the likelihood that a given census block group could experience multiple stages of development simultaneously, and sought to quantify this

possibility using PDF. Birch provided evidence against the concentric model, by interpreting means of PDFs as the "age" of a neighborhood on the stage scale, and mapping these ages across census tracts for the New Haven, Connecticut MSA. The results show topography of age with a structure more complex than Burgess's concentric model. Change is successfully given form in the Birch paper, however both lack of data availability as well as computational power are cited as serious limitations of the analysis (Birch, 1979).

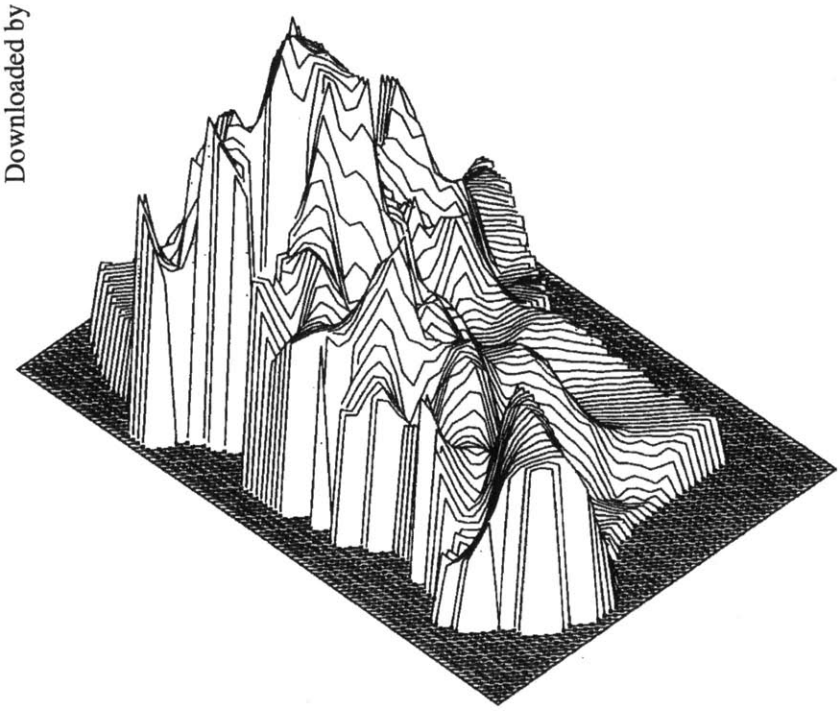

FIGURE 3 ***Three-Dimensional View of the New Haven SMSA***

Figure 1 Stage means plotted for census blocks in the New Haven, Connecticut MSA. Results reveal a topology of urban change that does not fit the traditional concentric model.

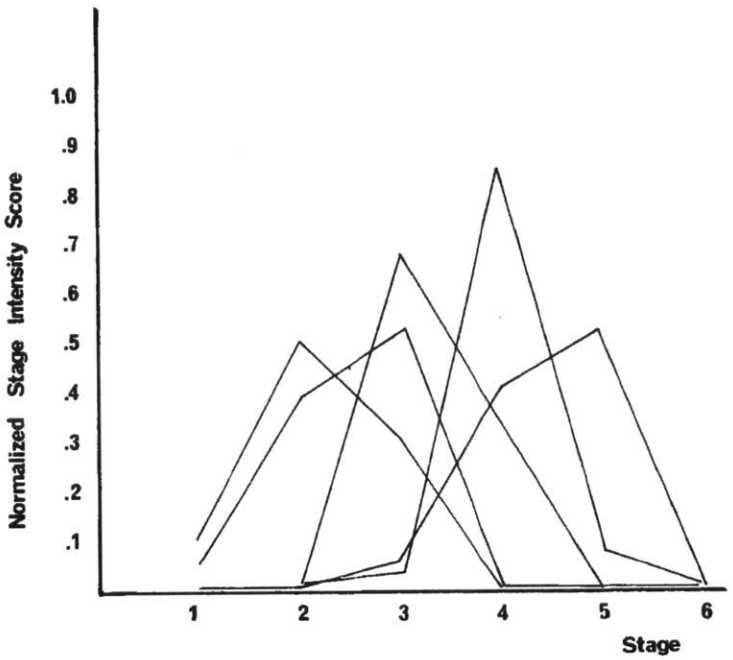

FIGURE 2 *Five Sample Neighborhood (Block Group) Stage Intensity Curves*

Figure 3 Probability density functions of stage scoring for five block groups from Birch (1979).

Life cycle and stage models of gentrification have recently been criticized for failing to explain or predict observed change, and for their inconsistency with observations of metropolitan cities (Marcuse, 1985). Marcuse calls for a more nuanced analysis of urban change that includes the interrelationships among social behaviors and environmental conditions. He identifies that household composition is itself linked to economic changes, and that changes in preferences in neighborhoods is not merely a spatial process, but is also driven by employment patterns. Understanding dimensions of relationships between variables that are drivers or consequences of urban change is the focus of the next generation of researchers.

2.5.2 Spatial Simulations

QP Researchers have also taken a simulation approach towards modeling gentrification and urban change, often borrowing from computational techniques in other fields. Jay Forrester's *Urban Dynamics,* uses logistic growth and diffusion processes for developing a simulation of urban dynamics (Forrester, 1969). Forrester emphasizes the important effect environmental change has on neighborhood change, and presents 'modes' of urban growth, survival and revival. Forrester concluded, "The city can change from the inside (Forrester, 128);" that processes of urban revitalization in an era of urban decline would emerge organically from the interior, and not depend on federally instituted urban

renewal. Forrester's work arguably set the groundwork for a complex systems theory of urban change.

Adopting a similar approach, Thomas Schelling, in 1971, showed that a preference for a neighbor's race could lead to starkly segregated environments. Schelling's model was based on cellular automata (CA), where spatial outcomes evolve through a series of discrete time steps according to a set of rules based on the states of neighboring cells. Likewise, O'Sullivan models gentrification based on a CA programmed using rules sourced from the rent gap hypothesis (O'Sullivan, 2002).

Cellular Automata (CA) is still widely used in QP models of urban change. Portugali (2006) notes how the shift from Newtonian conceptions of space in 1960s, to relativity-based perceptions of space in the 1980s contributed to the popularity of using CA models that do not assume a fixed identity of space:

> After the advent of the theory of relativity and quantum theory, this mechanics world-view came to be seen as an abstraction from a subtler reality in which space is only relatively independent from time and the bodies in it (653).

While CA adds an interesting layer of complexity to spatial simulations, there are three weaknesses in the CA approach to modeling gentrification. First, CA

produces a simulation based on rules programmed a-priori, and do not allow structures to emerge from data analysis, as machine learning techniques do. Second, transition rules play out in an abstract space, and it is difficult to apply simulation results to the actual physical boundaries of urban neighborhoods. Third, transition rules depend on the states of their physical neighbors, while the extent to which an adjacent block's gentrification-status actually influences your own are not well established in gentrification studies. While Michael Batty also makes use of CA analysis to model urban change, he stresses that theories of urban change must give equal weight to questions of socioeconomic dynamics as well as spatial form (Batty, 2005).

Multi-agent simulations (MAS) are sometimes employed to model the spatial outcomes of urban change. Torrens & Nara advocate a hybrid approach between MAS and CA (Torrens & Nara, 2006). This work focuses on households as "agents" making choices in dynamic property markets. The MAS approach to modeling urban change is problematic because it depends on pre-programmed relationships that generate a simulated outcome. In the context of gentrification, where causal factors are disputed if not unknown, simulating spatial outcomes of gentrification using pre-programmed dynamics is less relevant. Alternatively, an agent-based model (ABM)---where interactions of autonomous agents are assessed for their effects on an emergent outcome, may be more appropriate. In any case, models that do not assume system dynamics a-priori are likely to be to be more suitable for modeling what appears to be a black box of gentrification

dynamics.

Spatial models of gentrification increasingly pay more attention to the role of human perception. Recent work from Cesar Hidalgo at the Macro Connections Group in the MIT Media Lab, focus on measuring urban change and perception by combining machine learning and statistical techniques. In *"Do People Shape Cities or do Cities Shape People"* Hidalgo and Glaeser apply a machine learning algorithm to user-generated Google Street View data using safety perception rankings to obtain a "Street Score" of safety perception in urban neighborhoods (Hidalgo, 2014). Changes in scores of safety perception over time for individual blocks are used as indicators (UCC, or Urban Change Coefficient) of physical typology changes in the built environment. UCC values are used as a proxy for gentrification, and correlated with socioeconomic and demographic data from the American Community Survey (ACS) and U.S. Census data using multivariate regression (Hidalgo & Glaeser, 2015). Hidalgo and Glaeser find that population density and share of college-educated adults are strongly correlated with UCC values, suggesting these two factors may be significant drivers of gentrification.

2.6 QPH Models of Urban Change: The Place-based Perspective

Key Points

- *Place-based models of urban change focus on social factors underlying observed changes, and give broader consideration to social justice, human experience and oppressive political structures.*

- *Place-based models emerging from the QPH perspective are not often data-driven or quantitative.*

- *The Rent Gap model and Succession & Displacement studies are seminal models in QPH studies.*

QPH models of urban change focus on the social factors underlying urban change, and give broader consideration to social justice and equitable development. While approaches are sometimes quantitative, they are primarily used to uncover inefficiencies in the political structures that give rise to inequality, displacement and cultural deterioration.

2.6.1 Rent Gap Theory

Neil Smith in "Toward a Theory of Gentrification: A Back to the City Movement by Capital, not People," proposed the seminal "rent-gap" model arguing that capital flows are responsible for observed patterns of reinvestment in inner city neighborhoods. Smith observed what he called the "rent gap": the disparity between the potential ground rent and the actual ground rent capitalized under a parcel's present land use (Smith, 1979). According the Smith, gentrification is a structural product of land and housing markets, of which displacement and demographic changes are merely a symptom. For Smith, the "leading edge" of

gentrification occurs because capital flows to where the rate of return is highest, and inevitably a rent gap is generated as capital flows out of inner city districts. This depreciation of capital invested in inner-city districts produces economic conditions that in turn make a capital revaluation a rational market response (Smith, 1979). Smith says gentrification occurs when the gap between capitalized rent and potential ground rent is wide enough that developers can purchase shells cheaply and resell at their highest potential, i.e., "the highest and best use." Slater contends, perhaps more radically than Smith, that these shells are not the product of a natural phenomenon, but are deliberately generated through redlining, policies that displace residents, the withdrawal of public services and eminent domain (Slater, 2014).

Figure 2. The depreciation cycle of innercity neighborhoods.

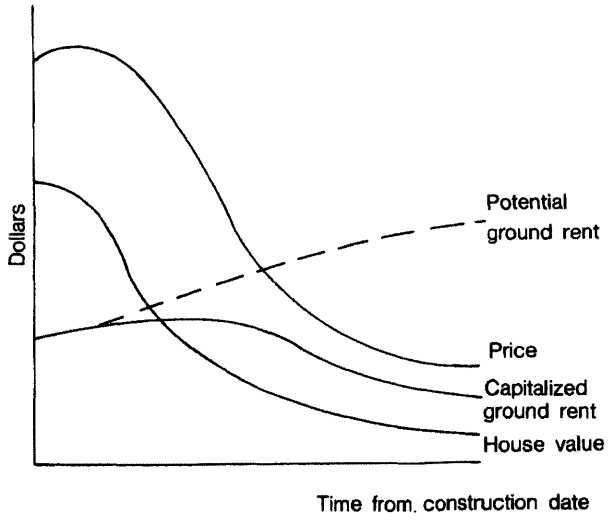

Figure 4 Graphic representation of Smith's Rent Gap hypothesis (Smith, 1979).

The Rent Gap model is an important concept in gentrification modeling, and has been used as the primary model for dynamics of urban change. Slater, a strong supporter of the model, has used it to explain gentrification as "a broader attempt to trace the circulation of interest-bearing capital in urban land markets, and to elaborate the role of the state in lubricating that circulation (Slater, 2014)." Slater describes the importance of power structures in "False Choice Urbanism," where gentrification is misconstrued by political observers as reinvestment and the public is forced to make a false choice between gentrification and disinvestment (Slater, 2014). Clark finds further evidence of Rent Gap in a case study of 125 years of rent fluctuations in Malmo, Sweden (Clark, 1988). However, Rent Gap has undergone criticism in the literature (Hamnett, 1984; Ley, 1986), as it fails to address the role of public and private institutional actors in neighborhood revitalization. A major criticism of the rent gap hypothesis is also that it favors a supply-side explanation of gentrification, over demand (Ley, 1986).

2.6.2 Social Statistical Modeling

Due to a lack of data availability from the 1930s to the early 1970s, modeling attempts of gentrification were largely theoretical. Reactions to these theory-based approaches emerged in the late 1970s and 80s, as data sets became more readily available, and statistical analyses became more solidified in the social sciences. Additionally, focus shifted from macro-scale analyses, to relationships between factors of finer-grained demographic and socioeconomic data. For modeling urban phenomena such as gentrification, a distinct procedure

of statistical analyses emerged.

David Ley makes the case for the operationalization of criteria used to describe gentrification, as well as the contextual differences of the gentrification process between nations. Ley established a common analytical procedure for gentrification analyses today: identify descriptive criteria in the literature, operationalize them, and then hold each defining factor as an independent variable in a sequence of multivariate analyses (Ley, 1986; Freeman & Braconi, 2004; Vigdor, 2002; Freeman, 2005). Because the factors shaping gentrification are diverse and complex, multivariate analyses are intuitively appropriate for isolating the effects of certain factors on gentrification. Ley creates a "gentrification index" using indicators of social status: the mean value of the percentage of the workforce employed in the quaternary sector, and the percentage of the population with a university education. The gentrification index is then treated as a dependent variable for simple correlation and regression studies. Multicollinearity among 35 predictor variables is treated using a Principal Components Analysis. Strong relationships are found between the gentrification index and variables representing urban amenity and economic dimensions; the highest being office space per capita (Ley, 1986). No significant correlation is found between housing market dimensions (starts, new household formation), despite strong housing demand from the (then young adult) baby-boomer generation.

There are two key issues with this approach, which become increasingly apparent as today's computational analyses are better able to handle complexity: 1) researchers are forced to find a proxy for gentrification to serve as a dependent variable in their analyses, and 2) lack of consensus about the definition of gentrification makes identification of such a proxy difficult. The results of this type of modeling approach is limited in that it can only draw conclusions about the dependent variable selected, which may or may not be a true proxy of gentrification. Consequentially, the results of several of these studies directly contradict each other, since the choice of proxy is subject to researcher bias. Indeed, researchers are able to find correlations to gentrification proxies, but these correlations may not truly reveal the nature of gentrification as an emergent phenomenon. Furthermore, it is notable that researcher choice of gentrification proxies varies with cultural and temporal perceptions, and that the results of multivariate analyses often reflect changes in urban dynamics characteristic of that time period. Despite changes in our statistical results regarding correlates to gentrification, the phenomenon persists, indicating that systemic relationships have yet to be identified.

2.6.3 Succession & Displacement Studies

A core debate emerging from QPH gentrification studies was whether gentrification was characterized by succession or displacement (Freeman, 2005). Theorists considered the importance of demographics, amenity access, and lifestyle changes (Ley, 1986; Beauregard, 1986; Hamnett, 1991), beyond the

market-based and physical dimensions explored primarily by stage-based analyses in the QP camp. Succession studies compared the characteristics of in-movers and out-movers (Henig 1980), while displacement studies surveyed respondents to identify displacement hotspots in a neighborhood or region (Grier, 1978). These studies relied primarily on survey responses and were ultimately flawed in that they were unable to correlate gentrification processes to displacement outcomes, or were unable to distinguish gentrification from other environmental factors contributing to displacement (Freeman, 2005). However, Slater recently advocated for reintroduction of displacement studies, citing how there is wide agreement that class should be the undercurrent in the study of gentrification (Slater, 2006). His recent work suggests that because researchers using quantitative analyses have not reached clear conclusions regarding gentrification processes, correlates or outcomes should instead focus on a value-based approach that examines the political structures responsible for observed uneven development.

2.7 Conclusion

The debate between Tom Slater and Philip Ball about whether or not gentrification is "natural" represents an important cultural divide between methods and perspectives of both the social and natural sciences when it comes to urban issues. This semantic debate comes at the expense of our communities' health and the preservation of unique, local cultures. The split between the

quantitative, positivist perspectives of communities (the "space" perspective) and qualitative, poststructuralist perspectives (the "place" perspective) resulted over the last several decades in a body of scholarly gentrification study that agrees on few aspects of the process. However gentrification persists globally, as the result of complex interactions between social, economic, physical and institutional forces. Meanwhile the development community has embraced and acted upon a simplified, space-based view of communities, resulting in the "colonization" of cultural capital and complex networks of financing and equity that are necessarily divorced from the communities they shape. One possible explanation of gentrification may be this separation between local cultural capital and global financial capital.

The CTC approach to cities, which to date has been vastly misunderstood by the urban theorist community may be useful for understanding the underlying constructs of gentrification, without taking a prescriptive view of what those underlying constructs may be. In this light a CTC approach to modeling gentrification may be able to narrow the gap between "space" and "place" perspectives by applying quantitative methods designed to capture the complexity produced by cultural systems, and observing their spatial manifestations. Pattern classification methods are uniquely suited for this purpose for reasons further discussed in the next chapter. Because the literature has produced conflicting narratives of gentrification, pattern classification methods are used to identify different typologies of gentrification, observe their

spatial patterning, and glean insight into the debate of whether gentrification is a "natural force" or the result of institutionalized cycles of disinvestment and reinvestment in urban communities.

3 Pattern Classification

This section provides a brief survey of the elements of pattern classification. Pattern classification refers to a suite of methods focused on identifying systematic and random patterns in static and dynamic events. In this chapter, I assess pattern classification methods in relationship to identifying patterns of urban change, isolate the elements of pattern classification methods highlighted in Duda et al. (2001) and Gershenfeld (1999), followed by a motivation of selected methods for subsequent analysis of gentrification.

3.1 Pattern Classification and Urban Change

Key Elements

- *Pattern classification identifies relationships between data factors that give rise to emergent phenomena.*
- *Pattern classification methods are data-driven, and can analyze relationships in multi-dimensional data sets, which is appropriate for studies that point to multiple indicators of urban change.*
- *Pattern classification methods may incorporate existing knowledge or intuition about how an observed pattern might be generated.*
- *Some pattern classification methods assume path dependence in time series, which may be appropriate for sequential studies of urban change.*

Pattern classification methods are useful for identifying patterns of urban change for several reasons. Pattern recognition is the act of taking raw data and producing an action based on the "category" of the pattern at hand (Duda, 2001). The ultimate objective of pattern classification is to recover the generating

parameters, or the underlying model that gives rise to an observed or sensed pattern. Many real world pattern recognition systems attempt to incorporate existing knowledge about a hypothesized model underpinning an observed pattern in order to ensure accurate representation. This aspect of the method is especially compatible with the study of urban change that has already been explored in depth through qualitative urban theory. In this way, pattern classification provides an opportunity to bridge the gap between "space" and "place" based perspectives in urban studies, by incorporating qualitative modes of inquiry into a computational process.

In the case of gentrification, complex interactions between social, environmental and economic factors produce emergent effects of urban change indicated by changes to specific data features such as property values, demographic and typological compositions, and amenity availability. Redundant, and specific relationships between these indicators result in different patterns or typologies of urban change. Pattern classification methods are often applied to dynamic, complex systems in order to better describe their components, and predict future outcomes. In many cases pattern recognition systems are trained from labeled "training" data (supervised learning), and can be used to discover previously unknown patterns when no labeled data is available (unsupervised learning). When combined with qualitative studies, pattern classification offers a suite of data-driven methods that can identify patterns of urban change as they relate to

gentrification.

As applied to studies of urban change, pattern classification methods can accommodate the vast categories of data hypothesized to be related to gentrification processes in the literature, and provide empirical insight into typologies of urban change emerging at the metropolitan scale. Additionally, the time-series component of pattern classification is particularly useful for exploring how features interact in dynamic systems to generate observable change over time. Because gentrification and urban change are considered time series processes, it is important to employ classification methods that handle sequential change. Some pattern classification methods such as Markov Models are designed to capture how indicators vary together or separately over time. These types of pattern classification methods assume path dependence of state changes. Path dependence refers to the fact that the condition of an observation (for example, a Census block group) at time t. is in part dependent on previously observed conditions. This assumption is useful for studying urban systems, where processes of change seldom occur in isolation.

3.2 Fundamentals of Pattern Classification

Pattern recognition is synonymous with Machine Learning, which evolved from computer sciences, rather than engineering (Bishop, 2006), and is a subfield of Decision Theory. The key objective of pattern classification methodology is to

recover generating parameters of an observed or even sensed pattern. Pattern classification typically results in a model that attempts to describe how observed inputs result in observed outputs. Pattern classification is a popular tool in Neuroscience to classify FMRI data derived from brain stimuli in order to process neural signaling patterns (known as brain decoding). It has also been used as a method to recognize patterns of consumer choice behavior, such as Netflix's genre recommendation feature. Ultimately, pattern recognition is about observing a signal, finding patterns within that signal, and learning from those patterns to predict future outcomes of the system. There are several elements of a basic pattern recognition procedure; they are outlined in the following section.

Feature Extractor: Feature extractors measure object properties that are useful for classification. Feature extractors identify which features are useful for distinguishing patterns in a data set. When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features or a "feature vector." The extracted features are a reduced representation of the complete dataset, and are expected to contain the relevant information from the input data, so that the desired classification task can be performed.

Classifier: Classifiers use features identified by a feature extractor as an input to assign an observation to a category. Perfect classification performance is

impossible, and instead probabilities are defined for classification decisions. The degree of difficulty of a classification problem depends on the variability in feature values for objects in the same category, relative to the difference between feature values for objects in different categories. This variability, typically due to the complexity of a system of study, is referred to as noise. The problem of classifier design is creating a classifier that handles noise and minimizes error as a result. The simplest measure of classifier performance is the classification error rate (the percentage of new patterns assigned to the wrong category).

Post-Processor: Post processors take into account other considerations during classification, such as the effect of context or the cost of errors in order to decide an appropriate classification action. A post-processor may also be able to exploit context, i.e., the input-dependent information other than from the target pattern itself.

State of Nature: The classification or the pattern that an observation falls into, labeled heretofore as w_x.

A Priori Probability or "prior": The probability of the next state of nature. If two states are equally likely, then this probability is 50%.

Decision Rule: The rule by which classification occurs; typically based on finding

the largest probability of the next state of nature, and minimizing the cost of misclassification.

Class-Conditional Probability Density Function: The probability distribution of a feature (of which there can be many, n features), given its state of nature.

Feature Vector: If an object is characterized by more than one feature, (for example a census tract can be characterized by rent, median household income, density, household race or age); then the values of each feature for that object is called a feature vector.

3.3 Fundamentals of Bayesian Decision Theory

Bayesian Decision Theory is the fundamental statistical approach to the problem of pattern classification. Bayesian decision theory asks what the probability is that an observation x, falls in the category of a given "state of nature," and uses a *class-conditional probability density function* to classify unknown observations into respective states. X is a continuous random variable, whose distribution depends on the state of nature, and the class-conditional probability density function shows the likelihood of measuring a particular feature value of x, given the state of nature or pattern category it is in.

Bayes formula states that by observing a continuous variable x, we can convert

the prior probability into a posterior probability (the probability the state of nature being w, given that the feature value x has been measured). The Bayes Decision Rule, minimizes the probability of error in classification by stating:

$$Decide\ w_1\ if\ P(w_1|x) > P(w_2|x);\ otherwise\ decide\ w_2$$

Meaning, "If the probability of this observation being in State 1 is greater than it being in State 2, then classify it as State 1."

The decision rule is a function that tells us which action to take for every possible observation. However, there are costs to making each decision, and the Loss Function computes how costly each action (of rejection or classification is).

3.3.1 Feature Extractors: Component Analysis and Discriminants

Component analyses try to handle the problem of excessive dimensionality (when there are many features being analyzed simultaneously in a data set), by either reducing dimensionality into components, or reducing dimensionality into components that are linearly uncorrelated with each other. Components are vectors that represent correlations between features in a data set, and describe the majority of these correlations. Analyses that reduce the data set into components that are statistically independent of each other are referred to as Discriminant Analyses. There are two types of Component Analysis and two

types of Discriminant Analyses. Component analyses include the following:

Principal Components Analysis (PCA)

PCA reduces dimensionality while preserving as much variance as possible. While PCA finds components that are useful in representing the data, there is no reason to assume that these components are useful for discriminating between data in different classes. This is because the directions that are discarded by the PCA could be useful for distinguishing between classes.

Independent Components Analysis (IDA)

IDA determines the directions in feature space that are statistically most independent. The following are major types of Discriminant Analyses:

1) Linear (Fischer) Discriminant Analysis (LDA): LDA reduces dimensionality while preserving as much of the class discriminatory information as possible. It maximizes the distance between class means. The LDA is the linear function yielding the maximum ratio of between-class scatter to within-class scatter. It can only be used for two states.

2) Multiple Discriminant Analysis (MDA): MDA seeks a projection that best separates the data according to least-squares. MDA is often used in preparation for classification, meaning that the components emerging from MDA are

statistically significantly different from each other and can be used to by a classifier. MDA differs from LDA because the feature becomes a categorical variable with *n* possible states, instead of just two.

