ORIGINAL ARTICLE

Journal Section

Neighborhood change in the United States - a comparison of sequence analysis methods

Wei Kang[1] | **Sergio Rey**[1] | **Levi John Wolf**[2] | **Elijah Knaap**[1] | **Su Yeon Han**[1]

[1]Center for Geospatial Sciences, University of California, Riverside

[2]School of Geographical Sciences, University of Bristol

Correspondence
Wei Kang PhD, Center for Geospatial Sciences, University of California Riverside, Riverside, California, 92521, USA
Email: weikang@ucr.edu

Funding information
National Science Foundation Award 1733705

There is a recent surge in research focused on urban transformations in the United States via empirical analysis of neighborhood sequences. The alignment-based sequence analysis methods have become the dominant techniques for the neighborhood sequence analysis. However, it is unclear to what extent these methods are robust in terms of producing consistent and converging sequence typologies. This article sheds light on this issue by applying five sequence analysis methods to the same data set - 50 largest Metropolitan Statistical Areas (MSAs) of the United States from 1970 to 2010.

KEYWORDS
Sequence Analysis, Neighborhood Change, Optimal Matching, Clustering, Sensitivity, Robustness

1 | INTRODUCTION

There is a recent surge in research in the United States focused on understanding urban transformations through empirical analyses of neighborhood sequences (Delmelle, 2016, 2017; Lee et al., 2017a,b; Li and Xie, 2018). Driven by an interest in the social and economic restructuring of cities and the associated consequences like gentrification and displacement, this work uncovers emergent patterns in the evolution of neighborhood socioeconomic characteristics over time. Typically, this work uses census tracts as proxies for neighborhoods and consists of two stages: the first stage classifies neighborhoods into a set of discrete types based on selected socioeconomic attributes, yielding for each neighborhood a temporal sequence of discrete types; the second stage employs sequence analysis (SA) methods to further investigate these neighborhood sequences, providing insights in neighborhood change. Two types of SA methods are at researchers' disposal: "stepwise approaches," such as Markov Chains, view the sequence as being generated stochastically and model the probability of transitions between neighborhood types over time; "whole sequence approaches" such as the optimal matching (OM) algorithm, meanwhile, view the sequence from a holistic perspective and evaluate the pairwise similarity between each neighborhood sequence in a study region (Abbott, 1995). The latter method produces a sequence similarity matrix, which can be further distilled with a clustering algorithm into a typology of prototypical neighborhood sequences.

The OM algorithm, originally developed for matching protein and DNA sequences in biology (Carrillo and Lipman, 1988; Wong et al., 2008) and used extensively for analyzing strings in computer science, has become the dominant SA technique in the neighborhood literature (Delmelle, 2016, 2017; Lee et al., 2017a,b; Zwiers et al., 2017; Li and Xie, 2018). It generally works by finding the minimum cost for aligning one sequence to match another using a combination of operations including substitution, insertion, deletion and transposition. The cost of each operation can be parameterized differently and may be theory-driven or data-driven. Applications in the neighborhood literature often adopt the data-driven approach based either on socioeconomic dissimilarities in contemporary experience or empirical transition probabilities between neighborhood types over two consecutive time points.

The fact that the OM algorithm relies on multiple assumptions about the evolution of the sequences makes it an easy target of criticism. In bioinformatics, Wong et al. (2008) shows that the alignment of genomic data and thus the resultant similarity values are greatly affected by small changes in the operation parameters such as substitution, insertion, and deletion costs. There is also an ongoing debate on the adequacy of the OM method in the life course research, and the social sciences more generally. Biemann (2011) argue that the direct application of OM analyses to life course data is inappropriate since the life course is an unfolding process, whereas DNA sequences for which OM was designed originally, share common ancestors. Variants of OM should be proposed which take account of characteristics specific to life courses. Several simulation studies have been conducted to shed light on the behavior of OM and its variants in terms of revealing differences of sequences in timing, duration and sequencing which are important in life course research (Robette and Bry, 2012; **?**; Studer and Ritschard, 2016; Ritschard and Studer, 2018). Though much could be borrowed from life course research when it comes to the application of the SA methods to neighborhood change research, it should be noted that the latter is usually concerned with a very short sequence (of length 5 at most in the case of the United States) due to data availability while the former deals with a longer sequence (sequences of length 20 are simulated in **?**). The other major difference is that the unit of study for the latter is census tract (or neighborhood), which is a spatial entity, posing potential issues of spatial aggregation, spatial autocorrelation and spatial heterogeneity.

This article focuses on the application of the SA method to neighborhood change research and explores two related issues. We examine the relationship between neighborhood sequence typology and operation costs as well as whether this relationship displays spatial disparity. We are particularly interested in the sensitivity of neighborhood

sequence typology to the choice of operation costs, that is, whether a small change in the operation costs will result in a much different typology. We also note that the current literature focuses solely on substitution costs while setting prohibitive costs of other operations so that they are unlikely to be chosen in the OM process. This means that current research considers only one sequence characteristic when determining the similarity of any two neighborhood sequences, that is, the year in which a specific neighborhood type appears. We argue that considering other characteristics, including the order in which successive neighborhood types appear and the duration of a neighborhood type, could help reveal interesting patterns that are critically important for understanding urban socioeconomic transformations. Therefore, incorporating other cost choices or SA methods that can identify these sequence characteristics provides a promising new direction for neighborhood change research.

We support these arguments through an empirical review of five SA methods applicable for uncovering neighborhood sequence patterns from different aspects. We do so by applying these methods to the same data set - the 50 largest Metropolitan Statistical Areas (MSAs) of the United States at census years 1970, 1980, 1990, 2000, and 2010. We have found that the neighborhood sequence typology varies with the choice of operation costs as well as the MSA under study. In other words, the typology of neighborhood sequence is sensitive to operation cost and this sensitivity displays spatial heterogeneity. The sensitivity is more severe in MSAs including the San Antonio-New Braunfels MSA in Texas, and less severe in MSAs including the San Francisco Metropolitan Area in California and the Providence Metropolitan Area in Rhode Island. In addition, a method or cost choice could be effective to reveal one particular characteristic of neighborhood evolution on the one side while failing to provide useful information on the other side. Thus, researchers should take caution when both adopting a method and interpreting results.

The rest of the article proceeds as follows. We provide a description of SA and a review of its application in neighborhood change research in Section 2. Section 3 introduces the longitudinal census data, the neighborhood segmentation method, five SA methods to be compared, and the sequence clustering method. We provide results of the neighborhood sequence typologies based on selected SA methods and the evaluation of the sensitivity in Section 4, and we conclude the article in Section 5 with future directions.

2 | NEIGHBORHOOD CHANGE AND SEQUENCE ANALYSIS

Urban researchers from across the social sciences have long sought to understand the social and political processes that delineate and modify conceptions of "neighborhoods". Such processes include not only those which *form* neighborhoods, like housing development and urban design, but those which transform and circumscribe neighborhoods through residential sorting and social exchange, like segregation, gentrification, and disinvestment. Given the considerable breadth of the urban studies, neighborhood research over the last 100 years has burgeoned, and is currently in a sort of renaissance, thanks to growing attention to the importance of neighborhood effects and the dramatic patterns of gentrification that are beginning to fundamentally reshape cities in many Western nations (Schwirian, 1983; Beauregard, 1990; Temkin and Rohe, 1996). Over the last few decades, a growing body of empirical work has attempted to provide insight into these important trends through a wide variety of modeling strategies, and in recent years these efforts have been bolstered by new computational methods and techniques from data science.

One particularly promising technique for modeling neighborhood change is the application of sequence analysis methods that consume time series of neighborhood data to examine how each neighborhood moves through a sequence of discrete "types" or "states" (Lee et al., 2017b; Delmelle, 2016, 2015). Although these methods rely on emerging analytical techniques, they are also motivated by longstanding theory in urban ecology originally posited by Chicago School sociologists in the early 1900s. Chicago School theorists posited that cities tend to fragment into

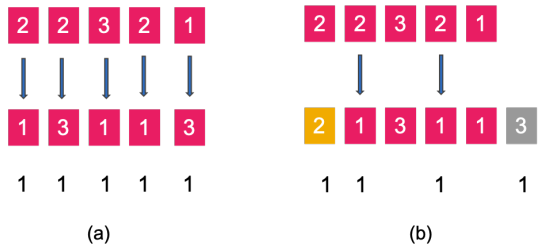

FIGURE 1 A small example of calculating OM distance between two short sequences. (a) The cost of substituting any number with a different one is 1 while the cost of inserting or deleting (indel) any number is 2. (b) The cost of substituting any number with a different one is 1 and the cost of inserting or deleting any number is also 1.

