Computers, Environment and Urban Systems 81 (2020) 101480

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Sensitivity of sequence methods in the study of neighborhood change in the United States

Wei Kang[a,*], Sergio Rey[a], Levi Wolf[b], Elijah Knaap[a], Su Han[a]

[a] Center for Geospatial Sciences, University of California, Riverside, USA
[b] School of Geographical Sciences, University of Bristol, UK

ARTICLE INFO

Keywords:
Sequence analysis
Neighborhood change
Optimal matching
Clustering
Sensitivity
Geodempraphics

ABSTRACT

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 seen many applications in urban neighborhood change research. However, it is unclear to what extent these methods are robust in terms of producing consistent and converging neighborhood sequence typologies. This article sheds light on this issue by applying four sequence analysis methods to the same data set – 50 largest Metropolitan Statistical Areas (MSAs) of the United States from 1970 to 2010, and finds that these methods do not provide converging neighborhood sequence typologies, and their behavior varies across MSAs, thus prohibiting meaningful comparisons of similar studies. MSAs with higher average household income in 1970 tend to be less sensitive to the choice of the SA methods. In other words, when investigating neighborhood change in these MSAs, different SA methods tend to produce a more converging neighborhood sequence typology. Comparatively, for MSAs hosting neighborhoods which have experienced frequent changes during the period 1970–2010, they are less likely to produce similar typologies with different SA methods. In addition, there is a big difference in the neighborhood sequence typology between applying the classic SA methods with varying costs and using the SA variant focusing on the second-order sequence property. After comparing the behavior of these methods, we highlight one method ("OMecenter") which leverages the socioeconomic similarities of neighborhoods and suggest researchers consider it as the building block towards designing a meaningful sequence analysis method for neighborhood change research.

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, 2015, 2016, 2017; Li & Xie, 2018; Ling & Delmelle, 2016; Patias, Rowe, & Cavazzi, 2019; Zwiers, Kleinhans, & Van Ham, 2017). 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 within a contextual mode of analysis. Instead of measuring and tracking the numerical changes to specific variables or composite indices, like median household income or education attainment, within the variables paradigm (Abbott, 1997), neighborhood changes are conceptualized as ensembles of variables, and are evaluated based on the change of type. As demonstrated by Spielman and Singleton (2015), this latter approach is especially promising to address statistical concerns introduced by small area estimates of commonly used census surveys (e.g. ACS 5-year estimates at the census tract level).

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.[1] Two types of SA methods are at researchers'

* Corresponding author at: Center for Geospatial Sciences, University of California, Riverside, Riverside, CA 92521, USA.
E-mail addresses: weikang@ucr.edu (W. Kang), sergio.rey@ucr.edu (S. Rey), levi.john.wolf@bristol.ac.uk (L. Wolf), elijah.knaap@ucr.edu (E. Knaap), suhan@ucr.edu (S. Han).

[1] Lee et al. (2017a, 2017b); Greenlee (2019) adopted a similar strategy to investigate the neighborhood sequences experienced by households. We note that in these studies, the sequence was not organized around a focal neighborhood whose location is "fixed" over time (place-based), but rather were organized based on which neighborhood the focal household was located at (or moved into if household experienced the displacement) (household-based). Our focus in this article is on place-based neighborhood change.

https://doi.org/10.1016/j.compenvurbsys.2020.101480
Received 13 February 2019; Received in revised form 16 January 2020; Accepted 1 March 2020
Available online 13 March 2020
0198-9715/

disposal: "stepwise approaches," such as Markov Chains, view the sequence as being generated stochastically and model the probabilities of transitions between neighborhood types over time (Delmelle, 2015); "whole sequence approaches," mainly the optimal matching (OM) analysis, meanwhile, view the sequence from a holistic perspective and evaluate the pairwise similarity between neighborhood sequences 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. Compared with the former, the "whole sequence approaches" fall within the pattern recognition data modeling tradition, and could identify predominant as well as irregular "outlier" neighborhood change pathways, which could have important implications for the development of the neighborhood change theory.

The OM method, originally developed for matching protein and DNA sequences in biology (Carrillo & Lipman, 1988; Wong, Suchard, & Huelsenbeck, 2008) and used extensively for analyzing strings in computer science, has become the dominant SA technique in the neighborhood literature (Delmelle, 2016, 2017; Li & Xie, 2018; Patias et al., 2019). 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 (Li & Xie, 2018) or empirical transition probabilities between neighborhood types over two consecutive time points (Delmelle, 2016, 2017; Patias et al., 2019).

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) showed that the alignment of genomic data and thus the resultant similarity values were 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) argued that the direct application of OM analyses to life course data was inappropriate since the life course was an unfolding process, whereas DNA sequences for which OM was designed originally, shared 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 (Ritschard & Studer, 2018; Robette & Bry, 2012; Studer & Ritschard, 2016; Studer, Ritschard, Tabin, & Perriard, 2014). 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 were simulated in Studer, Struffolino, and Fasang (2018)). The other major difference is that the neighborhood types constituting a sequence in neighborhood change research are constructs from a specified clustering algorithm with selected neighborhood-level attributes. Thus, valuable information such as the distance between clusters could be used as the prior knowledge for defining parameters in sequence analysis. This information is unavailable in life course studies as the sequence is comprised of natural states such as employment, unemployment, and school, the difference between which is hardly straightforward to quantify.

This article focuses on the application of the SA methods to neighborhood change research and explores two related issues. We examine the relationship between neighborhood sequence typology and operation costs. We are particularly interested in the sensitivity of neighborhood sequence typology to the choice of operation costs in the OM algorithm, that is, whether a small change in the operation costs will result in a much different typology. We are also interested in whether such sensitivity displays spatial heterogeneity. In other words, whether cities with certain characteristics are less sensitive to the choice of operation costs.

We examine these issues through an empirical analysis of four SA methods, the cost of substitution, insertion or deletion for each of which is varied, which are considered applicable for uncovering neighborhood sequence patterns from different aspects. Three of them fall within the classic OM scheme with different choices of cost operations while the fourth operationalizes the OM scheme on sequences of transitions of neighborhoods, utilizing the second-order property of a sequence. We applied 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 the operation cost and this sensitivity displays spatial heterogeneity. MSAs with higher average household income in the beginning year (1970) tend to be less sensitive while if they hosted neighborhoods experiencing frequent changes during the period 1970–2010, they are less likely to reach a converging typology with different SA methods. There is a big difference in the neighborhood sequence typology between applying the classic OM methods with varying costs and using the OM variant focusing on the second-order sequence property. While the former reflects contemporaneous experiences and/or the order, the latter emphasizes the stability characteristic of the neighborhood sequence. In terms of selecting the cost of substituting one neighborhood type with another, we recommend researchers to utilize the valuable information provided by the continuous tract-level variables – more specifically, basing the substitution costs on (at least partially) distances between neighborhood segmentation (cluster) centers.