3.3.2 Classifiers

Classifiers use features identified by a feature extractor as an input to assign an observation to a category. Classifier design attempts to handle noise and minimize error as a result. The simplest measure of classifier performance is the classification error rate (the percentage of new patterns assigned to the wrong category). There are three main types of classifiers, Markov Models (MM), Support Vector Machines (SVM) and Neural Networks (NM). Markov Models are suited for time-series classifications, exhibiting minimal path dependence where the next classification depends on the previous one. SVM uses training data to assign observations into one of two categories. NM models are a form of network-based pattern recognition, where the parameters governing a non-linear mapping are learned at the same time as those governing the linear discriminant.

Markov Models (MM)

Markov Models make a sequence of classification decisions, where the state at time *t* is directly influenced by the previous state. There are many types of machine learning problems that Markov Models can solve. Three of these problems are evaluation, decoding and learning problems. Evaluation problems

occur when a researcher has an existing Hidden Markov Model (HMM) complete with transition probabilities *a priori*, and needs to determine the probability that the model generated a particular sequence of observed symbols. The decoding problem occurs when the researcher is in possession of HMM as well as a set of observations, and needs to determine the most likely sequence of hidden states that led to those observations. Learning problems arise when the researcher only has the number of states, and the number of visible symbols, but not the transition probabilities. Given a set of training observations of visible symbols the model must learn these parameters.

There are two main types of Markov Models:

1) *First Order Discrete Markov Models*: First Order Discrete MMs state that the probability of observing state (w) at $(t)+1$ depends on the current state at t. These are the most common type of Markov Models.

2) *First Order Hidden Markov Models (HMM)*: Here the researcher assumes that for every time step t, the system is in a state $w(t)$, but this state is hidden and instead a continuous visible symbol $v(t)$ is emitted. HMM attempts to uncover hidden states $w(t)$ using the emitted symbol $v(t)$. Figure 5 displays a Markov model.

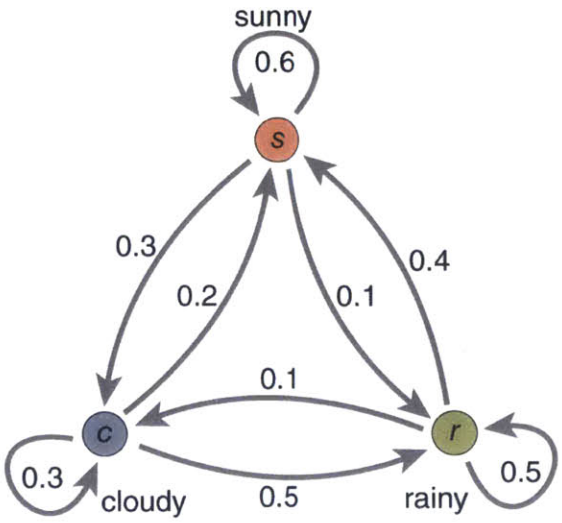

Figure 5 Markov Model with transition probabilities for three states, "sunny," (s) "cloudy," (c) and "rainy" (r). Training data learns the probability that an observation will transition from being characterized as one state to another. Transition probabilities are obtained by tracking transitions throughout the entire data set.

Support Vector Machines (SVM)

SVM are a probabilistic binary classifier. Given a set of training examples, each bucketed into one of two categories; SVM ultimately builds a model or classifying engine that assigns new examples into binary categories. SVM is inappropriate for a model of gentrification because gentrification patterns do not fall into two explicit categories.

Multi-layer Neural Networks or Multi-Layer Perceptrons

Neural Networks and Perceptrons are powerful forms of network-based pattern recognition that map a set of input data onto a set of outputs in order to distinguish patterns in data that are non-linearly separable. An MNN consists of multiple layers of processing elements (linear classifiers) or neural nodes in a

directed graph. Here, the output of one linear classifier becomes the input of another in a designed network arrangement. This network of linear classifiers is referred to as the "hidden layer" which occurs in between the input features and the final, output classifier. In feed-forward networks, information flows from the input features, through the hidden layers to the output classifier. There can be more than one possible output classifier, which is useful in classification for non-binary targets. Figure 6 illustrates a common MNN network.

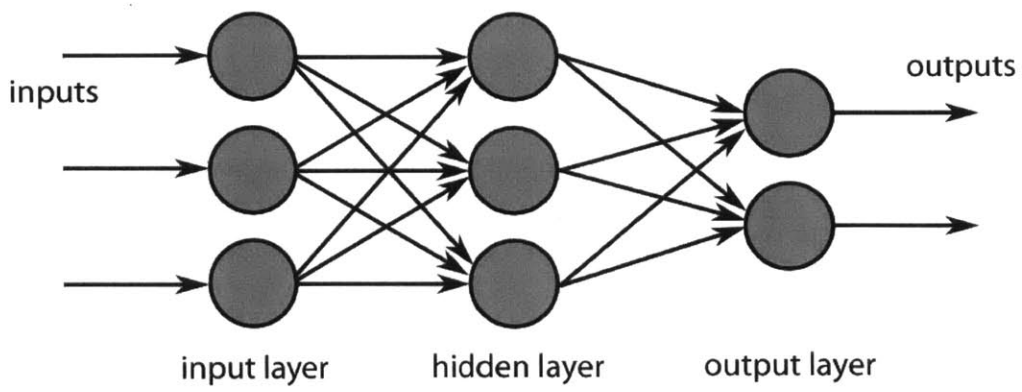

Figure 6 A Multi-Layer Neural Network with three inputs, three linear classifiers in the hidden layer, and two output classifiers. Information travels from left to right, feedforward, in this model. Inputs are observations from data, which are fed through classifiers (hidden layer) that result in a classification output (output layer). The hidden layer is referred to as hidden, often because the model is being generated at this stage; that is, the researcher has inputs and outputs, but seeks to understand what types of classifiers could have resulted in the observed classification.

In MNN, a supervised learning technique called back-propagation is used to train the network. The training problem is to select weights for all layers that minimize mean squared error. Because linear classifiers are networked in the MNN, it can perform classifications on more complex data. A common problem for neural networks involves selecting or adjusting the complexity of the network, called

"regularization."

3.3.3 Unsupervised Learning: Clustering to Label Unlabeled Data

Unsupervised processes use unlabeled samples. Unlabeled samples are not described, tagged, or categorized. Typically, a machine-learning model is trained on labeled samples, and then validated by classifying unknown samples. Unsupervised learning refers to the process of classifying unlabeled samples. There are five reasons to use these: 1) when the classifier can be designed on a small set of labeled samples and then 'tuned up' without supervision on a large data set, 2) to train with large, unlabeled data and then use a supervised method to label the found groups, 3) when characteristics of patterns change over time, changes can be tracked by an unsupervised classifier 4) when features need to be found in order to classify and 5) the discovery of distinct subclasses in the data suggest we alter our classifier. Clustering methods are one form of unsupervised learning. The following are three popular types of clustering methods:

K-means Clustering

A method to obtain maximum-likelihood estimates for the means. Random cluster centers are chosen (Voronoi cells), and used to calculate new cluster centers until the algorithm has converged. Each data point is assumed to be in exactly one cluster. K-means clustering is a partitioning method that treats observations

in your data as objects having locations and distances from each other (Mathworks Documentation, 2014). It partitions the objects into K mutually exclusive clusters, so that objects within each cluster are as close to each other as possible (determined by either Euclidean, city-block or mahalanobis distances), and as far from objects in other clusters as possible. Each cluster is characterized by its centroid, or center point.

Fuzzy K-means Clustering

This method does not assume that a data point must be in exactly one cluster, and that there are "fuzzy" memberships of points across clusters. Incorporating probabilities, sometimes improves the convergence of k-means clustering, but the correct choice of k is much more important for the algorithm to be successful.

Hierarchical Clustering

Acknowledges that there may be sub-clusters within each cluster. There are two types of hierarchical clustering methods that uncover sub-clusters; the first is Agglomerative or "bottom-up clumping", and the second is Divisive or "top-down splitting." While this technique may be appropriate for gentrification studies, ACS data may not have enough resolution to support this complex analysis.

3.3.4 How to Evaluate Clustering

This is one of the most important questions when it comes to k-means clustering.

Choice of the number of clusters is key to partitioning data in such a way that captures the majority of the dataset. To verify that the clustering method chosen is an accurate representation of the data, one first measures the degree of "similarity" between samples that is, the distance between samples *within* the same cluster should be significantly less than the distance of samples *between* clusters. Second, if clusters are meaningful, they should be invariant to scale transformations that are 'natural to the problem (Duda pg. 539)'. Data should always be normalized before clustering so that distances between points generate an appropriate cluster output.

3.3.5 Choice of Clusters

Choice of the number of clusters is typically verified using a criterion function. One popular type of criterion function is the sum-of-squared-error criterion where for a given cluster the mean vector is the best representative of the samples in that cluster because it minimizes the sum of the squared lengths of the 'error' vectors in the cluster. Criterion functions are used in two different ways in order to address cluster validity (the appropriate number of clusters):

1) By repeating the clustering procedure for $c = 1, c = 2, c = 3$ etc., to see how the criterion function changes with the number of clusters c. For example, the 'sum-of-squared-error' criterion J, must decrease with c, because squared error can be reduced each time c is increased. J decreases rapidly until it converges to zero at

$c = n$.

2) By hypothesis testing. In this case, the null hypothesis is that there is some number of c clusters. Then compute the sampling distribution for a criterion function J, for $J(c+1)$. This distribution tells us what kind of improvement to expect (Duda, 557). Then determine if the reduction of the cluster criterion due to the addition of a new cluster is significant. The null hypothesis is accepted if the observed value of $J(c+1)$ falls within limits corresponding to an acceptable probability of false rejection. That is, the likelihood that there are clusters in the data is determined by the probability that the criterion function improves with additional clusters.

3.4 Selected Methods

Cities are intricate artifices housing the products of human culture, economy, innovation and expression. Likewise gentrification as an urban phenomenon is best described in multidimensional terms, as the data sets describing it contain social, demographic, economic, physical and environmental layers that change and interact over time. Gentrification is observed as a collection of symptoms that are time and space varying, each having context-specific impacts. Unfortunately, gentrification typically described through the lens of a single "symptom," such as rental markets or displacement. Alternatively, a review of the literature points to several significant factors relating to gentrification ranging from income to age of building stock. This poses a challenging computational problem, as these factors

are hypothesized to vary in context-specific, complex ways to yield gentrification as an emergent effect. As Jacob's suggests, measuring these features at the surface level, does little to extract principles of urban change. Instead, analyzing discrete data sources to obtain latent *principles* of change can help planners avoid topical interventions. Capturing this kind of complexity in urban data is therefore the primary motivation of the descriptive portion of the analysis presented here. We sought to observe both spatial and temporal variation in our data spanning the five boroughs, resulting in the discretization of key components of variation.

Pattern classification learns how hidden data structures change over time. To identify developmental states of census block groups in ACS data, we employed the k-means clustering algorithm (MacQueen, 1967) as our feature extractor, where the resulting features (*k* clusters) were treated as input states into a Markov Chain Model. As discussed above, k-means is an unsupervised machine learning technique and identifies unknown patterns in data that have not been assigned a category or label. This is useful for describing multi-dimensional data sets, and reducing their dimensionality into distinct typologies. The Markov Chain Model is applicable for two key reasons. First, it handles time series data and makes a sequence of classification decisions, where the state at time *t* is directly influenced by the previous state. This is a key assumption of our model: that discrete urban conditions are observable and depend on immediately previous conditions. Second, the Markov Chain Model is useful for predicting future states,

and given current conditions can be run to convergence several years into the future. This is an interesting proposition for studying patterns of urban change, where future conditions are of primary concern to gentrification stakeholders. Ultimately, the k-means clustering and Markov Chain methods combined create a pattern classification process that enables the description of multi-dimensional data, extraction of "features" or states, and incorporation of states into a time-series model of urban change---and possibly gentrification.

4 Descriptive Statistics for the New York MSA

4.1 American Community Survey Data: Advantages and Limitations

The literature review on gentrification motivates several features for gentrification analysis: income, (Brookings Institution, 2011; Smith 1979), education level (Clay, 1979; Freeman, 2005; Ley, 1986), density and percentage family households (Hoover & Vernon, 1959; Birch, 1979, Marcuse 1985, Kolko, 2010), structure age (Hidalgo, 2015; Birch, 1979), land use (Smith, 1979; Burgess, 1923), employment status (Ley, 1986) and race (Beauregard 1986; Hamnett, 1991; Slater, 2006; Schelling, 1971). Using American Community Survey Data, six features representing these categories were selected at the Census Tract level: percentage white, number of households, percentage of family households (an approximation of density), education level, income level, and structure age. Land use, and employment status were not available in the data set used, and these were therefore removed from the analysis.

We obtained five-year estimates from the American Community Survey (ACS) between 2005 and 2013 via *socialexplorer.com*, a common data resource for Census and ACS data in the United States. Four counties were selected for our analysis at the block group level: Bronx, Kings, New York, and Queens. These counties were chosen due to their spatial proximity, data set size, and consistency in data sampling across the region. The final data set consisted of 29,058 observations (i.e. census block group five-year estimates) per region (see Table 1), characterized by 32 fields.

The primary advantages of ACS data are its accessibility and the availability of data at the census block group-level. ACS samples nearly 3 million addresses each year (US Census Bureau, 2008), through survey and interviews. Data at the census block group level or smaller is appropriate for modeling processes like gentrification, are visible and have effects at the neighborhood scale. Additionally, the variety and amount of data available through ACS is appropriate for the clustering technique proposed here, and relevant to the study of urban change and gentrification. Furthermore, the overlapping sampling method improves the statistical reliability data (US Census Bureau, 2008), though it is a significant limitation for our model as discussed below.

There are notable limitations of the ACS data. First, ACS only provides five-year, multi-year estimates at the census block group level. Five-year estimates are updated annually by removing the earliest year of the estimate and replacing it with the latest one (US Census Bureau, 2008). For example, following collection of data for 2013, data estimates from 2007 will be dropped to create a 2008-2013 estimate. Therefore, multi-year estimates represent a period, rather than a specific year, and estimates overlap across periods. In our analysis, census blocks are characterized according to feature vectors by year. Due to the overlap of the ACS sampling method, a single year actually represents an interval estimate (data collected over a 60 month period). This overlap is a significant limitation for our model, as the overlapping estimates might significantly

underestimate short-term variations occurring in vivo. Furthermore, overlapping estimates greatly reduce the variation in data making it less suitable for the assessment of urban change (US Census Bureau, 2008). Following the guidelines for determining the suitability of comparing overlapping estimates for features in ACS data provided by Appendix 4, "Making Comparisons" in *A Compass for Understanding and Using American Community Survey Data* (US Census Bureau, 2008), to determine whether observed changes in features of overlapping estimates were due to chance or are statistically significant. Table 1 reports the results of significance tests across estimates for the six features, for Bronx County.

Table 1 Significance test for observed differences in overlapping estimates from 2009-2013, for the six features studied. Table values were obtained by calculating the square root of the sum of squared Standard Errors between observations, and dividing by the difference between observations. An approximation for the Standard Error was used that included the fraction of overlapping years per the ACS User Guide Appendix 4.

Years Compared	White%	Households	Family%	Income	Education	YrStruct
2009-10	0.240	1.319	0.330	0.317	0.277	0.240
2010-11	0.047	0.018	0.024	0.054	0.002	0.047
2011-12	0.047	0.018	0.024	0.054	0.002	0.047
2012-13	0.031	0.048	0.032	0.009	0.032	0.031

The following approximation for Standard Error was used:

$$SE(\hat{X}_1 - \hat{X}_2) \cong \sqrt{(1-C)}\sqrt{SE_1^{\,2} + SE_2^{\,2}}$$

Where C is the fraction of overlapping years. For example, the periods between

2009-2010 represent an overlap of 4/5 years = 0.8. Comparing the results to critical values shows that the observed differences between observations in subsequent years is not highly significant, and can be attributed to chance.

For our model, single-year estimates at the census block group level would clearly be more ideal, since we aim to capture time-series variation in data. Additionally, the ACS data collection process is relatively new, having started only in 2006 and the five-year estimates beginning as late as 2009. Consequently, there are inconsistencies in earlier five-year estimates across regions; some fields are missing or unavailable and other fields were collected under varying conditions. Specifically, our data includes estimates made after the 2008 financial crisis and may not be representative of typical trends or patterns occurring under normal economic conditions. The sample size of 5-year estimates is smaller than the long-form sample in decennial census, resulting in larger standard errors (US Census Bureau, 2008). Finally, the period of the data is relatively short, encompassing only five years between 2009 and 2013, obscuring long-term patterns and trends.

4.2 Data Processing & Preparation

Before submitting the data to the clustering algorithm, we took several steps to increase its suitability for clustering. To reduce the dimensionality of the data, improve clustering speed and intelligibility we turned 108 fields from the ACS data into five features that were more closely associated with gentrification

symptoms observed in the literature. Census block groups without population were discarded, assuming that uninhabited areas such as industrial compounds and natural reserves display patterns of development that are different from inhabited and urbanized areas.

The pre-preprocessing steps of the fields involved converting some of the fields into percentages, consolidating several fields into a single feature, and calculating a weighted average from several other fields. We took four features directly from the ACS data (total population, number of households, number of housing units, and the median year of structures built).

To convert fields into percentages, we divided the value of more specific fields, such as the number of vacant housing units, by an appropriate more general value (in this case the number of all housing units). Such percentages are more suitable for clustering since they allow a better comparison of relative values. We included absolute values, for example the total number of housing units, as separate features. Other fields denoting specific categories or brackets were summed together, and the result converted into a percentage. For example, we simplified the sixteen categories of household income into five features (based on the definition of middle class by Thompson and Hickey, 2004).

To improve clustering results, we further reduced these categorical features into indices. This was necessary for both the income and educational features of ACS

data. A weighted income index was computed according to the following formula:

$$Income = (c_i/t)x_i + (c_{i+1}/t)x_{i+1}$$

Where c_i is the count of households of income bracket *i*, x_i is the maximum income value of income bracket *i*, and t is the total number of households within the census block group for which the index is computed. Likewise, education was also converted into an index:

$$Education = (c_i/t)w_i + (c_{i+1}/t)x_{i+1}$$

Where c_i is the count of households of education bracket *i*, w_i is a weighted value for the education bracket *i*, ranging from 1 (less than high school) to 6 (PhD) and t is the total number of households within the census block group for which the index is computed.

Weighted average of housing units per building as an approximation of density, were consolidated by converting brackets or categories into a weighted average. For this calculation we assumed a hypothetical average value for each bracket as the mean value of the bounds of the bracket, and calculated an overall, weighted average based on the brackets' sizes. For example, we assumed that the average age of the men in the 18 to 24 year age bracket is 21.5 (since the next bracket starts at 25) and included this value in the overall average, weighted according to the number of men in this age bracket. Note that this technique

requires the assumption of an upper bound for the highest open-ended bracket, which includes values such as the number of men of "85 years and above", or the number of units with a rent of "2000 USD or more".

Finally, we normalized the values for every feature to be between zero and one to ensure an equal weightage in terms of the clustering algorithm. In other words, we created a broad selection of potentially relevant features, and refrained from a-priori assessing the relative importance of these features. The various pre-processing steps described above yielded 29,058 observations across all boroughs varying along 5 features for inclusion in the clustering.

4.3 Selected Method: K-means Clustering

Our study area included four New York City boroughs for which we obtained American Community Survey data across several features for the years 2005-2013. The analyses occurs in three steps 1) The procurement and preparation of socio-economic data at a census block group level and over four time-periods (for the four counties of Bronx, Kings, New York, and Queens), 2) the clustering of these block groups into states (separately for each county), and 3) the derivation of four Markov models by tracking transitions between states over time for each county. American Community Survey data for the New York CBSA between 2009 and 2013 were obtained and tested along five dimensions in agreement with the literature review on gentrification: race, household density, family composition, income, education and structural age.

The k-means cluster was computed as follows: For a set of observations $(x_1, \ldots, x_n)$ where each observation is a *d*-dimensional real vector, k-means clustering partitions *n* observations into sets $S = \{S_1, S_2, \ldots, S_k\}$ where $k (\leq n)$, so as to minimize the within-cluster sum of squared distance functions (our criterion function) of each point in the cluster to the k center:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

where μ_i is the mean of points in S_i.

K-means requires the choice of the number of k clusters a-priori. Because the algorithm will produce results regardless of the number of k selected, it is important to identify which number of clusters is most appropriate for the data. There are several proposed methods for this (Sugar and James, 2011), but choice of k is typically determined on a case-by-case basis. The choice of the number of clusters is further constrained by the fact that increasing numbers of k reduces the accuracy of classification. It is therefore advised to choose lower values for k, however this may be less descriptive of the data. The k-means algorithm is therefore subject to a perpetual tradeoff between robust classification and accurate data description.

A common method for determining the accuracy of the k-means clustering method is to obtain silhouette values $s(i)$, for each cluster. Silhouette values are a

ratio of within-cluster similarity and between cluster similarity between observations (Rousseeuw, 1987), as measured by Euclidean distance between observations. For each observation i *in* the dataset, a silhouette value $s(i)$ is assigned:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

Where given clusters A, B and C (See Figure 8) $a(i)$ is the average dissimilarity of each observation i to all other observations of cluster A, $d(i,C)$, that is the average dissimilarity of i to all observations of cluster C, and $b(i)$ is the minimum of $d(i,C)$ which we call the neighbor of observation i.

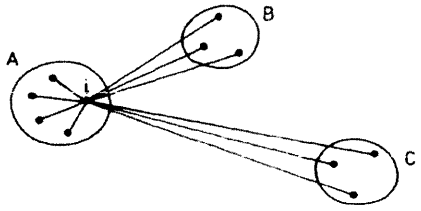

Fig. 1. An illustration of the elements involved in the computation of $s(i)$, where the object i belongs to cluster A.

Figure 7 An illustration of clusters and the computation of s(i) where object I belongs to cluster A.

Silhouette values are evaluated on a scale of -1 to 1 where:

$$-1 \leqslant s(i) \leqslant 1$$

for each object i.

4.4 Inside the Clusters

Table 2 reports silhouette plots and values for selections of the best classification of five features selected from ACS data *(k=2)*, and next best classification obtained through silhouette value comparison, for each borough. Selections of k with the second best silhouette value were chosen as inputs into the Markov model. In all cases $k=2$ was the optimal classification, which may be robust as the data is only partitioned into two categories, but may not capture latent correlations in the data. The second best silhouette value occurred for selections of k ranging between 3 and 5.

Table 2 Average Silhouette Values s(i) for all observations (i) in cluster selections k=2 through 10 by borough. Next best s(i) values are highlighted. Silhouette values drop for higher numbers of k.

Borough	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=10**
Bronx	0.53	0.39	0.39	0.34	0.31	0.30
Kings	0.40	0.39	0.31	0.34	0.32	0.29
New York	0.59	0.49	0.50	0.45	0.36	0.34
Queens	0.34	0.30	0.32	0.30	0.31	0.29

The selection of k was confirmed through visualization of clusters using silhouette plots visualize clustered data. The following silhouette plots illustrate clusterings for each choice of k by borough.

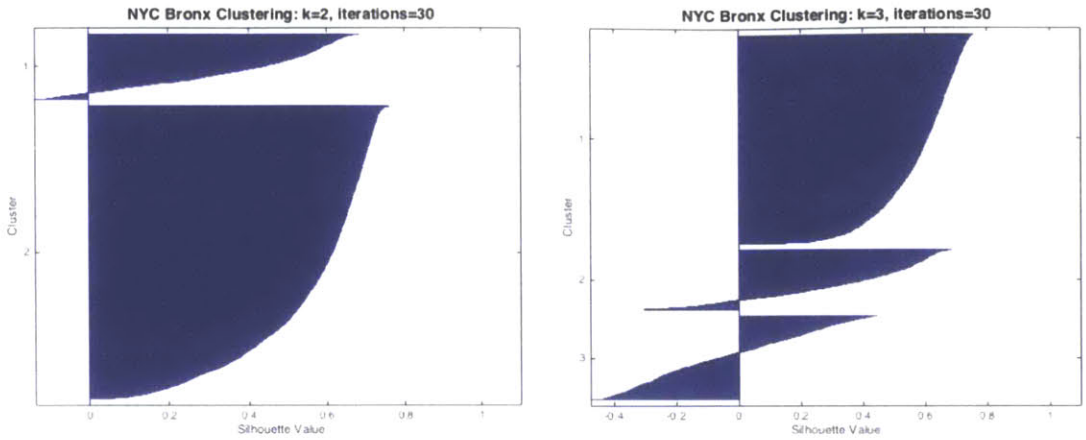

Figure 7 Silhouette Plots for Bronx County for selections of k=2 (left) and k=3 (right). Left: Average s(i) = 0.53, classification appears to fit most observations into a cluster 2. Right: Average s(i) = 0.39, a portion of observations in Cluster 2 and 3 appear to be misclassified, though the average silhouette value is comparable to K=2, and better than values of k=5 through 10.