"natural" areas delineated by race and class, and that urban dynamics can be understood as the process by which households translate socioeconomic gains into spatial advantages. Put differently, urban space is partitioned into areas that indicate the social status of their residents, and as city dwellers climb the social hierarchy, they tend to move into correspondingly higher "social areas" of the city (Schwirian, 1983). Neighborhood sequence analysis is designed to help shed light on these processes by examining how *places* move through the social hierarchy over time.

Sequence analysis consists of two general types, "stepwise" and "whole sequence", each of which views and models the sequence from a different perspective. In this article, we focus on the latter, which holds a holistic perspective by considering the sequence as a whole and attempting to measure the distance between every pair of sequences. Based on whether the computation of the distance requires sequence alignment, the "whole sequence approaches" can be further divided into alignment-free (Vinga and Almeida, 2003; Cha, 2007; Zielezinski et al., 2017) and alignment-based methods. The former consists of the distance measures between longitudinal distributions such as Euclidean distance and χ^2 distance which focus on the frequency of each type while neglecting sequencing and exact timing of the neighborhood type. The latter consists of the OM method and its variants. Aisenbrey and Fasang (2010) and Studer and Ritschard (2016) provide comparative surveys of these methods in the study of life courses such as professional careers and distinguish them from those applied to other domains including biology and computer science. Since OM has become the dominant approach in the research of neighborhood change, we focus on OM and its variants in the rest of the paper.

OM measures the distance between two sequences as the minimum cost of transforming one sequence to be one exactly like the other. The operations involved in the transformation are substitution, insertion, and deletion, each of which is parameterized with a prior cost–the values of which are vital to the algorithm's performance (Hollister, 2009). For example, if we are to calculate the OM distance between two short sequences - '2,2,3,2,1' and '1,3,1,1,3' as shown in Figure 1, we could arrive at two divergent matching processes ((a) and (b)) and thus different resultant OM distances by giving different substitution and/or insertion/deletion (indel) costs. For both of them, the cost of substituting any number with a different number is 1, while (a) has a larger cost of inserting or deleting (indel) any number - 2, and (b) has a smaller cost - 1. Because of the large indel cost, matching process (a) does not involve operations of insertion and deletion, and the OM distance is 5. In contrast, (b) shifts the sequence '1,3,1,1,3' slightly to the right, insert '2' to the left, and delete the rightmost '3'. With a combination of 2 substitutions, 1 insertion and 1 deletion, (b) arrives at the OM distance of 4, which is smaller than (a). It is obvious that a change in the indel cost makes a difference to the OM process and distance, and it should also be noted that the alignment involved in (b) reflects a distortion in time and by doing so it allows for the matching of two sequences experiencing similar development stages but at different time periods. Comparatively, (a) focuses solely on the contemporaneous experience.

In implementation practice, OM is usually stated as a dynamic programming problem. Through a series of simulation experiments, Studer and Ritschard (2016) shows that specific characteristics of a sequence can be picked up by appropriately selecting the operation costs, including contemporaneous similarity, sequencing, and duration of a state. Naturally, if the research focus lies in the contemporary similarity between sequences, a very large value for the insertion and deletion costs should be selected so that only substitutions are possible in the OM process. Even so, the selection of the substitution costs is still a serious issue as different values could lead to divergent results. The extent to which OM-based SA methods are robust techniques in their ability to produce consistent and converging results has been a pervasive issue in the literature (Robette and Bry, 2012) and is also the focus of this article.

There have been a series of studies employing the OM algorithm to analyze neighborhood sequences which could provide insights into neighborhood change from a holistic perspective compared with the stochastic Markov Chains approaches (Schwirian, 1983). More specifically, SA methods are used to assess the similarity between each pair of neighborhood sequences based on socioeconomic characteristics. Together with cluster analysis, the research is aimed at identifying the predominant sequences in which neighborhoods change as well as producing a typology of neighborhood sequences (Delmelle, 2017). To date, the selection of operation costs is mostly data-driven. For example, in a study of neighborhood sequences in Chicago and Los Angeles from 1970 to 2010, Delmelle (2016) bases substitution costs on empirical transition rates across census years. If the empirical transition rate between two neighborhood types is large, the cost of substituting one with the other is small. Later, Delmelle (2017) employs a variant of OM which focuses on sequences of transitions between neighborhood types in 50 U.S. MSAs from 1980 to 2010. Other neighborhood research in the U.S. (Lee et al., 2017a,b) and the Netherlands (Zwiers et al., 2017) adopt another variant of OM which leads to a subsequence based distance measure and is more sensitive to differences in the order of neighborhood types.

Despite a growing body of research, the application of SA methods to the study of neighborhood evolution is not straightforward and involves another layer of uncertainty. Unlike life course research where the life states constitute a sequence directly, neighborhood “types" (or “states") are unknown and are usually determined by employing multivariate clustering algorithms approaches in a process known as “geodemographic segmentation" (Rey et al., 2011; Reibel, 2011; Singleton and Spielman, 2014). Uncertainty comes from the geodemographic cluster assignment process where various clustering algorithms could lead to different results (Singleton et al., 2016). We do not intend to investigate this uncertainty, but rather produce an baseline neighborhood segmentation scheme which will be used for the comparison between several SA methods.

3 | DATA AND METHODS

To examine how neighborhood change classification is sensitive to the choice of the SA method, we selected five SA methods and applied each to a decennial census data set in the United States from 1970 to 2010. Several evaluation measures were employed to compare the neighborhood sequence clustering results to shed light on the sensitivity of each SA method as well as the spatial variation of such sensitivity. In this section, we introduce the complete workflow of the empirical comparisons including the census data set, the neighborhood segmentation algorithm, the five SA approaches measuring the pairwise similarity of neighborhood sequences as well as the subsequent sequence clustering algorithm, and the final evaluation indices.

TABLE 1 List of fourteen variables to depict neighborhoods.

Category	Variable	Description
Demographic	CHILD	% persons who are children under 18 years old
	OLD	% persons who are 65+ years old
	SHRWHT	% white population
	SHRBLK	% black/African American population
Socioeconomic	UNEMPRT	% persons 16+ years old who are in the civilian labor force and unemployed
	PRFE	% persons 16+ years old employed in manufacturing, transportation, and public administration
	POVRAT	% total persons below the poverty level last year
	EDUC	% persons 25+ years old with at least a 4-year degree
Housing	BL30OLDPRO	% total housing units built MORE than 30 years ago
	TTMULTI	% total multiunit structures
	YRMV10PRO	% occupied housing units where household heads moved in less than 10 years ago
	MNVALHS	Mean value of specified owner-occupied housing units
	OWNO	% total owner-occupied housing units
	VACHUPRO	% total vacant year-round housing units

3.1 | Study Area and Data

Following many existing neighborhood segmentation and neighborhood change analyses (Mikelbank, 2011; Wei and Knox, 2014; Delmelle, 2015, 2016, 2017; Lee et al., 2017a; Li and Xie, 2018), we adopt the census tract as the primitive unit in constructing neighborhood definitions. We expected to compare the SA methods based on a large spatial and temporal extent, but the limited availability of census tract data in earlier years such as 1970 and 1980 prevented us from a consideration of all urban areas in the United States. Therefore, we selected 50 MSAs with the largest population in 2010 as reported by the U.S. census bureau in September 2012 [1] to ensure that most tracts can be traced back to the decennial censuses in earlier years.

Because the boundaries of many census tracts changed between decennial censuses due to population change, a comparison across various years to reveal neighborhood change cannot be made directly. To overcome this challenge, we use the Geolytics Neighborhood Change DataBase 2010 (NCDB 2010) [2] which provides census tracts in 1970, 1980, 1990, and 2000 with boundaries and attributes recalculated and normalized to 2010. The 2010 sources are 2010 long-form census and 2006-2010 American Community Survey (ACS) estimates. The latter has high degree of uncertainty in some estimates (Folch et al., 2016).

Following earlier studies on geodemographics (Singleton and Longley, 2009; Singleton and Spielman, 2014), we selected fourteen variables covering demographic, socioeconomic, and housing characteristics as shown in Table 1 to depict neighborhoods. Some of these variables were directly extracted from NCDB 2010 including CHILD and OLD, while others were constructed from relevant variables available in NCDB 2010 such as BL30OLDPRO.

