

A Scalable Analytical Framework for Spatio-Temporal Analysis of Neighborhood Change: A Sequence Analysis Approach

Nikos Patias , Francisco Rowe and Stefano Cavazzi

Abstract Spatio-temporal changes reflect the complexity and evolution of demographic and socio-economic processes. Changes in the spatial distribution of population and consumer demand at urban and rural areas are expected to trigger changes in future housing and infrastructure needs. This paper presents a scalable analytical framework for understanding spatio-temporal population change, using a sequence analysis approach. This paper uses gridded cell Census data for Great Britain from 1971 to 2011 with 10-year intervals, creating neighborhood typologies for each Census year. These typologies are then used to analyze transitions of grid cells between different types of neighborhoods and define representative trajectories of neighborhood change. The results reveal seven prevalent trajectories of neighborhood change across Great Britain, identifying neighborhoods which have experienced stable, upward and downward pathways through the national socioeconomic hierarchy over the last four decades.

Keywords Neighborhood change, Sequence analysis, Spatio-temporal data analysis, Classification, Population dynamics

N. Patias

Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool
email: n.patias@liverpool.ac.uk

F. Rowe

Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool
email: F.Rowe-Gonzalez@liverpool.ac.uk

S. Cavazzi

Ordnance Survey Limited
email: stefano.cavazzi@os.uk

1 Introduction

Changes over space and time reflect the complexity and evolution of demographic and socio-economic processes (Miller, 2015). Yet measuring the magnitude, location and temporal frequency of these changes is challenging. Using traditional forms of data (i.e. census and survey data), demographic and socio-economic changes have often been captured at a very coarse temporal levels i.e. every month, year or decade. Also, these data are also normally available at some spatially aggregated level. Administrative boundaries have traditionally been the default spatial framework for census and survey data collection and analysis (Goodchild, 2013), and these areas are usually affected by boundary changes over time, particularly splitting an area in two (Casado-Díaz *et al.*, 2017; Rowe *et al.*, 2017). So, a ‘freeze history’ approach has been generally employed to develop a consistent geography by freezing the zonal system at a certain point in time and systematically tracking subsequent alterations in geographical boundaries to amalgamate subsequently created areas (Rowe, 2017).

Different levels of spatial aggregation can however produce different representations of a socio-economic process as a result of the Modifiable Areal Unit Problem (MAUP). MAUP refers to the statistical sensitivity and variability of results relating to the spatial framework of analysis (Openshaw, 1983; Fotheringham and Wong, 1991). The most appropriate spatial framework of analysis may thus differ according to the process in study (Prouse *et al.*, 2014). MAUP can create ‘unreal’ spatial patterns which are caused by loss of information (Hayward and Parent, 2009). Choosing areal units based on geographical coordinates, rather than aggregation of administrative boundaries, could help to tackle this issue by offering the possibility to analyze temporal data regardless of changes in geographic boundaries.

An increasing number of methods for spatio-temporal data analysis have been developed to study complex demographic and socio-economic processes, namely space-time point pattern, probabilistic time geography and latent trajectory models (An *et al.*, 2015). Clustering techniques are often employed on space-time data, identifying patterns (Warren Liao, 2005; Aghabozorgi *et al.*, 2015; Arribas-Bel and Tranos, 2018). There is also a wide variety of spatio-temporal statistical techniques in current literature where traditional deterministic trend models, stochastic trend models and stochastic residual models have been generalized to capture spatiotemporal processes using individual level data (Kyriakidis and Journel, 1999), as well as spatial and temporal correlation using Spatio Temporal Autoregressive Regression (STAR) and Bayesian hierarchical models on areal data (Huang, 2017). Yet, these models are often restricted on specific situations namely particular data format or geometry types and are not flexible or adaptable to contribute in scalable space-time analysis frameworks. Spatio-temporal processes involve measurement of four dimensions namely occurrence, timing, order and duration of events or transitions and while the aforementioned methods can provide useful information about movements and points of interest, they only capture some dimensions of spatio-temporal processes. The integration of multiple approaches can provide context ‘aware’ data

and expose patterns based on analysis of the sequencing of events, rather than comparison of discrete points in time, capturing the full range of dimensions of spatio-temporal processes.