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, four 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 a discussion of the key implications of our findings and the identification of directions for future research.

2. Neighborhood change and sequence analysis

2.1. Urban neighborhood change and the theory

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 (Beauregard, 1990; Schwirian, 1983; Temkin & 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" (Delmelle, 2015, 2016; Patias et al., 2019; Wei & Knox, 2014). 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 "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.

2.2. Sequence analysis

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 (Cha, 2007; Vinga & Almeida, 2003; Zielezinski, Vinga, Almeida, & Karlowski, 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) provided comparative surveys of these methods in the study of life courses such as professional careers and distinguished 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 Fig. 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.

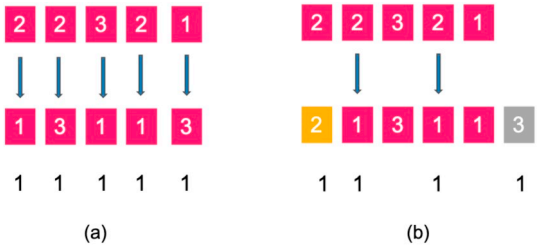


Fig. 1. A small example of calculating OM distance between two short sequences. (a) The cost of substituting any number with a different one is 1 while the cost of inserting or deleting (indel) any number is 2. (b) The cost of substituting any number with a different one is 1 and the cost of inserting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

In practice, OM is usually stated as a dynamic programming problem. Through a series of simulation experiments, Studer and Ritschard (2016) showed that specific characteristics of a sequence could be picked up by appropriately selecting the operation costs or the OM variants, 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 & 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) basesd substitution costs on empirical transition rates across census years. If the empirical transition rate between two neighborhood types was large, the cost of substituting one with the other was small. Later, Delmelle (2017) employed a variant of OM which focused on sequences of transitions between neighborhood types in 50 U.S. MSAs from 1980 to 2010. Other similar research in the U.S. (Lee, Smith, & Galster, 2017a, 2017b) and the Netherlands (Zwiers et al., 2017) adopted another variant of OM which led to a subsequence based distance measure and was 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 in a process known as "geodemographic segmentation" (Reibel, 2011; Rey et al., 2011; Singleton & Spielman, 2014). Uncertainty comes from the geodemographic cluster assignment process where various clustering algorithms could lead to different results (Singleton, Pavlis, & Longley, 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. On the other hand, we do note that this specific aspect of neighborhood research provides abundant information for selecting the operation costs, which life course research does not afford. In other words, the distances between neighborhood segmentations are very meaningful indicators of costs taken to substitute one neighborhood type with another.

3. Data and methods

To examine how neighborhood change classification is sensitive to the choice of the SA method, we selected four 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 four SA approaches measuring the pairwise similarity of neighborhood sequences as well as the subsequent sequence clustering algorithm, and the final evaluation indices.

3.1. Study area and data

Following many existing neighborhood segmentation and neighborhood change analyses (Delmelle, 2015, 2016, 2017; Li & Xie, 2018; Mikelbank, 2011; Wei & Knox, 2014), we adopted 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[2] 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)[3] which provides census tracts in 1970, 1980, 1990, and 2000 with boundaries and attributes recalculated and normalized to 2010. The 2010 sources are a mixture of 2010 long-form census and 2006–2010 American Community Survey (ACS) estimates.

Following earlier studies on geodemographics (Li & Xie, 2018; Singleton & Longley, 2009; Singleton & 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.

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 tracts, we limited 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 contained 25,961 census tracts for each of the 5 census years. The analysis, therefore, proceeded 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 was 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.[4] We applied the clustering algorithm to all 129,805 tracts at once to produce k neighborhood types which were consistent and comparable across space and time. Since feature scaling can impact clustering results significantly, we transformed each variable using z-score standardization relative to each census year.

Performance of the *k*-means clustering algorithm is contingent on the choice of k – number of clusters. We relied on the average silhouette coefficient to select an "appropriate" number of clusters. This coefficient is defined as follows:

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

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 maximized 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 constituted our observations for sequence analyses. We select ed four SA methods, or more specifically four 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. A small value of the insertion/deletion cost allows for a certain level of time distortion while heterogeneous substitution costs across different pairs of neighborhood types indicate researchers' belief of different dissimilarity across pairs. For instance, a downtown neighborhood where poor black live could be considered more different from a suburb where white middle class concentrate than a multiethnic neighborhood.

3.3.1. Hamming

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 infinitely large cost for insertion or deletion. The application of this OM distance metric to neighborhood sequences assumes that the distance between any pair of distinct neighborhood types is identical with a strict focus on contemporaneous similarity between neighborhood sequences. It should be noted that the infinitely large cost for insertion or deletion is equivalent to any value no less than twice the constant substitution cost.

3.3.2. OMtranr

We also examined the "OMtranr" method in which the substitution costs are based on and usually negatively correlated with empirical transition rates between neighborhood types over time. For example, the empirical transition rate from neighborhood type i to j is computed by dividing the number of transitions from i to j by the total number of transitions from i. The transition rate from i to j (p_{ij}) is usually different from the p_{ji} – the empirical rate of transitioning from j to i. To arrive at a symmetric substitution cost matrix which requires the cost of

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

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

[4] We used the k-means algorithm for the initial round of clustering simply because it is a convenient and scalable clustering algorithm. Other multivariate clustering algorithms are equally applicable here.

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

Table 2
Selected sequence analysis approaches for empirical comparison.

Index	Name	Substitution costs	Insertion/ deletion costs
1	Hamming	1	$+\infty$
2	OMtranr	$s_{ij} = 1 - \frac{p_{ij}+p_{ji}}{2}$	1
3	OMecenter	Euclidean distance between cluster centers d_{ij}	$\max(d_{ij})$
4	OMstran	stable-stable = 0, change-change = 0, stable-change = 1	$+\infty$

substituting i with j to be identical to the cost of substituting j with i, common practice is to average p_{ij} and p_{ji} and then subtract it from 1. That is, the substitution cost between i and j is $s_{ij} = 1 - \frac{p_{ij}+p_{ji}}{2}$. The cost of inserting or deleting any neighborhood type is set to be 1, which is close to the maximum of the varying substitution costs to allow for a certain degree of time warping. 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 & Ritschard, 2016). We consider it here regardless because it has been used elsewhere for similar work (Delmelle, 2016; Patias et al., 2019).