[1] https://www.census.gov/library/publications/2012/dec/c2010sr-01.html

[2] http://www.geolytics.com/USCensus,Neighborhood-Change-Database-1970-2000,Products.asp

3.1.1 | Data Cleaning

The total number of census tracts within the 50 largest MSAs based on the 2010 boundaries is 38,453. Because the Decennial U.S. Censuses in 1970 and 1980 do not cover all the 38453 tracts, we limit our analysis to include only the tracts whose data have been consistently collected since 1970. Further, following the strategy of Wei and Knox (2014), tracts with a population less than 500 were excluded to avoid bias from small samples. After dropping miscoded or missing values, our final dataset contains 25,961 census tracts for each of the 5 census years. The analysis, therefore, proceeds with 129,805 total observations in the initial geodemographic segmentation, yielding 25,961 neighborhood trajectories of length 5 to enter the SA process.

3.2 | Neighborhood Segmentation

Geodemographic segmentation is based on the k-means clustering algorithm to assign each census tract at each of five decennial census years to one of k neighborhood types. We apply the clustering algorithm to all 129,805 tracts at once to produce k neighborhood types which are consistent and comparable across space and time. Since feature scaling can impact clustering results significantly, we transform each variable using z-score standardization relative to each census year.

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{(d_i - c_i)}{\max(d_i, c_i)} \tag{1}$$

Performance of the k-means clustering algorithm is contingent on the choice of k - number of clusters. We rely on the average silhouette coefficient to select an "appropriate" number of clusters. This coefficient is defined in Equation (1) where n is the number of observations, d_i is the shortest average distance of observation i to all points in each of other clusters to which i does not belong, and c_i is the average distance between i and any other observations within the same cluster. S lies within the range $[-1, 1]$. A larger S indicates a better clustering. We calculated average silhouette coefficients for clustering results with k ranging from 2 to 15 and selected the number which maximizes the coefficient. We note that this process does not necessarily result in the "optimal" or "correct" neighborhood classification, but rather produces a set of neighborhood labels as the basis of the further sequence analysis and comparison.

3.3 | Neighborhood Sequence Analysis

After neighborhood segmentation, we obtained one categorical cluster label for each census tract at each census year. We then organized labels for each tract into a chronological sequence, resulting in 25,961 neighborhood sequences of length 5. These constitute our observations for sequence analyses.

We select five SA methods, or more specifically five global alignment methods, for the empirical comparison displayed in Table 2. They differ in either the choice of the operation costs, or the formation of the sequence. For the former, we pay special attention to the insertion/deletion costs as small values for these costs could imply a distortion of time. Some of these methods have been applied to studies of neighborhood change while others have not.

Our first SA method uses the classic Hamming edit distance to evaluate sequence similarity. It can be viewed as a classic OM approach with a constant substitution cost (=1) and an infinite cost for insertion or deletion. The application of this OM distance metric to neighborhood sequences assumes that the distance between any pair of distinct

TABLE 2 Selected sequence analysis approaches for empirical comparison.

Index	Name	Substitution costs	Insertion/deletion costs
1	Hamming	1	$+\infty$
2	OMecenter	Euclidean distance between cluster centers	Maximum of substitution costs
3	OMtranr	Based on empirical transition rates	1
4	OMarbitr	0.5	1
5	OMstran (Sequence of transitions)	stable-stable=0, change-change=0, stable-change=1	2

neighborhood types is identical with a focus on contemporaneous similarity between neighborhood sequences.

Since we do not expect the similarity/distance between any two neighborhood types to be identical, our second approach relaxes this assumption. One natural choice is the Euclidean distances between cluster centers which can be easily obtained from the previous neighborhood segmentation step. We also slightly adjust the emphasis of contemporaneous similarity and allow for a low degree of insertion and deletion. Here, the largest Euclidean distance between any two neighborhood cluster centers is adopted as the cost of insertion and deletion. This novel OM variant is named "OMecenter".

We also examined the "OMtranr" method in which the substitution costs are based on empirical transition rates between neighborhood types over time. Although this method has been criticized on the grounds that temporal transition rates may not be a good proxy for the similarity between two types (Studer and Ritschard, 2016), we consider it here because it has been used elsewhere for similar work (Delmelle, 2016).

The fourth method, "OMarbitr", deviates slightly from the Hamming method. It employs an "arbitrary" set of choices for operation costs and is useful when a clear theory informing the choice of edit costs is unavailable. Specifically, a constant 0.5 is set for the substitution cost between any two neighborhood types. The insertion/deletion cost is set as 1, making it possible to match subsequences at different temporal periods.

The last method, "OMstran", views neighborhood change as an unfolding process explicitly, which is different from the common ancestor view of DNA sequences (Biemann, 2011). Rather than aligning sequences of neighborhood types, "OMstran" attempts to align sequences of *transitions*, pairs of neighborhood types over two consecutive periods. Each sequence of neighborhood types of length 5 is transformed into a sequence of neighborhood transitions of length 5. For example, sequence '1, 1, 1, 1, 1' is transformed into 'S1,11,11,11,11' where 'S' represents the start of a sequence. The (k, k) substitution cost matrix for classic OM algorithms is extended in this case to $(k(k+1), k(k+1))$, in which each element represents the cost of substituting a transition (e.g. '11') in one sequence with a transition (e.g. '21') in another sequence.

To illustrate, assume that we have two other sequences '3, 2, 3, 3, 3' and '1, 2, 3, 1, 2', and we would like to calculate the respective distances from the focal sequence '1, 1, 1, 1, 1'. We first transform them into sequence of transitions 'S3, 32, 23, 33, 33' and 'S1, 12, 23, 31, 12'. As we focus on whether the neighborhood has been stable over time, we define the substitution costs in such a way that there is no cost of matching two 'stable' transitions of neighborhood types (e.g. '11' and '33') and two 'unstable' transitions of neighborhood types (e.g. '12' and '32'), while the cost of matching a 'stable' transition with a 'unstable' transition (e.g. '11' and '32') is 1. Based on the "OMstran" method, the distance between neighborhood sequences '1, 1, 1, 1, 1' and '3, 2, 3, 3, 3' is 3, which is larger than the distance 4 between '1, 1, 1, 1, 1' and '1, 2, 3, 1, 2'. Comparatively, the Hamming distance will produce distances of 5 for the former and 3 for the latter.

3.4 | Classifying Neighborhood Sequences

The distance matrix between neighborhood sequences produced by each of the five SA methods was fed into the agglomerative hierarchical clustering for acquiring clusterings of neighborhood sequences. Compared with the *k*-means clustering algorithm used for neighborhood segmentation, the agglomerative hierarchical clustering algorithm starts by considering each observation (a neighborhood sequence) as a cluster and merges clusters at each step based on distances as well as a selected criterion. Here, Ward's minimum variance criterion was adopted which is aimed at minimizing the total within-cluster variance at each merging step (Ward, 1963). The hierarchical clustering process can be visualized by a dendrogram which also displays the distances between merged clusters. Since a large jump in distance is typically related to distinct clusters, an appropriate number of clusters could be obtained based on the selection of a distance cutoff by inspecting the dendrogram. It should be noted that the resulting number of neighborhood trajectory clusters can vary across five SA methods.

3.5 | Evaluation Measures for Sequence Clusterings

The Rand index assesses the similarity of two clusterings by counting all pairs of observations whose assignments agree between the two clusterings (Rand, 1971). If for n observations, a is the number of pairs of observations which are in the same cluster in both clusterings and b is the number of pairs of observations which are in different clusters in both clusterings, then Rand Index (RI) is defined as follows (Equation (2)):

$$\mathrm{RI} = \frac{2(a+b)}{n(n-1)}. \tag{2}$$

We adopted an extension of RI, the adjusted Rand Index (ARI) (Hubert and Arabie, 1985) which is corrected for chance as an evaluation measure for the neighborhood sequence clusterings based on five SA methods. ARI is defined Equation (3):

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{E(RI)}}{\max(\mathrm{RI}) - \mathrm{E(RI)}}, \tag{3}$$

where $\mathrm{E(RI)}$ is the expectation of RI and $\max(\mathrm{RI})$ is the maximum of RI. $\mathrm{ARI} = 1$ means the two clusterings under comparison are identical, whereas ARI being close to 0 suggests the two clusterings are far from identical and can be considered as independent of each other. A large ARI value is an indication of a high level of robustness of the SA methods under comparison. It suggests that these SA methods find similar neighborhood sequence characteristics. In addition to calculating one ARI value for the study area (all the 50 MSAs), we applied the index to individual MSAs to look at the spatial variations in this index. It is possible that some MSAs present very similar neighborhood sequence clusterings based on different SA methods and thus it does not matter much in terms of the SA method selection, while other MSAs are very sensitive to the choice of the SA method. We also exploited the structure of the confusion matrix for every two neighborhood sequence clusterings to match clustering labels towards more meaningful visualizations across five clusterings.