Sequence analysis provides a useful framework to integrate various analytical approaches and capture the four key dimensions of spatio-temporal processes i.e. occurrence, timing, order and duration of events or transitions. Sequence analysis was originally developed for analyzing DNA sequences (Sanger and Nicklen, 1977; Bailey, 2017), and theoretically introduced in the Social Sciences in the 1980s (Abbott, 1983). Sequence analysis has recently been widely applied to analyze longitudinal individual family, migration and career trajectories (e.g. Rowe, Corcoran, *et al.*, 2017; Backman *et al.*, 2018).

Sequence analysis can also be applied to better understand the evolution of places. Conceptually, neighborhoods for example are assumed to progress through a number of pre-determined stages, transitioning through phases of development, growth, stability and decline (Hoover and Vernon, 1959). However, prior empirical work has employed a static cross-sectional framework to explore these transitions between two points in time (e.g. Teernstra and Van Gent, 2012) and assumed all neighborhoods undergo the same rigid pathway of change. These shortcomings partly reflect the lack of consistent spatial data over a longer window of time, but also the absence of an analytical approach to study these transitions in a temporally dynamic framework.

Only recently, empirical analyses have recognized the diversification in neighborhood transitions and enabled exploration and quantification of neighborhood change over a long period of time by using sequence analysis. Delmelle (2016) conducted a first study using sequence analysis for Chicago and Los Angeles, expanding her focus on a subsequent investigation to US 50 metropolitan areas over a 50 year period (Delmelle, 2017). These studies contributed in providing a general approach for analyzing differentiating pathways of neighborhoods namely upgrading, downgrading or stable trajectories in the socio-economic hierarchy as well as gentrification processes. Yet they focused only on urban and metropolitan areas, missing the interaction between urban and rural continuum.

While these studies have advanced our understanding of neighborhood change in particular urban settings, significant gaps remain to be addressed. First, gridded data generation is needed to address the lack of consistent geographical boundaries over time (Janssen and Ham, 2019). Second, the use of gridded data offers the potential to perform analyses at various geographical levels through aggregation of grids at particular administrative or functional areas. This opportunity provides a flexible dataset and a scalable approach for the use of purpose-built areas. Third, weighted clustering of the sequences provides a scalable approach on analyzing big datasets by separating the unique sequences matrix and their frequency in different vectors. This addresses the lack of information by using clustering approaches based on 'prototype' sequences where their frequency is not captured (i.e. how many neighborhoods followed the same sequence). Fourth, the aforementioned gaps are partly the result of the absence of a workflow that addresses the lack of temporally consistent geographical units and offers a way to effectively capture the key

elements of changes in space and time (occurrence, timing, order and duration). This limitation is addressed in this study by the integration of different approaches (i.e. population grid surface estimation, clustering analysis and optimal matching).

This paper aims to develop a scalable analytical framework for spatio-temporal data analysis addressing all four identified gaps. By doing so, it contributes to the current literature on spatio-temporal data analysis in three key ways:

1. By providing geographical consistent gridded data over a 40-year period for Great Britain;
2. By developing a scalable analytical framework in two ways: (i) offering a flexible dataset which can be aggregated at various geographical levels; and (ii) employing a weighted clustering approach to measure dissimilarity between individual sequences;
3. By formulating a workflow to effectively capture the key elements of changes in demographic and socio-economic process across space and over time through the integration of multiple approaches.

The remainder of this article is organized as follows. Section two describes the dataset and methods used in this research project, followed by results and discussion that are presented in section three. Finally, section four provides some concluding remarks and suggestions for further research.