3.3.3. OMecenter

Since we do not expect the similarity/distance between any two neighborhood types to be identical, our third approach relaxes this assumption in a different manner than “OMtranr”. One natural choice of assessing the distance is the Euclidean distances between cluster centers which can be easily obtained from the previous neighborhood segmentation step. Natural choice as it seems, it has been rarely used in the neighborhood change research. One reason is that the object of DNA or life course research where SA methods have been widely applied is sequence of discrete values where an extraction of similarity between discrete values is hardly possible. Aside from utilizing very useful information from neighborhood socioeconomic differences, 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”.

3.3.4. OMstran

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 sequences 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. Assuming the deletion and insertion of a transition pair is infinitely large, the “OMstran” method produces 3 for the distance between neighborhood sequences ‘1, 1, 1, 1, 1’ and ‘3, 2, 3, 3, 3’, and 4 for the distance 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.

Among the four methods, “OMecenter” and “OMtranr” are data-dependent, meaning that the costs of substitution between neighborhood types, as well as the insertion and deletion are determined by the available neighborhood sequences. We will elaborate on the costs after introducing the method for segmenting neighborhood trajectories.

3.4. Classifying neighborhood sequences

The distance matrix between neighborhood sequences produced by each of the four 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 together with the average silhouette coefficient S. It should be noted that the resultant “optimal” number of neighborhood sequence clusters can vary across four SA methods, but to reach a fair comparison we restricted this number to be identical.

3.5. Evaluation measures

We adopted two indices to evaluate the differences/similarities in the distances between neighborhood sequences and the sequence clustering assignments between the four SA methods respectively.

3.5.1. Mantel test

Mantel test is a commonly used statistical test of the correlation between a pair of distance matrices (Guillot & Rousset, 2013; Mantel, 1967; Robette & Bry, 2012). For two 25,961 × 25,961 distance matrices of neighborhood sequences (e.g. $\boldsymbol{X}$ and $\boldsymbol{Y}$) based on two SA methods, we first vectorize them into two vectors $\boldsymbol{vecX}$ and $\boldsymbol{vecY}$ and then calculate the Pearson correlation coefficient as follows:

$$\rho = \frac{\mathrm{cov}(\boldsymbol{vecX}, \boldsymbol{vecY})}{\sigma_{\boldsymbol{vecX}}\,\sigma_{\boldsymbol{vecY}}}, \tag{2}$$

where cov is the covariance and σ is the standard deviation. ρ is within the range of $[-1, 1]$ with negative values indicating negative correlation between two distance matrices and positive values indicating positive correlation. If $\rho > 0$, a larger ρ indicates a higher similarity between two distance matrices of neighborhood sequences. Mantel test adopts a random permutation scheme to evaluate the significance of ρ while accounting for the fact that each matrix is symmetric and the matrix elements (distances) are dependent on each other – meaning that shortening the distance between two sequences might incur an increase in other distances.[5]

3.5.2. Adjusted Rand index

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 (Eq. (3)):

$$\mathrm{RI} = \frac{2(a+b)}{n(n-1)}. \tag{3}$$

We adopted an extension of RI, the adjusted Rand Index (ARI) (Hubert & Arabie, 1985) which is corrected for chance as an evaluation measure for the neighborhood sequence clusterings. ARI is defined in Eq. (4):

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{E(RI)}}{\max(\mathrm{RI}) - \mathrm{E(RI)}}, \tag{4}$$

where E(RI) is the expectation of RI and max(RI) is the maximum of RI. 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.

[5] For a detailed explanation of the permutation-based inference of Mantel test, refer to Mantel (1967); Robette and Bry (2012).

Table 3
Neighborhood classifications and compositions.

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

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 obtained 14 clusterings, each of which could be the potential neighborhood segmentation scheme. When $k = 2$ and 3, the resulting clusterings gave the largest average silhouette coefficients, 0.276 and 0.216 respectively. The coefficient dropped to 0.148 and 0.147 for $k = 4$ and 5; as k continued to increase, the coefficient increased – 0.156, 0.156 and 0.157 for $k = 6,7,8$. For $k = 9$, the coefficient dropped to 0.128 and failed to increase to 0.15 as k continued to increase. Based on the pattern of the average silhouette coefficients, we considered $k = 8$ as the "appropriate" number of clusters for the neighborhood segmentation since it was a local maxima for the average silhouette coefficient and offered more details than $k = 2$ or 3. Fig. A1 displays the median z-scores of all 14 variables for each of the 8 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 Fig. 2, we observe that *White Laborer* and *Newer Suburban* are more common in the 50 MSAs under study from 1970 to 2010 while *Aging Suburban* is the least common.

As mentioned in Section 3, "OMecenter" and "OMtranr" are data-dependent. After completing the neighborhood segmentation, we calculated the substitution costs for each of them. These costs are displayed in Fig. 3(a) and (b). It is obvious that both "OMecenter" and "OMtranr" allow for certain levels of heterogeneity in the substitution costs. For example, "OMecenter" makes it hard to substitute *White Elite* with *Black Poor* than with *White Laborer*. While this is also the case with "OMtranr", the difference is much smaller. The fact that all the substitution costs between different types for "OMtranr" are close to 1 makes "OMtranr" more similar to "Hamming".

4.2. Neighborhood sequence patterns

4.2.1. Descriptive statistics

Since there were 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 examined, we observed only 2,958 unique sequences, meaning that only 9% potential unique sequences were realized. Fig. 4 shows the histogram of the top 20 most common neighborhood sequences: 4 are sequences exempt from any change meaning that the tract remained in the same neighborhood type across all four decades (*White Laborer* ('3'), *Newer Suburban* ('8'), *Poor Mixed Race* ('7') and *New Renters* ('5')), and the other 16 experienced two neighborhood types. 14