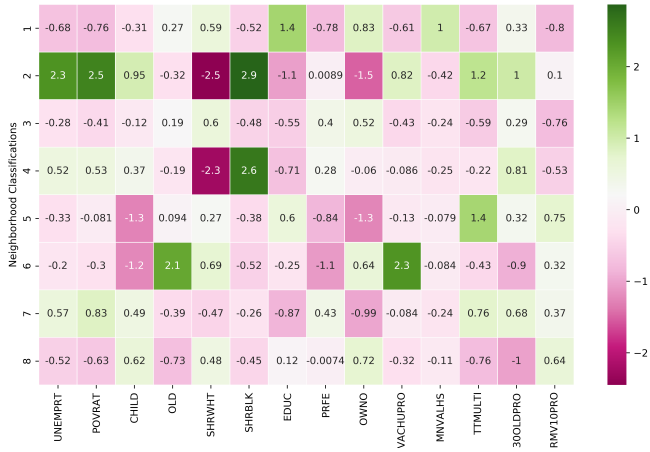

FIGURE 2 Heatmap of median z-scores for each neighborhood type.

4 | RESULTS

4.1 | Neighborhood Types and Compositions

After applying the k-means clustering to $25,961 * 5$ census tracts with 14 variables while varying the number of clusters $k = 2, 3, \ldots, 15$, we obtain 14 clusterings each of which could be the potential neighborhood segmentation scheme. When $k = 2$ and 3, the resulting clusterings give the largest average silhouette coefficients, 0.276 and 0.216 respectively. The coefficient drops to 0.148 and 0.147 for $k = 4$ and 5; as k continues to increase, the coefficient increases - 0.156, 0.156 and 0.157 for $k = 6, 7, 8$. For $k = 9$, the coefficient drops to 0.128 and fails to increase to 0.15 as k continues to increase. Based on the pattern of the average silhouette coefficients, we consider $k = 8$ as the "appropriate" number of clusters for the neighborhood segmentation since it is a local maxima for the average silhouette coefficient and gives more details than $k = 2$ or 3. Figure 2 displays the median z-scores of all 14 variables for each of the eight neighborhood clusters. It should be noted that the ordering of the neighborhood clusters (or types) is arbitrary and clusters with numerically closer labels should not be interpreted as being more similar. A descriptive summary of the composition of each neighborhood type is given in Table 3. Looking at the histogram of the neighborhood classifications per census year in Figure 3, we observe that neighborhood types 3 and 8 are more common in the 50 MSAs under study from 1970 to 2010 while type 6 is the least common.

4.2 | Neighborhood Sequence Patterns

4.2.1 | Descriptive Statistics

Since there are eight unique neighborhood types over five periods, potentially there could be $8^5 = 32,768$ unique neighborhood sequences of length 5. However, for $25,961$ sequences within the 50 largest U.S. MSAs we examine, we observe only 2,853 unique sequences, meaning that only 8.7% potential unique sequences are realized. Figure 4 shows the histogram of the top 20 most common neighborhood sequences: 4 are sequences exempt from any change

TABLE 3 Eight neighborhood classifications and compositions.

Classification	Composition
1	White, educated, wealthy, owners
2	Black, high poverty, high unemployment, renters, older homes
3	White, less educated, blue collar
4	Black, medium poverty and unemployment, less vacant and older homes
5	Few kids, multiunit housing, renters, recent in-movers
6	Old residents, white, vacant homes
7	Mixed race, blue collar
8	Kids, owners, single-family homes, new homes

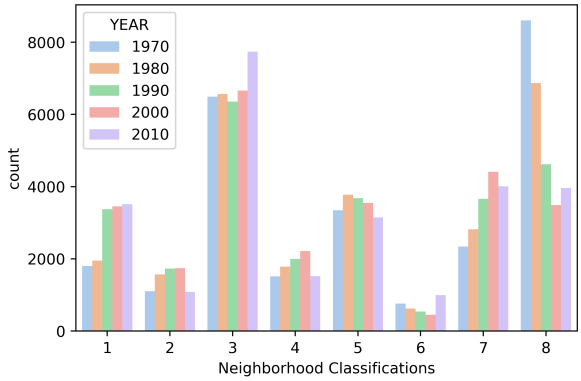

FIGURE 3 Histogram of Neighborhood Classifications per Census Year.

meaning that the tract remains in the same neighborhood type from 1970 to 2010 ('3,3,3,3,3', '8,8,8,8,8', '7,7,7,7,7' and '5,5,5,5,5'), and the other 16 experienced two neighborhood types and are characterized by one state change - 14 changed in 2010 ('8,8,8,8,3', '5,5,5,5,3' etc.), 1 in 1980 ('8,3,3,3,3') and 1 in 1990 ('8,8,3,3,3'). At first sight, it appears that the census tracts were quite stable in terms of the neighborhood composition. However, the top 20 most common sequences account for only 4% of the 25,961 sequences. Meanwhile, 2,117 out of 2,853 unique sequences contain less than 2 successive identical values (e.g. '8,3,1,1,3', '5,5,1,1,3') which we interpret as having experienced "frequent" changes. For example, the spatial distributions of neighborhood types within the San Francisco Metropolitan Area from 1970 to 2010 are visualized in Figure 5. It is obvious that the tract in the southeastern corner has experienced drastic changes ('3,8,1,8,6'). Together, these results suggest that, because neighborhood sequences are quite diverse, it is possible that different sequence distance metrics (or SA methods) produce divergent pairwise distances, thus resulting in divergent clusterings of neighborhood sequences.

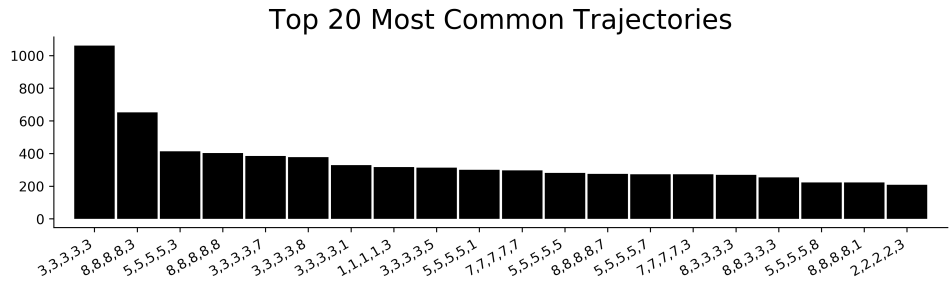

FIGURE 4 Histogram of top 20 most common neighborhood trajectories 1970-2010.

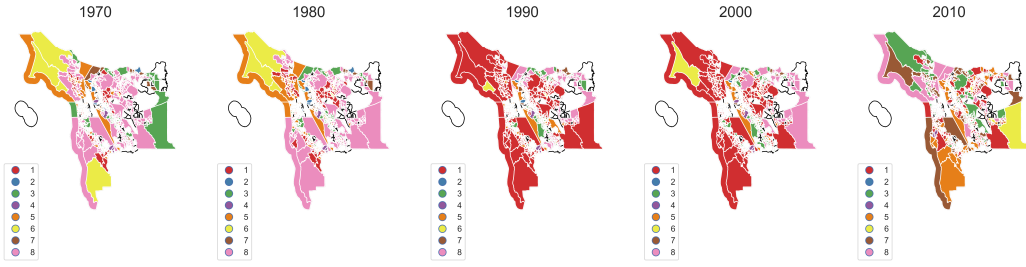

FIGURE 5 Neighborhood types over 1970-2010 in San Francisco Metropolitan Area.

4.2.2 | Clusterings of Neighborhood Sequences

The five SA methods were then applied to the neighborhood sequences to acquire five sequence distance matrices. Each distance matrix was then used in the agglomerative hierarchical clustering with Ward's minimum variance criterion. We obtained the appropriate number of clusters by visually inspecting the hierarchical clustering dendrogram and truncating the dendrogram with a distance cutoff where there is a large gap in the tree. An example of the truncated dendrogram based on "OMecenter" is given by Figure 6 in which a distance cutoff 350 was selected to truncate the dendrogram and form clusters. For the first four SA methods "Hamming", "OMecenter", "OMtranr" and "OMarbitr", a nine-cluster solution was deemed to be appropriate while a six-cluster solution was deemed to be appropriate for the OM variate - "OMstran". The five clusterings of neighborhood sequences within the San Francisco Metropolitan Area are visualized in Figure 7 [3]. It should be noted that the neighborhood sequence cluster labels across five maps are not perfectly comparable as they come from clustering processes based on different SA methods. Though we have exploited the confusion matrix for attempting to match sequence clusters across clusterings, we cannot say that the same sequence cluster label across different clusterings in Figure 7 represents the same neighborhood sequence compositions, and this is precisely what we aim to investigate in this paper - to examine whether different SA methods produce converging neighborhood sequence clustering results.