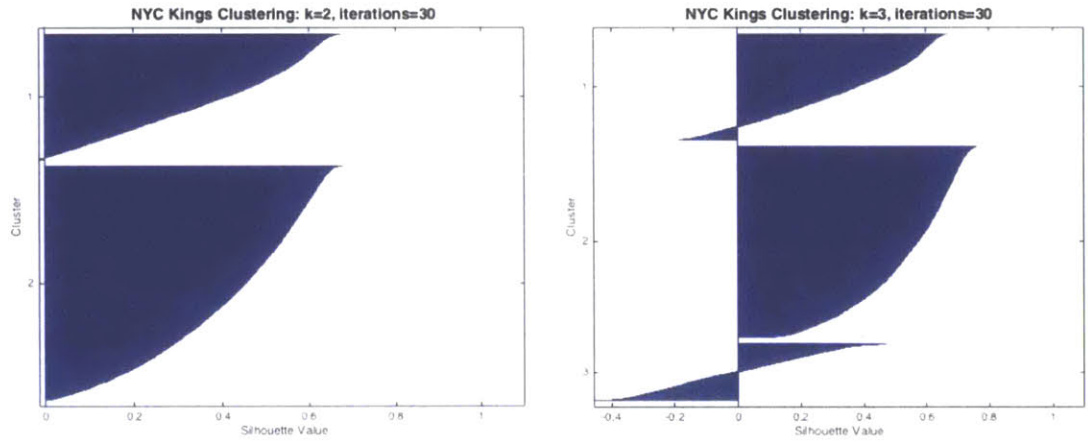

Figure 8 Silhouette Plots for Kings County for selections of k=2 (left) and k=5 (right). Left: Average s(i) = 0.40, classification appears to distribute over half of the total observations into a single cluster. Right: Average s(i) = 0.39, a small portion of observations in Cluster 1, and half of Cluster 3 appear to be misclassified, though the average silhouette value is comparable to K=2, and better than values of k=4 through 10.

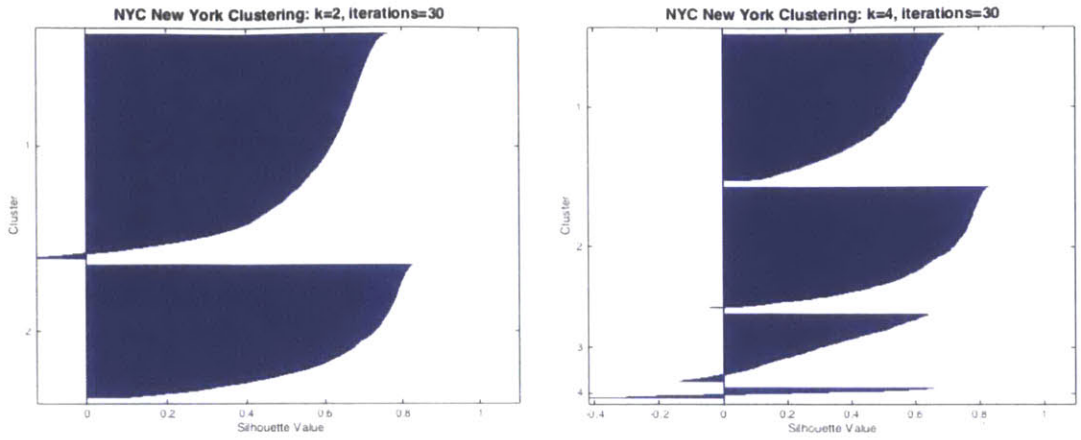

Figure 9 Silhouette Plots for New York County for selections of k=2 (left) and k=3 (right). Left: Average s(i) = 0.59, classification appears to distribute observations unevenly into two clusters. Right: Average s(i) = 0.50, nearly half the observations in Cluster 4 appear to be misclassified, though the average silhouette value is comparable to K=2, and better than values of k=3 and 5 through 10.

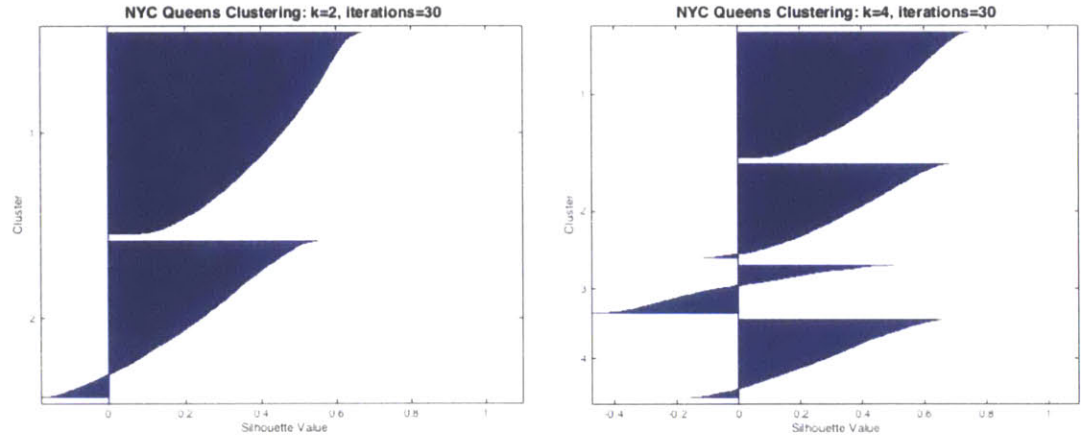

Figure 10 Silhouette Plots for Queens County for selections of k=2 (left) and k=4 (right). Left: Average s(i) = 0.34, classification appears to fit most observations into a single category, and misclassifies a small proportion of cluster 2. Right: Average s(i) = 0.32, over half the observations in Cluster 3 appear to be misclassified, however very few observations are misclassified in other clusters. The average silhouette value is comparable to K=2, and better than values of k=3, and 5-10.

Descriptive statistics were calculated for each cluster in a set of *k*, in order to

provide a better picture of the contents of each clustering. Maximum, minimum, mean, standard deviations, kurtosis and skew were obtained for each feature classified within a particular cluster. As anticipated given the nature of the clustering algorithm, kurtosis and skew for these features do not reflect normal distributions, however means and ranges across the five dimensions provides important context for understanding the differences between "states" obtained by the algorithm.

Table 3 Descriptive statistics for Bronx County. Statistics were calculated for a total of two clusters each of which is represented as k=2(1) and k=2(2), respectively.

k=2(1)	**White%**	**Households**	**Family%**	**Income**	**Education**	**YrStruct**
Max	1.00	2430.00	1.00	200000.00	5.41	2000
Min	0.00	5.00	0.00	19896.99	1.74	1939
Mean	0.61	423.61	0.61	84554.42	3.02	1952
StDev	0.24	224.04	0.17	20649.49	0.57	10.17
Kurtosis	2.57	16.00	3.11	4.08	4.16	4.66
Skew	-0.53	2.36	-0.23	0.35	1.08	0.81
k=2(2)						
Max	0.82	11266.00	1.00	123647.65	3.74	2005
Min	0.00	4.00	0.00	10000.00	0.00	1939
Mean	0.15	445.82	0.68	45812.59	2.20	1950
StDev	0.12	323.68	0.14	15676.56	0.36	14.35
Kurtosis	5.27	343.28	3.81	3.55	3.45	4.53
Skew	1.26	12.67	-0.43	0.74	0.13	1.38

Table 4 Descriptive statistics for Bronx County. Statistics were calculated for a total of three clusters each of which is represented as k=3(1), k=3(2) and k=3(3), respectively.

k=3(1)	**White%**	**Households**	**Family%**	**Income**	**Education**	**YrStruct**
Max	0.82	1817.00	1.00	115290.26	3.74	1967
Min	0.00	7.00	0.00	10000.00	0.00	1939
Mean	0.15	410.11	0.69	48594.36	2.25	1943
StDev	0.11	181.63	0.13	15818.30	0.36	6.21
Kurtosis	5.51	6.45	3.95	3.85	3.77	3.08
Skew	1.29	1.19	-0.42	0.83	0.08	1.14
k=3(2)						
Max	1.00	2430.00	1.00	200000.00	5.41	1996
Min	0.00	5.00	0.00	19896.99	1.74	1939
Mean	0.64	421.68	0.61	85235.33	3.04	1952
StDev	0.22	212.29	0.17	20709.45	0.58	9.58
Kurtosis	2.97	14.91	3.12	4.19	3.90	4.15
Skew	-0.61	2.10	-0.22	0.34	1.00	0.62
k=3(3)						
Max	0.82	11266.00	1.00	123647.65	3.51	2005
Min	0.00	4.00	0.08	11874.81	1.13	1946
Mean	0.18	533.00	0.66	40725.87	2.13	1969
StDev	0.13	513.75	0.15	15932.34	0.38	12.70
Kurtosis	4.34	177.93	3.47	4.08	2.78	3.04

Skew	1.09	10.12	-0.40	1.03	0.35	0.89

Table 5 Descriptive statistics for Kings County. Statistics were calculated for a total of two clusters each of which is represented as k=2(1) and k=2(2) respectively.

k=2(1)	**White%**	**Households**	**Family%**	**Income**	**Education**	**YrStruct**
Max	1.00	6602.00	1.00	122499.00	3.98	2005
Min	0.00	6.00	0.00	0.00	0.00	1939
Mean	0.30	444.25	0.70	56706.68	2.38	1947
StDev	0.29	221.55	0.14	17322.66	0.38	13.71
Kurtosis	2.56	101.66	3.53	3.19	3.57	6.18
Skew	0.83	5.04	-0.50	0.33	-0.13	1.89
k=2(2)						
Max	1.00	3006.00	1.00	199999.00	6.00	2006
Min	0.00	9.00	0.00	0.00	2.11	1939
Mean	0.75	448.21	0.55	88024.63	3.40	1943
StDev	0.21	201.02	0.17	24639.02	0.51	10.25
Kurtosis	3.94	19.05	2.66	3.23	3.26	15.00
Skew	-1.16	2.38	0.02	0.36	0.67	3.16

Table 6 Descriptive statistics for Kings County. Statistics were calculated for a total of five clusters each of which is represented as k=5(1) through k=5(5), respectively.

k=5(1)	**White%**	**Households**	**Family%**	**Income**	**Education**	**YrStruct**
Max	1.00	1827.00	1.00	199999.00	6.00	2006
Min	0.00	9.00	0.00	0.00	2.32	1939
Mean	0.76	437.29	0.54	90750.88	3.47	1942
StDev	0.21	170.59	0.17	24629.43	0.50	8.28
Kurtosis	4.00	6.42	2.72	3.21	3.27	19.80
Skew	-1.16	1.10	0.10	0.29	0.67	3.49
k=5(2)						
Max	1.00	1200.00	1.00	122499.00	3.98	1962
Min	0.00	6.00	0.00	0.00	0.00	1939
Mean	0.35	412.65	0.70	59239.37	2.43	1941
StDev	0.30	149.57	0.14	16167.85	0.38	4.44
Kurtosis	2.13	4.51	3.36	3.39	3.90	6.77
Skew	0.58	0.75	-0.48	0.40	-0.24	2.16
k=5(3)						
Max	1.00	6602.00	1.00	116026.67	4.31	2005

Min	0.00	8.00	0.03	0.00	1.26	1939
Mean	0.29	571.83	0.69	52666.03	2.40	1969
StDev	0.31	370.68	0.16	21287.96	0.47	13.66
Kurtosis	2.81	49.17	3.56	2.71	3.44	2.97
Skew	1.08	3.93	-0.48	0.52	0.60	0.79

Table 7 Descriptive statistics for New York County. Statistics were calculated for a total of two clusters each of which is represented as k=2(1) and k=2(2) respectively.

k=2(1)	**White%**	**Households**	**Family%**	**Income**	**Education**	**YrStruct**
Max	1.00	8434.00	1.00	200000.00	7.00	2005
Min	0.00	9.00	0.00	24277.10	1.91	1939
Mean	0.80	767.90	0.35	114911.42	4.21	1951
StDev	0.13	446.67	0.15	24101.47	0.41	16.70
Kurtosis	4.07	47.06	3.28	3.04	5.07	3.88
Skew	-0.91	4.38	0.43	-0.16	-0.18	1.30
k=2(2)						
Max	0.97	2389.00	1.00	124999.00	4.53	2005
Min	0.00	9.00	0.00	13556.78	0.00	1939
Mean	0.26	625.28	0.55	52336.48	2.52	1948
StDev	0.15	280.02	0.15	16969.40	0.53	13.99
Kurtosis	3.72	7.99	2.95	3.35	3.19	4.89
Skew	0.83	1.38	-0.17	0.58	0.21	1.55

Table 8 Descriptive statistics for New York County. Statistics were calculated for a total of three clusters each of which is represented as k=3(1), k=3(2) and k=3(3) respectively.

k=3(1)	**White%**	**Households**	**Family%**	**Income**	**Education**	**YrStruct**
Max	1.00	1572.00	1.00	200000.00	7.00	1960
Min	0.00	12.00	0.00	24277.10	1.91	1939
Mean	0.80	687.48	0.34	111347.96	4.18	1941
StDev	0.13	255.07	0.16	25078.35	0.42	5.06
Kurtosis	4.23	3.40	3.27	2.95	4.81	5.61
Skew	-0.97	0.44	0.50	-0.07	-0.20	2.01
k=3(2)						
Max	0.95	1935.00	1.00	124999.00	4.53	2001
Min	0.00	9.00	0.00	13556.78	0.00	1939
Mean	0.25	614.80	0.56	51143.40	2.48	1947
StDev	0.15	256.07	0.15	16080.89	0.51	12.41

Kurtosis	3.69	4.76	2.95	3.39	3.39	4.49
Skew	0.82	0.81	-0.19	0.54	0.18	1.45
k=3(3)						
Max	1.00	1535.00	0.92	200000.00	5.94	2005
Min	0.11	9.00	0.00	49273.00	2.32	1956
Mean	0.76	724.73	0.37	119476.45	4.20	1972
StDev	0.16	282.92	0.14	23655.17	0.48	13.17
Kurtosis	4.31	2.99	3.16	3.18	4.89	2.74
Skew	-1.15	0.26	0.23	-0.37	-0.55	0.92
k=4(4)						
Max	0.99	8434.00	0.70	174904.36	4.78	2005
Min	0.00	1415.00	0.11	28791.52	1.92	1939
Mean	0.74	2209.87	0.35	108925.73	4.06	1966
Stdev	0.20	857.12	0.12	25573.76	0.52	18.80
Kurtosis	6.51	23.10	3.04	4.25	7.37	2.21
Skew	-1.76	3.70	0.61	-0.92	-2.02	0.40

Table 9. Descriptive statistics for Queens County. Statistics were calculated for a total of two clusters each of which is represented as k=2(1) and k=2(2) respectively.

k=2(1)	**White%**	**Households**	**Family%**	**Income**	**Education**	**YrStruct**
Max	1.00	2933.00	1.00	141598.30	3.68	2005
Min	0.00	8.00	0.31	17516.86	1.00	1939
Mean	0.26	439.60	0.77	73666.72	2.51	194
StDev	0.21	209.35	0.12	18965.09	0.36	10.58
Kurtosis	2.76	12.07	3.10	2.98	2.80	6.72
Skew	0.72	1.85	-0.51	0.21	0.00	1.49
k=2(2)						
Max	1.00	5652.00	1.00	176344.16	6.00	2005
Min	0.00	8.00	0.00	10000.00	2.00	1939
Mean	0.69	505.99	0.62	85420.46	3.13	1948
StDev	0.19	305.10	0.16	20683.99	0.43	9.94
Kurtosis	2.94	56.20	2.75	3.49	3.47	7.38
Skew	-0.49	4.85	-0.15	0.41	0.40	1.49

Table 10 Descriptive statistics for Queens County. Statistics were calculated for a total of five clusters each of which is represented as k=4(1) through k=4(4), respectively.

k=4(1)	White%	Households	Family%	Income	Education	YrStruct
Max	1.00	939.00	1.00	111979.84	3.36	1969
Min	0.00	8.00	0.41	17516.86	1.00	1939
Mean	0.27	399.48	0.79	70849.66	2.40	1945
StDev	0.22	141.13	0.10	14462.48	0.31	6.94
Kurtosis	2.68	3.22	3.02	3.05	2.99	2.65
Skew	0.71	0.45	-0.40	-0.27	-0.20	0.76
k=4(2)						
Max	1.00	1483.00	0.87	124952.79	4.53	1973
Min	0.00	8.00	0.00	19999.00	1.82	1939
Mean	0.73	509.55	0.55	75613.17	3.01	1946
StDev	0.17	198.73	0.13	14832.49	0.43	7.23
Kurtosis	3.14	4.01	3.36	3.12	2.81	2.35
Skew	-0.59	0.74	-0.36	-0.05	0.28	0.68
k=4(3)						
Max	1.00	5652.00	1.00	133043.09	4.22	2005
Min	0.00	17.00	0.00	19288.87	1.66	1939
Mean	0.34	777.45	0.63	65551.57	2.70	1964
StDev	0.23	428.54	0.13	17797.21	0.44	11.74
Kurtosis	2.50	34.21	3.42	3.54	3.34	5.35
Skew	0.48	4.02	-0.46	0.21	0.54	1.49
k=4(4)						
Max	1.00	1197.00	1.00	176344.16	6.00	2005
Min	0.00	9.00	0.18	10000.00	2.02	1939
Mean	0.46	338.09	0.80	103778.80	3.17	1949
StDev	0.31	124.75	0.10	15421.80	0.40	8.12
Kurtosis	1.93	4.48	3.73	4.63	5.08	5.30
Skew	0.01	0.65	-0.43	0.41	0.87	0.77

Means across the six features were visually compared to understand the contents of each cluster resulting from classification. Visual comparison of means revealed distinct states typically distinguished by race, income and education levels. Similar states were observed across the five boroughs, however the number of states obtained by the clustering varied across boroughs. This

indicates that there may be distinct typologies of urban change unique to geographical contexts. Additionally, it is observed that race correlates strongly with income across choices of k, regardless of borough.

Figure 11 Comparison of means for Bronx County across six measured dimensions for clusters K=1 through K=3.

Figure 12 Comparison of means for Kings County across six measured dimensions for clusters K=1 through K=5.

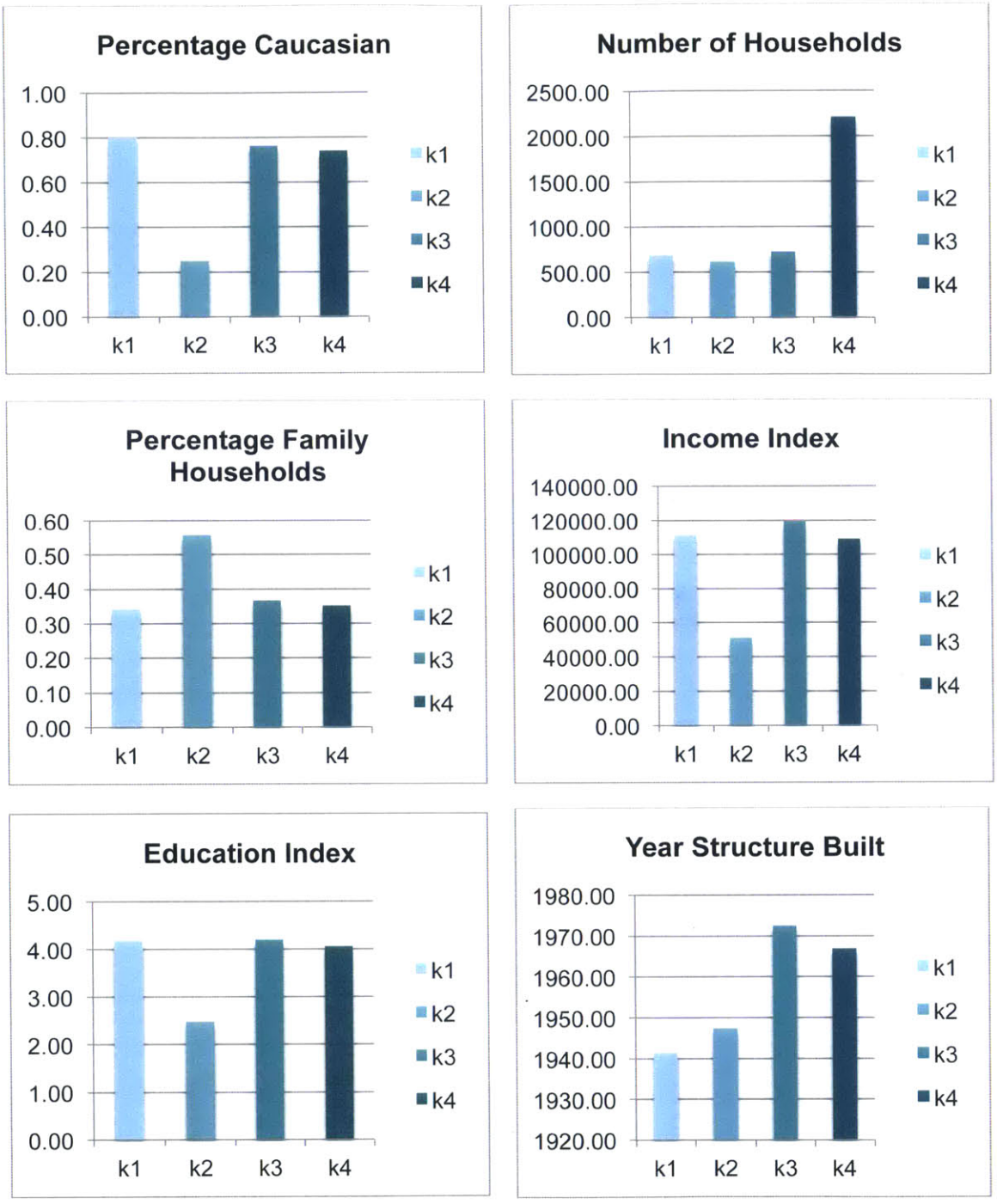

Figure 13 Comparison of means for New York County across six measured dimensions for clusters K=1 through K=3.

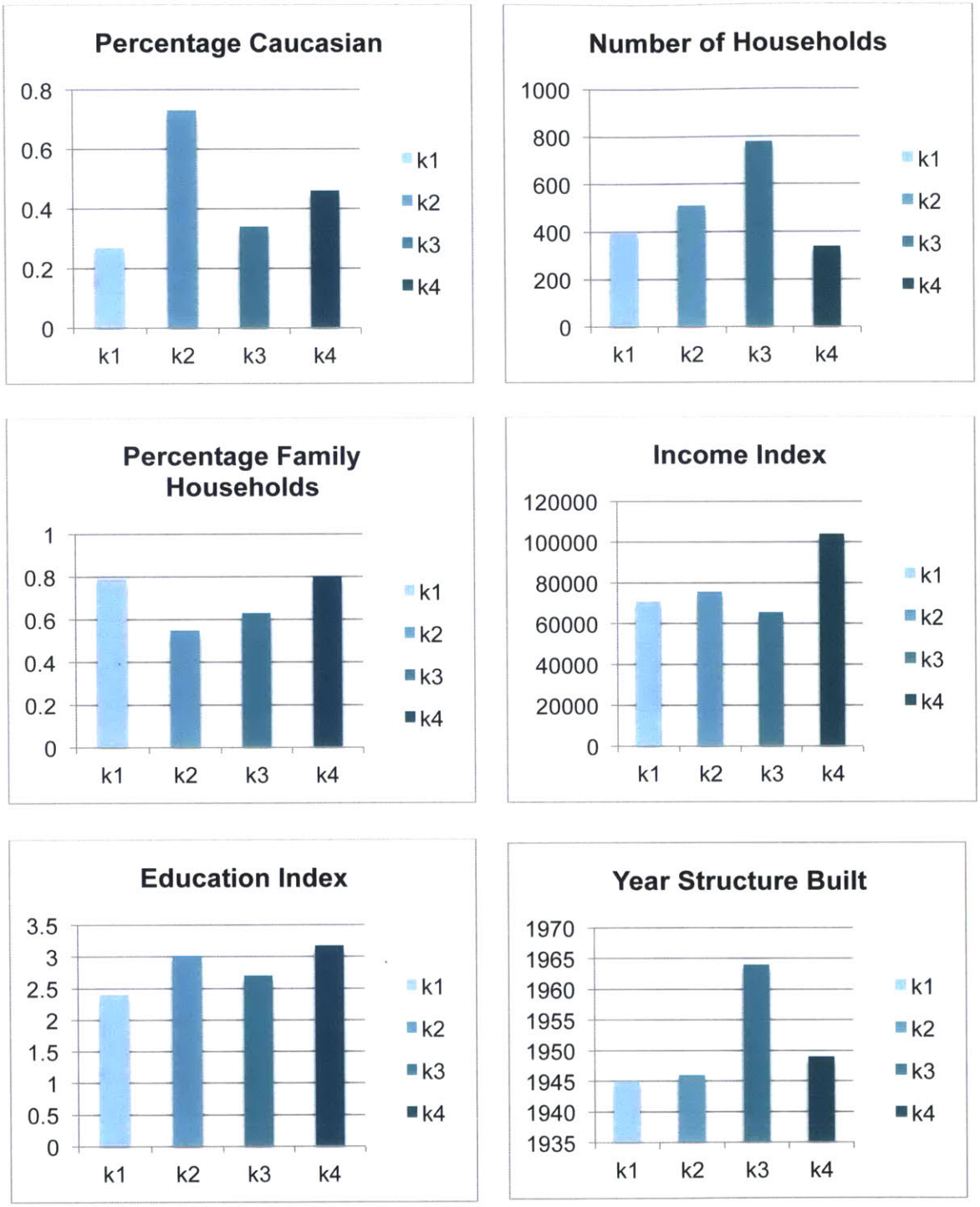

Figure 14 Comparison of means for Queens County across six measured dimensions for clusters K=1 through K=4.