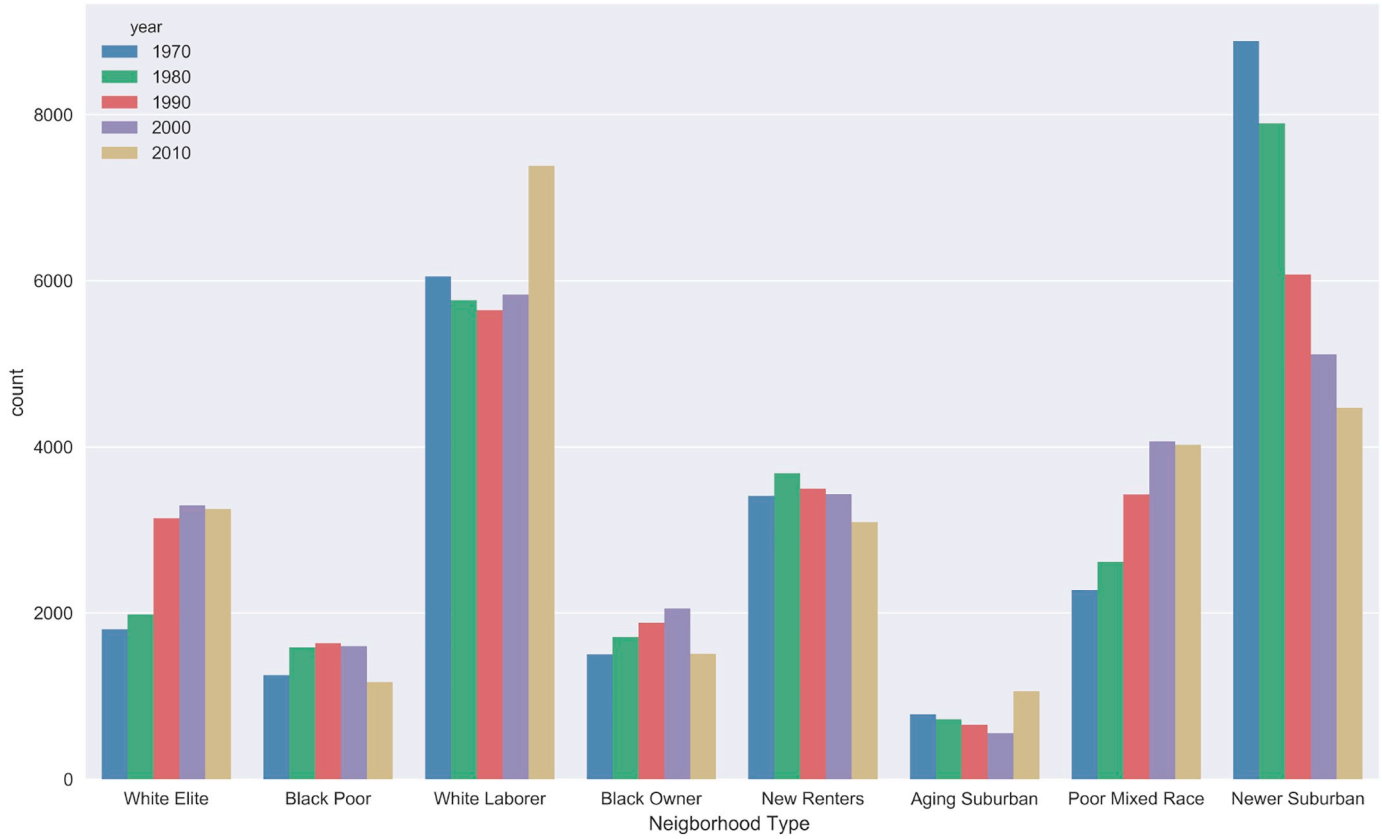


Fig. 2. Histogram of neighborhood segmentations per census year.

neighborhood sequences are characterized by one state change in the last decade 2010. For instance, about 900 neighborhoods started from *Newer Suburban* in 1970, stayed there for four decades, and then transitioned to *White Laborer* in 2010. Another set of transition happened in 1980 (from *Newer Suburban* to *White Laborer*) instead of changing in 2010. The last set of neighborhood sequences ranking top 20 involves two changes, one happened in 1980 from *White Laborer* to *Newer Suburban* while the other happened in 2010 involving transitioning back to *White Laborer*. 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 accounted for only 3% of the 25,961 sequences. Meanwhile, 2117 out of 2853 unique sequences contained less than 3 successive identical values which we interpret as having experienced "frequent" changes.

4.2.2. Clusterings of neighborhood sequences

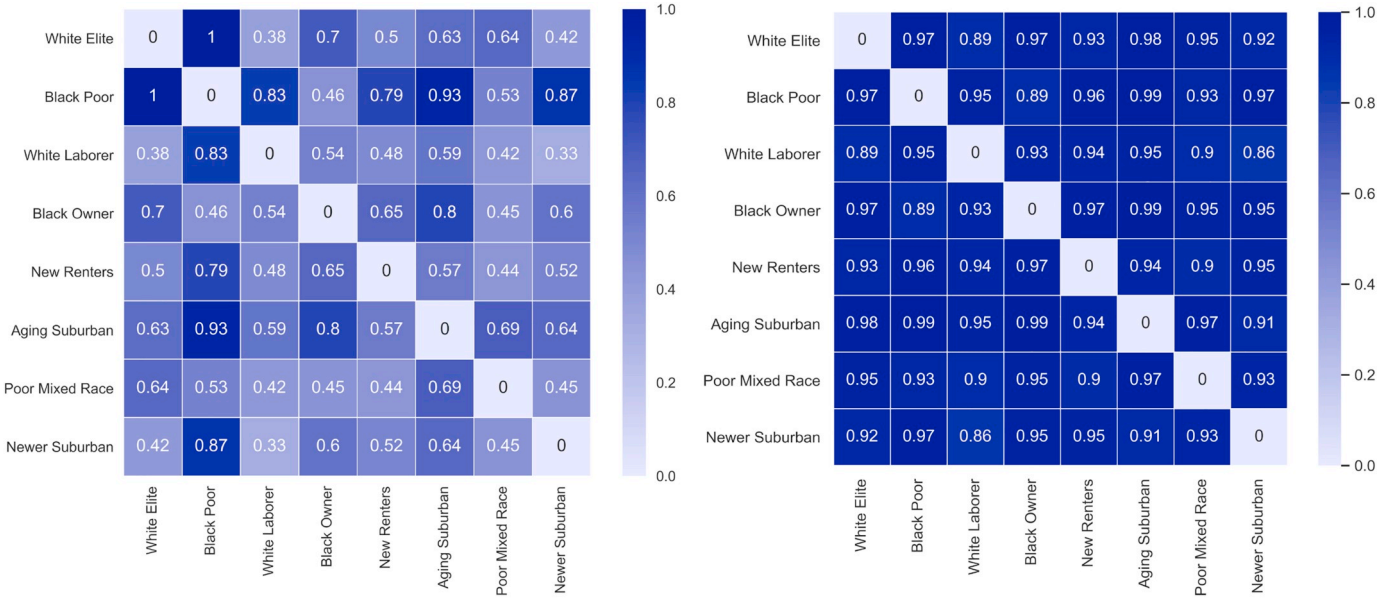
The four SA methods were then applied to the neighborhood sequences to acquire four sequence distance matrices. Each distance matrix was then used in the agglomerative hierarchical clustering with Ward's minimum variance criterion. We obtained the appropriate number of clusters by visually inspecting the hierarchical clustering dendrogram and truncating the dendrogram with a distance cutoff where there is a large gap in the tree together with the help of the average silhouette coefficient. To reach a reasonable comparison between four sequence methods, we adopted an eight-cluster solution which was deemed to be appropriate for most of them.[6]