2 Data and methods

2.1 Data

The original data used in this study is drawn from five decennial Censuses for Great Britain (i.e. England, Scotland and Wales) covering the period from 1971 to 2011 with 10-year intervals. The five Censuses were conducted in 1971, 1981, 1991, 2001, 2011. The data was downloaded from the Office of National Statistics (ONS) portal (<http://casweb.ukdataservice.ac.uk> & <http://infuse.ukdataservice.ac.uk>).

Census administrative boundaries are not consistent over time. To this end, this paper uses an approach of recalculating Census counts from administrative boundaries to gridded data using *Popchange* project algorithm. *Popchange*, is a tool that provides population surfaces across Great Britain but also provides the algorithm which calculates correspondence between low-level Census administrative geographies and 1km² grids (Lloyd *et al.*, 2017). For this project, raw Census data covering a range of demographic, socio-economic and housing variables were downloaded in low-level Census administrative geographies (i.e. enumeration districts

for 1971, 1981 and 1991 and output areas for 2001 and 2011) and *Popchnage* algorithm used to convert Census counts to 1km² grid counts.

These grid counts data correspond to estimates of census variables. As they are generated from a coarser level of geography, there is certain degree of uncertainty around these estimates. However, they offer two key advantages. Firstly, they provide a consistent level of geography to make comparisons of spatial units over a period of time. Secondly, they provide an effective tool to address the MAUP in a spatio-temporal context by providing a spatial framework based on geographical coordinates (i.e. 1km² grids), rather than some arbitrary level of geographical aggregation. Grids can be aggregated to create purpose-built geographical systems depending on the process under analysis.

A drawback of grids is that administrative areas in rural and remote areas are often larger than a grid. Thus, population counts that are split between two or more grids in an administrative area, resulting a small number of population counts per grid. In this study percentages of the variables were calculated by grid and given the small number of counts per grid, the accuracy of the variables' estimation is low. To overcome this issue, only grids which encompass multiple small areas were considered. To this end, the 1km² grid layer overlaid over the 2011 census Output Area boundaries for Great Britain. The final output is 16,035 grid cells covering the whole Great Britain. The grid cells containing zero values can be removed to aid visualization and mapping of the data.

This study measures neighborhood change across three dimensions: demographic, socio-economic and housing. Table 1 lists the set of census variables used to capture these dimensions, all of which are measured as percentages for each grid cell i.e. the grid-specific population aged 0-14 over the grid-specific total population across all age groups.

Table 1 Variables used in the analysis

Broad category	Specific cate-	1971	1981	1991	2001	2011
	Variable					
Demographic	Children: 0 to 14 years old	✓	✓	✓	✓	✓
	Young persons: 15 to 29 years old	✓	✓	✓	✓	✓
	Age structure of population					
	Middle aged adults: 30 to 44 years old	✓	✓	✓	✓	✓
	Older adults: 45 to 64 years old	✓	✓	✓	✓	✓
	Retired: 65+ years old	✓	✓	✓	✓	✓
Place of Birth	Born in United Kingdom (UK) and Republic	✓	✓	✓	✓	✓

		of Ireland (ROI)				
		Born in Europe	✓	✓	✓	✓
		Born in Rest of the World	✓	✓	✓	✓
	Students	Proportion of students	✓	✓	✓	✓
		Managerial occupations	✓	✓	✓	✓
	Socio-economic Group	Non-Manual Workers	✓	✓	✓	✓
		Manual and other Workers	✓	✓	✓	✓
		Private mode	✓	✓	✓	✓
		Public Transport	✓	✓	✓	✓
	Mode of Travel to Work	Active mode	✓	✓	✓	✓
		Other mode (i.e. other and work from home)	✓	✓	✓	✓
	Unemployment	Unemployment rate	✓	✓	✓	✓
		Own occupied housing	✓	✓	✓	✓
	Home Ownership	Private rented housing	✓	✓	✓	✓
		Council rented (social) housing	✓	✓	✓	✓
Housing	Housing vacancy	Vacancy rate	✓	✓	✓	✓

There is variation in the number of categories across census years. For example, a greater number of categories is available for socio-economic status in the 2001 and 2011 Census compared to earlier years. So, data have been aggregated to broader categories which are consistent through time. Also, note that information on students was not available in 1971; nonetheless, it is considered as an important variable and is therefore included for the analysis.