1.1 Bronx County

For Bronx county, the k-means clustering algorithm returned k=2, s(i) = 0.53, as the best clustering and k=3, s(i) = 0.39 as the second best clustering. The results of the k=3 clustering trial was selected as the optimal feature extraction mechanism, since it represented a balance of robust classification and accuracy. In this clustering, high income-earning Caucasians classified into one cluster (State 2, 64% white, with an average income of $85,235) while minorities were split across two clusters (State 1, 15% white and State 3, 18% white), varying primarily along dimensions of education and housing stock (structural age). In the Bronx, White demographics lived in mid-century building stock (1952 on average), while minorities were found on the extremes of early to late building stock with centroid averages of 1943 and 1969, respectively. Minorities occupied high-density areas, while whites occupied medium density areas. Education levels for State 2 were significantly higher (3.04 compared to 2.13 and 2.25), while percentage of family households were comparable (61%) to the other two clusters (61% and 69%). Income levels for State 2 were nearly double that of State 1 and State 3 ($85,235 versus $48,594 and $40,725, respectively).

1.2 Kings County

For Kings county, the k-means clustering algorithm returned k=2, s(i) = 0.40, as the best clustering and k=3, s(i) = 0.39. In Kings County, high income and education levels and non-family households correlated strongly with Caucasian

demographics, suggesting a class of educated, high-earning whites who may be younger or elderly. A distinct "yuppy" class emerged from this clustering classification in State 1, characterized by a primarily white population (76%) with relatively low percentages of family households (54%), high education (3.47) and high income levels ($90,750.88) in relatively aged building stock (1942). Tracking the volatility of this state; that is, the degree to which census blocks transition in and out of this state in Kings county may provide key insights into gentrifying areas. States 2 and 3 showed a lower population of whites (35% and 29%, respectively) and diverged primarily along dimensions of structural age and density. State 2 had the highest percentage of family households out of the classification (70%), a low education index (2.43), moderate-income level ($59,239.37), and occupied relatively old building stock (1941). State 3 residents resided in high-density areas (571 households), with a high percentage of family households (69%), low education index (2.40), low-income levels ($52,666), and occupied relatively recent, mid-century building stock (1969).

1.3 New York County

In New York county, the k-means clustering algorithm returned k=2, s(i) = 0.59, as the best clustering and k=4, s(i) = 0.50. For New York County, non-minority groups were split into three of four States, varying primarily along dimensions of building stock age and density. State 2 was strongly characterized as a minority group (25% white), relatively low density (614 households), high percent family levels, low incomes (nearly less than half of the other States at $51,143) and

relatively low education levels (2.48). Both income and education levels were around half of the centroid averages for the other states, while percentage of family households was nearly twice the average of other states. States 1, 3 and 4, were characterized by high rates of non-minorities (80%, 76% and 74%, respectively). These clusters displayed the highest income brackets of all observations across boroughs, ($111,347, $119,476 and $108,925, respectively). These classes diverged primarily around structural age, ranging from 1941 for State 1 to 1972 for State 3, and density where State 4 shows density levels nearly twice the size as States 1,2 and 3 (2209). Non-minority states on average are more educated, richer and exhibit lower percentages of family households (34%, 37% and 35% compared to 56% for State 2).

1.4 Queens County

In Queens county, the k-means clustering algorithm returned k=2, s(i) = 0.34, as the best clustering and k=4, s(i) = 0.32. Race maps on to three of the four clusters somewhat evenly for Queens county (27%, 34% and 46%) while State 2 is classified by a higher rate of non-minorities (73%). State 2 is also classified by older building stock (1946), the third lowest income level ($75,613), medium density levels (509 households) and a relatively low percentage of family households (55%). State 1, with a 27% non-minority rate, has the second highest rate of family households (79%) and relatively high income levels ($70,849) compared to the other States. State 4, which is less than 50% white, possesses the highest income bracket for the borough ($103,778), highest income levels

(3.17), low density (338 households) and a high percentage of family households (80%). State 3, characterized by high minority levels (34% white), high density, low income ($65,551) and low rates of family households (63%) occupies the most recent building stock of the group (1964 compared to 1945, 1946 and 1949 centroid averages for States 1, 2 and 4, respectively).

4.5 Advantages and Disadvantages of k-means Clustering

There are key disadvantages to the k-means clustering algorithm, namely the tradeoff between robust classification and classification accuracy. As the number of *k* increases, classification may be better representative of the data set, but robustness is compromised. This imposes a subjective selection of *k* onto the researcher. Furthermore, k-means clustering does not obtain maximum likelihood for classification, nor does it provide a probability of classification for each observed feature. It is understood that the features given by the ACS data are continuous variables, variables that could theoretically take up any value along a continuum between specified ranges. Features like income level, for example vary continuously along a scale. The k-means clustering algorithm treats feature observations as discrete variables, using Euclidean distance to specify correlations between feature vectors. While this is convenient in terms of feasibility, and provides a generally reliable description of the data, it may not be the best method for obtaining states as inputs into the Markov model.

One suggestion, and future motivation of this research is to utilize the Latent

Class Modeling (LCM) as the means of obtaining state classifications. LCM relates a set of observed (often discrete) multivariate variables to a set of latent variables that capture the residual error and greatest amount of variance in the data. Classes are characterized by a pattern of conditional probabilities that indicate the likelihood that variables will take on certain values. Latent Class Analysis (LCA) performs the same function for categorical data. A combination of these methods may be more applicable for both 1) urban data and 2) the process of obtaining states as inputs into a Markov model in order to observe transitions between latent classes over time. Opportunities for this type of modeling technique are discussed further in the Discussion.

5 Journal Submission

Towards an Epidemiology of Gentrification: Modeling Urban Change as a Probabilistic Process using k-Means Clustering and Markov Models

Emily Royall and Thomas Wortmann

Abstract

Gentrification is viewed as both as a tool and a force--as a systematized vehicle for class-based oppression and racism, and an empirical force of change based on social, environmental and economic interactions. This complexity makes it challenging for researchers to study the impact of gentrification, for planners to anticipate the effects of gentrification with planning policy, and for developers to foresee investment outcomes. Current planning policy addresses the symptoms of gentrification, without defining the underlying construct of the process. This paper examines latent constructs of gentrification through a data-driven process that identifies emergent states of change and assigns them to a Markov process, i.e. a process that assigns probabilities to potential "state" changes over time. For census block groups in four boroughs of New York City, our model takes three steps: 1) cluster census block groups into latent states defined by ACS socioeconomic and demographic data, 2) derive a Markov model by tracking transitions between states over time, and 3) validate the model by testing predictions against historic data and qualitative documentation.

E. Royall (Corresponding Author)
Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139
Email: eroyall@mit.edu

T. Wortmann
Architecture and Sustainable Design, Singapore University of Technology and Design, 487372 Singapore
Email: thomas_wortmann@mymail.sutd.edu.sg

1. Introduction

Recent scholarly debate reveals two cultures of gentrification theory emerging from studies in domains of the natural and social sciences. Whether gentrification can be viewed as a natural force examined through quantitative modeling processes, or a political structure understood through capital flows and public testimony, defines contemporary debate around an emerging interdisciplinary approach to gentrification studies (Ball, 2014; Slater, 2014; Veneradi, 2014; Portugali, 2009). Since the term's introduction by Ruth Glass in the 1960s, gentrification has been described and explored primarily in relation to its symptoms: racial discrimination, displacement, rising rents and changes to physical and community attributes (Lees, 2008; Smith, 2002; Clay 1989). However, these symptoms cannot be confused with the underlying processes causing the disease suffered by countless communities, in America and abroad. Indeed, gentrification literature draws no consensus that any of these symptoms can be confidently assumed as causal of gentrification. However, throughout the literature, gentrification is consistently described as a *process* of socioeconomic and demographic *change* in urban areas. Understanding, conceptualizing and modeling this process of change is the first step towards an epidemiology of gentrification.

In epidemiology, symptoms are defined as subjective experiences observed by the patient that occur as the result of a disease process, while diseases are disorders in the objective structure or function of a healthy system (CDC, 2012). Diseases are understood as the latent processes that symptoms only describe. This paper adapts this analogy to design a quantitative methodology which identifies latent constructs of neighborhood-level, socioeconomic/ demographic conditions, and tracks how those conditions change over time in both gentrifying and non-gentrifying neighborhoods. This data-generating process combines pattern classification and machine learning methods to explore potential disease typologies of gentrification in the four NYC boroughs.

Our method identifies emergent patterns or "states" of correlated socioeconomic conditions in census block groups, and monitors how block groups transition through these states. Using k-means clustering to identify common socio-economic "states" through which urban census blocks transition, we represent neighborhood change as a probabilistic process of state transformations over time, i.e. a Markov process (Rabiner, 1989). Using American Community Survey (ACS) data for four counties in New York (Bronx, Queens, Kings and New York) between 2009 and 2013 (including demographic, economic, geographic, and physical characteristics of census block groups), we create a Markov model in three steps: 1) clustering census block groups into "states" defined by ACS socioeconomic and demographic data, 2) deriving a

Markov model by tracking transitions between "states" over time, and 3) validating the model by generating predictions for un-tested data and comparing them against qualitative documentation of neighborhood change and gentrification. Our method allows the empirical study of neighborhood-level urban development by condensing complex urban data into latent profiles that attempt to describe and predict urban change. Our findings indicate that machine learning and pattern recognition processes hold promise for understanding patterns of urban change associated with gentrification. Tests of the predictive capacity of the Markov models for the four counties show the promise of our method as a tool for planning agencies to model urban changes in a metropolitan area, and as an opportunity to refine the approach of planning policy by targeting symptoms of gentrification in support of negatively impacted communities. The k-means clustering method successfully identifies states, i.e., patterns of urban development. However robustness of this method could be improved. States emerging from ACS data map consistently to neighborhood boundaries, and enable the spatial comparison of neighborhood areas exhibiting similar socio-economic and demographic properties. The degree to which census block groups are spatially grouped into state categories may be an indicator of relative levels of socio-economic segregation; i.e., areas where census block groups fall into a single state represent pockets of homogenous characteristics, while regions with a patchwork distribution of states show variation of socio-economic conditions.

2. Concepts and Causes of Gentrification

2.1 The Gentrification Debate

In 2014, Tom Slater, Reader of Urban Geography at the University of Edinburg published a seething critique of an article written by Philip Ball, the editor of the American scientific journal, Nature. "Gentrification is a Natural Evolution," described gentrification as a natural force underpinning the evolution of cities. The piece reviewed a recent paper "The Form of Gentrification" (Veneradi, 2014) published earlier that year in *Physics and Society*, which took a Complexity Theory of Cities (CTC) approach to identifying emergent properties of neighborhood characteristics and statistically correlated them to gentrification. As Ball put it, the work suggested that "cities obey laws beyond the reach of planning," likening the study of cities to that of the evolution of biological organisms[10]. Slater responded to Veneradi's work and Ball's support of it, arguing that if an urban process resulting in the destruction of communities is understood to be "natural," it will be justified by the political elite to further marginalize and exploit minorities and low income communities (Slater, 2014). In his blog, "Homunculus" Ball responds to Slater's attack charging that the urban theorist had confused urban science with Social Darwinism (Ball, 2014). Ball's response echoes the broader position of CTC theorists: that Complexity offers methodologies designed to capture 'non-intuitive' or latent socioeconomic structures of urban processes that are assumed to behave as complex systems. The attraction of these modeling capacities according to CTC theorists is the possibility that they may uncover hidden insights into a living system. CTC theorists, to the chagrin of their poststructuralist counterparts, take a weak positional stance regarding the political frameworks surrounding their work.

This public debate between a scientist and humanist about cities reflects a critical rupture in today's urban development community. At the core of this rift is gentrification, viewed both as a tool and a force---as a systematized vehicle for class-based oppression and racism, and an empirical force of change based on social, environmental and economic interactions[11]. Is gentrification a political or natural process? Are political and natural processes necessarily different? This question has long motivated researchers of urban change and gentrification studies, and continues to stimulate heated contemporary debate. The two sides of

[10] See Gentrification is a Natural Evolution: http://www.theguardian.com/commentisfree/2014/nov/19/gentrification-evolution-cities-brixton-battersea

[11] This is a Newtonian reading of "force" whose third law describes a force as an interaction between different bodies. The analogy of "force" as applied to gentrification is intended to reflect the recent contributions of physicists to the debate on gentrification studies and urban change.

the gentrification debate reflect a broader, historic divide between quantitative, positivists and qualitative, poststructuralists (Portugali, 2009). The split between the quantitative, positivist perspectives of communities (the "space" perspective) and qualitative, poststructuralist perspectives (the "place" perspective) resulted in a body of gentrification literature that can't agree on virtually any aspect of gentrification, much to the detriment of bringing positive change to residents in gentrifying neighborhoods.

2.2. Conflicting Definitions

Urban change has a long history of empirical study (Du Bois, 1899; Weber, 1899). Gentrification, a specific typology of urban change was documented relatively recently however, with the language to describe it formally coined by Ruth Glass in 1964. The emergence of gentrification in both idea and form during the mid 1960s posed a challenge to traditional ideas about how urban change worked. Observations of gentrification described by inner-city reinvestment, socioeconomic and demographic compositional changes, and rising property values at the urban core challenged the conventional wisdom that urban change was a process restricted to the periphery. Historically, neighborhood change was viewed as a process of perpetual expansion on the periphery, driven by class and capital in which the upper classes were thought never to return to older neighborhoods (Hoyt, 1939). Likewise, Burgess's traditional concentric model predicted growth and change on the urban fringe (Burgess, 1923). These early perspectives understood urban change as a symptom of suburban growth in a post-war housing market. Scholars of urban change in the 1960s and early 1970s began documenting how instances of gentrification deviated from convention and these cases demonstrate some of the early frameworks shaping gentrification studies of this period (Freeman, 2005).

Late 20th century gentrification studies confirmed that a process of urban change was occurring in many inner-city communities across the US during the 1970s (Clay 1979; Sumka 1979; National Urban Coalition; 1977). Seminal views emerging from these observations have characterized gentrification as urban re-investment resulting in displacement of the poor (Lees et al., 2008, Clay 1979), a result of the "rent-gap" hypothesis (Smith, 1979) and a symptom of regional economic or demographic change (Clay, 1989). While these scholars agreed that a unique pattern of urban change was emerging from America's inner city neighborhoods---they were unable to settle on causes or definitions of the process. As an interdisciplinary problem gentrification was difficult to define and both the social and natural sciences had their own views. The widening cultural divide between natural and social scientists further aggravated this lack of consensus in the late 20th century (Portugali, 2006).

A robust definition of gentrification does not exist today. The U.S. Department of Housing and Urban Development currently defines gentrification as "the process by which a neighborhood occupied by lower-income households undergoes revitalization or reinvestment through the arrival of upper-income households (U.S. Department of Housing and Urban Development, 1979.)." This definition does not include displacement as a necessary outcome of gentrification. Alternatively, the Brookings Institution cites gentrification as "the process of neighborhood change that results in the replacement of lower income residents with higher income ones (Brookings Institution, 2001. pg 11)." The Brookings Institution draws a clear connection between gentrification and displacement caused by the arrival of a specific actor: a high-income demographic. However, a measure of gentrification relying on solely income could dismiss gentrifying neighborhoods that experience a high influx of an educated, but not necessarily high-income class (Clay, 1979; Freeman, 2005). Alternatively, the Encyclopedia of Housing (Smith 1998) defines gentrification as "the process by which central urban neighborhoods that have undergone disinvestments and economic decline experience a reversal, reinvestment, and in-migration of a relatively well-off, middle and upper middle class population." Unlike the Brookings Institution, the Encyclopedia of Housing does not ascribe agency to any one actor as responsible for gentrification, and dismisses the typology of gentrification that is documented in well-invested, but relatively low-income neighborhoods. While these authorities agree that gentrification is fundamentally a "process," the causes and effects remain open for interpretation, and are often in conflict with each other.

Quantitative approaches to defining urban change also reflect the cultural split. As a result, models of urban change and gentrification often fall into binary categories of causality (government-assisted vs. market-driven, spatial vs. social), outcomes (displacement vs. succession) and modes (political agents vs. natural forces). Two important perspectives emerge from the division: the "space" perspective and "place" perspective (Portugali, 2006), where models emphasize data-driven spatial outcomes, or humanistic experiences and political outcomes. Overall, researchers over the decades have drawn little consensus about the causes and outcomes of gentrification, and models have disappeared and re-emerged in the popular discourse.

2.3 Space and Place-based Models

Modeling of gentrification falls into two categories that reflect the cultural split. Place and Space based models reflect emphasis on institutions, people and experiences, or physical forces, geographies, and environmental factors respectively. The "space" based perspective sees urban change in terms of spatial actors and outcomes, and uses modeling tools that tie data to spatial boundaries for the purpose of organizing information in a computationally friendly way. This view was adapted

from location theory, which was the foundation of quantitative, positivist approaches to urban geography (Portugali, 2006). From this perspective spatial interaction between bodies and settlements, central places and demand, is governed by spatial forces such as distance measured by transportation costs (Portugali, 2006). Largely informed by spatial modeling in ecology, these early models of urban form saw space as a landscape of physical forces and largely neglected dimensions of human experience and culture. Alternatively, Place-based models focus on the social factors underlying urban change, and give broader consideration to social justice and equitable development. The seminal Rent Gap Theory is one such example (Smith, 1979). While these approaches are sometimes quantitative, they are primarily used to uncover inefficiencies in the political structures that give rise to inequality, displacement and cultural deterioration.

Neighborhood Life Cycles and Stage-models

Several scholars have framed urban change and gentrification in terms of lifecycles, stages of development, and evolution (Burgess, 1929; Hoyt, 1939; Smith, 1979; Birch, 1971; Hoover & Vernon, 1959; Vernon, 1959). For decades Burgess's concentric model of urban growth was the conventional wisdom regarding urban change, and several urban change researchers built on the principles of his work. Burgess's model showed concentric rings of growth and disinvestment, suggesting that cities expanded on the periphery, as higher-income classes developed or purchased new property, and lower-income classes would move into the neighborhoods left behind.

The concentric, stage model of urban development proved fruitful for quantitative analysis and policymaking for several decades. In "The Changing Economic Function of the Central City," Vernon identifies an "early" and "late" state of gentrification, which would later inform the seminal "rent-gap" hypothesis (Vernon, 1959). Hoover and Vernon further elaborated on this model in *Anatomy of a Metropolis*, where they identified five distinct stages of development contributing to waves of concentric change (Hoover & Vernon, 1959). In "Towards a Stage Theory of Urban Growth," David Birch used Hoover and Vernon's stages to measure change at the census block group level. Birch acknowledged the likelihood that a given census block group could experience multiple stages of development simultaneously, and sought to quantify this possibility using probability density functions (Birch, 1979). The results show topography of age with a structure more complex than Burgess's concentric model. Life cycle and stage models of gentrification have recently been criticized for failing to explain or predict observed change, and for their inconsistency with observations of metropolitan cities (Marcuse, 1985).

Spatial Simulations

QP Researchers have also taken a simulation approach towards modeling gentrification and urban change, often borrowing from computational techniques in other fields. Jay Forrester's *Urban Dynamics*, uses logistic growth and diffusion processes for developing a simulation of urban dynamics (Forrester, 1969). Forrester emphasizes the important effect environmental change has on neighborhood change, and presents 'modes' of urban growth, survival and revival. Forrester's work arguably set the groundwork for a complex systems theory of urban change.

Adopting a similar approach, Thomas Schelling, in 1971, showed that a preference for a neighbor's race could lead to starkly segregated environments. Schelling's model was based on cellular automata (CA), where spatial outcomes evolve through a series of discrete time steps according to a set of rules based on the states of neighboring cells. Likewise, O'Sullivan models gentrification based on a CA programmed using rules sourced from the rent gap hypothesis (O'Sullivan, 2002).

Cellular Automata (CA) is still widely used in space-based models of urban change. While CA adds an interesting layer of complexity to spatial simulations, there are three weaknesses in the CA approach to modeling gentrification. First, CA produces a simulation based on rules programmed a-priori, and do not allow structures to emerge from data analysis, as machine learning techniques do. Second, transition rules play out in an abstract space, and it is difficult to apply simulation results to the actual physical boundaries of urban neighborhoods. Third, transition rules depend on the states of their physical neighbors, while the extent to which an adjacent block's gentrification-status actually influences your own are not well established in gentrification studies. While Michael Batty also makes use of CA analysis to model urban change, he stresses that theories of urban change must give equal weight to questions of socioeconomic dynamics as well as spatial form (Batty, 2005).

Multi-agent simulations (MAS) are sometimes employed to model the spatial outcomes of urban change. Torrens & Nara advocate a hybrid approach between MAS and CA (Torrens & Nara, 2006). This work focuses on households as "agents" making choices in dynamic property markets. The MAS approach to modeling urban change is problematic because it depends on pre-programmed relationships that generate a simulated outcome. In the context of gentrification, where causal factors are disputed if not unknown, simulating spatial outcomes of gentrification using pre-programmed dynamics is less relevant. Alternatively, an agent-based model (ABM)---where interactions of autonomous agents are assessed for their effects on an emergent outcome, may be more appropriate. In any case, models that do not assume system dynamics a-priori are likely to be to be more suitable for modeling what appears to be a black box of gentrification dynamics.

Spatial models of gentrification increasingly pay more attention to the role of human perception. Recent work from Cesar Hidalgo at the Macro Connections Group in the MIT Media Lab, focus on measuring urban change and perception by combining machine learning and statistical techniques. In *"Do People Shape Cities or do Cities Shape People"* Hidalgo and Glaeser apply a machine learning algorithm to user-generated Google Street View data using safety perception rankings to obtain a "Street Score" of safety perception in urban neighborhoods (Hidalgo, 2014).

Rent Gap Theory

Neil Smith in "Toward a Theory of Gentrification: A Back to the City Movement by Capital, not People," proposed the seminal "rent-gap" model arguing that capital flows are responsible for observed patterns of reinvestment in inner city neighborhoods. Smith observed what he called the "rent gap": the disparity between the potential ground rent and the actual ground rent capitalized under a parcel's present land use (Smith, 1979). According the Smith, gentrification is a structural product of land and housing markets, of which displacement and demographic changes are merely a symptom. For Smith, the "leading edge" of gentrification occurs because capital flows to where the rate of return is highest, and inevitably a rent gap is generated as capital flows out of inner city districts. This depreciation of capital invested in inner-city districts produces economic conditions that in turn make a capital revaluation a rational market response (Smith, 1979).

The Rent Gap model is an important concept in gentrification modeling, and has been used as the primary model for dynamics of urban change. Slater, a strong supporter of the model, has used it to explain gentrification as "a broader attempt to trace the circulation of interest-bearing capital in urban land markets, and to elaborate the role of the state in lubricating that circulation (Slater, 2014. Online)." Clark finds further evidence of Rent Gap in a case study of 125 years of rent fluctuations in Malmo, Sweden (Clark, 1988). However, Rent Gap has undergone criticism in the literature (Hamnett, 1984; Ley, 1986), as it fails to address the role of public and private institutional actors in neighborhood revitalization. A major criticism of the rent gap hypothesis is also that it favors a supply-side explanation of gentrification, over demand (Ley, 1986).

Social Statistical Modeling

David Ley makes the case for the operationalization of criteria used to describe gentrification, as well as the contextual differences of the gentrification process between nations. Ley established a common analytical procedure for gentrification analyses today: identify descriptive criteria in the literature, operationalize them, and then hold each defining factor as an independent variable in a sequence of multivariate analyses (Ley, 1986; Freeman & Braconi, 2004; Vigdor,

2002; Freeman, 2005). Because the factors shaping gentrification are diverse and complex, multivariate analyses are intuitively appropriate for isolating the effects of certain factors on gentrification. Ley creates a "gentrification index" using indicators of social status that is then treated as a dependent variable for simple correlation and regression studies. Multicollinearity among 35 predictor variables is treated using a Principal Components Analysis. Strong relationships are found between the gentrification index and variables representing urban amenity and economic dimensions; the highest being office space per capita (Ley, 1986).

There are two key issues with this approach, which become increasingly apparent as today's computational analyses are better able to handle complexity: 1) researchers are forced to find a proxy for gentrification to serve as a dependent variable in their analyses, and 2) lack of consensus about the definition of gentrification makes identification of such a proxy difficult. The results of this type of modeling approach is limited in that it can only draw conclusions about the dependent variable selected, which may or may not be a true proxy of gentrification. Consequentially, the results of several of these studies directly contradict each other, since the choice of proxy is subject to researcher.