Compositions of the nine clusters of neighborhood sequences for all the 50 MSAs based on "OMecenter" are visualized in Figure 8 in which 8 different colors represent 8 neighborhood types identified in Table 3. The colors are consistent with Figure 5 which displays spatial distributions of neighborhood types at census years 1970, 1980, 1990,

[3]The clusterings for all 50 MSAs are not visualized in the article because the census tracts are not visually discernible on the article page, but it is available upon request.

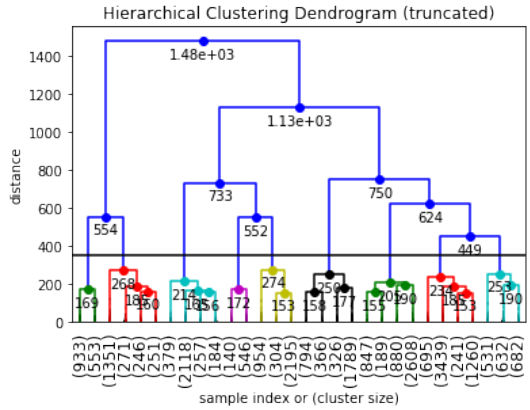

FIGURE 6 Truncated dendrogram based on "OMecenter".

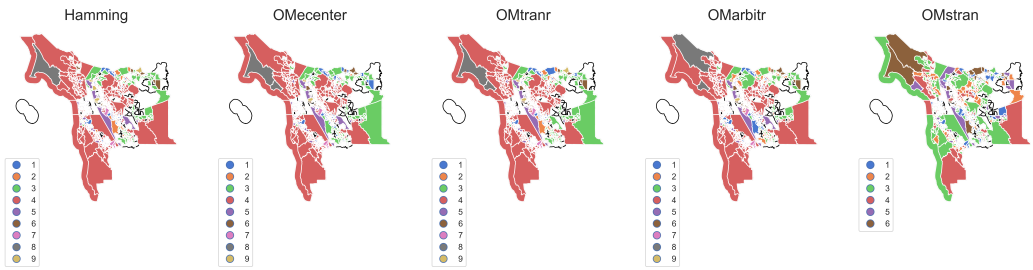

FIGURE 7 Five clusterings of neighborhood sequences in San Francisco Metropolitan Area.

2000, and 2010. It is immediately clear that each of Neighborhood Sequence Clusters 1, 3-9 are mostly comprised of sequences dominated by one neighborhood type. For example, Sequence Cluster 1 is dominated by neighborhood sequences experiencing type 3 "White, less educated, blue collar" for at least two consecutive census years while Neighborhood Sequence Cluster 3 is dominated by sequences experiencing type 8 "Kids, owners, single-family homes, new homes" for at least two consecutive census years. We found similar patterns for neighborhood sequence clusterings based on "Hamming", "OMtranr", and "OMarbitr". In contrast, the clustering based on "OMstran" which evaluates the distance between sequences of transitions across neighborhood types over time produces very different compositions as displayed in Figure 9. Here, none of the six neighborhood sequence clusters is dominated by sequences experiencing one neighborhood type for several census years. The clusters are differentiated by the frequency of changes in neighborhood types over time as well as the timing of the changes: Neighborhood Sequence Cluster 1 is primarily comprised of sequences which were stable from 1970 to 2000 but experienced a change in 2010 irrespective of the stable neighborhood type in the initial census year (1970) and the type in 2010, while Neighborhood Sequence Cluster 3 is mainly comprised of sequences experiencing more changes - in both 1980 and 2010. As it stands, the interpretation of the clustering based on "OMstran" is considerably different from the other four, and the choice of the method should be guided by the research question. We shall also see the difference in the sequence clusterings based on using the first four SA methods in the next subsection.

4.2.3 | Similarity between Neighborhood Sequences Clusterings

To further quantify the difference in neighborhood sequence clusterings based on five different SA methods, we have calculated ARIs between any pair of clusterings which are displayed in Figure 10. It turns out that the most similar clusterings are those based on "Hamming" and "OMarbitr". The ARI value for the pair is 0.76. Both methods set an arbitrary constant value to the substitution cost as well as the insertion/deletion cost. The two data-driven SA methods "OMecenter" and "OMtranr" are somewhat concordant with an ARI value of 0.6. This is the smallest value, if not taking "OMstran" into account. Comparing compositions of sequence clusters based on "OMecenter" (Figure 8) and "OMtranr" (Figure 11), we see Sequence Cluster 1 in "OMecenter" consists of sequences experiencing several periods of neighborhood type 8 while "OMtranr" does not. The reason is due to the smaller cost of substituting type 3 with type 8 based on the Euclidean distance between cluster centers than that based on the empirical transition rates, where transitions between 3 and 8 are rare.

A relatively high value (0.72) of ARI is observed between "OMtranr" (Figure 11) and "Hamming" (Figure 12). This may seem a little surprising as the substitution costs for "OMtranr" are based on empirical transition rates across time while "Hamming" simply sets all substitution costs to be a constant 1. Therefore, we would expect the former to be more informative. It turns out that the observed transition rates between distinct neighborhood types are pretty small, resulting in similar costs of substituting between any pair, which is also part of the reason why transition rates-based costs are not suggested in empirical studies such as the life course research (Studer and Ritschard, 2016).

As expected, because "OMstran" reformulates sequences of neighborhood types into sequences of transitions of neighborhood types and the operation costs are set accordingly, the resultant clustering is very different from the other four. For example, based on the "OMstran", the neighborhood sequence '5,5,1,1,8' is assigned to the same cluster as '8,8,1,1,7' while '3,5,1,1,1' is assigned to a different cluster since the first two experienced more frequent changes. The clusterings are just the opposite based the other four methods.

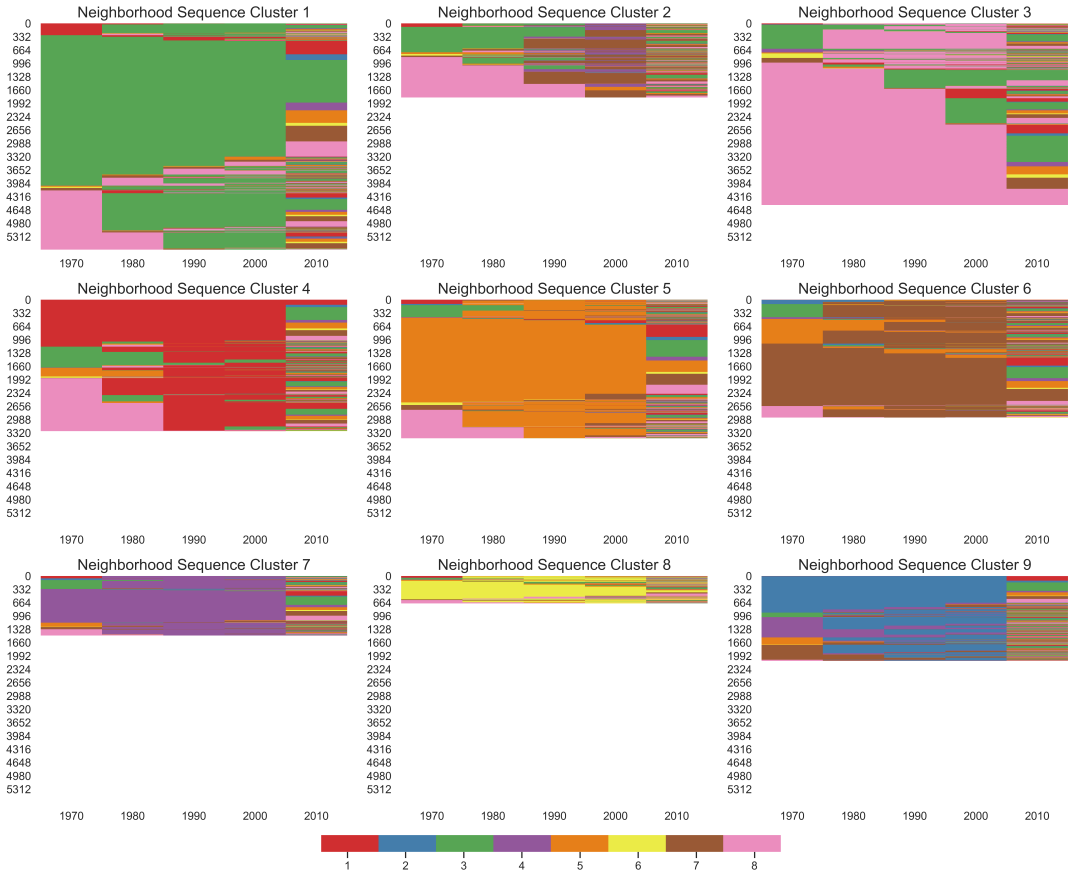

FIGURE 8 Compositions of nine neighborhood sequence clusters based on "OMecenter" for the overall 50 largest MSAs. Each of eight colors represents one neighborhood type identified in Table 3 and the colors are consistent with Figure 5 which displays spatial distributions of neighborhood types at census years 1970, 1980, 1990, 2000, and 2010. For each Neighborhood Sequence Cluster, y axis shows the accumulated number of sequences in that cluster.