2.2 Methods

The methodological framework developed in this study involves four main stages which can be divided into six steps, as presented in Figure 1. In general, the

first stage involves the production of gridded population data. These data is used in a second stage to create a geodemographic classification of neighborhoods based on the variables listed in Table 1, using k-means clustering. This classification provides representative types of neighborhoods. In a third stage, the classification is used to analyze the year-to-year transition of individual grids between neighborhood types and measure their similarity via optimal matching. In a final stage, this measure of similarity is employed to define a typology of representative neighborhood trajectories based on a k-medoids clustering. Details on each of these stages are provided next.

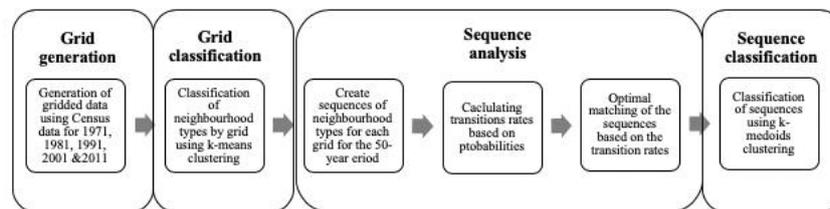


Figure 1 Methodological framework workflow

Stage 1. The official administrative boundaries used to collect the census data are not consistent over time. Boundary changes hamper temporal comparability of this data. Harmonization of these boundaries is needed to effectively track changes over time. To this end, *Popchange* algorithm was used to generate gridded data using raw data drawn from five decennial Censuses for Great Britain as described in section 2.1.

Stage 2. Gridded population data is then used to create a geodemographic classification using a k-means algorithm. The input data is a pooled dataset of grids covering the whole Great Britain for all five census periods i.e. 80,175 grids (=16,035 grids * 5 years). A cluster analysis is performed on a pooled dataset including all five census periods to ensure consistency and comparability of cluster membership in the resulting partitioning solution. These are important elements for the longitudinal analysis of spatial data. For the k-means clustering algorithm, the number of k partitions, which define the number of cluster groups, need to be set prior to performing the analysis (Gentle *et al.*, 1991). This has been set to eight performing 1,000 iterations. The approach used to specify the optimal number of clusters is a two-step sequential process. First, the sum of distances of each observation to their closest cluster center was calculated for a range of cluster options, from 3 to 15, creating an elbow curve. In an elbow curve, the sum of distances tends to decrease towards 0 as the k increases (the score is 0 when k is equal to the number

of data points in the dataset, because then each data point belongs to its own cluster, with no error between the cluster and the center of the cluster). The goal is to determine the smallest number of k partitions that minimizes the sum of distances, and the elbow represents the point at which diminishing returns by increasing k are achieved. Second, the k number at which these diminishing returns are achieved is used as the seed number of partitions. Various clustering partitions around this point were analyzed and mapped to determine the optimal number of clusters for this study i.e. eight. The output from the cluster analysis is temporally consistent geodemographic classification in which each year-specific grid cell is assigned to a neighborhood type.

Stage 3. This geodemographic classification is then used as input for sequence analysis. A key aim of this analysis is to define trajectories that characterize the ways in which the internal demographic and socio-economic structure of neighborhoods have changed over time. To this end, sequence analysis was used. Sequence analysis is built to analyze longitudinal categorical data and enables identification of representative patterns over a period of time. In the current study, the key aim is to identify a small number of representative trajectories of neighborhood change, and the application of sequence analysis involves three key steps. For the implementation of these steps, the TraMineR package in the R programming language was used (Gabadinho *et al.*, 2009).