Succession & Displacement Studies

A core debate emerging from place-based gentrification studies was whether gentrification was characterized by succession or displacement (Freeman, 2005). Theorists considered the importance of demographics, amenity access, and lifestyle changes (Ley, 1986; Beauregard, 1986; Hamnett, 1991), beyond the market-based and physical dimensions explored primarily by stage-based analyses in the space-based camp. Succession studies compared the characteristics of in-movers and out-movers (Henig 1980), while displacement studies surveyed respondents to identify displacement hotspots in a neighborhood or region (Grier, 1978). These studies relied primarily on survey responses and were ultimately flawed in that they were unable to correlate gentrification processes to displacement outcomes, or were unable to distinguish gentrification from other environmental factors contributing to displacement (Freeman, 2005). However, Slater recently advocated for reintroduction of displacement studies, citing how there is wide agreement that class should be the undercurrent in the study of gentrification (Slater, 2006). His recent work suggests quantitative analyses have not reached clear conclusions regarding gentrification processes, correlates or outcomes and researchers should instead focus on a value-based approach that examines the political structures responsible for observed uneven development.

In contrast to the spatial, rule-based models described above, we propose a data-driven model of gentrification as a probabilistic process in time.

Two factors support this modeling choice. First, the lack of consensus regarding the outcomes of gentrification (displacement, environmental change, social reorganization or property valuation changes) suggests that less emphasis should be placed on the result of the modeling or simulation process. This definitional uncertainty makes agent-based modeling, where interactions are pre-defined to achieve desired effects, less applicable to the modeling of gentrification. Second, the view that gentrification is a temporal process and not a static condition demands the modeling of a process in time, which is a powerful feature of the Markov Model described in section 0.

3. Materials and Methods

Our analyses occurs in three steps 1) The procurement and preparation of socio-economic data at a census block group level and over four time-periods (for the four counties of Bronx, Kings, New York, and Queens), 2) the clustering of these block groups into states (separately for each county), and 3) the derivation of four Markov models by tracking transitions between states over time for each county.

3.1 Procuring and Preparing ACS Data

We obtained five-year estimates from the American Community Survey (ACS) between 2005 and 2013 via socialexplorer.com, a common data resource for Census and ACS data in the United States. Four counties were selected for our analysis at the block group level: Bronx, Kings, New York, and Queens. These counties were chosen due to their spatial proximity, data set size, and consistency in data sampling across the region. The final data set consisted of 29,058 observations (i.e. census block group five-year estimates) per region (see Table 1), characterized by 32 fields.

Table 1. Number of census block groups 5-year estimates per county and year

County	Total	2009	2010	2011	2012	2013
Bronx	5400	925	1116	1119	1121	1119
Kings	10182	2031	2037	2038	2038	2038
New York	5164	854	1076	1078	1078	1078
Queens	8312	1571	1685	1685	1685	1686
Total	29058	5381	5914	5920	5922	5921

5.1.1

3.1.1 Advantages and Limitations of American Community Survey (ACS) Data

The primary advantages of ACS data are its accessibility and the availability of data at the census block group-level. Data at the census block group level or smaller is appropriate for modeling processes like gentrification, are visible and have effects at the neighborhood scale. Additionally, the variety and amount of data available through ACS is appropriate for the clustering technique proposed here.

There are notable limitations of the ACS data. First, ACS only provides five-year, multi-year estimates at the census block group level. Five-year estimates are updated annually by removing the earliest year of the estimate and replacing it with the latest one (US Census Bureau, 2008). For example, following collection of data for 2013, data estimates from 2007 will be dropped to create a 2008-2013 estimate. Therefore, multi-year estimates represent a period, rather than a specific year, and estimates overlap across periods. This overlap is a significant limitation for our model, as the overlapping estimates might significantly underestimate short-term variations occurring in vivo. For our model, single-year estimates at the census block group level would be more ideal, since we aim to capture time-series variation in data. Additionally, the ACS data collection process is relatively new, having started only in 2006 and the five-year estimates beginning as late as 2009. Consequently, there are inconsistencies in earlier five-year estimates across regions; some fields are missing or unavailable and other fields were collected under varying conditions. Specifically, our data includes estimates made after the 2008 financial crisis and may not be representative of typical trends or patterns occurring under normal economic conditions. Finally, the period of the data is relatively short, encompassing only five years between 2009 and 2013, obscuring long term patterns and trends. For further discussion regarding the pre-processing of ACS data for clustering see Appendix 5.1.

3.2 Clustering ACS Data

To identify developmental states of census block groups in ACS data, we employed the K-means clustering algorithm (MacQueen, 1967). K-means is an unsupervised machine learning technique. The algorithm aims to find previously unknown patterns in data that have not been assigned a category or label. In this section, we address the k-means clustering algorithm, our application of the algorithm to the ACS data, and our choice of the number of clusters k.

3.2.1 The k-means Algorithm

Given a set of multi-dimensional data points, k-means partitions the set into k clusters, while aiming to minimize the difference between the data points in each cluster. Mathematically, this difference is computed as the sum of the distances from the data points in each cluster to the center

point (or centroid) of the respective cluster. This sum of distances is the *objective function* that the algorithm attempts to minimize.

Starting with k randomly chosen cluster centers, each data point is assigned to the cluster center that is closest to it. In a second step, a new center point can be computed for each cluster by finding the center of mass (i.e. the average) of the data points that belong to the cluster. This procedure is repeated until the clustering no longer improves, i.e. until the cluster centers stop to change. The procedure can be summarized as follows:

1. *Choose k random cluster centers.*
2. *Assign each data point to the cluster whose center point is closest to it.*
3. *Recalculate the position of each cluster center as the average of the cluster's members.*
4. *Repeat steps 2 and 3 until the cluster centers no longer improve.*

3.2.2 Applying k-means to the ACS Data

For the k-means algorithm, every five-year estimate for every census block group is represented as a 32-dimensional data point (since, as described above, our data set has 32 normalized features). Most census block groups appear several times in our data set, representing change in a census block group over time. In other words, the same census block group can occupy different positions in the 32-dimensional space of the data set due its developmental changes over time.

We clustered the pre-processed data described in section 3.1.2 with the k-means clustering algorithm included in the Statistics and Machine Learning Toolbox of MATLAB. To mitigate the effect of the random choice for the first cluster centers, we computed 20 clusters with different starting points for each of the four counties, minimizing the sum of squared distances criterion function:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Where d represents Euclidean distance. The smallest distance was selected as the final clustering for that block group. The criterion function was compared across several selections of the number of clusters k, so that minimum error was achieved.

This procedure assigned each data point, i.e. each five-year estimate of a census block group, into a category or state, based on its socio-economic data. Note that, although the number of states had to be decided a-priori, the properties of the states emerge from the clustering process itself. (The states are characterized in more detail below.) By computing a single k-means clustering for the census block groups of each county over several

years (from 2009 – 2013), we could assign a category to every census block group at every time step, resulting in a series of state changes over time. After addressing the issue of choosing the number of clusters k in the following section, we describe how we developed a probabilistic model of urban change based on these state changes.

3.2.3 Cluster Size Selection

As mentioned above, k-means requires its user to choose the number of clusters k a-priori. How can one determine the "true" number of clusters in a data set? No straightforward answer exists, although many different methods have been proposed (e.g. Sugar and James, 2011).

Mardia et al. (1980) propose to choose k as the square root of half of the number of data points as a rule of thumb. According to this rule, in our case we would have around 60 clusters, based on an average sample size of 7.265. However, due to its inherent complexity, a model based on such a large number of states would contributed little in terms of understanding gentrification.

Instead, the statistical properties of cluster sizes k = 3, 6, 9, and 12 were investigated. k = 6 had the largest and most evenly distributed cluster size, and the greatest number of fields displaying low dispersion rates (as measured by a coefficient of variation) across clusters. At k = 9, clusters appeared to me more random in composition, and at k = 3, not enough variation appeared between clusters to enable meaningful comparison. Further comparison of the criterion function identified k = 6, as the most appropriate cluster size.

3.3. Creating a Markov Model from the ACS Data

As previously discussed, each census block group was assigned to one of six clusters by applying the k-means clustering algorithm. As estimates for most blocks groups were consistently available for each period in our data, we were able to track the states, and thus the state changes, of the census block groups over time. From these state changes, we derived a Markov model of socio-economic change at the block group level.

3.3.1 Markov Chains

A Markov chain is a mathematical model that describes a probabilistic process of changes over time (Durret, 2010). As such, Markov chains have found wide applications in the natural and social sciences. Generally, a Markov chain is a system defined by a set of states N (i.e., the state space), and a matrix of transition probabilities P. N contains all the possible states n of the system, while P assigns probabilities to the transitions between these states. (See table 5 for an example of a transition probability matrix.) Given N and P, one can simulate the trajectory of a system by generating a random number p and letting the

system change to the new state $n(t+1)$ defined by P for the current state $n(t)$ in case of p:

$$n(t+1) = P(n, p)$$

By repeating this process, a Markov chain model traverses a sequence, or chain, of states over time. Note that a key modeling assumption of Markov chains is their *memoryless* quality. That is, the next state only depends on the current state and the transition probabilities for that state. For our model we assume that P is *time-homogenous*, i.e. that the transition probabilities remain stable over time. In the following sections, we discuss how observed state changes in the clustered ACS data are modeled as a Markov chain.

3.3.2 A Markov Model of Urban Change

Given our consideration of clusters as states of urban development, it is natural to regard these states as defining the state space of a Markov process. Assuming that the processes of urban development are probabilistic and further assuming that the probabilities behind these processes are fixed in time are major abstractions from reality. However, we regard these abstractions as valid in the context of largely unplanned, emergent urban phenomena such as gentrification and especially on the scale of a neighborhood or borough, which is large compared to our unit of analysis. We further believe that the advantages of our model, which include the representation of urban change in both time and space, the absence of any a-priori assumptions about urban dynamics, and the inclusion of socio-economic data, outweigh the cost of these abstractions.

3.3.3 Calculating Transition Probabilities

Since the clustering algorithm assigns census block groups five-year estimates to one of six states, one only needs to define the transition probabilities between these states to complete the Markov model. We derived these transition probabilities by counting the transitions from one state to another, and dividing them by the total number of transitions.
Using the method described above, we calculated transition probabilities for the four counties included in our analysis for the period of 2009-2012. (To conserve space, we only include values for Bronx County, see table 2). We calculated the transition probabilities for 2013 separately, in order to assess the predicative capacity of the four Markov models.

4. Results and Discussion

In the following discussion, we characterize the states we found in our cluster analysis both quantitatively and spatially. We also discuss the results and predictive capacity of the Markov models resulting from such analysis.

4.1 Characterizing States

We performed k-means clustering separately for each county (Bronx, Kings, New York, and Queens), and compared the content of these clusters or states statistically to determine whether the clustering method achieved significantly different states. We first identified features that were tightly distributed around the mean, showing strong clustering within states. These features were further tested using paired t-tests to determine whether the difference between means for each of these features differ significantly between states in each county.

4.1.1 Identifying Significant Features by Coefficient of Variation

A tight clustering around the mean of a feature, i.e. a standard deviation of less than 50% of its mean, or a coefficient of variation less than 0.5, suggests that a feature differs significantly between states and therefore is an important indicator of a state's composition.

Table 3. Mean, Standard Deviation, and Coefficient of Variation for low-dispersion features of a cluster from Bronx County.

Field	Mean	STD DEV	COEFF VAR
Male %	0.47	0.077	0.164
Avg Male Age	35.752	8.454	0.236
Female %	0.53	0.077	0.145
Avg Female Age	39.005	8.865	0.227
Family HH %	0.644	0.156	0.242
Nonfamily HH %	0.356	0.156	0.438
High School %	0.328	0.123	0.374
Some College %	0.148	0.068	0.462
Income < $35.000 %	0.445	0.216	0.486
$35.000-$75.000 %	0.284	0.127	0.446
Median year structure built	1941.457	140.94	0.073
HU Renter Occupied %	0.641	0.256	0.400
Average Rent	1131.239	410.261	0.363
Transportation %	0.651	0.203	0.312

We identified several significant features for each county using this method. (Table 3 displays features with a coefficient of variation less than 50% within a single cluster for Bronx County.) Five features displayed low variance, and thus high significance, across all counties and states. These features were:

- percentage of tracts reported as "family households"

- percentage of respondents walking, cycling or taking public transportation to work, or working at home
- education level reported as "high school" or below
- income reported below $30,000
- property value

4.1.2 Testing for Significance using Paired t-tests

We used a paired t-test to determine the statistical difference between means of these five features within clusters 1-6 for each county. Comparisons between six clusters in each county resulted in 20 comparisons for each county. Sample sizes of each feature ranged from 40 to 3002 observations. (As an example, see table 4 for t-test results for the percentage of reported family households in Bronx County.)

The majority of feature means varied significantly ($t > 2$) across states with each county. According to this finding, k-means successfully clustered the ACS block groups into statistically different groups, varying primarily by features describing household structure, transportation modes, education levels, household income and home value.

Of the five significant fields, the percentage of reported family households varied most significantly across clusters for each county. In other words, this feature showed the fewest number of insignificant comparisons between states. The finding suggests that of the four counties sampled, the percentage of reported family households varies the most both between the states within each county and across counties (reported family household percentages ranged from 25% to 76%). The percentage of family households in a region may be an important indicator of a region's current or future gentrification status. Another important factor appears to be transportation mode to work (car or public transport, walking or biking). Like percentage of reported family households, this factor appears to vary most both within and across counties (The range of reported transportation mode varies from 89.7% to 42%). People in New York County (Manhattan Island) reported the highest percentages of walking, cycling or taking public transportation to work, while those in the Bronx County reported the lowest. Accordingly, there may be a correlation between gentrification and proximity to jobs that enable walking or cycling, although this finding may also be due to the relative presence of biking infrastructure, access to public transportation, or other culturally mediated behaviors.

Table 4. Paired t-test results for the percentage of family households for clusters 1-6 (k1 – k6) within Bronx region. Highlighted comparisons are not significant.

% Family Households					
	k1	**k2**	**k3**	**k4**	**k5**
k1	3.5				
k2	8.84	-10.26			
k3	***1.37***	4.14	-7.07		
k4	-2.94	13.23	***1.73***	8.58	
k5	35.49	29.13	25.08	41.98	23.51

4.1.3 Understanding Complex Relationships

The value of the k-means clustering method lies in the ability to draw associations between patterns within states. For example in Bronx County, because state 1 differs significantly from state 4 for all five features, we can assume the relationships between these features are non-trivial. Therefore the mean values for education in state 1, (32% reported a high school education or below) are associated with mean values for income in state 1 (28% report income levels below $30,000), and that this relationship is distinctly different from that occurring between the same variables in state 4 (where larger percentages of both high school education and income are reported). The clustering method therefore provides simultaneous insight into relationships between several variables. Additionally, the Markov Model examines how these complex relationships evolve over time. The consistent appearance of five significant features across states and counties suggests that k-means clustering has captured at least some of the complexity of urban development. The spatial analysis provided in the following section further reinforces this impression.

4.2 Visualizing States

To study our clustering results qualitatively, we visualized the location of census tracts in terms of their state (cluster) identity. Initial observation reveals little variation between states, that is, the majority of census block groups rarely transition between states across the period studied (2009-2013, see figure 1). This lack of variation is to be expected given the lack of variation in the ACS census data, and the ACS estimation methodology (see section 3.1.1). However, the k-means clustering does result in visible spatial groupings that appear to correlate with neighborhood boundaries. These spatial groupings are notable considering that the input data used in the clustering did not contain any spatial indicators.

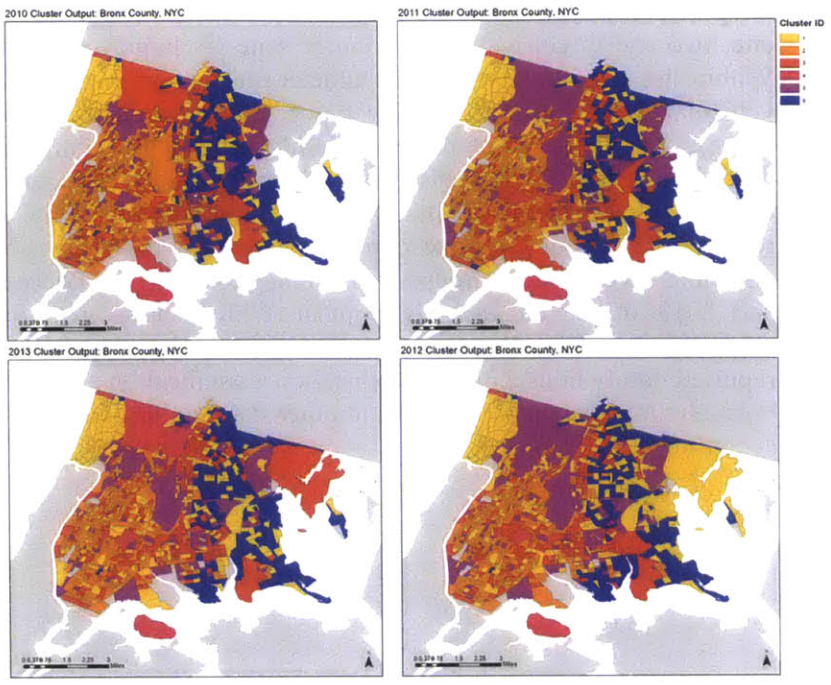

Figure 1. State visualizations of Bronx County from 2010-13 (clockwise from upper left).

For example, a visualization of states identified in Bronx County shows evident spatial clustering of states one (yellow) and six (blue). Examination of these states' composite ACS field data reveals differences between education attainment level and household composition. State six has 10% fewer family households, a 3% higher education attainment level, lower poverty levels (18% fewer residents report an income less than $35,000, newer housing stock (10%), Higher average rent (by an average of $246), and a lower percentage of residents that report non-vehicular travel to work (45% compared to 65%). Comparison of groups of census block groups falling within specific states to neighborhood boundaries in Bronx County shows that state one maps tightly to the affluent Riverdale neighborhood, while state six encompasses a belt of several neighborhoods across East Bronx.

Additionally, the clustering method identifies groups of census tracts with similar properties that may not be adjacent to each other. For example, both Co-op City (one of the largest cooperative housing developments in New York) and Kingsbridge (a westerly working class community) fall into state five (purple), but are separated geographically. State five is characterized by very low education attainment levels (18% Some College), an evenly split distribution of family and non-family households, lower rents on average ($939) and a low-income bracket, (47% report making less than $35,000 annually.).

Kings County (see figure 2) exhibits tight spatial clustering for states one, two and three, with dispersion of state six largely near coastlines. Within this county, state six is characterized by very high rates of non-vehicular transportation (walking or cycling to work or working from home), and low rates of reported family households, compared to the other five states. These areas may be highly affluent, particularly along coastline developments. Notably, one of these clustered state six areas represents a recent coastline condo-development in Williamsburg. We also note several block groups transitioning to state six in between 2010 and 2013 in the increasingly popular Williamsburg area. State one, mapping onto East Brooklyn, is characterized by a high instance of reported family households and education attainment levels, but low non-vehicular transport levels relative to other states in the county.

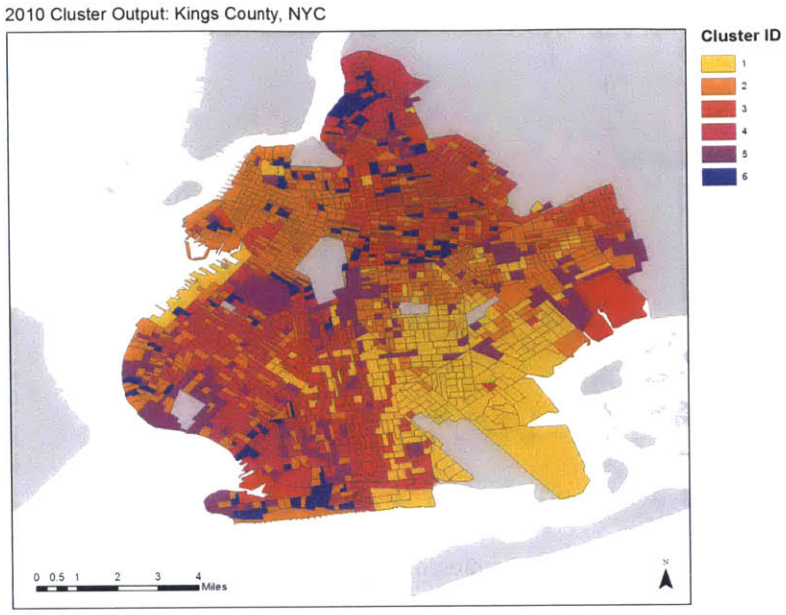

Figure 2. State visualization of Kings County for 2010.

New York County (see figure 3) shows three stable states over time, mapping onto upper (state one) and lower (state six) Manhattan and the areas surrounding central park (state five). State six, mapping largely onto lower Manhattan shows fewer family households and a larger share of the population reporting ownership of dwelling units, while state five is characterized by higher percentages of family households and renter occupancy.

Finally, Queens County (see figure 4) shows tight clustering of states four and five, which appear similar in composition except for one important factor: average value of owner-occupied units. Average home value in state four is significantly lower than state five by a difference of over $100,000 (401,253 vs. 556,677). Unlike previous counties, these states do not appear to map clearly onto neighborhood boundaries. Inquiry into the reason for these tight spatial clusters should be the subject of further research.

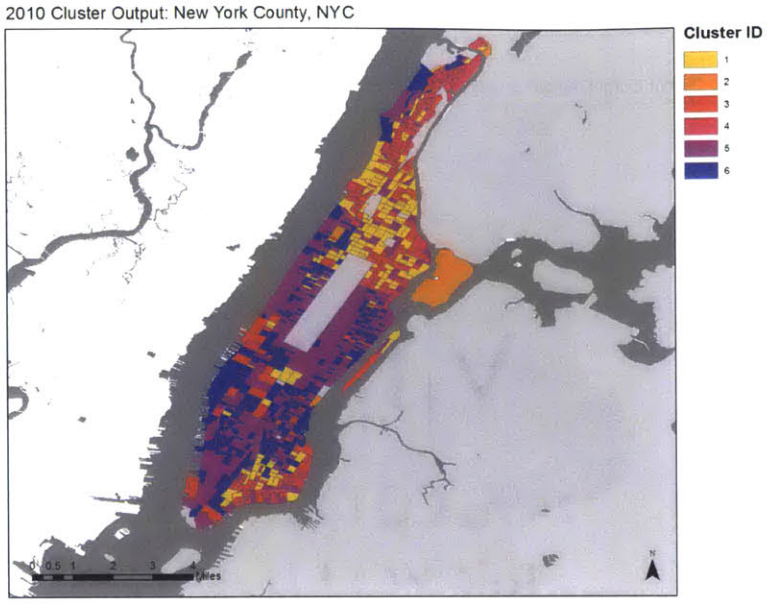

Figure 4. State visualization of New York County for 2010.

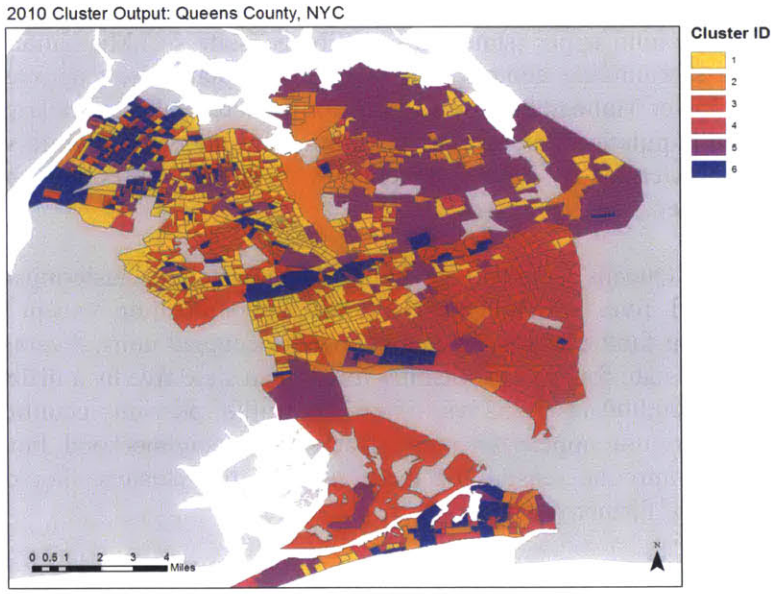

Figure 3. State visualization of Queens County for 2010.

4.3 Predicting State Transitions

The four Markov Models described in section 0 track an individual census block group's path through the state space in order to identify transition probabilities between states over time. These transition probabilities can then be applied to previously untested block data.

Specifically, we derived transition probabilities for 2009→2010, 2010→2011, and 2011→2012, and compared them with the transition probabilities for 2012→2013. (See table 5 for the transition probabilities from Bronx County. See table 6 for the average errors of predicted transition probabilities per county and state.)

For all four counties, percentages along the diagonal of the transition matrix (non-transitions) make up the largest percentage of observations, which is the probability of a census block not transitioning to a different state was greater than transitioning to a different state. The average probability in the period from 2009 to 2012 for a census block group to retain its state is 76%. This means that, statistically, during the transitions of 2009→2010, 2010→2011, and 2011→2012, 44% of census block groups retained their states. We attribute this finding to the lack of variation in the ACS census data, ACS sampling methodology, and the short time frame (2009-13) for which ACS data was available. Apart from this finding, transition probabilities vary markedly between the four counties. For example the transition probability from state six to state five are 1.86% for Bronx, 1.47% for Kings, 12.94% for New York, and 5.14% for Queens County. This suggests that the patterns through which block groups change may vary significantly across counties.