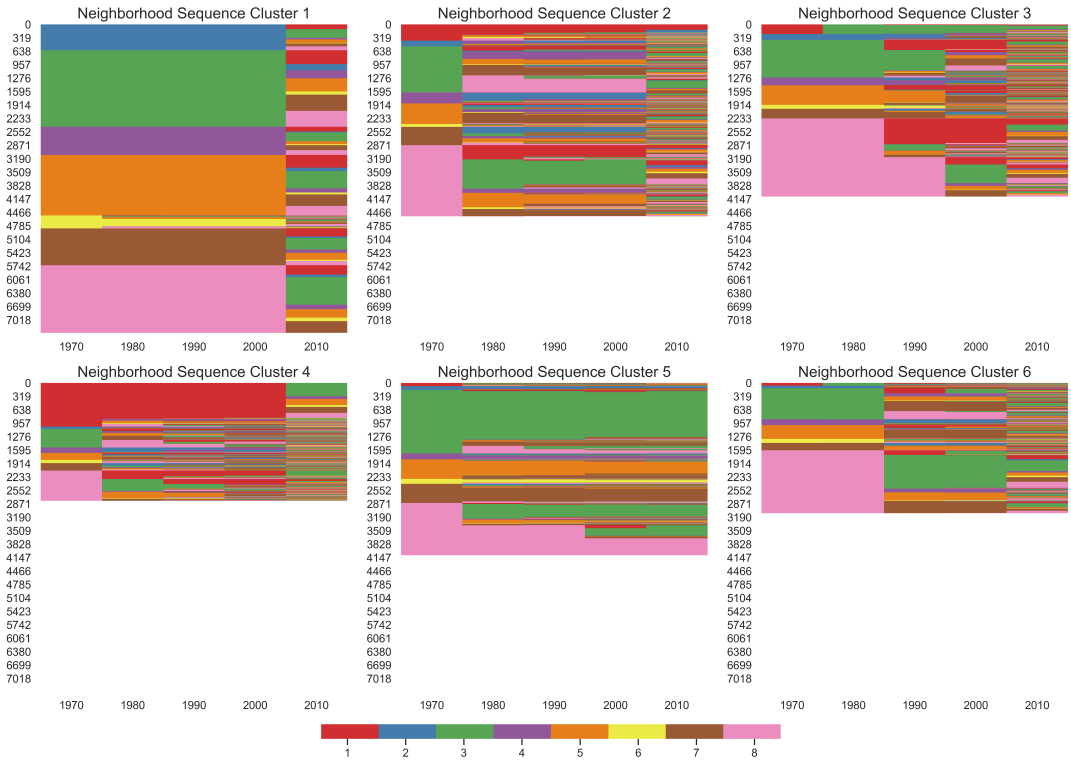

FIGURE 9 Compositions of six neighborhood sequence clusters based on "OMstran" for the 50 largest MSAs. Each of eight colors represents one neighborhood type identified in Table 3 and the colors are consistent with Figure 5 which displays spatial distributions of neighborhood types at census years 1970, 1980, 1990, 2000, and 2010. For each Neighborhood Sequence Cluster, y axis shows the accumulated number of sequences in that cluster.

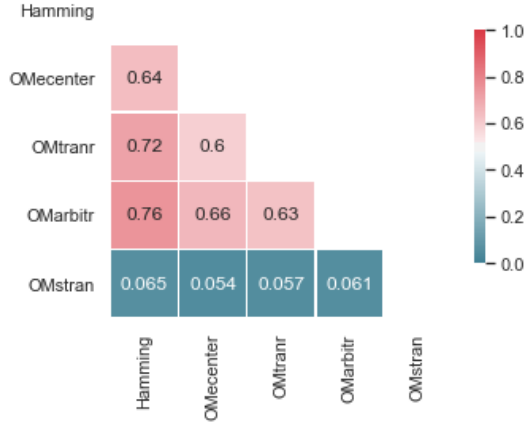

FIGURE 10 Pairwise similarities between cluster assignments (ARI) for the overall 50 largest MSAs.

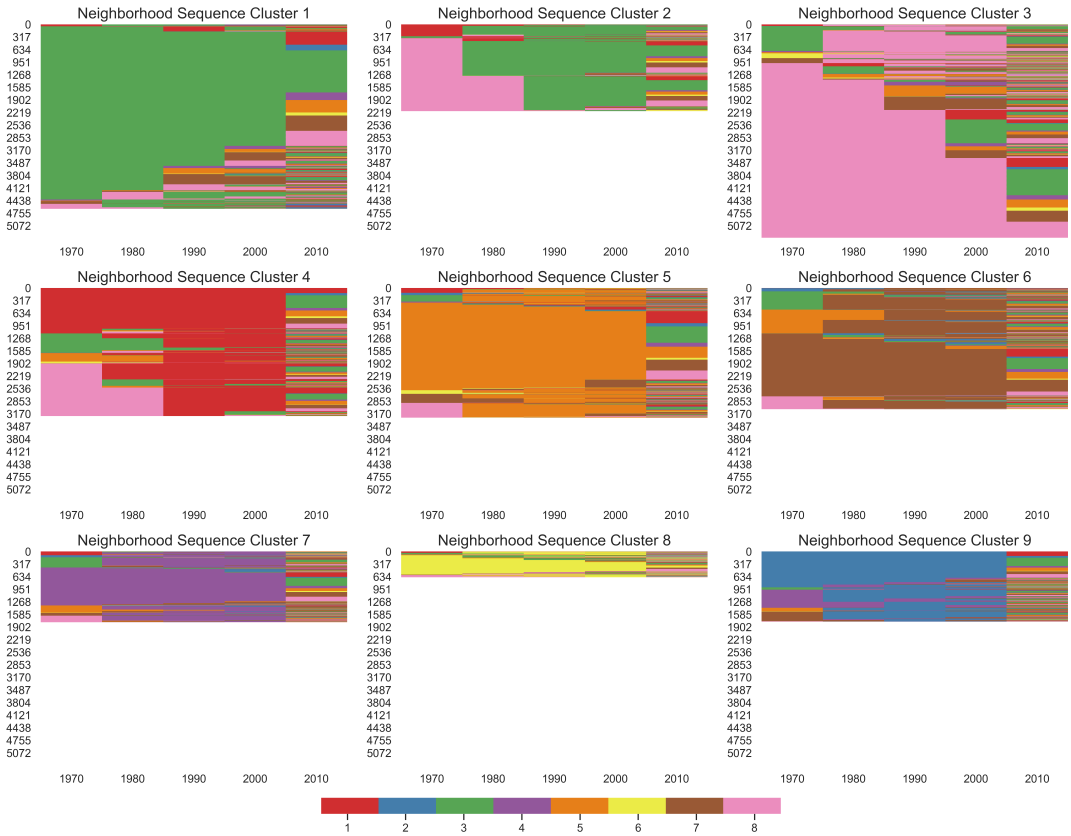

FIGURE 11 Compositions of nine neighborhood sequence clusters based on "OMtranr" for the 50 largest MSAs. Each of eight colors represents one neighborhood type identified in Table 3 and the colors are consistent with Figure 5 which displays spatial distributions of neighborhood types at census years 1970, 1980, 1990, 2000, and 2010. For each Neighborhood Sequence Cluster, y axis shows the accumulated number of sequences in that cluster.