For "Hamming", "OMtranr", and "OMecenter", each of the eight clusters of neighborhood sequences is dominated by one neighborhood type though there are some small variations in terms of which cluster a weakly dominated sequence belongs to. For instance, while sequence "*Newer Suburban – Newer Suburban – White Laborer – White Laborer – White Elite*" is assigned to the cluster dominated by stable *Newer Suburban* with "Hamming", it belongs to the cluster dominated by stable *White Laborer* with the latter two methods. In contrast, the clusterings based on "OMstran" which evaluates the distance between sequences of transitions across neighborhood types over time produces very different compositions.[7] Here, none of the eight 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. For instance, one cluster 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 another cluster is mainly comprised of sequences experiencing more changes – in both 1980 and 2010. As it stands, the interpretation of the clustering based

[6] As the reviewers noted, the current neighborhood sequence typologies are coarse and contain much heterogeneity. We do want to point out that there is not a perfect index based on which the "best" number of clusters can be selected. A larger number is favorable to reduce heterogeneity but the interpretation of a lot more neighborhood sequence types could be formidable. We have experimented with a range of numbers for sequence clusters (k = [6, 50])

(*footnote continued*)
with the pairwise ARI values shown in Fig. A2. If a small number of clusters is selected, "OMstran" is very dissimilar to the other three methods, while as the number increases, it becomes more and more similar. It should be noted that when the number of clusters is pretty large (> 30) meaning that neighborhood sequence clusters are more internally homogeneous, "OMecenter" becomes one least similar to the others potentially due to its heterogeneous substitution costs. This further corroborates our advice of utilizing the valuable information offered by the continuous tract-level socioeconomic variables when setting the OM operation costs.

[7] These neighborhood sequence clusters are visualized in a set of index plots and are available upon request.



(a) Substitution costs for "OMecenter". (b) Substitution costs for "OMtranr".

Fig. 3. Substitution costs between every pair of neighborhood types for two data-dependent sequence approaches: (a) "OMecenter" (b) "OMtranr".

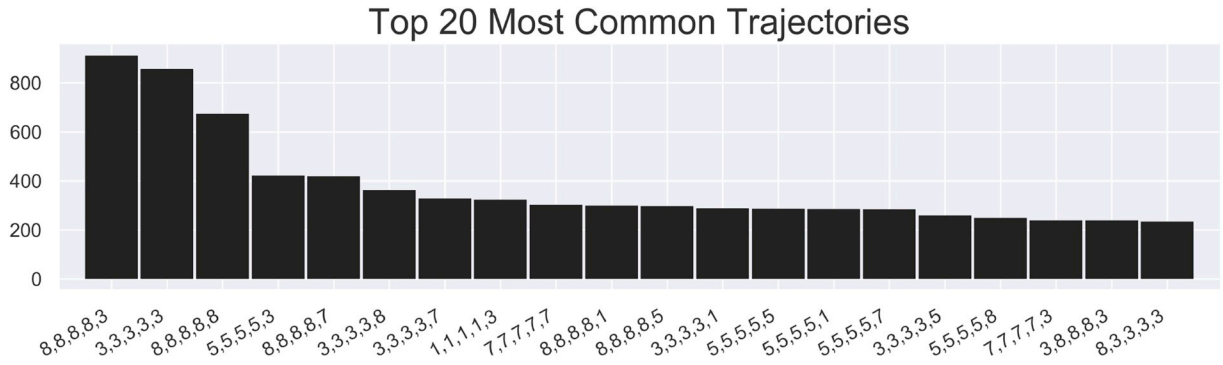


Fig. 4. Histogram of top 20 most common neighborhood trajectories 1970–2010.

on "OMstran" is considerably different from the others, and the choice of the method should be guided by the research question. We shall adopt two indices to quantify the difference in the sequence clusterings in the next subsection.

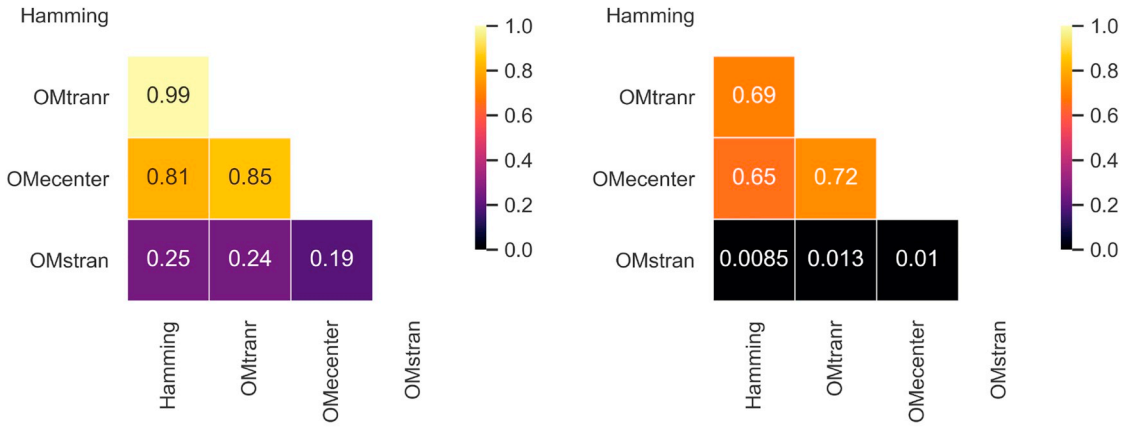
4.2.3. Similarity between neighborhood sequences clusterings

We applied the Mantel test to each pair of neighborhood sequence distance matrices to test for similarity. It turns out that all tests are rejected at the 1% significance level and the test statistics are positive as shown in Fig. 5(a). Two distance matrices of neighborhood sequences produced by applying "Hamming" and "OMtranr" are very similar. The very high similarity is a result of "OMtranr"'s highly homogeneous and close to 1 substitution costs which were constructed based on empirical transition rates as shown in Fig. 3(b). Because the observed transition rates between distinct neighborhood types are pretty small, substitution costs between any pair are quite similar which turns out to be quite uninformative. This is also part of the reason why transition rates-based costs are not suggested in empirical studies such as the life course research (Studer & Ritschard, 2016). The difference in the insertion/deletion costs does not seem to produce a huge difference though "OMtranr" allows for a mild level of time warping by setting the cost to be 1 while "Hamming" does not allow time warping at all. "OMecenter", which utilizes the Euclidean distances between neighborhood cluster centers, gives a much different distance matrix. The OM variant "OMstran" leads to a distance matrix very different from the others as expected.