In order to assess the predictive capacity of the four Markov models, we compare the transitions probabilities computed for 2009→2012, with the probabilities for 2012→2013. These probabilities are remarkably similar for each county, with an average error of 5.9%. (See table 6 for the error in the predicted transition probabilities for each county and state.) This average includes outliers with errors of around 20%: these are due to very small cluster sizes leading to big variations of the transition probabilities. For example, the error for state four of Bronx County is 16.67%, corresponding to a cluster with only three to four members (see table 5.).

These outliers indicate that adapting our method of analysis to maintain relatively equal cluster sizes as an area for future research. Also, note that due to the probabilistic nature of Markov models, the predictions apply strictly to the county level, and not to individual census block groups. Allowing for more fine-grained predictions, perhaps including interactions between neighboring states, is another interesting direction for future research.

Table 5. Transitions probabilities for Bronx County from 2009 to 2012, and from 2012 to 2013. Starting states are according to row, and end states according to column.

2009→ 2012	**→ State 1**	**→ State 2**	**→ State 3**	**→ State 4**	**→ State 5**	**→ State 6**
State 1 →	93%	0%	0%	0%	2%	5%
State 2 →	0%	86%	9%	0%	5%	0%

State 3 →	0%	18%	71%	0%	7%	3%
State 4 →	0%	30%	10%	50%	10%	0%
State 5 →	2%	12%	12%	0%	71%	3%
State 6 →	1%	0%	8%	0%	2%	88%
2012→2013	**→ State 1**	**→ State 2**	**→ State 3**	**→ State 4**	**→ State 5**	**→ State 6**
State 1 →	90%	0%	5%	0%	5%	0%
State 2 →	0%	89%	6%	0%	4%	0%
State 3 →	0%	9%	88%	0%	2%	1%
State 4 →	0%	0%	0%	100%	0%	0%
State 5 →	2%	6%	5%	0%	86%	2%
State 6 →	0%	0%	4%	0%	0%	96%

Table 6. Predictive average error ε per county and state, comparing transition probabilities from 2009-2012 with 2012-13

County	Total ε	ε 1	ε 2	ε 3	ε 4	ε 5	ε 6
Bronx	6%	3%	1%	6%	17%	5%	3%
Kings	5%	19%	1%	1%	1%	1%	3%
New York	9%	3%	17%	26%	1%	3%	2%
Queens	5%	1%	17%	1%	2%	5%	3%

5. Conclusion

The research presented here proposes modeling neighborhood change as a transition between meaningful states that emerge empirically from socio-demographic datasets. We identify states as profiles of complex relationships between socio-economic and demographic factors that may not be clear to the researcher a priori, and may not be identifiable through traditional forms of linear regression analysis.

The k-means clustering method proves to be a successful methodology for identifying these states, i.e., patterns of urban development. States emerging from ACS data map consistently to neighborhood boundaries, and enable the spatial comparison of neighborhood areas exhibiting similar socio-economic and demographic properties. The degree to which census block groups are spatially grouped into state categories may be an indicator of relative levels of socio-economic segregation; i.e., areas where census block groups fall into a single state represent pockets of homogenous characteristics, while regions with a patchwork distribution of states show variation of socio-economic conditions. Because state identification is not dependent on spatial information, other

regions, which may not fall into traditional neighborhood boundaries, but exhibit homogenous data characteristics, are identifiable. Such visualizations enable researchers to visualize emergent trends in census data without restricting their analysis to existing neighborhood boundaries. The tests of the predictive capacity of the Markov models for the four counties show the promise of our method as a tool for planning agencies to model urban changes in a metropolitan area, and as an opportunity to refine the approach of planning policy by targeting symptoms of gentrification in support of negatively impacted communities.

Data collected over a longer period, and using a different sampling method than that performed by ACS would significantly enhance the results presented here. Future applications of this research include investigating the application of the method to larger and more robust data sets, as well as different regions and context. Another ambition is the development of better and more fine-grained predictions, which could possibly achieved by including interactions between neighboring states. We observe that machine learning and other pattern-recognition techniques host a wealth of possible applications for model development in urban analytics. Machine learning methods are able to handle large and complex data sets, such as those that characterize urban environments. However, it is important to acknowledge that such an approach depends heavily on data quality and responsible algorithm development. Furthermore, evidence-based planning must always be supplemented by qualitative observation.

Modeling gentrification as a probabilistic process of state changes in time and space provides insights into the dynamic nature of this complex phenomenon. While mathematical models may be unable to account for many of the social and cultural intimacies of a particular site, they can generate more refined research questions for this important driver of urban regeneration. However, the research does not present a complete model of gentrification. Rather, our method allows the empirical study of neighborhood-level urban development by condensing complex urban data into latent profiles that both describe and predict urban change.

References

Barton, M. (2014) An exploration of the importance of the strategy used to identify gentrification. Urban Studies. http://usj.sagepub.com/content/early/2014/12/03/0042098014561723.abstract. Accessed May 2015.

Beauregard, R. "Trajectories of Neighborhood Change: The Case of Gentrification." Environment and Planning 22 (n.d.): 855–74.

Birch, David. "Toward a Stage Theory of Urban Growth." Journal of the American Institute of Planners 37, no. 2 (1971): 78–87.

Bryson, J. (2013) The Nature of Gentrification. Geography Compass, 7(8), 578-587.

Carbonell, Jaime. "Machine Learning: A Historical and Methodological Analysis." AI Magazine 4, no. 3 (1983).

Clay, P. (1990). Choosing Urban Futures: The Transformation of American Cities. Stanford Law and Policy Review, 1(1), 28-39.

Crang, Mike. "SENTIENT CITIES Ambient Intelligence and the Politics of Urban Space." Information, Communication & Society 10, no. 6 (n.d.).

Duany, Andres. "Three Cheers for Gentrification." American Enterprise, 2001.

Durrett, R. (2010). Probability: Theory and Examples, 4th ed. Cambridge, UK: Cambridge University Press.

Freeman, Lance. "Displacement or Succession? Residential Mobility in Gentrifying Neighborhoods." Urban Affairs Review 40, no. 4 (March 2005): 463–91.

Freeman, Lance, and Frank Braconi. "Gentrification and Displacement: New York City in the 1990s." Journal of the American Planning Association 70, no. 1 (2004).

Glaeser, Edward. "Consumer City." Harvard Institute of Economic Research, June 2000.

Glaeser, Edward. "Housing Booms and City Centers." National Bureau of Economic Research: Working Paper Series, no. 17914 (March 2012).

"Housing Dynamics." Harvard Institute of Economic Research, March 2007.

Graham, Stephen. "Software-Sorted Geographies." Progress in Human Geography 29, no. 5 (2005): pp. 562–80.

Guerrieri, Veronica. "Endogenous Gentrification and Housing Price Dynamics." University of Chicago: Working Papers, 2011.

Hidalgo, Cesar. "Streetscore - Predicting the Percieved Safety of One Million Streetscapes," n.d.

Hidalgo, Cesar, and Edward Glaeser. "Do People Shape Cities or Do Cities Shape People? The Co-Evolution of Physical, Social, and Economic Change in Five Major U.S. Cities." National Bureau of Economic Research: Working Paper Series, October 2015.

Kolko, Jed. "The Determinants of Gentrification." Public Policy Institute of California, n.d.

Lees, Loretta. "Rethinking Gentrification: Beyond Positions of Economics or Culture." Progress in Human Geography 18, no. 2 (1994): 137–50.

Lees, L., Slater, T., & Wyly E. K. (2008), Gentrification. Growth and Change, 39(3), 536-539.

Lees, L., Slater T., & Wyly, E. K. (Eds.) (2010). The Gentrification Reader. London, UK: Routledge.

Ley, David. "Alternative Explanations for Inner-City Gentrification: A Canadian Assessment." Annals of the Association of American Geographers 76, no. 4 (n.d.): 521–35.

Light, Jennifer. "Discriminating Appraisals: Cartography, Computation and Access to Federal Mortgage Insurance in the 1930s." Technology and Culture 52, no. 3 (July 2011): 485–522.

Maciag, Mike (2015). Gentrification in America. Governing DATA, http://www.governing.com/gov-data/census/gentrification-in-cities-governing-report.html. Accessed May 2015.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297.

Marcuse, Peter. "Gentrification, Abandonment, and Displacement: Connections, Causes, and Policy Responses in New York City." Urban Law Annual ; Journal of Urban and Contemporary Law 28 (1985): 195–240.

Mitchell, Tom. Machine Learning: A Guide to Current Research. Kluwer

Academic Publishers, 1986.

O'Sullivan, D. (2002) Toward Micro-scale Spatial Modeling of Gentrification. Journal of Geographical Systems, 4, 251–274.

Oswalt, P., Overmeyer, K. & Misselwitz P. (2013) Urban Catalyst: The Power of Temporary Use. Berlin, DE: Dom.
Porta, Sergio. "The Form of Gentrification." University of Strathclyde Glascow: Working Papers, 2014.

Portugali, Juval. "Complexity Theories of Cities: Achievements, Criticisms, Potential." In Complexity, Cognition and the City. Springer, 2010.
———. "Complexity Theory as a Link Between Space and Place." Environment and Planning A 38 (2006): 647–64.

Rabiner, L. (1989). Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Proceedings of the IEEE, 77(2), 257-286.
Robinson, Ira. "A Simulation Model For Renewal Programming." JAPA 31, no. 2 (1965): 126–34.

Rousseeuw, Peter. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." Journal of Computational and Applied Mathematics 20 (1987): 53–65.

Samuel, Arthur. "Some Studies in Machine Learning Using the Game of Checkers." IBM Journal 3, no. 3 (July 1969).

Shmueli, Galit. "To Explain or Predict?" Statistical Science 25, no. 3 (2010): 289–310.

Slater, Tom. "Planetary Rent Gaps." Antipode, 2014.
———. "The Eviction of Critical Perspectives from Gentrification Research." Urban Journal of Urban and Regional Research 30, no. 4 (2006): 737–57.

Smith, Neil. "Toward a Theory of Gentrification A Back to the City Movement by Capital, Not People Neil Smith." Journal of the American Planning Association 45, no. 4 (1979): 538–48.

Smith, N. (2002). New globalism, new urbanism: gentrification as global urban strategy. Antipode 34(3), 428-450.

Snow, C.P. The Two Cultures. 1st ed. Cambridge University Press, 1959.

Sugar, C. A. & James, G. M. (2003). Finding the number of clusters in a data set: An information theoretic approach. Journal of the American Statistical Association, 98(1), 750–763.

Thompson W. E. & Hickey J. V. (2004). Society in Focus: An Introduction to Sociology, 5th Ed. Boston, MA: Allyn & Bacon

Torrens, P. M. & Nara A. (2007). Modelling Gentrification Dynamics: A Hybrid Approach. Computers, Environment and Urban Systems, 31(3), 337-361.

Tweedie S. P. & Meyn R. L. (2009). Markov Chains and Stochastic Stability. Cambridge, UK: Cambridge University Press.

U.S. Census Bureau (2008). A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know. Washington, DC:
Government Printing Office.

Vernon, Raymond. The Changing Economic Function of the Central City. The Area Development Committee of CED, 1959.

Wacquant, Loic, and Tom Slater. "Territorial Stigmatization in Action." Environment and Planning A 46 (2014): 1270–80.

White, Kirk. "Who Gentrifies Low-Income Neighborhoods?" National Bureau of Economic Research: Working Paper Series, May 2008.

Wright, Melissa. "The Gender, Place and Culture Jan Monk Distinguished Annual Lecture: Gentrification, Assassination and Forgetting in Mexico: A Feminist Marxist Tale." Gender, Place and Culture 21, no. 1 (2014): 1–16.

Wyly, Elvin. "The Right To Stay Put, Revisited: Gentrification and Resistance to Displacement in New York City." Urban Studies 43, no. 1 (January 2006): 23–57.

Appendix

5.1 Preprocessing for Clustering

Before submitting the data to the clustering algorithm, we took several steps to increase its suitability for clustering. To reduce the dimensionality of the data, improve clustering speed and intelligibility we turned 108 fields from the ACS data into 32 features that were more closely associated with gentrification symptoms observed in the literature. Census block groups without population were discarded, assuming that uninhabited areas such as industrial compounds and natural reserves display patterns of development that are different from inhabited and urbanized areas.

The pre-preprocessing steps of the fields involved converting some of the fields into percentages, consolidating several fields into a single feature, and calculating a weighted average from several other fields. We took four features directly from the ACS data (total population, number of households, number of housing units, and the median year of structures built).

To convert fields into percentages, we divided the value of more specific fields, such as the number of vacant housing units, by an appropriate more general value (in this case the number of all housing units). Such percentages are more suitable for clustering since they allow a better comparison of relative values. We included absolute values, for example the total number of housing units, as separate features. Other fields denoting specific categories or brackets were summed together, and the result converted into a percentage. For example, we simplified the sixteen categories of household income into five features (based on the definition of middle class by Thompson and Hickey, 2004). In the same manner, we summed and converted 23 fields into 18 additional features.

Five other fields were consolidated by converting brackets or categories into a weighted average. For this calculation we assumed a hypothetical average value for each age bracket as the mean value of the bounds of the bracket, and calculated an overall, weighted average based on the

brackets' sizes. For example, we assumed that the average age of the men in the 18 to 24 year age bracket is 21.5 (since the next bracket starts at 25) and included this value in the overall average, weighted according to the number of men in this age bracket. This method was also used to calculate the weighted average of housing units per building, the weighted average of owner-occupied housing units, and the weighted average cash rent of renter-occupied housing units. Note that this technique requires the assumption of an upper bound for the highest open-ended bracket, which includes values such as the number of men of "85 years and above", or the number of units with a rent of "2000 USD or more". (See table 2 for the fields we employed to calculate the weighted averages and the hypothetical average values.)

Table 2. Features from ACS fields converted to weighted averages

New Feature	Original ACS Field(s)	Upper Bound
Average Male Age	SE_T005_003 - SE_T005_014	92.5 Years
Average Female Age	SE_T005_016 - SE_T005_027	92.5 Years
Average Housing Units	SE_T097_002 - SE_T097_010	75 Units
Average Value For Owner-occupied housing units	SE_T100_002 - SE_T100_010	1.500.000 USD
Average Rent for Renter-occupied housing units	SE_T102_002 - SE_T102_012	3500 USD

Finally, we normalized the values for every feature to be between zero and one to ensure an equal weightage in terms of the clustering algorithm. In other words, we created a broad selection of potentially relevant features, and refrained from a-priori assessing the relative importance of these features. The various pre-processing steps described above yielded 29.058 observations with 32 features for inclusion in the clustering.

6 Results and Discussion

Urban change and gentrification are dynamic, complex events. As such the results described here scratch at the surface of a challenging set of patterns and dynamics. Our results demonstrate that there is structure in the ACS data, and that structure varies predictably over time. The value of the k-means clustering method lies in the ability to draw associations between patterns within states, while the Markov Model examines how these complex relationships evolve over time. Overall, the five indicators selected from a literature review of gentrification show distinct spatial clustering patterns across the five boroughs. In many cases the extreme states (those with very high or very low values for the six features) exhibit few transitions, while states with more intermediary values are more likely to transition. We identify spatial clustering of States, as well as levels of dispersion. As input data into the k-means clustering included no spatially identifying features, it is remarkable that the algorithm was able to reproduce neighborhood boundaries. In some cases, States were distinctly spread out through a borough, not clustering spatially. These states typically represented socioeconomic groups that were subject to "churn," often occurring on the fringes of neighborhood boundaries. Different boroughs display different "personalities" of spatial state clustering and transitions. For example, Kings County shows very tight spatial clustering of States, with transitions occurring on the fringes

between clusters. Alternatively, other boroughs like the Bronx showed more fluid transitions at the interiors (away from coast lines). Indeed, State transitions also correlated to physical features such as high ways or public transportation nodes. Individual census blocks also exhibited "State Paths" the sequence of states through which an individual Census block passed between 2009 and 2013. State paths showed shared sequential patterns in individual boroughs, further pointing to structure in how census blocks experience urban change over time. This is largely a condition of punctuated stability, as there are very few "complex" transitions that are observed. The following spatial analysis in this section further reinforces these impressions.

6.1 Visualizing States

To study clustering results qualitatively, locations of census block groups in terms of their State (cluster) identity were visualized. Initial observation reveals little variation between States, that is, the majority of census block groups rarely transition between states across the period studied (2009-2013, see figure 1). This lack of variation is to be expected given the lack of variation in the ACS census data, and the ACS estimation methodology. However, the k-means clustering does result in visible spatial groupings that appear to correlate with neighborhood boundaries---in some cases transcending them. States that "transcend boundaries" or do not cluster spatially, are also those that appear to demonstrate the most upward

mobility. That is, they are highly educated, high-income, non-family households. Associating transition probabilities with State identities is a key contribution of this model towards gaining insight into spatial dynamics. These spatial groupings are also notable considering that the input data used in the clustering did not contain any spatial indicators.

6.1.1 Bronx County

A visualization of states identified in Bronx County shows evident spatial clustering of all three states. Examination of these states' composite ACS field data reveals differences between education attainment level and household composition. State 2 (64% white, with an average income of $85,235 and education index of 3.04) appears to cluster spatially to the northwestern region and southeastern regions of the Bronx. Very few transitions are visible in areas where this State clusters. The affluent North Riverdale and Riverdale neighborhoods map tightly onto State 2. Alternatively, State 1 (minority 15% white, low-income $48,594, less-educated 2.13) appears to bleed into other neighborhood areas, and clusters near industrial sites or infrastructure. Additionally, the clustering method identifies groups of census tracts with similar properties that may not be adjacent to each other. For example, both Co-op City (one of the largest cooperative housing developments in New York) and Kingsbridge (a westerly working class community) fall into State 3, but are separated geographically. State 3, a state that does not exhibit distinct clustering, also appears to transition to State 1, in the lower Bronx representing a

transition in this region to demographics with fewer families and higher income levels, a possible indicator of gentrification.

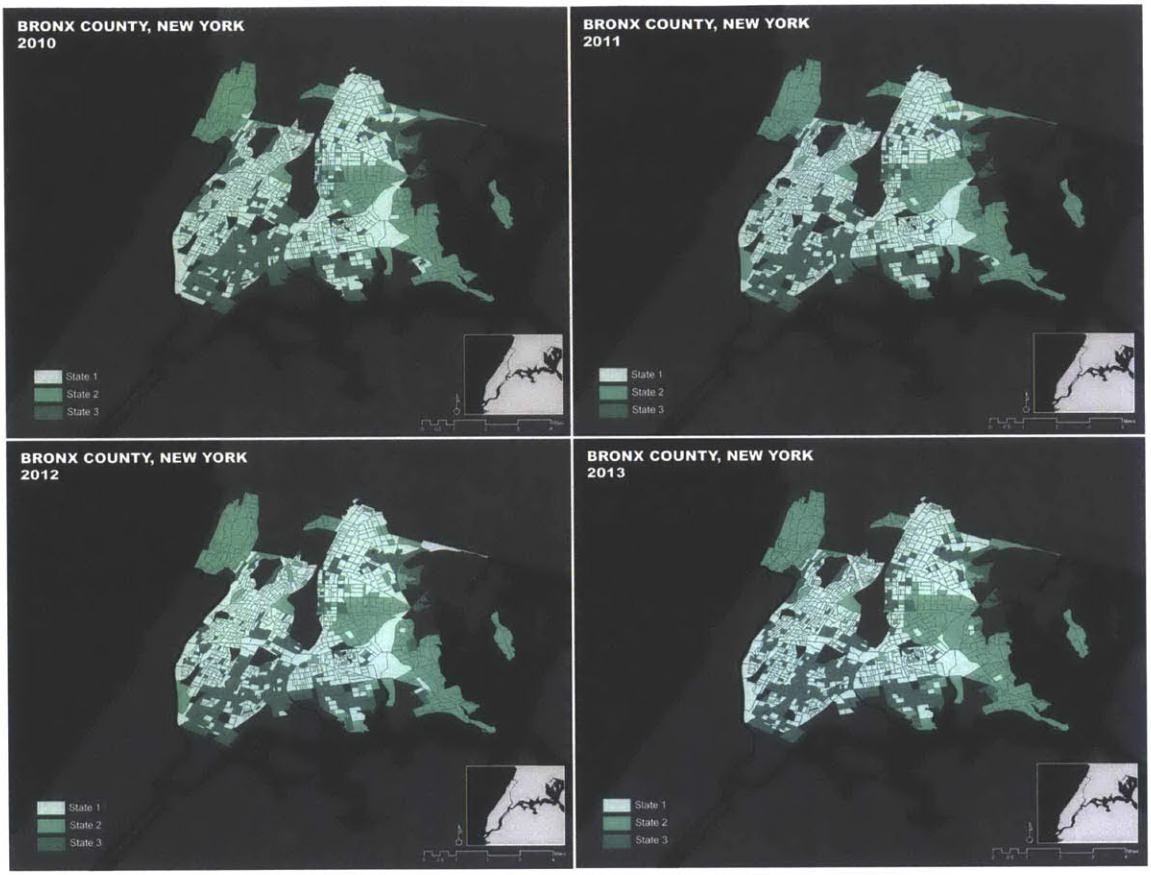

Figure 15 State Clusterings for Bronx County NYC, years 2010 (top left) to 2013 (bottom right).

Figure 16 Street View of Riverdale, Bronx. A visibly highly affluent, less dense, single-family neighborhood, characterized as strongly as State 2.

Figure 17 Street View of Kingsbridge neighborhood in the Bronx. Kingsbridge is characterized as State 3, along with Co-op City (below).

Figure 18 Street View of Co-op City, Bronx. Typology and building stock similarities are visible. Despite being in very different regions of the Bronx, both Co-op City and Kingsbridge were characterized as State 3.

Figure 19 Streetview of Melrose, Bronx. Melrose is classified as State 3, representing a low-income, uneducated demographic in the Lower Bronx. State 3 is susceptible to transition to State 1, characterized by higher incomes and education levels.

6.1.2 Kings County

Kings County (see figure 2) exhibits tight spatial clustering of three states, with most variation occurring on the fringes of boundaries between spatial clusters. State 1 (76% white, non-family, high education and income levels in aged building stock) maps largely onto coastlines, while State 2 (family households, moderate income and education levels) maps heavily onto interior, high-density parcels. State 3 primarily occurs in the far Eastside and lower portion of Kings County characterized by low income and education levels, and a high percentage of family households. Visually, the majority of block groups appear to be classified as State 2, occupying the interior of Kings. Visually, Kings appears to be in a state of tension between these states, with a large amount of transitioning activity occurring on the fringes between state clusters. The majority of Brooklyn falls into State 1 and 2 categories. Redhook, Gowanus and Park Slope, all appear to be classified by the affluent State 1. Many of the State 1 clusters represent recent coastline condo-developments in Williamsburg. However, Williamsburg itself appears to be more diverse, with Census blocks falling into all state categories. West 37th Street appears to directly split States 1 and 3 (the most and least affluent states), at Seagate a wealthy, gated community. "Seagateness" can also be observed in Bay Ridge, Carrol Gardens and Greenpoint Neighborhoods, as they all group to State 1. Alternatively, State 3 maps onto Canarsie, East New York and Brighton

Beach and areas less accessible through public transportation. One particularly notable pocket of census blocks classifying as State 3, is surrounded on all sides by State 1. Taking a second look on Street View, we observe a striking image of gentrification (see Figure 21). This points to the potential specificity of our model in identifying spatial transitions as indicators of gentrifying areas.

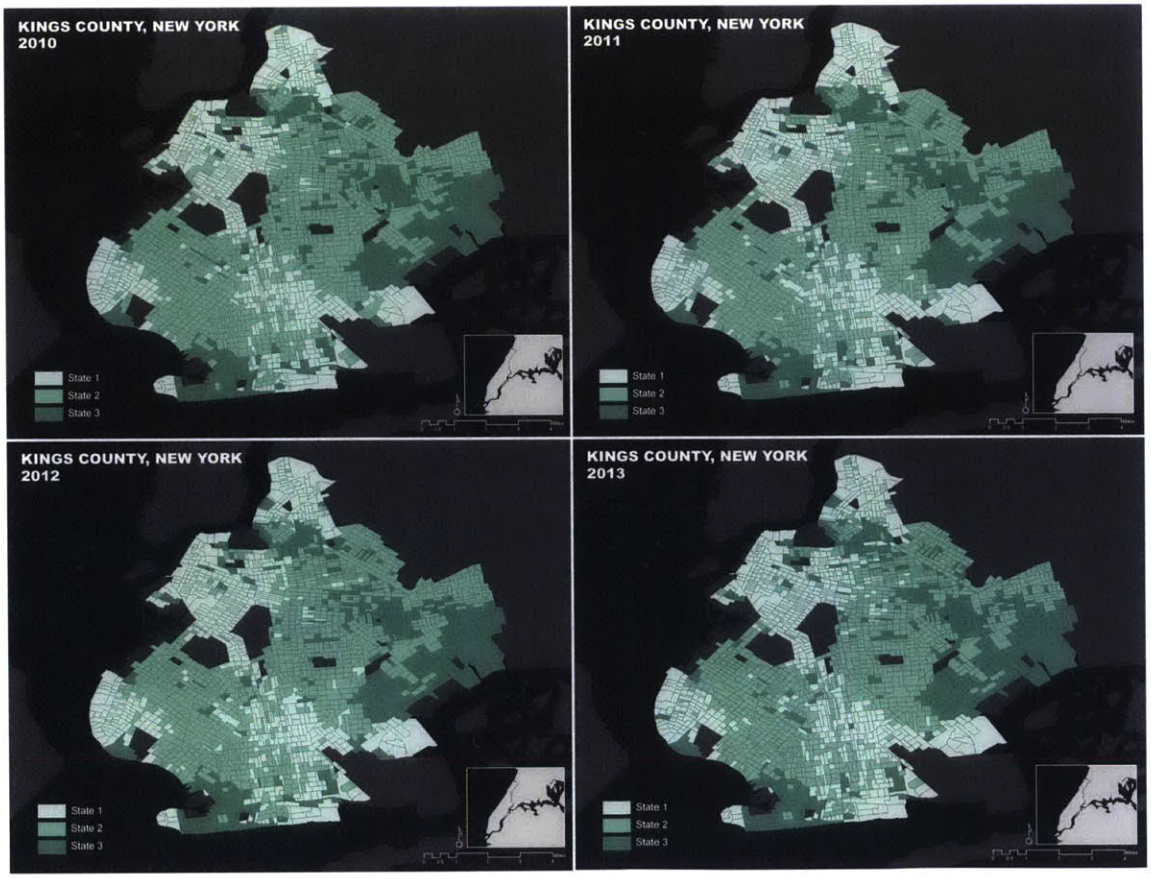

Figure 20 State Clusters for Kings County NYC, years 2010 (top left) to 2013 (bottom right).