Step 1. The starting point is the creation of a sequence state object. A sequence state object refers to a dataset arranged in a wide format with rows identifying each spatial unit, columns identifying each time point, and individual cells indicating a specific state. To create a sequence state object, the geodemographic classification was used. Rows identify each geographical grid. Columns identify each of the five census years and each individual cell contains their corresponding year-specific neighborhood type. So, horizontally, each row provides a sequence of transition between different neighborhood type over the five census years.

Step 2. Sequences comparison requires a measure of the minimal cost of transforming one sequence to another. The operations can be used are insertion/deletion (i.e. indel) cost where a single value is specified to reflect how many insertions/deletions need to be made so that the two sequences match. But there is also the option of substitution cost matrix which a square matrix of $s \times s$ dimensions, where s is the number of neighborhood types. So each (i, j) matrix element is the cost of substituting neighborhood type i with neighborhood type j . These elements called transition rates and are calculated based on the probability of transitioning from one neighborhood type to another. Then the optimal matching can be performed (i.e. measuring the similarity of those sequences) which is the sum of those rates for a given sequence.

Step 3. A key innovation of this study is the scalability of the developed framework to build and analyze sequences of neighborhood change. The calculation of dissimilarity between individual sequences is computationally intensive as it involves the use of substitution operations for each pair sequence in the dataset which increase proportionally with the number of spatial units and time points in the analysis.

The analysis of this paper involves the calculation of a dissimilarity matrix for 16,035 grids over 5 years; that is, a resulting matrix of 257,121,225 entries. In order to provide a scalable analytical framework, the unique sequences were identified and their frequencies were calculated and stored in different vectors. Then the unique (1,112) individual sequences used to compute the dissimilarity matrix of 1,236,544 entries. The idea behind this is that the dissimilarity matrix between all pairs of sequences has identical pairs (i.e. many grids that display the same transition, for example, from affluent to thriving neighborhoods). So, if only one pair is considered for the calculation and then it is expanded by the number of similar pairs in the dataset makes the computation less intensive. The use of the proposed approach can be applied to very large datasets for which the resulting dissimilarity matrix can go beyond the storage memory limits of R.

Timing of transitions between neighborhood types was considered a critical element for the definition of sequences as it helps discriminating between transitions resulting from structural economic changes and localized socio-economic shifts. To this end, substitution costs have been used capturing the temporal variation of transitions rather than indel costs which is static cost measure. The substitution cost between neighborhood types i and j for $i \neq j$ is computed by:

$$4 - p(i | j) - p(j | i) \quad (1)$$

Where $p(i | j)$ is the transition rate between states i and j between neighborhood types i and j . This probability is assumed to be dynamic reflecting the year to year transition between neighborhood types. So, a dynamic method of optimal matching was used which updates the substitution costs year to year to calculate distances between individual sequences. This method is referred as to Dynamic Hamming method in the literature (Lesnard, 2009).

Stage 4. The last stage involves producing a typology of neighborhood trajectories using the resulting sequence dissimilarity matrix from Stage 3. Partitioning Around Medoids (PAM) clustering method was selected for classifying sequences. It was preferred over hierarchical clustering methods because, although the PAM algorithm is similar to k-means, it is considered more robust than k-means as it can accept a dissimilarity matrix as an index and its goal is to minimize the sum of dissimilarities compared to k-means that it tries to minimize the sum of squared Euclidean distances (Gentle *et al.*, 1991). The PAM is based on finding k representative objects or medoids among the observations and then k clusters (that should be defined as in k-means) are created to assign each observation to its nearest medoid.