In addition, we attempted to quantify the differences in neighborhood sequence clusterings conditional on the cluster scheme (eight-cluster solution) introduced in Section 4.2.2. We calculated ARIs between any pair of clusterings as displayed in Fig. 5(b). We observe that the similarity for any pair dropped at this cluster scheme, which is especially true for that between "Hamming" and "OMtranr". The most similar pair is "OMtranr" and "OMecenter", both of which allow for a certain level of time warping though the latter comes with more heterogeneous substitution costs. Looking at Fig. A2 which visualizes the relationship between a wide range of sequence clusters ($k \in [6, 50]$) and pairwise ARI values, we notice a change in the dominant power from the difference in insertion/deletion costs (time warping) at a smaller k to the difference in substitution costs at a larger k.

4.2.4. Spatial variations and determinants

We further investigated the spatial variations of pairwise similarity between neighborhood sequence clusterings across 50 MSAs at the eight-cluster scheme. It appears that ARI varies substantially as shown in Fig. 6. For instance, the ARI between clusterings based on "Hamming" and "OMtranr" reaches as high as 0.9 for the Raleigh-Cary MSA, and as low as 0.41 for the Milwaukee MSA. Similarly, the ARI between clusterings based on "OMtranr" and "OMecenter" has a wide range



(a) Pairwise similarities between sequence distance matrices (Mantel Test).

(b) Pairwise similarities between cluster assignments (Adjusted Rand Index (ARI)).

Fig. 5. Pairwise similarities of neighborhood sequence: (a) Mantel test (sequence distance matrices); (b) ARI (sequence cluster assignments).

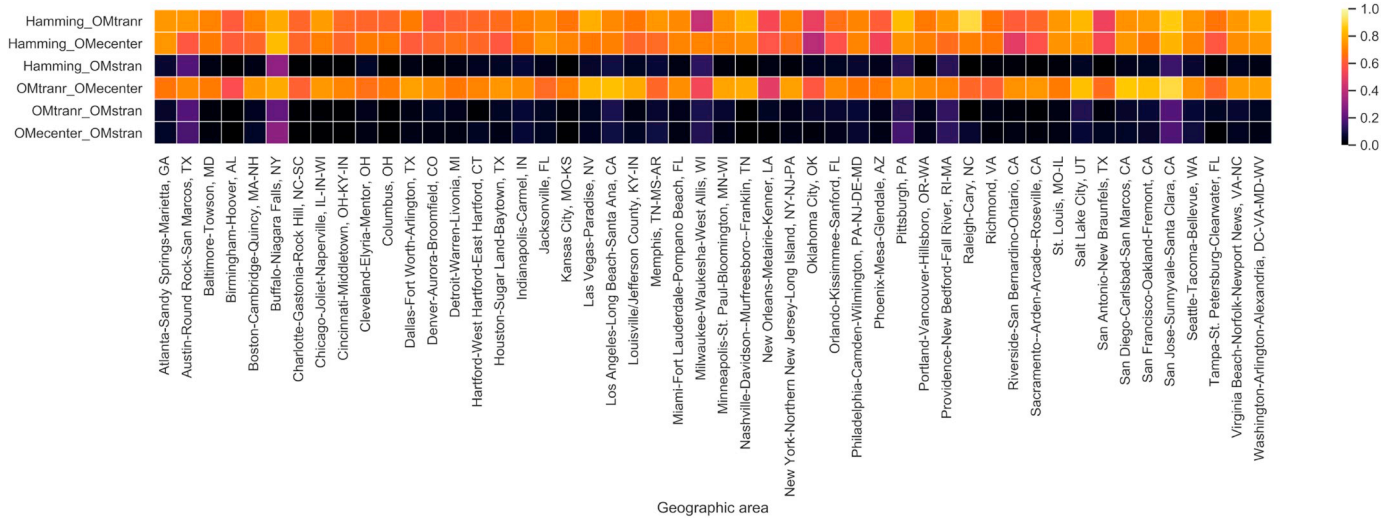


Fig. 6. Pairwise similarities between cluster assignments (ARI) of neighborhood sequences for each MSA. Each row represents a pair of sequence methods and each column represents a MSA.

[0.49, 0.9], while the smallest ARI (-0.028) is found between "Hamming" and "OMstran" in the Raleigh-Cary MSA, which indicates two divergent neighborhood sequence typologies.[8]

We further conducted a regression analysis to explore potential factors which could help explain the observed variations in ARI values. We estimated a fixed effect regression model where SA pair fixed effects were included to control for any unobserved method pair-specific characteristics:

$$\text{ARI} = \beta_0 + \sum_i X_i + \sum_{j=1}^{5} \delta_j \text{SP}_j + \varepsilon. \tag{5}$$

Here, the dependent variable is the ARI between a pair of SA methods for each MSA, SP_j is the dummy variable constructed for each pair of four SA methods (e.g., Hamming_OMecenter, OMtranr_OMarbitr). The dummy variable for the SA pair – OMtranr_OMstran was dropped to avoid perfect multicollinearity. $\Sigma_i X_i$ is a bunch of predictor variables at the MSA level which could help explain the variations in observed ARI. It is possible that characteristics of neighborhood sequences experienced by each MSA are important. One such characteristic is the sequence complexity which could be measured by the entropy of the state distribution in the sequence, the number of transitions (changes) in the sequence, or a combination of these two indices (Gabadinho, Ritschard, Müller, & Studer, 2011). Another characteristic is the sequence turbulence proposed by Elzinga and Liefbroer (2007). Generally, a sequence which has more distinct states and more state changes is considered more turbulent. Since these measures are sequence-wise, we obtained the MSA averages which were used as predictor variables in the regression model. In addition to the sequence-wise characteristics, we also explored aggregate sequence properties at the MSA level, including the number of unique sequences within each MSA. Several socioeconomic characteristics of each MSA at the initial census year (1970) were also included to help explain the variations, such as the population, the average median household income, the income inequality among neighborhoods (measured by Gini), and the education attainment proxied by the college educated residents. The increments of these variables across the study period (1970–2010) were included as predictor variables as well.

We first removed variables which induced multicollinearity

[8] The five maps of neighborhood types from 1970 to 2010 for all 50 MSAs together with their five clusterings of trajectories are available upon request.

Table 4
MSA-level ARI regression results.