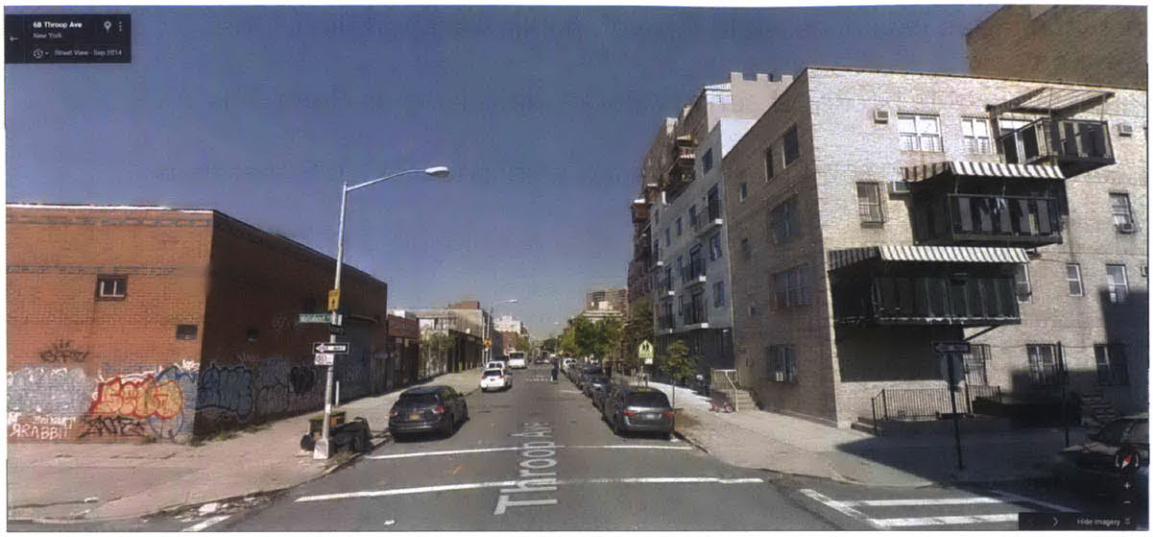

Figure 21 Street View of Flushing and Broadway Triangle—visibly gentrifying from a formerly industrial landscape (left) to a high income residential compound (right).

Figure 22 Carrol Gardens, an instance of State 1, characterized by high income and education levels.

Figure 23 Street View of Brighton Beach, an instance of State 3, characterized by low education and income levels.

6.1.3 New York County

New York County census blocks fall into four State Categories, clustering tightly to Upper Manhattan, the Upper East and West Sides, East Village and SoHo. State 1 maps onto some of the most affluent areas of Manhattan, typically bordering coastlines or Central Park. State 2 (35% white, high percent family levels and incomes nearly less than half of the other States) maps onto Two Bridges, Harlem and Washington Heights. Some portions of these areas are visibly transitioning from State 2 to the other more "affluent" states, as particularly evident near Two Bridges. State 1, notably with the highest percentage of Caucasians, and second highest income level, maps onto census blocks surrounding Central Park and Hudson Heights. This State also appears to be the most likely to

"consume" other neighboring State identities, and is more volatile. This may be due to the fact that of the three "affluent" states, State 1 has the lowest percentage of family households, indicated a more mobile upper-class condition. States 3 and 4 often map onto large business districts.

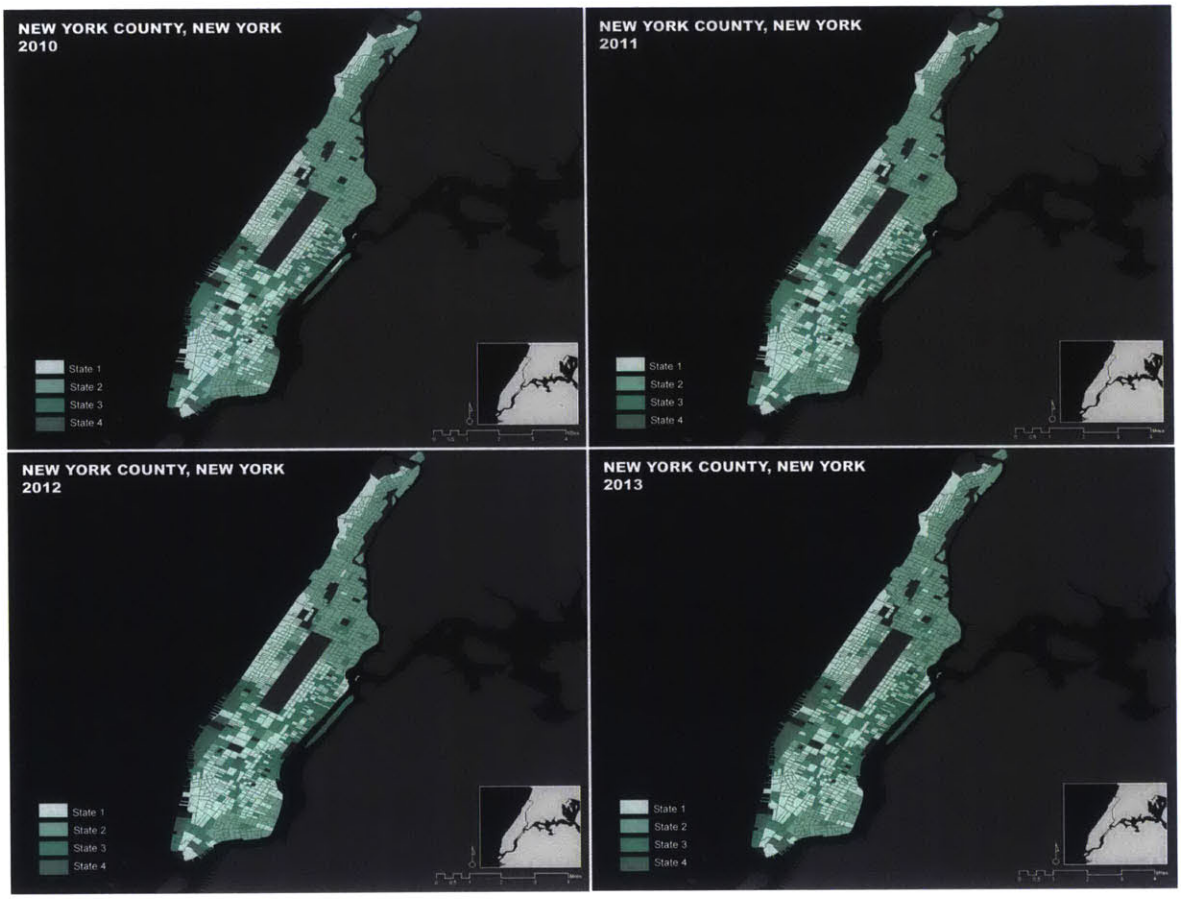

Figure 24 State classifications visualized for New York County, NY for years 2010-13.

Figure 25 Street View of Harlem, classified as State 2.

Figure 26 Street View of Hudson Heights, classified as State 1.

Figure 27 Street View of Upper West Side, classified as State 1.

6.1.4 Queens County

The census blocks of Queens County fall into four States. States 1, 2 and four exhibit tight spatial clustering while State 3's spatial distribution is more scattered. State 1 has a 27% non-minority rate, the second highest rate of family households (79%) and relatively high income levels ($70,849) compared to the other States. This State maps onto census blocks in Ozone Park and Jackson Heights. Fresh Meadows and Murray Hill map strongly onto State 4, the most affluent state characterized by the lowest percentage of minorities, high income and high education levels. State 2 maps primarily onto Astoria, and neighborhoods bounding major thoroughfares. State 2 classifies older building stock, medium density levels, low percentages of family households and moderate incomes. State two also represents a high rate of non-minorities. As it represents

minorities living in affordable, but not impoverished neighborhoods near transportation access, it may be a reasonable indicator of gentrifying areas in Queens County. Indeed, the Astoria neighborhood has recently been identified as an emerging hotbed for young professionals (Lu, 2015).

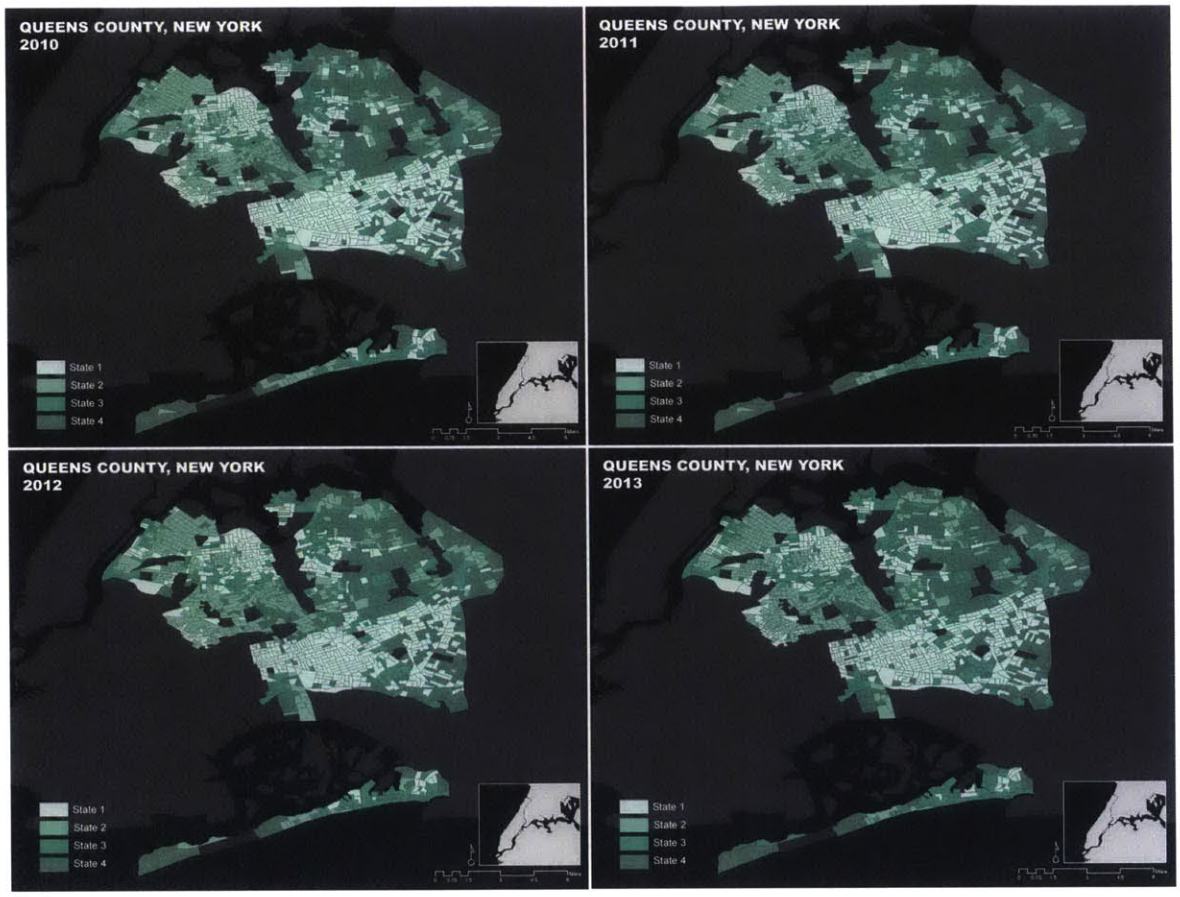

Figure 28 State classifications visualized for Queens County, NY for years 2010-13.

Figure 29 Street View of Astoria, classified as State 2, and a possibly gentrifying neighborhood.

Figure 30 Fresh Meadows, classified as State 4.

6.2 Identifying & Predicting State Transitions

The four Markov Models described in Chapter 4 track an individual census block group's path through the state space in order to identify transition probabilities between states over time. These transition probabilities can then be applied to previously untested block data. Specifically, transition probabilities were derived for 2009-10, 2010-11, and 2011-12. These were compared with the transition probabilities for 2012-13.

For all four counties, percentages along the diagonal of the transition matrix (non-transitions) make up the largest percentage of observations, which is the probability of a census block not transitioning to a different state was greater than transitioning to a different state. The average probability in the period from 2009 to 2012 for a census block group to retain its state is 89% (Bronx), 91% (Kings), 85% (Queens), 78% (New York). We attribute this finding to the lack of variation in the ACS census data, ACS "oversampling" methodology, and the short time frame (2009-13) for which ACS data was available. Apart from this finding, transition probabilities vary markedly between the four counties. Transition probabilities for the four counties are reported in Tables 11 through 14 as both summary and yearly tables. This suggests that the patterns through which block groups change may vary significantly across counties.

Table 11 Transition Probabilities for State transitions (represented here as States "1 to 1" etc.), by above) overall transition probabilities for all four and below) year from 2009-2013 for Bronx County. Transition probabilities are highlighted according to relative size, where large probabilities are green and low probabilities are red. Transitions from 1-2 are second most likely after the diagonal.

	From 1	From 2	From 3
To 1	0.92	0.06	0.02
To 2	0.17	0.82	0.01
To 3	0.05	0.02	0.92

	2009-10	2010-11	2011-12	2012-13
1 to 1	0.84	0.94	0.94	0.93
1 to 2	0.13	0.05	0.04	0.05
1 to 3	0.03	0.01	0.02	0.02
2 to 1	0.30	0.16	0.12	0.14
2 to 2	0.69	0.83	0.86	0.85
2 to 3	0.01	0.01	0.02	0.02
3 to 1	0.08	0.03	0.04	0.07
3 to 2	0.05	0.02	0.01	0.02
3 to 3	0.86	0.94	0.95	0.91

Table 12 Transition Probabilities for State transitions (represented here as States "1 to 1" etc.), by above) total transition probabilities for all four and below) year from 2009-2013 for Kings County. Transition probabilities are highlighted according to relative size, where large probabilities are green and low probabilities are red. Transitions from 2-3 are second most likely after the diagonal. 5% of tracts transition from State 1 to 2.

	From 1	From 2	From 3
To 1	0.93	0.06	0.01
To 2	0.05	0.92	0.04
To 3	0.02	0.11	0.87

	2009-10	2010-11	2011-12	2012-13
1 to 1	0.91	0.94	0.94	0.92
1 to 2	0.07	0.05	0.05	0.06
1 to 3	0.02	0.01	0.01	0.01
2 to 1	0.05	0.04	0.05	0.05
2 to 2	0.90	0.93	0.93	0.92
2 to 3	0.05	0.03	0.03	0.04
3 to 1	0.02	0.01	0.03	0.03
3 to 2	0.22	0.08	0.10	0.05
3 to 3	0.76	0.91	0.88	0.92

Table 13 Transition Probabilities for State transitions (represented here as States "1 to 1" etc.), by above) total transition probabilities for all four and below) year from 2009-2013 for Queens County. Transition probabilities are highlighted according to relative size, where large probabilities are green and low probabilities are red. Transitions between State 1 and other states are more likely than the majority of other transitions not along the diagonal.

	From 1	From 2	From 3	From 4
To 1	0.86	0.03	0.04	0.06
To 2	0.04	0.87	0.03	0.06
To 3	0.08	0.05	0.84	0.02
To 4	0.09	0.07	0.02	0.82

	2009-10	2010-11	2011-12	2012-13
1 to 1	0.86	0.87	0.85	0.88
1 to 2	0.05	0.02	0.05	0.02
1 to 3	0.02	0.04	0.04	0.05
1 to 4	0.07	0.08	0.06	0.06
2 to 1	0.07	0.04	0.03	0.04
2 to 2	0.79	0.87	0.91	0.89
2 to 3	0.04	0.03	0.03	0.01
2 to 4	0.11	0.06	0.03	0.06
3 to 1	0.16	0.06	0.06	0.08
3 to 2	0.11	0.07	0.03	0.03
3 to 3	0.65	0.86	0.90	0.88
3 to 4	0.08	0.01	0.01	0.01
4 to 1	0.12	0.07	0.07	0.10
4 to 2	0.10	0.07	0.08	0.06
4 to 3	0.01	0.01	0.01	0.03
4 to 4	0.76	0.85	0.85	0.81

Table 14 Transition Probabilities for State transitions (represented here as States "1 to 1" etc.), by above) total transition probabilities for all four and below) year from 2009-2013 for New York County. Transition probabilities are highlighted according to relative size, where large probabilities are green and low probabilities are red. Transitions between State 3 and 4 are likely, as well as between 1 and 4.

	From 1	From 2	From 3	From 4
To 1	0.93	0.02	0.05	0.00
To 2	0.02	0.96	0.02	0.00
To 3	0.08	0.01	0.89	0.01
To 4	0.26	0.08	0.32	0.35

	2009-10	2010-11	2011-12	2012-13
1 to 1	0.83	0.94	0.96	0.94
1 to 2	0.06	0.01	0.01	0.02
1 to 3	0.10	0.05	0.03	0.04
1 to 4	0.01	0.00	0.00	0.00
2 to 1	0.04	0.01	0.02	0.02
2 to 2	0.95	0.97	0.95	0.97
2 to 3	0.01	0.02	0.02	0.01
2 to 4	0.00	0.00	0.00	0.00
3 to 1	0.32	0.06	0.05	0.05
3 to 2	0.03	0.01	0.00	0.02
3 to 3	0.62	0.92	0.93	0.91
3 to 4	0.03	0.00	0.01	0.01
4 to 1	0.37	0.00	0.12	0.00
4 to 2	0.11	0.00	0.06	0.00
4 to 3	0.46	0.00	0.06	0.07
4 to 4	0.06	1.00	0.76	0.93

In general, transition probabilities reflect a majority of non-transitions. However, all possible transitions are not observed in equal quantities. Because the variation is reduced by the ACS sampling method, in addition to the clustering algorithm itself, the effects of these state transitions may be underestimated. Transitions occurring from States 1 to 2 are second

most likely after the diagonal in the Bronx. Since State 1 is characterized by low-income minorities, while State 2 is representative of an upwardly mobile, white, high-income earning class, regions of "churn" become visible in the Bronx. In Kings, transitions from States 2 to 3 are second most likely after the diagonal, and 5% of tracts transition from State 1 to 2. Once again, this reflects churn in Kings County, as the income, education, family household rates and races are dramatically different between States 1 and 2. States 2 and 3 are both minority, low-income and education States, however State 3 represents more recent building stock. This points to possible construction areas where more recent building stock raises the neighborhood average. Transitions between State 1 and other states are more likely than the majority of other transitions not along the diagonal in Queens. State 1 in Queens has a 27% non-minority rate, has the second highest rate of family households (79%) and relatively high income levels ($70,849) compared to the other States. This suggests a middle-income minority group that may be an indicator of transitioning areas in Queens. For New York, transitions between State 3 and 4 are likely, as well as between 1 and 4. State 4 is primarily characteristic of high-density, minority areas and may reflect reservoirs of displacement for development activity in New York County.

In order to assess the predictive capacity of the four Markov models, we compare observed transition probabilities computed for 2009, 2010, 2011 and 2012, with the observed probabilities for 2013. This provides a rough indicator of the extent to which the observed probabilities of 2013 could be predicted by the model. These probabilities are relatively high for each borough, with average prediction rates of 79% (Bronx), 63% (Kings), 79% (Queens), and 83% (New York), resulting in an overall prediction rate of 76% accuracy. Note that due to the probabilistic nature of Markov models, the predictions apply strictly to the county level, and not to individual census block groups. Allowing for more fine-grained predictions, perhaps including interactions between neighboring states, is another interesting direction for future research.

6.3 State Paths

Block groups were analyzed for path dependency, i.e., to determine the extent to which certain state paths were more likely than others. 1,100 State path combinations were possible. However, the majority of paths occurred once or not at all. The following charts display the number of census block groups falling into each state path category, where paths were observed in 5 or more cases. State paths show a tendency for a single transition, either preceding or following a series of stable states. Further examination of these patterns may provide additional insight into

the dynamics of urban change, where we can identify precisely what types of transitions are more likely under specific circumstances. In the following charts, the number of the state a census block group falls in for each of the 5 sampled years notes state paths. For example, "22211" refers to a census block group that stayed in State 2 for 2009, 2010 and 2011, and transitioned to State 1 for 2012 and 2013.

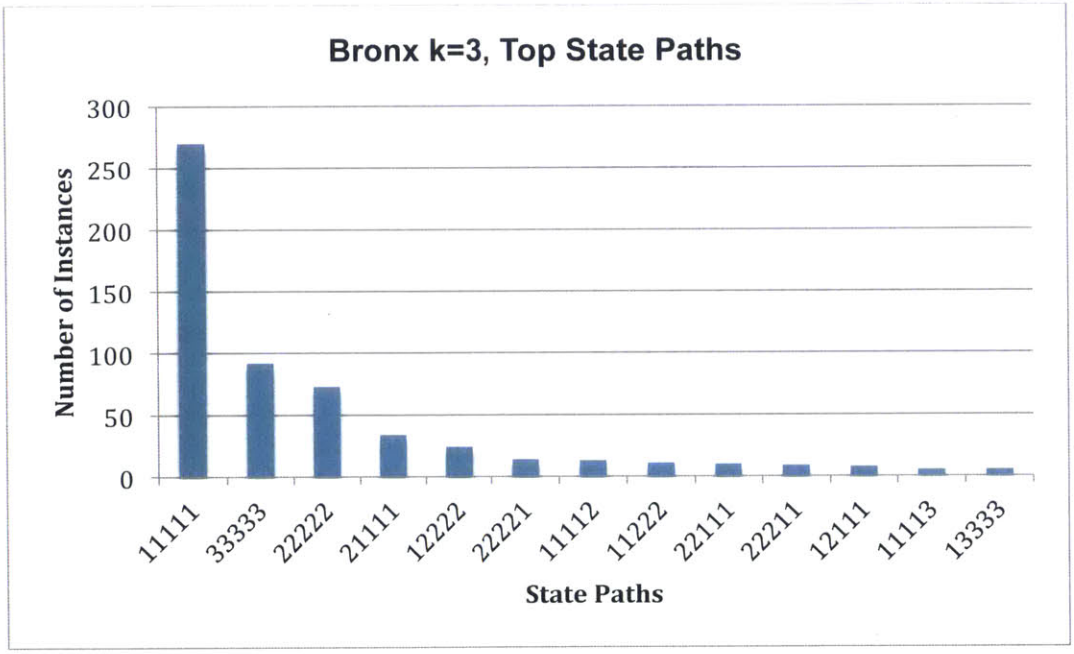

Figure 31 Count of top State Paths (five or more instances) for Bronx County. The majority of paths exhibit no transition (11111), however several block groups show patterns of a single transition followed by a sequence of stability. Transitions typically occur between States 1 and other States.

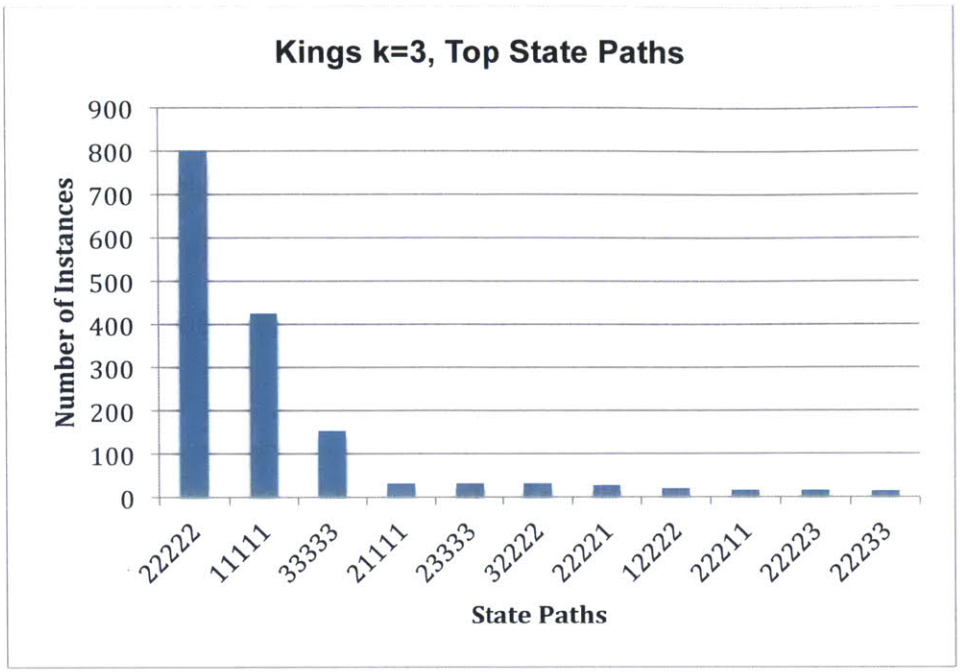

Figure 32 Count of top State Paths (five or more instances) for Kings County. The majority of paths exhibit no transition, however several block groups show patterns of a single transition followed by a sequence of stability.

