Finding the State Space of Urban Regeneration: Modeling Gentrification as a Probabilistic Process using k-Means Clustering and Markov Models

Emily Royall and Thomas Wortmann

Abstract

Gentrification is a dynamic, globalized urban process whose complex definition varies with stakeholder perspectives. This complexity makes it challenging for researchers to study the impact of gentrification, and difficult for planners to anticipate the effects of gentrification with planning policy. This paper proposes to model gentrification as a Markov process, i.e. a process that assigns probabilities to potential "state" changes over time (Rabiner, 1989). Using American Community Survey (ACS) data for four boroughs of New York City between 2009 and 2013 (including demographic, economic, geographic, and physical characteristics of census block groups), we develop our 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 testing predictions against historic data and comparing them with 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

Gentrification is a dynamic, globalized urban process whose complex and much debated definition varies with stakeholder perspectives. This complexity makes it challenging for researchers to study the impact of gentrification, and difficult for planners to anticipate the effects of neighborhood change with planning policy.

The definition of gentrification is widely disputed in both academic and media circles. Since the term's introduction by Ruth Glass in the 1960s, gentrification has been characterized as both a cause and a symptom of social injustice. Seminal views have characterized gentrification as urban re-investment resulting in displacement of the poor (Lees et al., 2009), a consequence of the transition to post-industrial cities (Lees, 2008), a generalized global urban strategy replacing liberal urban policy in an increasingly capitalistic society (Smith, 2002), and a symptom of regional economic or demographic change (Clay, 1989). Varying academic interpretations of, and professional experiences with gentrification make quantitative analysis of this social process a difficult endeavor. Recent quantitative studies disagree on how to characterize gentrification as a spatial, temporal or socio-economic process. However, throughout the literature, gentrification is described as a process of socioeconomic and demographic *change* in urban areas.

This research proposes a quantitative methodology that identifies neighborhood-level, temporal patterns of socioeconomic and demographic change. Our method identifies patterns or "states" of social and economic conditions in neighborhoods, and tracks how these states change over time. Using k-means clustering to identify common socio-economic "states" through which urban areas transition over time, 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.

2. Concepts and Causes of Gentrification

2.1 Gentrification as a Complex Process

The dynamic relationships that characterize gentrification call for a mode of scientific inquiry that recognizes and engages their complexity. Concepts of gentrification vary depending on their political contexts. However, regardless of the political environments, gentrification is usually described as a temporal process. We understand a process as a series of actions or operations that result in a particular outcome. Certainly, the "particular outcome" most often associated with the gentrification process is displacement. Glass (1964) observed gentrification primarily as displacement of the lower class, citing that "once this process of 'gentrification' starts in a district, it goes on rapidly until all or most of the original working-class occupiers are displaced". Lees (2008) characterizes gentrification as a self-organized urban reinvestment process that uniquely results in the displacement of the lower class. However, following a recent report "Gentrification in America" from *Governing Magazine* (Maciag, 2015), pundits assert that gentrification might not be characterized by displacement at all. City Observer's Joe Cortright (2015) suggests that the tendency of higher-income families to isolate themselves in suburbs and gated communities creates pockets of spatial income segregation that are not necessarily a consequence of gentrification. Alternative views see gentrification as residential ethnicization (Hwang et. al (2014) and landscape aestheticization (Bryson, 2013). In European contexts, gentrification is sometimes viewed as a socio-economic process that can be proactively harnessed for urban revitalization (Oswalt, 2013).

2.2 Modeling Gentrification

Gentrification remains difficult to model quantitatively, in part due to its wider political, social and cultural aspects (O'Sullivan, 2002), although multi-agent systems (MAS) and cellular automata (CA) models are commonplace in several areas of urban systems modeling (Torrens, 2006). CAs map agents onto a grid and model state changes in time relative to a set of rules based on the states of neighboring cells. MAS modeling and simulation attempt to model the complex interactions of individual agents in relation to each other and environmental conditions. Both MAS and CA models attempt to describe system-wide behaviors based on pre-defined relationships between actors and environmental features of a system.