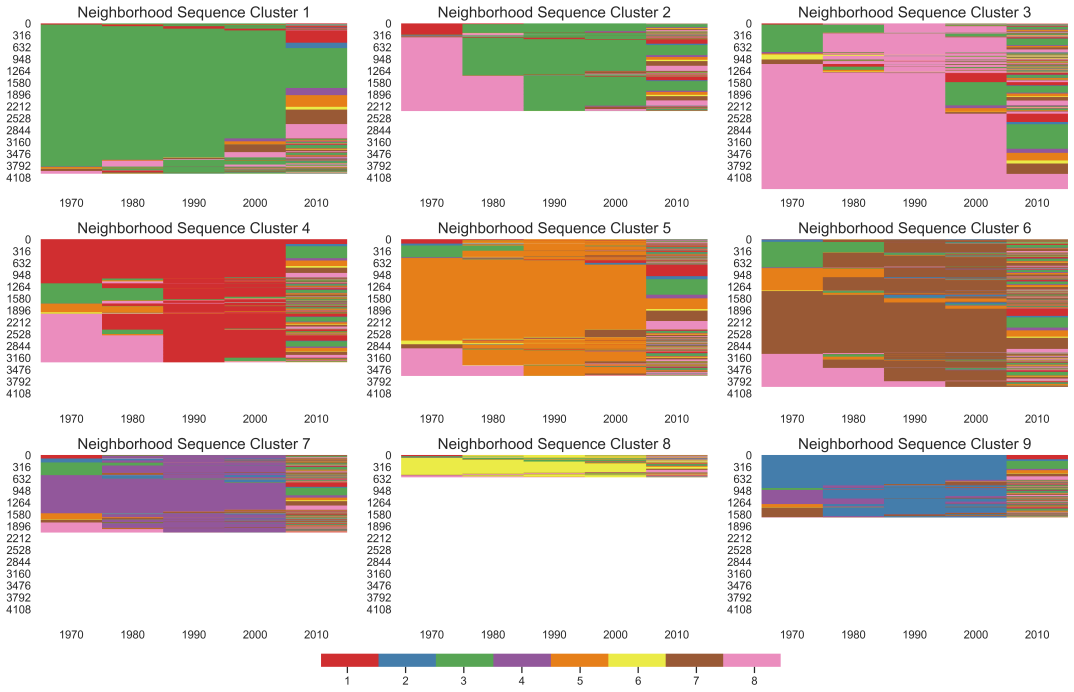

FIGURE 12 Compositions of nine neighborhood sequence clusters based on "Hamming" for the 50 largest MSAs. Each of eight colors represents one neighborhood type identified in Table 3 and the colors are consistent with Figure 5 which displays spatial distributions of neighborhood types at census years 1970, 1980, 1990, 2000, and 2010. For each Neighborhood Sequence Cluster, y axis shows the accumulated number of sequences in that cluster.

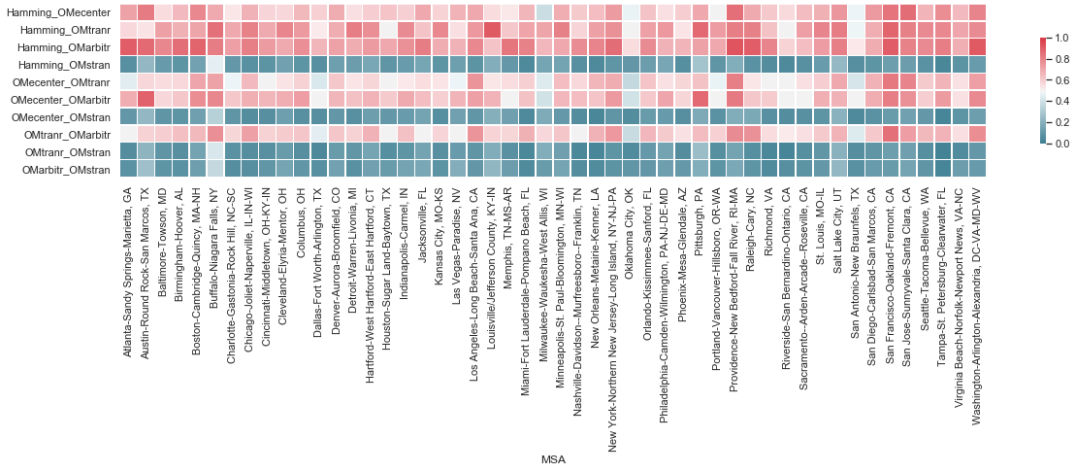

FIGURE 13 Pairwise similarities between cluster assignments (ARI) for each MSA.

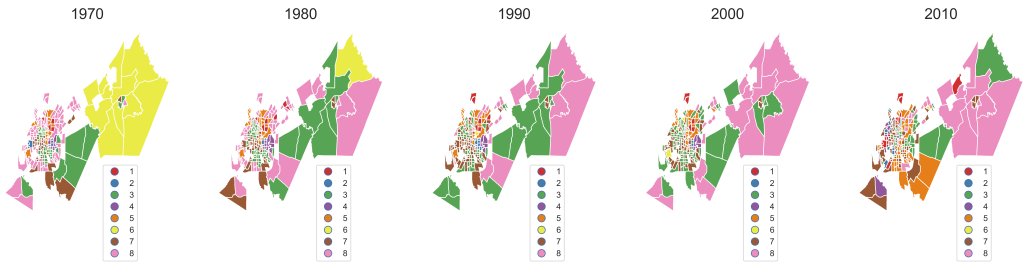

FIGURE 14 Neighborhood types over 1970-2010 in the San Antonio-New Braunfels MSA.

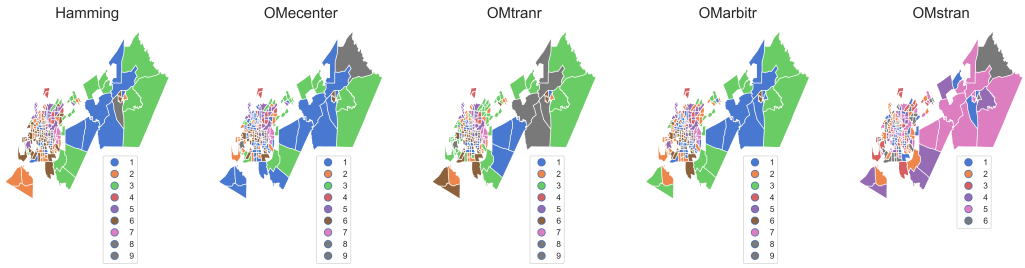

FIGURE 15 Five clusterings of neighborhood sequences in the San Antonio-New Braunfels MSA.

4.2.4 | Spatial Variations

We further investigate the spatial variations of pairwise similarity between clusterings across 50 MSAs. It turns out that ARI varies substantially as shown in Figure 13. For instance, the ARI between clusterings based on "Hamming" and "OMtranr" reaches as high as 0.91 for the Louisville Metropolitan Area, and as low as 0.41 for the San Antonio-New Braunfels MSA. The neighborhood sequence clustering based on 'OMtranr" for the former MSA is more similar to that based on "Hamming", and the clustering given by "OMstran" is also more similar to the other four. The five maps of neighborhood types from 1970 to 2010 for the San Antonio-New Braunfels MSA are visualized in Figure 14 together with the five clusterings of trajectories in Figure 15. It is visually obvious that the five clusterings are more different (especially for those based on "Hamming" and "OMtranr") than what we observe from the San Francisco Metropolitan Area in Figure 7.

Another robust case is presented by the Providence Metropolitan Area. The neighborhood segmentations from 1970 to 2010 are visualized in Figure 16 and the five clusterings for neighborhood trajectories are visualized in Figure 17. As is shown in Figure 13, the ARI value is generally higher for any pair of neighborhood sequence clusterings.

5 | DISCUSSION AND CONCLUSION

The alignment-based sequence analysis (SA) methods represent a useful toolkit for uncovering emergent patterns in the evolution of neighborhood socioeconomic characteristics and recently have been applied widely to neighborhood research. However, the fact that these methods rely on multiple assumptions about the evolution of the sequences makes them subject to potential criticism. This article attempts to shed light on the extent to which several SA methods

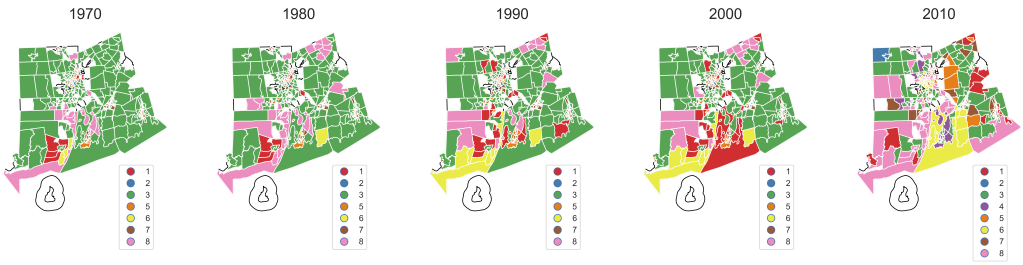

FIGURE 16 Neighborhood types over 1970-2010 in the Providence Metropolitan Area.