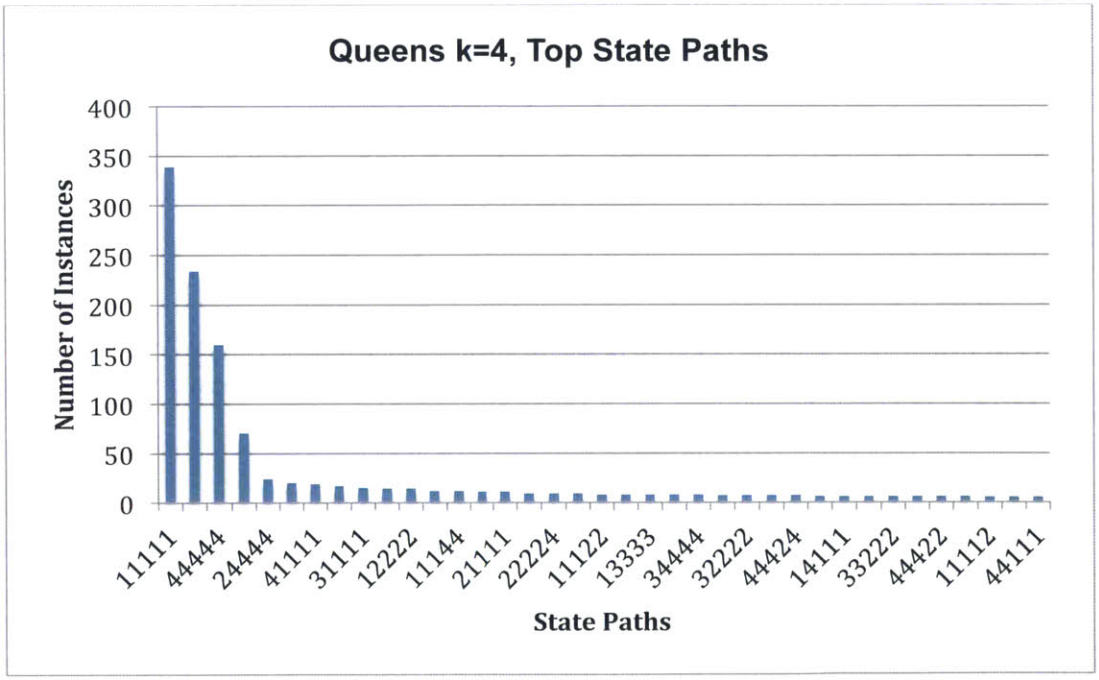

Figure 33 Count of top State Paths (five or more instances) for Queens County. The majority of paths exhibit no transition, however several block groups show patterns of a single transition followed by a sequence of stability. States 2 and 4 appear to transition

the most, and Queens notably has several more observed transition typologies than the other Counties.

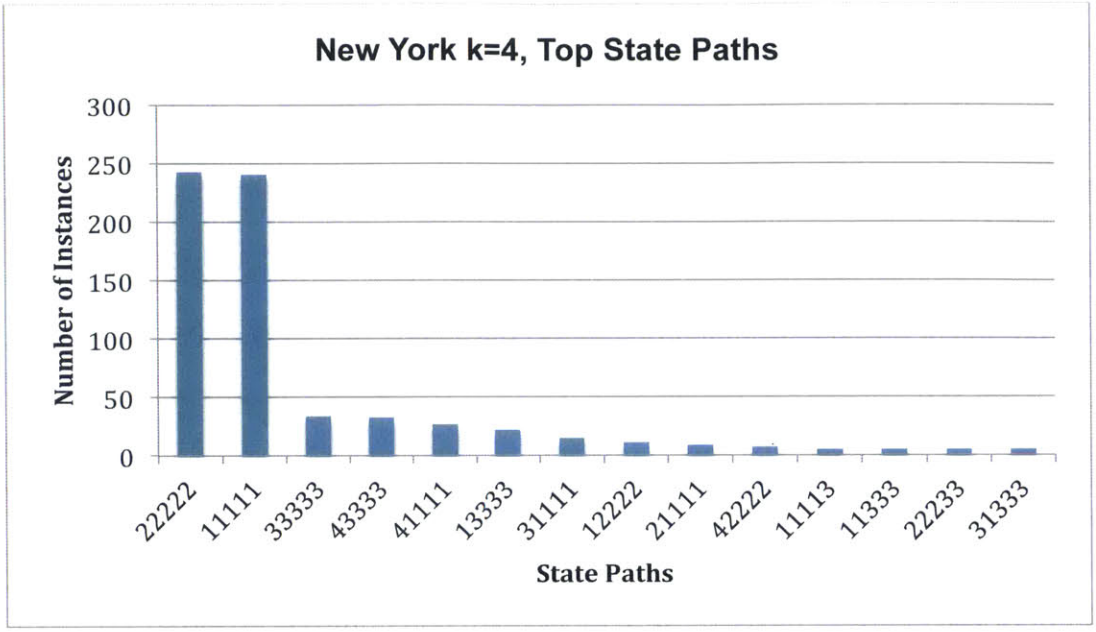

Figure 34 Count of top State Paths (five or more instances) for New York County. The majority of paths exhibit no transition, however several block groups show patterns of a single transition followed by a sequence of stability. New York appears to be substantially more stable than other counties, rarely displaying complex transition patterns.

6.4 Discussion

This research proposes modeling neighborhood change as a transition between meaningful "states" that emerge empirically from socio-demographic datasets. We identify states as profiles of complex relationships between socio-economic and demographic factors that may not be clear to the researcher a priori, and may not be identifiable through traditional forms of linear regression analysis. This is an advantage in the case of gentrification studies, where traditional models of the phenomenon reflect conflicting assumptions from a variety of disciplines. Machine

Learning techniqes, k-means clustering and markov chain modeling are applied to find emergent states of socioeconomic conditions and observe how those states change over time in the five New York boroughs. Modeling gentrification as a probabilistic process of state changes in time and space provides insights into the dynamic nature of this complex phenomenon. While mathematical models may be unable to account for many of the social and cultural intimacies of a particular site, they can generate more refined research questions for this important driver of urban regeneration. However, the research does not present a complete model of gentrification. Rather, this method allows the empirical study of neighborhood-level urban development by condensing complex urban data into latent profiles that both describe and predict urban change.

6.4.1 Key Findings

The following findings are significant observations as a result of this research. Ultimately, more questions are asked than answered, pointing towards promising opportunity for machine learning applications in urban change studies.

Education, Race and Income are State Drivers

Regardless of borough, state classifications differed primarily along lines of education, race and income. Family households, density and structural age were less powerful indicators of state identities. In New York City,

race appears to be a key factor determining access to opportunity. However, possible endogeneity of race as a correlate to income and education has not been accounted for in this research, and would be the subject of future studies. If race was excluded from the analysis, would the same results emerge? Perhaps education levels correlate strongly to race, and therefore education is a stronger determinant of a neighborhood's socioeconomic state. Such a finding could be informative for gentrification studies predicated primarily on divisions of race, pointing towards solutions rooted in improving access to education for susceptible communities rather than providing them with quick fixes to ameliorate the effects of displacement.

Transitions are observed

Perhaps the most simple and significant finding is that transitions were observed in the selected features using the coupled k-means clustering and markov method. This implies that the six features selected (income, education, race, structural age, density and family households) change over time substantially enough to be detected by our methods. Such temporal variation in demographics appears to be spatially dependent. However, when census blocks fall into categorical extremes (typically either very high or low income and education levels), census blocks rarely transition. This finding suggests that "pockets" of urban conditions exist where change rarely occurs. The degree to which census block groups

are spatially grouped into state categories may be an indicator of relative levels of socio-economic segregation; i.e., areas where census block groups fall into a single state represent pockets of homogenous characteristics, while regions with a patchwork distribution of states show variation of socio-economic conditions.

Neighborhood Boundaries are Located without Spatial Input

Although no spatial information was included in the features selected for analysis, states consistently map on to neighborhood boundaries. This suggests that administrative boundaries may determine socioeconomic conditions to a large extent. However, some states do not visibly cluster to neighborhood boundaries. Therefore, some socioeconomic conditions are tied to administrative boundaries while others are not. In most cases, these "volatile" states are those representing high income earning, educated, Caucasian non-family households. Perhaps the privilege of choice, marked by an association with race, education and domestic independence makes this population more mobile. Furthermore, transitions to this state from lower-class states are observed in high frequencies across all boroughs, making visible a phenomenon of spatial cannibalization between classes.

Profiles of Change are Context Specific

Different boroughs show different spatial configurations of change.

Queens saw change occurring primarily on the fringes of neighborhoods and exhibited much more volatility than Manhattan. Other boroughs experienced state transitions primarily alongside major transportation arteries and public transportation routes. Different patterns and profiles of change varied by borough, suggesting that "one size fits all" policies for gentrification and urban change are not likely to be successful. This prospect can influence hierarchies of decision making for gentrification policy, placing more power in the hands of boroughs as opposed to the municipality.

6.4.2 Advantages of Machine Learning in Urban Planning

The gentrification debate in Chapter 1, illustrates the lack of consensus about the definition, origins or mechanisms of gentrification. The fact that machine learning and pattern classification techniques are predicated on using emergent patterns in data to inform the model building process is a key advantage in this context. Machine learning processes note not only key or secondary variables, but also every relationship that takes part in an observed pattern. This is an advantage for planners incorporating nuanced tradeoffs into land management, zoning or social policy decisions. Urban planning needs more quantitative tools that can learn the complex patterns and behaviors that make cities so dynamic and unique. Understanding the mechanisms behind these patterned observations points towards smarter policymaking. However, the complexity of urban

dynamics as accommodated by a machine learning model can still be difficult for the researcher to interpret, leading to the use of methods that reduce variation in the data and filter a complex story into simple, interpretable elements. Traditional regression analyses are the most common method of reducing complexity for the sake of interpretability. However, Machine Learning techniques are capable of finding those elements that truly matter (correlate in a statistically significant way) without excluding all possibilities. In this way, data driven conclusions are "emergent," and less susceptible researcher bias pending a robust and viable data set.

6.4.3 Reflections on Data

The weakest point in any quantitative analysis is its data source. In the case of this research, the oversampling method used by ACS significantly reduced variation making the identification of states and their transitions difficult. Continuous, time series data in single year increments at the census block or building level with geographic identifiers is ideal for machine learning analysis in planning. However outside of the private sector, this information is difficult to come by. Data sets collated between ACS and other sources, such as AHS and other open source entities may be useful for such analyses, but could lack specificity and resolution. Furthermore, concerns about bias in sampling and survey methods may

limit the reliability of a machine learning study. Partnerships between municipalities and private sector sources such as Reonomy, or Google Flux are encouraged for municipalities looking to incorporate machine learning studies in to their operations.

6.4.4 The Role of Political Institutions

The idea of using machine learning in Urban Planning is not new. In the early 1970s, the little-known Los Angeles Community Analysis Bureau (CAB) employed machine learning techniques to understand shifting demographics and housing quality in order to allocate resources to alleviate blight and poverty. However these tools were situated in political frameworks and cultural institutions that compromised their execution. The bureau's analysis proved so useful in securing federal grant money that the city focused its activities on grant development and administration, using data analysis to justify funding. As a result, innovative technology research no longer guided the city's actions; rather the bureau was reduced to supporting predetermined goals as outlined by funding applications with data analysis. This failure suggests that data analysis must be better connected to planning, politics, policy and advocacy. In the CAB case, the bureau was not closely allied with agencies that would have pushed forward their findings, nor was it integrated into decision making structures. The role of the "technocrat" silos data analytics into

supporting roles. Today's changing urban and digital landscapes demand interdisciplinary teams with a semblance of political will and executive authority. Data analysis cannot only support social objectives, but it can also steer them. Machine learning is poised for such a role in social stewardship as it accommodates complexity, is flexible and robust, and finds emergent "pattern languages" that capture the nuances of dynamic relationships in social environments.

6.4.5 Future Research Opportunities

This research scratches at the surface of machine learning opportunities in gentrification studies and studies of urban change. Ideally, this research would be repeated with a better data source that does not use oversampling methodologies. A more thorough analysis of transition probabilities and state paths is needed, with close examination of the conditions affecting particular state transitions. Additionally the analysis should be repeated following tests for possible endogeneity between selected features, notably race and education or race and income. As it is likely that the features selected for this analysis are correlated it may be suitable to replace the k-means clustering methodology with a latent class model. Advantages to this model include the ability to create continuous states, that is census blocks are defined by probabilities of being associated with particular states, rather than forced into discrete

categories. Following an analysis of endogeneity, additional features could be explored and added to the analysis, including notably, the presence of small businesses. The predictive power of this model could be improved by targeting preprocessing methods that do not reduce variation in the dataset. Furthermore, allowing the model to run several years in the future would generate possible future states resulting from current conditions. An analysis of state transitions in relation to transportation infrastructure could also be insightful for tying gentrification trends and real estate markets to transportation. A similar analysis could be performed in relation to cultural institutions. Pairing this research with ethnographic studies of community perceptions of gentrification is important for integrating computational analysis with public experiences. This "groundtruthing" is essential for integrating local knowledge into model assumptions. Furthermore, developing new data sets for integration into the model could direct the role of machine learning in urban planning. For example, collecting data about community perceptions on various features of urban change in time series would produce state-transition maps reflecting cultural perceptions rather than historic accuracies. This poses an interesting prospect as a design tool.

6.4.6 Towards Computational Ethics for Cities

Finally, we return to the question posed in the foreword: what would a computational ethics of urbanism look like? This research is a first step

towards formulating an answer to this question. In the design of a machine learning framework for gentrification I identified the importance of using algorithms that *learn* rather than reproduce assumptions, the value of distilling large and complex data relationships into nuanced intuitions, and the important role political institutions and data sources play in the validity of arriving at conclusions that can inform policy on complex issues like gentrification. Connecting computational analysis with social movements, indeed using data driven research to define objectives for social movements---is an exciting opportunity for planners. By including quantitative tools in our scope of urban intuition, we can excavate systemic patterns of urban change to inform policy that changes mechanisms, rather than simply treating the symptoms of problems like housing affordability and gentrification. A computational ethics of urbanism would support cultures and communities by securing robust and comprehensive data sets, supporting a suite of analytics that captures the complexity of human systems and informing algorithmic design through engagement and diverse participation. As we face new challenges in the urban environments of the future, we must understand and replicate the organic infrastructure that keeps our cities healthy, inclusive and humane.

7 References

Alexander, C. (1979). *The timeless way of building*. New York: Oxford University Press.

Batty, M. (2005). *Cities and complexity: Understanding cities with cellular automata, agent-based models, and fractals*. Cambridge, MA: MIT Press.

Ball, P. (2015, November 19). Gentrification is a natural evolution. *The Guardian*.

Ball, P. (n.d.). The science and politics of gentrification [Web log post]. Retrieved from http://philipball.blogspot.co.uk/2014/11/the-science-and-politics-of.html

Bettencourt, L. M. (2009). The Rules of Information Aggregation and Emergence of Collective Intelligent Behavior. *Topics in Cognitive Science, 1*(4), 598-620. doi:10.1111/j.1756-8765.2009.01047.x

Beauregard, R. (n.d.). Trajectories of Neighborhood Change: The Case of Gentrification. *Environment and Planning, 22*, 855–874.

Birch, D. (1971). Toward a Stage Theory of Urban Growth. *Journal of the American Institute of Planners, 37*(2), 78–87.

Blokland-Potters, T. (2003). *Urban bonds*. Cambridge, UK: Polity Press.

Brenner, N., Marcuse, P., & Mayer, M. (2012). *Cities for people, not for profit: Critical urban theory and the right to the city*. London: Routledge.

Bryson, J. (2013) The Nature of Gentrification. Geography Compass, 7(8), 578-587.

Carbonell, J. (1983). Machine Learning: A Historical and Methodological Analysis. *AI Magazine, 4*(3).

Clay, P. (1990). Choosing Urban Futures: The Transformation of American Cities. Stanford Law and Policy Review, 1(1), 28-39.

Crang, M. (n.d.). SENTIENT CITIES Ambient intelligence and the politics of urban space. *Information, Communication & Society, 10*(6).

Duany, A. (2001). Three Cheers for Gentrification. American Enterprise.

Durrett, R. (2010). Probability: Theory and Examples, 4th ed. Cambridge, UK: Cambridge

University Press.

Forrester, J. W. (1969). *Urban dynamics*. Cambridge, MA: M.I.T. Press.

Freeman, L. (2005). Displacement or Succession? Residential Mobility in Gentrifying Neighborhoods. *Urban Affairs Review*, *40*(4), 463–491.

Freeman, L., & Braconi, F. (2004). Gentrification and Displacement: New York City in the 1990s. *Journal of the American Planning Association*, *70*(1).

Frey, W. H., Edmondson, B., & DeWitt, J. P. (2008). *A compass for understanding and using American community survey data*. Washington, D.C.: U.S. Dept. of Commerce, Economics and Statistics Administration, U.S. Census Bureau.

Fullilove, M. T. (2004). *Root shock: How tearing up city neighborhoods hurts America, and what we can do about it*. New York: One World/Ballantine Books.

Glaeser, E. (2007). Housing Dynamics. *Harvard Institute of Economic Research*.

Glaeser, E. (2012). Housing Booms and City Centers. *National Bureau of Economic Research: Working Paper Series*, (17914).

Goodspeed, R. (2014). Smart cities: Moving beyond urban cybernetics to tackle wicked problems. *Cambridge Journal of Regions, Economy and Society CAMRES*, *8*(1), 79-92. doi:10.1093/cjres/rsu013

Graham, S. (2005). Software-Sorted Geographies. *Progress in Human Geography*, *29*(5), pp. 562–580.

Green, H. W., & Hoyt, H. (1941). The Structure and Growth of Residential Neighborhoods in American Cities. *American Sociological Review*, *6*(3), 445. doi:10.2307/2086238

Guerrieri, V. (2011). Endogenous Gentrification and Housing Price Dynamics. *University of Chicago: Working Papers*.

Hidalgo, C. (n.d.). Streetscore - Predicting the Percieved Safety of One Million Streetscapes.

Hidalgo, C., & Glaeser, E. (2015). Do People Shape Cities or Do Cities Shape People? The Co-Evolution of Physical, Social, and Economic Change in Five Major U.S. Cities. *National Bureau of Economic Research: Working Paper Series*.

Jacobs, J. (1977). *The death and life of great American cities*. New York: Vintage Books.

Johnston, R. (n.d.). Explanation in Geography (1969): David Harvey. *Key Texts in Human Geography*, 25-32. doi:10.4135/9781446213742.n4

Kolko, J. (n.d.). The Determinants of Gentrification. *Public Policy Institute of California*.

Lees, L. (1994). Rethinking Gentrification: beyond positions of economics or culture. *Progress in Human Geography*, *18*(2), 137–150.

Lees, L., Slater, T., & Wyly, E. K. (2008). *Gentrification*. New York: Routledge/Taylor & Francis Group.

Ley, D. (n.d.). Alternative Explanations for Inner-City Gentrification: A Canadian Assessment. *Annals of the Association of American Geographers*, *76*(4), 521–535.

Light, J. S. (2003). *From warfare to welfare: Defense intellectuals and urban problems in Cold War America*. Baltimore: Johns Hopkins University Press.

Light, J. (2011). Discriminating Appraisals: Cartography, Computation and Access to Federal Mortgage Insurance in the 1930s. *Technology and Culture*, *52*(3), 485–522.

Machine Learning: What it is and why it matters. (n.d.). Retrieved from https://www.sas.com/en_us/insights/analytics/machine-learning.html

Maciag, Mike (2015). Gentrification in America. Governing DATA, http://www.governing.com/gov-data/census/gentrification-in-cities-governing-report.html. Accessed May 2015.

Marcuse, P. (1985). Gentrication, Abandonment, and Displacement: Connections, Causes, and Policy Responses in New York City. *Urban Law Annual ; Journal of Urban and Contemporary Law*, *28*, 195–240.

Mitchell, T. (1986). *Machine Learning: A Guide to Current Research*. Kluwer Academic Publishers.

O'Sullivan, D. (2002) Toward Micro-scale Spatial Modeling of Gentrification. Journal of Geographical Systems, 4, 251–274.

Oswalt, P., Overmeyer, K. & Misselwitz P. (2013) Urban Catalyst: The Power of Temporary Use.

Berlin, DE: Dom.

Park, R. E., Burgess, E. W., McKenzie, R. D., & Wirth, L. (1925). *The city*. Chicago, IL: University of Chicago Press.

Porta, S. (2014). The Form of Gentrification. *University of Strathclyde Glascow: Working Papers*.

Portugali, J. (2006). Complexity Theory as a Link Between Space and Place. *Environment and Planning A*, *38*, 647–664.

Portugali, J. (2010). Complexity Theories of Cities: Achievements, criticisms, potential. In *Complexity, Cognition and the City*. Springer.

Rabiner, L. (1989). Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Proceedings of the IEEE, 77(2), 257-286.

Rascoff, S. (2015, January 28). Confirmed: Starbucks Knows the Next Hot Neighborhood Before Everybody Else Does. *Quartz*. Retrieved from www.qz.com.

Robinson, I. (1965). A Simulation Model For Renewal Programming. *JAPA*, *31*(2), 126–134.

Rosenbluth, A. (n.d.). Behavior, Purpose and Teleology. *Philosophy of Science*, *10*, 18–24.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Samuel, A. (1969). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal*, *3*(3).

Sassen, S. (2000). *Cities in a world economy*. Thousand Oaks, CA: Pine Forge Press.

Shmueli, G. (2010). To Explain or Predict? *Statistical Science*, *25*(3), 289–310.

Simon, H. A. (1990). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Slater, T. (2006). The Eviction of Critical Perspectives from Gentrification Research. *Urban Journal of Urban and Regional Research*, *30*(4), 737–757.

Slater, T. (2012, October 4). Rose Street and Revolution: A Tribute to Neil Smith (1954-2012) [Web log post]. Retrieved from http://www.geos.ed.ac.uk/homes/tslater/tributetoNeilSmith.html

Slater, T. (2014, April 11). Unravelling false choice urbanism [Web log post]. Retrieved from http://crisis-scape.net/conference/item/180-unravelling-false-choice-

urbanism#sdfootnote13sym

Slater, T. (2014, November 24). There is Nothing Natural about Gentrification. Retrieved from http://www.newleftproject.org/index.php/

Slater, T. (2014). Planetary Rent Gaps. *Antipode*.

Smith, N. (1979). Toward a Theory of Gentrification A Back to the City Movement by Capital, not People Neil Smith. *Journal of the American Planning Association*, *45*(4), 538–548.

Smith, N. (2002). New globalism, new urbanism: gentrification as global urban strategy. Antipode 34(3), 428-450.

Snow, C. P. (1959). *The Two Cultures* (1st ed.). Cambridge University Press.

Sugar, C. A. & James, G. M. (2003). Finding the number of clusters in a data set: An information theoretic approach. Journal of the American Statistical Association, 98(1), 750–763.

Thompson W. E. & Hickey J. V. (2004). Society in Focus: An Introduction to Sociology, 5th Ed. Boston, MA: Allyn & Bacon

Tweedie S. P. & Meyn R. L. (2009). Markov Chains and Stochastic Stability. Cambridge, UK: Cambridge University Press.

Torrens, P., & Nara, A. (2007). Modeling Gentrification Dynamics: A Hybrid Approach. *Computers, Environment and Urban Systems*, *31*, 337–361.

U.S. Census Bureau (2008). A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know. Washington, DC: Government Printing Office.

Vallianatos, M. (n.d.). Uncovering the early history of "Big Data" and the "Smart City" in Los Angeles. *Boom: A Journal of California*.

Vernon, R. (1959). *The Changing Economic Function of the Central City*. The Area Development Committee of CED.

Wacquant, L., & Slater, T. (2014). Territorial Stigmatization in Action. *Environment and Planning A*, *46*, 1270–1280.

White, K. (2008). Who Gentrifies Low-Income Neighborhoods? *National Bureau of Economic Research: Working Paper Series*.

Wright, M. (2014). The Gender, Place and Culture Jan Monk Distinguished Annual Lecture: Gentrification, assassination and forgetting in Mexico: a feminist Marxist tale. *Gender, Place and Culture*, *21*(1), 1–16.

Wyly, E. (2006). The Right To Stay Put, Revisited: Gentrification and Resistance to Displacement in New York City. *Urban Studies*, *43*(1), 23–57.