Variables	ARI
Average Household Income 1970	0.0145***
Average Household Income Change 1970–2010	−0.0066
Spatial Gini Change 1970–2010	0.0515
Population in 1970	−0.0044
Population Change 1970–201	0.0012
Average number of transitions in a Neighborhood Sequence	−0.5128***
Hamming_OMecenter	0.6057***
Hamming_OMstran	−0.0069
Hamming_OMtranr	0.6377***
OMecenter_OMstran	−0.0009
OMtranr_OMecenter	0.6584***
Intercept	0.2640***
Observations	300
F-statistic	637.7***
Adj. R-squared	0.959

Note: ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively.

problem. We did so by inspecting the variance inflation factor for each predictor variable, a widely used indicator for multicollinearity. We ended up with six variables as shown in Table 4. We then estimated the fixed effect model, the result of which is displayed in the table. It turns out that the MSA population, population change, spatial income inequality change, as well as the income change are insignificant. In contrast, the initial average household income is highly significant and positive, indicating that MSAs where richer households resided in 1970 tend to be less sensitive to the choice of the SA methods. In other words, when investigating neighborhood change in these MSAs, different SA methods tend to produce a more converging neighborhood sequence typology. Comparatively, for MSAs hosting neighborhoods which have experienced frequent changes during the period 1970–2010, they are less likely to produce similar typologies with different SA methods.

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 seen many applications to neighborhood change 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 are robust in their ability to provide consistent and converging neighborhood sequence typology and how this robustness (or non-robustness) presents spatial heterogeneity.

We applied four alignment-based SA methods to a common longitudinal census data set in the U.S. – the 50 largest MSAs at census years 1970, 1980, 1990, 2000 and 2010. While three of them fall within the classic OM scheme focusing on the contemporaneous similarity with a certain level of time warping with the difference only in the choice of operations costs such as the cost of substituting one neighborhood type with another, and the cost of inserting or deleting a neighborhood type, the fourth method is a OM variant focusing on the second-order property of the neighborhood sequence. We demonstrate that the neighborhood sequence typology is generally sensitive to the choice of OM method and the operation costs, and the sensitivity demonstrates heterogeneity across MSAs. MSAs with higher average household income in 1970 tend to be less sensitive to the choice of the SA methods. In other words, when investigating neighborhood change in these MSAs, different SA methods tend to produce a more converging neighborhood sequence typology. Comparatively, for MSAs hosting neighborhoods which have experienced frequent changes during the period 1970–2010, they are less likely to produce similar typologies with different SA methods. In addition, there is a big difference in the neighborhood sequence typology between applying the classic OM methods with varying costs and using the OM variant focusing on the second-order sequence property. While the former reflects contemporaneous experiences with some extent of time warping if the cost of insertion and deletion is not expensive, the latter emphasizes the stability of the neighborhood sequence. In terms of selecting the cost of substituting one neighborhood type with another, we recommend researchers to utilize the valuable information provided by the continuous tract-level socioeconomic variables – more specifically, basing the substitution costs on (at least partially) distances between neighborhood segmentation (cluster) centers.

Our findings suggest a number of important challenges to the practice of neighborhood change analysis. First, the growing number of applications of sequence based methods in urban dynamics may appear to offer a large body of results on typologies of neighborhood change at first glance. However, the sensitivity of sequence identification to the choice of operational parameters suggests that studies adopting alternative implementations are no longer directly comparable, and thus pooling trajectories from different studies to develop a holistic understanding of urban dynamics would be misguided. Second, we suggest urban researchers utilize the valuable information from geodemographic clustering to determine operation costs for the SA method. The distance between neighborhood segmentations/clusters is readily available, quite unique to the neighborhood research, and should be used to guide the researchers to better quantify the similarity between pairs of neighborhood types.

One limitation of the current research design pertains to the scope of factors which could contribute to the final neighborhood sequence typology and thus the sensitivity of such typology to the choice of the SA method. In addition to the choice of the SA method for measuring the distance between neighborhood sequence, it should be noted that many other parameters, particularly those related to the clustering algorithm used extensively in this type of analysis, could make a difference to the final sequence typology. For instance, the harmonization strategy used for obtaining temporally consistent neighborhood boundaries, the choice of the neighborhood socioeconomic variables, the normalization strategy for these variables, the choice of the cluster algorithm for neighborhood and sequence segmentations, as well as the number of clusters adopted. While we recognize the potential impact from these factors, we decided to isolate them so that we can specifically look at the influence from the choice of the SA method. Future research should be directed to extend the scope of the involved factors to better understand the underlying mechanism.

An interesting and valuable research direction would be to interrogate the spatial nature of the unit of study for neighborhood change – which is usually census tract, a spatial aggregate whose boundary is administratively defined and prone to change over time. Neighborhoods as spatial entities thus could be spatially autocorrelated, meaning that neighborhoods could be more similar to those close by than those farther away – as coined by “First Law of Geography” (Tobler, 1970), invalidating the independence assumption underlying classic statistics; each neighborhood could host households with heterogeneous characteristics; neighborhood analysis could suffer from modifiable areal unit problem (MAUP) which is an issue common to studies of spatial aggregates and the problem was stated by Openshaw (1984) as “the areal units used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating.” Thus, work that develops spatially explicit forms of sequence analysis would be an important contribution.

Acknowledgements

We greatly appreciate the constructive comments from the editor and reviewers. US NSF SES Divn Of Social and Economic Sciences Award #1733705.