As described in Stage 3 two vectors were created. One stores the dissimilarity matrix of the unique sequences and the other stores its sequence's frequency. The last issue that had to be tackled was the use of both vectors in a clustering algorithm, avoiding the creation of 'prototype' clusters but considering the whole dataset. In some hierarchical clustering methods (i.e. single linkage and complete linkage), the

frequency of unique sequences does not affect the resulting partition of the data. But in the PAM algorithm, the number of observations in the matrix plays a role in the final result as it attempts to minimize the distance between each data point. Large datasets (e.g. of 47,000) result in a dissimilarity matrix of large dimensions (e.g. more than 2 billion), which cannot be handled in R where the storage memory limit is 2.1 billion.

An approach to overcome this problem is data weighting. We applied a weighted version of the PAM clustering algorithm. The functionality of weighted PAM clustering method is the same as using the usual PAM clustering but reduces the amount of memory needed to perform calculations over large datasets. To implement this method, a vector of the number of each unique sequence in the dataset was created and then used to weight dissimilarity matrix of these sequences when applying the PAM clustering method. In this way, the complete dataset (i.e. 16,035 sequences) was used in a scalable, faster and less computationally intense process. To implement this approach, we used the R ‘WeightedCluster’ package developed by (Studer, 2013). The standard objective function for PAM is:

$$\text{minimize } \sum_{i=1}^n \sum_{j=1}^n d(i,j)z_{ij} \quad (2)$$

Where d is the dissimilarity between each pair of sequences and z is a variable ensuring that only the dissimilarity between entities from the same cluster is computed.

For the Weighted version of PAM, the following function is minimized:

$$\text{minimize } \sum_{i=1}^n \sum_{j=1}^n w_{ij}d(i,j)z_{ij} \quad (3)$$

The term w is the weight parameter, which in this study is the frequency of each unique sequence. Consequently, in the weighted PAM method the dimensionality of the full dissimilarity matrix is reduced by creating a vector that contains the frequency of each unique sequence.

At this point it is worth briefly mentioning two alternative approaches considered for Stage 2 and their limitations. The first approach was cluster creation for each year individually and matching these for the 50-year period. This approach would not explicitly consider the temporality of neighborhood differences that may occur overtime (i.e. a cluster type may include observations for specific years which show the changes of neighborhood processes between Census years). The other drawback of this approach is that requires manual matching of the clusters for each period to make them comparable overtime. The second option considered was the high

dimensional space of the dataset to be reduced into two or three dimensions using Principal Component Analysis (PCA) (Mardia *et al.*, 1979) or t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten and Hinton, 2008) and then perform clustering analysis. While the results of these two alternative methods did not drastically diverge from the proposed method, both options were rejected mainly due to the fact that the dimensionality reduction is ‘smoothing’ the data whilst preserving trends but also missing some information. To be more precise PCA finds a linear transformation of the data to minimize the sum of squared errors between the pairwise distances of high dimensional space to their low dimensional, while t-SNE is the most favorable technique for data visualization but it is unclear how it performs when used for other tasks (i.e. such as clustering) (Van Der Maaten and Hinton, 2008). For the purposes of this study data differentiation consider very important and was decided to keep the full dimensions of the dataset.

3 Results and discussion

This section illustrates the results of Stages 2, 3 and 4 described in section 2 in sequential order, starting by the geodemographic classification before discussing the sequence analysis and clustering of sequences to create representative neighborhood trajectories.

3.1 Temporal clustering

As described in Section 3, k-means clustering was performed to create a geodemographic classification of neighborhoods, considering a 40-year period from 1971 to 2011 for Great Britain. Twenty-one variables were included in the analysis, covering demographic, socio-economic and housing characteristics. Eight clusters were returned as the optimal solution.

Table 2 & Figure 2 report the mean variable values for each cluster. The name and key features of the eight neighborhood types are described below and displayed in Figure 3:

- **Affluent:** These are the most affluent areas with most of the population belonging to the managerial socio-economic group with high proportion of population from abroad (10%). These areas are usually suburban and their populations mainly travel to work via private cars (53%). Public transport mode to work is used by 25% reflecting good public transport connections to workplace areas. These areas also have a high proportion of students (4.5%) and owner-occupied houses (76%).
- **Mixed workers suburban:** This group of neighborhoods is characterized by a mixture of people in manual (46%) and non-manual (43%) socio-economic

groups with only a few students (3%). Their residents are largely UK and Republic of Ireland born (96%). There is high proportion of people travel to work with private mode of transport (70%) and finally high proportion of owner-occupied housing (70%).