In part, the appeal of CA and MAS models is their ambition to capture complex inter-relationships between agents and their environments, which is an evident characteristic of gentrification (see section 1). O'Sullivan (2002) demonstrates a spatial CA model of gentrification based on individual decision-making. Here, cells are characterized by abstract states such as "Not for Sale", "For Sale", "Seeking Tenants" and "Rented." Depending on the conditions of neighboring cells, cells transition between states according to a set of pre-defined rules. Underlying this model is the Rent Gap theory, which describes dynamic interactions between residential home value and capitalized ground rent for a property over time (Smith, 1979). Similarly, Torrens and Atushi (2006) propose a hybrid CA/MAS approach, using dynamic property markets as a theoretical basis for their models. These models emphasize spatial effects of agent-based decision-making, reflecting the common practice among developers and real-estate agents to determine property value based on location.

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 3.3.

3. Materials and Methods

In the following sections, we describe the three steps of our analysis: 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. We selected five-year estimates because they are the most fine-grained analysis available from ACS, providing data at the level of census block groups. 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 distinct 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

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. This fine-grained sampling allows a higher resolution for models that is appropriate for processes like gentrification, which are often visible 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.

3.1.2 Pre-processing the ACS Data

Before submitting the data to the clustering algorithm, we took several steps to increase its suitability for clustering. To increase the speed, quality, and intelligibility of the clustering, we turned 108 fields from the ACS data into 32 features that captured a broad picture of urban development in a more compact fashion. 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 those of 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.

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, choosing the one with the smallest total distance, i.e. the smallest objective value, as the final clustering for each county.

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.

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.

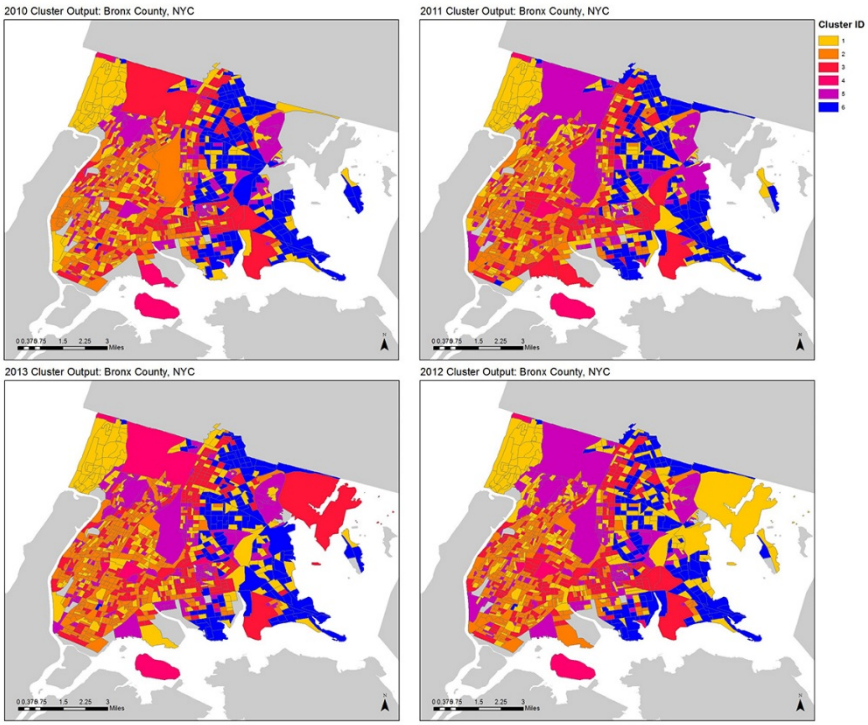

Figure 1. State visualizations of Bronx County from 2010-13 (clockwise from upper left).

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.