Appendix A. Appendix

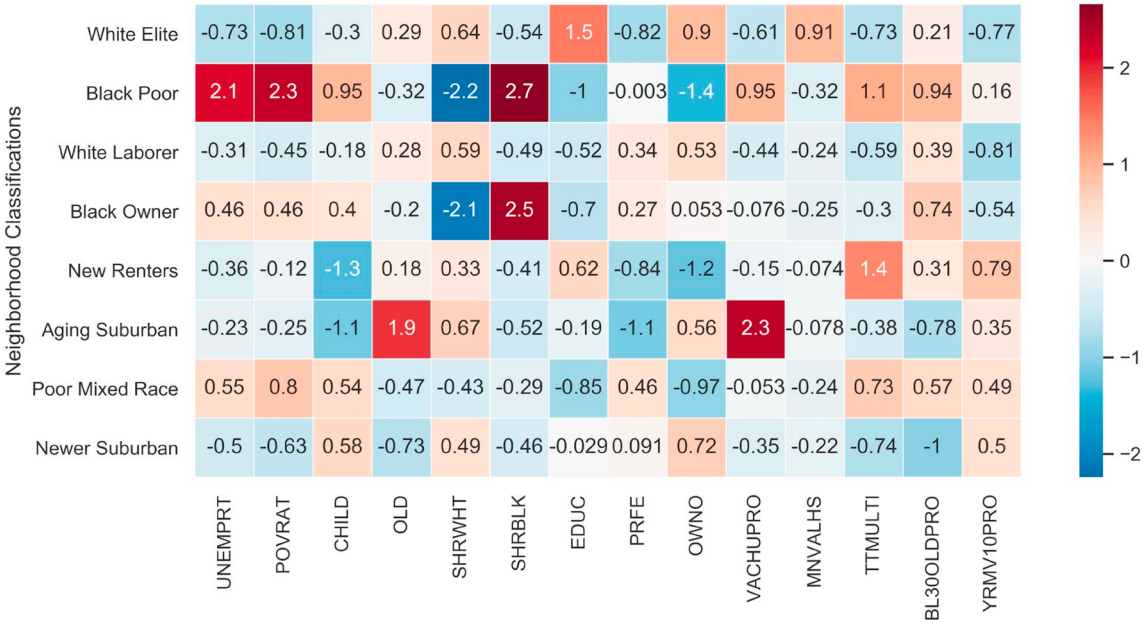


Fig. A1. Heat map of median z-scores for eight neighborhood clusters.

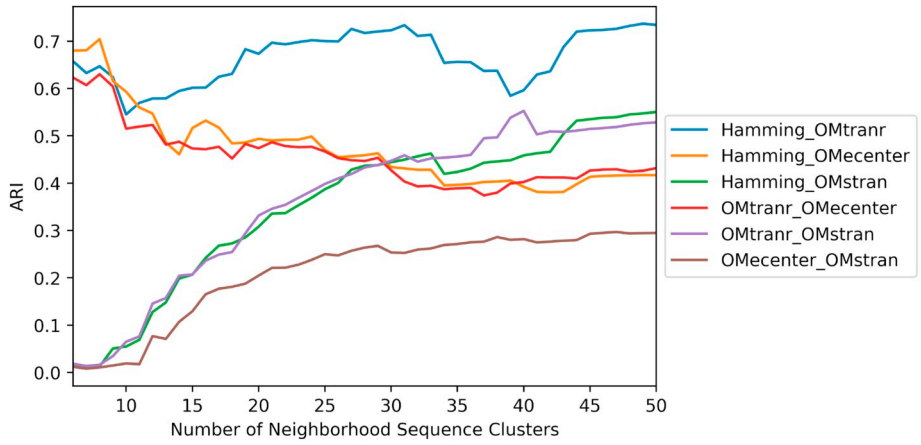


Fig. A2. Pairwise similarities (ARI) between cluster assignments of neighborhood sequences with varying number of sequence clusters.

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.

Abbott, A. (1997). Of time and space: The contemporary relevance of the Chicago school. *Social Forces, **75***, 1149–1182. URL: http://www.jstor.org/stable/2580667.

Aisenbrey, S., & 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. 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. https://doi.org/10.1068/a220855.

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

Carrillo, H., & Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics, 48*, 1073–1082. 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/S0143622814002860doi:10.1016/j.apgeog.2014.12.002.

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

Delmelle, E. C. (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.

Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population/Revue européenne de Démographie, 23*, 225–250.

Gabadinho, A., Ritschard, G., Müller, N., & Studer, M. (2011). Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software, Articles, **40***, 1–37. URL: https://www.jstatsoft.org/v040/i04.

Greenlee, A. J. (2019). Assessing the intersection of neighborhood change and residential mobility pathways for the Chicago metropolitan area (2006-2015). *Housing Policy Debate, 29*, 186–212. https://doi.org/10.1080/10511482.2018.1476898.

Guillot, G., & Rousset, F. (2013). Dismantling the mantel tests. *Methods in Ecology and Evolution, **4***, 336–344. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210x.12018.

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

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

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

Lee, K. O., Smith, R., & Galster, G. (2017b). Subsidized housing and residential trajectories: An application of matched sequence analysis. *Housing Policy Debate, 27*, 843–874. https://doi.org/10.1080/10511482.2017.1316757.

Li, Y., & 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, **108***, 1313–1337. https://doi.org/10.1080/24694452.2018.1433016.

Ling, C., & Delmelle, E. C. (2016). Classifying multidimensional trajectories of neighbourhood change: A self-organizing map and k-means approach. *Annals of GIS, 22*, 173–186. https://doi.org/10.1080/19475683.2016.1191545.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, **27***, 209–220. URL: http://cancerres.aacrjournals.org/content/27/2_Part_1/209.

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

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and techniques in modern geography*.

Patias, N., Rowe, F., & Cavazzi, S. (2019). A scalable analytical framework for spatio-temporal analysis of neighborhood change: A sequence analysis approach. *Geospatial Technologies for Local and Regional Development,* 223–241. https://doi.org/10.1007/978-3-030-14745-7_13.

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. 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., & Interlante, L. (2011). Measuring spatial dynamics in metropolitan areas. *Economic Development Quarterly, 25*, 54–64. https://doi.org/10.1177/0891242410383414.

Ritschard, G., & Studer, M. (Eds.). (2018). *Sequence analysis and related approaches.* Springer.

Robette, N., & 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. https://doi.org/10.1146/annurev.so.09.080183.000503.

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

Singleton, A. D., & 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., & 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. https://doi.org/10.1080/00330124.2013.848764. PMID:25484455.

Spielman, S. E., & Singleton, A. (2015). Studying neighborhoods using uncertain data from the american community survey: A contextual approach. *Annals of the Association of American Geographers, 105*, 1003–1025. https://doi.org/10.1080/00045608.2015.1052335.

Studer, M., & 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. https://doi.org/10.1111/rssa.12125.

Studer, M., Ritschard, G., Tabin, J., & Perriard, A. (2014). *A comparative review of sequence dissimilarity measures.* Lausanne.

Studer, M., Struffolino, E., & Fasang, A. E. (2018). Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. *Sociological Methodology, 0*, 1–34. https://doi.org/10.1177/0081175017747122.

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

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*, 234–240. http://www.jstor.org/stable/143141.

Vinga, S., & Almeida, J. (2003). Alignment-free sequence comparison—A review. *Bioinformatics, 19*, 513–523. https://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., & Knox, P. L. (2014). Neighborhood change in Metropolitan America, 1990 to 2010. *Urban Affairs Review, 50*, 459–489. https://doi.org/10.1177/1078087413501640.

Wong, K. M., Suchard, M. A., & 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., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology, 18*, 186. https://doi.org/10.1186/s13059-017-1319-7.

Zwiers, M., Kleinhans, R., & 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. https://doi.org/10.1007/s12061-016-9189-z.