- Families in council rent: These neighborhoods are predominantly occupied by UK and Republic of Ireland born people (96%). There is high unemployment rate (11%), with high proportion of people staying in council rented housing (77%). Finally, well connected or close to workplace areas as people use public (36%) and active mode of travel to work (22%).
- Blue collar families: These areas characterized by high proportion of manual workers (66%) owing a house (41%) that are predominantly UK and Republic of Ireland born (94%). Close to workplace areas with high proportion of people using active mode to travel to work (38%). This cluster appears in the earlier census years.
- Thriving suburban: These neighborhoods are quite similar to the Affluent areas with the difference of less people belonging to managerial socio-economic Group (18%) and higher ratio of owner-occupied houses (87%). Mainly using private mode to travel to work (74%) and low vacancy rate (4%) which shows that the demand for housing is high.
- Older striving: These neighborhoods are occupied by older people. Mainly manual workers (52%) but with few non-manual (38%) occupations too. There is high vacancy rate (7%) which represents low demand and thus people can afford to buy properties in these areas. The name of the cluster is Older striving but there are people from higher socio-economic Groups (i.e. non-manual and managerial occupations) living in these areas due to the affordability of housing.
- Struggling: Young and middle-aged families UK and Republic of Ireland born (96%) with high unemployment rate (10%) and an even split of people living in council rented (47%) or owner-occupied housing (46%). These neighbourhoods consist of -mainly- manual workers (56%) with few people in non-manual (37%) occupations.
- Multicultural urban: The two main characteristics of these neighborhoods are the high proportion of young people (29%) and high ratio of people born abroad (30%), which makes them highly ethnically diverse. There is a mixture of socio-economic Groups and high ratio of people relying on public (40%) or private (34%) transport to travel to work. It is also worth mentioning the high vacancy rate (7%) of these locations which are predominantly in city centers of urban areas, not the most 'desired' locations for housing in Great Britain.

Table 2 Mean values of each variable by cluster

	Affluent	Mixed workers suburban	Families in council rent	Blue col-lar families	Thriving suburban	Older striving	Struggling	Multicultural urban
Children %	18.5%	19.4%	24.8%	23.0%	18.6%	20.1%	21.5%	17.7%

Middle aged %	21.1%	21.3%	18.1%	16.8%	21.9%	18.8%	20.0%	22.1%
Retired %	16.3%	16.4%	11.7%	14.7%	16.7%	17.1%	14.8%	11.6%
Young Persons %	20.2%	19.2%	22.8%	21.5%	17.1%	21.0%	21.3%	29.2%
Older adults %	23.9%	23.8%	22.6%	24.0%	25.7%	23.0%	22.5%	19.3%
UK/ROI born %	89.8%	96.0%	96.3%	94.3%	95.8%	94.6%	96.1%	70.5%
European born %	3.7%	1.4%	0.8%	1.2%	1.5%	1.4%	1.1%	8.5%
Rest of World %	6.6%	2.5%	2.9%	4.5%	2.8%	4.0%	2.8%	20.9%
Unemployed %	4.7%	6.3%	11.2%	5.7%	3.9%	6.3%	9.7%	8.7%
Students %	4.5%	3.3%	1.8%	0.7%	3.7%	2.3%	3.0%	8.8%
Owner occupied housing %	75.9%	69.5%	17.4%	41.2%	86.7%	66.5%	46.3%	43.4%
Private rented housing %	12.9%	7.9%	5.6%	28.2%	6.6%	16.6%	6.6%	31.5%
Council rented housing %	11.2%	22.6%	77.1%	30.5%	6.7%	16.9%	47.1%	25.0%
Managerial SEG %	20.1%	11.2%	5.4%	6.7%	18.3%	10.2%	8.0%	18.7%
Non-Manual SEG %	52.7%	42.9%	31.4%	27.7%	51.6%	38.2%	36.7%	46.9%
Manual & Others SEG %	27.2%	45.9%	63.2%	65.6%	30.1%	51.6%	55.4%	34.4%
Private TTWM %	53.0%	69.8%	36.9%	30.4%	74.3%	48.2%	57.9%	33.6%
Public transport TTWM %	26.3%	10.4%	35.7%	22.7%	9.1%	15.3%	18.3%	39.7%
Active TTWM %	13.4%	13.4%	22.3%	38.4%	9.4%	28.3%	18.1%	19.5%
Other TTWM %	7.3%	6.3%	5.1%	8.6%	7.2%	8.1%	5.7%	7.2%
Vacancy rate %	4.9%	4.1%	4.8%	6.2%	3.9%	6.9%	4.4%	7.0%