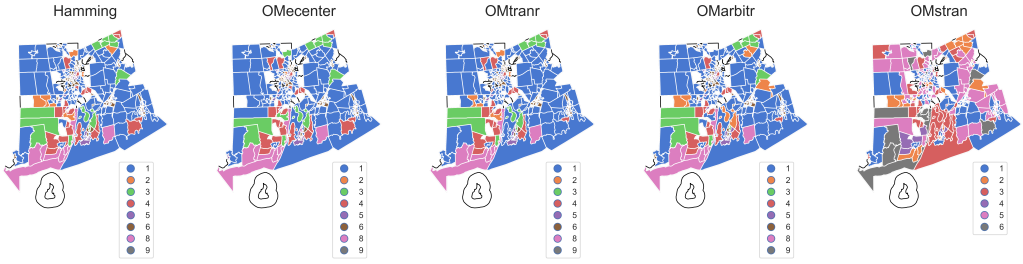

FIGURE 17 Five clusterings of neighborhood sequences in the Providence Metropolitan Area.

are robust in their ability to provide consistent and converging neighborhood sequence typology, and which sequence pattern could be easily revealed by each method.

We applied five alignment-based SA methods to a common longitudinal census dataset in U.S. - the 50 largest MSAs at census years 1970, 1980, 1990, 2000 and 2010. We demonstrate that the neighborhood sequence typology is sensitive to the choice of OM method and the operation costs, and the sensitivity demonstrates spatial heterogeneity. The typology is more sensitive to the method and cost choice in some MSAs such as the San Antonio-New Braunfels MSA in Texas, and less sensitive in MSAs including the San Francisco Metropolitan Area in California and the Providence Metropolitan Area in Rhode Island. In addition, a method or cost choice could be effective in revealing one particular characteristic of neighborhood evolution while failing to provide useful information in other aspects, and thus, researchers should take caution both when adopting a method and interpreting results.

Future work should provide additional testing beyond the five SA methods examined in this paper to include more variants of the OM method. It is also important to delve deep into the compositions of neighborhood sequence typology in the hope of providing more insights into the interpretation of the typology for each SA method employed. Another direction would be to propose extensions to existing OM methods which would be sensitive to the spatial context and take account of potential spatial dependence.

references

Abbott, A. (1995) Sequence Analysis: New Methods for Old Ideas. *Annual Review of Sociology*, **21**, 93–113. URL: `http://www.annualreviews.org/doi/10.1146/annurev.so.21.080195.000521`.

Aisenbrey, S. and Fasang, A. E. (2010) New life for old ideas: The "second wave" of sequence analysis bringing the "course" back

into the life course. *Sociological Methods & Research*, **38**, 420–462. URL: https://doi.org/10.1177/0049124109357532.

Beauregard, R. A. (1990) Trajectories of neighborhood change: The case of gentrification. *Environment and Planning A: Economy and Space*, **22**, 855–874. URL: https://doi.org/10.1068/a220855.

Biemann, T. (2011) A transition-oriented approach to optimal matching. *Sociological Methodology*, **41**, 195–221. URL: https://doi.org/10.1111/j.1467-9531.2011.01235.x.

Carrillo, H. and Lipman, D. (1988) The Multiple Sequence Alignment Problem in Biology. *SIAM Journal on Applied Mathematics*, **48**, 1073–1082. URL: http://epubs.siam.org/doi/10.1137/0148063.

Cha, S.-H. (2007) Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical models and Methods in Applied Sciences*, **1**, 300–307.

Delmelle, E. C. (2015) Five decades of neighborhood classifications and their transitions: A comparison of four US cities, 1970–2010. *Applied Geography*, **57**, 1–11. URL: http://linkinghub.elsevier.com/retrieve/pii/S0143622814002860http://dx.doi.org/10.1016/j.apgeog.2014.12.002.

— (2016) Mapping the DNA of urban neighborhoods: Clustering longitudinal sequences of neighborhood socioeconomic change. *Annals of the American Association of Geographers*, **106**, 36–56.

— (2017) Differentiating pathways of neighborhood change in 50 U.S. metropolitan areas. *Environment and Planning A*, **49**, 2402–2424. URL: http://journals.sagepub.com/doi/10.1177/0308518X17722564.

Folch, D. C., Arribas-Bel, D., Koschinsky, J. and Spielman, S. E. (2016) Spatial variation in the quality of American Community Survey estimates. *Demography*, **53**, 1535–1554. URL: http://dx.doi.org/10.1007/s13524-016-0499-1.

Hollister, M. (2009) Is Optimal Matching Suboptimal? *Sociological Methods & Research*, **38**, 235–264. URL: http://journals.sagepub.com/doi/10.1177/0049124109346164.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of classification*, **2**, 193–218.

Lee, K. O., Smith, R. and Galster, G. (2017a) Neighborhood trajectories of low-income U.S. households: An application of sequence analysis. *Journal of Urban Affairs*, **39**, 335–357. URL: https://doi.org/10.1080/07352166.2016.1251154.

— (2017b) Subsidized housing and residential trajectories: An application of matched sequence analysis. *Housing Policy Debate*, **27**, 843–874. URL: https://doi.org/10.1080/10511482.2017.1316757.

Li, Y. and Xie, Y. (2018) A new urban typology model adapting data mining analytics to examine dominant trajectories of neighborhood change: A case of Metro Detroit. *Annals of the American Association of Geographers*, **0**, 1–25. URL: https://doi.org/10.1080/24694452.2018.1433016.

Mikelbank, B. A. (2011) Neighborhood dã©jã vu: Classification in metropolitan cleveland, 1970-2000. *Urban Geography*, **32**, 317–333. URL: https://doi.org/10.2747/0272-3638.32.3.317.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, **66**, 846–850.

Reibel, M. (2011) Classification approaches in neighborhood research: Introduction and review. *Urban Geography*, **32**, 305–316. URL: https://doi.org/10.2747/0272-3638.32.3.305.

Rey, S. J., Anselin, L., Folch, D. C., Arribas-Bel, D., Sastré Gutiérrez, M. L. and Interlante, L. (2011) Measuring spatial dynamics in metropolitan areas. *Economic Development Quarterly*, **25**, 54–64. URL: http://dx.doi.org/10.1177/0891242410383414.

Ritschard, G. and Studer, M. (eds.) (2018) *Sequence Analysis and Related Approaches*. Springer.

Robette, N. and Bry, X. (2012) Harpoon or bait? a comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, **116**, 5–24.

Schwirian, K. P. (1983) Models of neighborhood change. *Annual Review of Sociology*, **9**, 83–102. URL: https://doi.org/10.1146/annurev.so.09.080183.000503.

Singleton, A., Pavlis, M. and Longley, P. A. (2016) The stability of geodemographic cluster assignments over an intercensal period. *Journal of Geographical Systems*, **18**, 97–123. URL: https://doi.org/10.1007/s10109-016-0226-x.

Singleton, A. D. and Longley, P. A. (2009) Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, **29**, 289 – 298. URL: http://www.sciencedirect.com/science/article/pii/S0143622808000726.

Singleton, A. D. and Spielman, S. E. (2014) The past, present, and future of geodemographic research in the United States and United Kingdom. *The Professional Geographer*, **66**, 558–567. URL: https://doi.org/10.1080/00330124.2013.848764. PMID: 25484455.

Studer, M. and Ritschard, G. (2016) What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **179**, 481–511. URL: http://dx.doi.org/10.1111/rssa.12125.

Temkin, K. and Rohe, W. (1996) Neighborhood change and urban policy. *Journal of Planning Education and Research*, **15**, 159–170. URL: https://doi.org/10.1177/0739456X9601500301.

Vinga, S. and Almeida, J. (2003) Alignment-free sequence comparison—a review. *Bioinformatics*, **19**, 513–523. URL: http://dx.doi.org/10.1093/bioinformatics/btg005.

Ward, J. H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

Wei, F. and Knox, P. L. (2014) Neighborhood change in Metropolitan America, 1990 to 2010. *Urban Affairs Review*, **50**, 459–489. URL: https://doi.org/10.1177/1078087413501640.

Wong, K. M., Suchard, M. A. and Huelsenbeck, J. P. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476. URL: http://science.sciencemag.org/content/319/5862/473.

Zielezinski, A., Vinga, S., Almeida, J. and Karlowski, W. M. (2017) Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, **18**, 186. URL: https://doi.org/10.1186/s13059-017-1319-7.

Zwiers, M., Kleinhans, R. and Van Ham, M. (2017) The path-dependency of low-income neighbourhood trajectories: An approach for analysing neighbourhood change. *Applied Spatial Analysis and Policy*, **10**, 363–380. URL: https://doi.org/10.1007/s12061-016-9189-z.