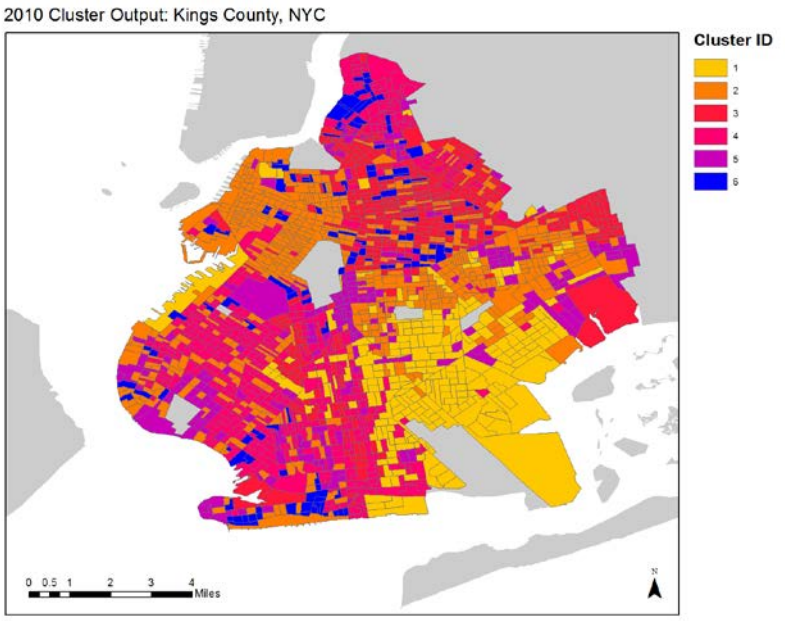

Figure 2. State visualization of Kings County for 2010.

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.

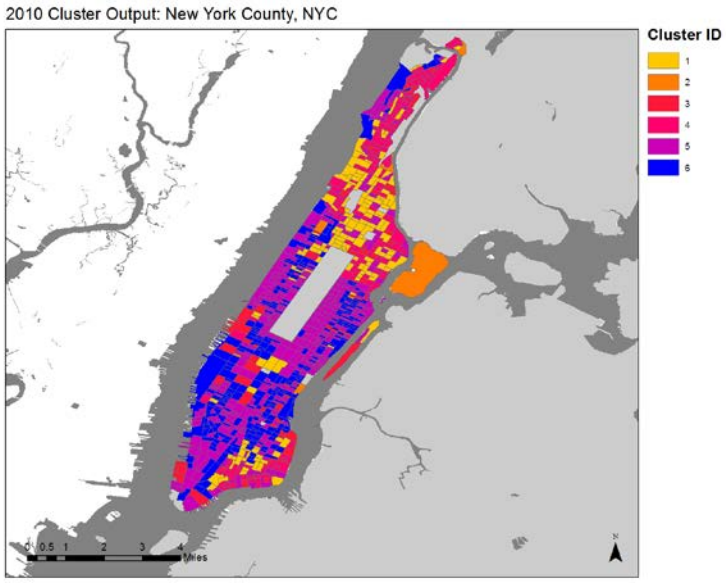

Figure 3. State visualization of New York County for 2010.

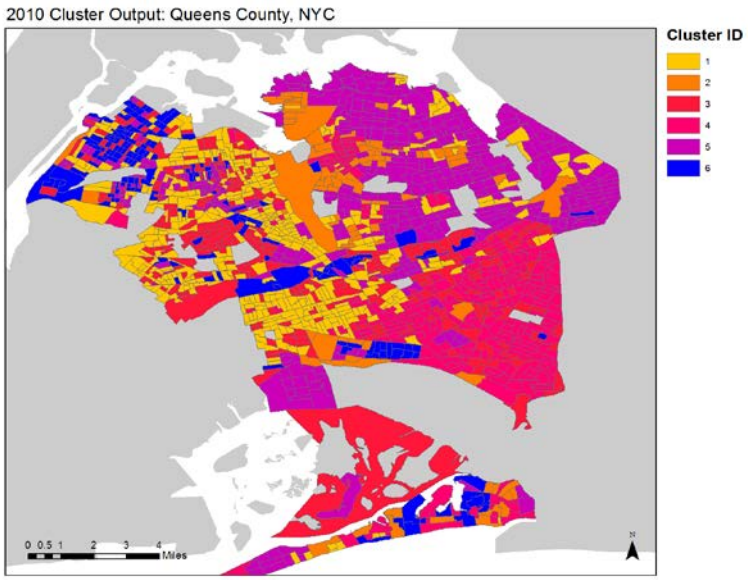

Figure 4. State visualization of Queens County for 2010.

4.3 Predicting State Transitions

The four Markov Models described in section 3.3 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.

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

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

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

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.

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.

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.

Portugali, J., Self-Organization and the City (2000). Berlin, DE: Springer.

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

Smith, N. (1979). Toward a theory of gentrification: a back to the city movement by capital not people. Journal of the American Planning Association, 45(4), 538-548.

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

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.