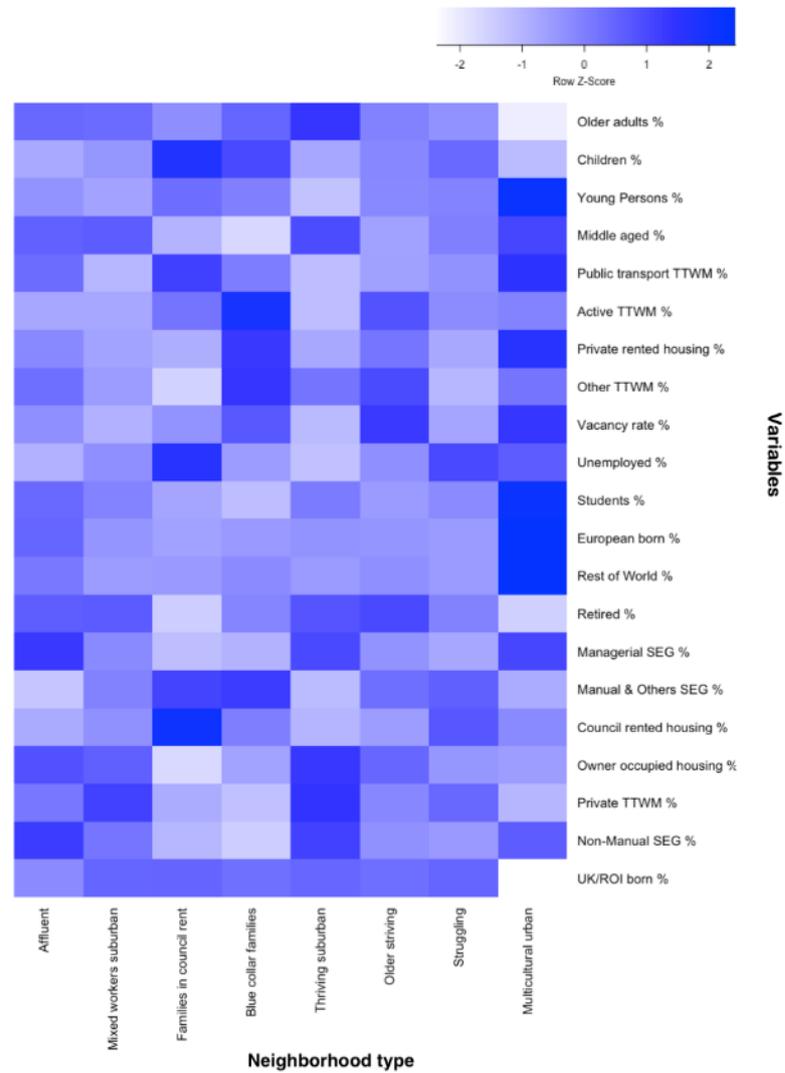
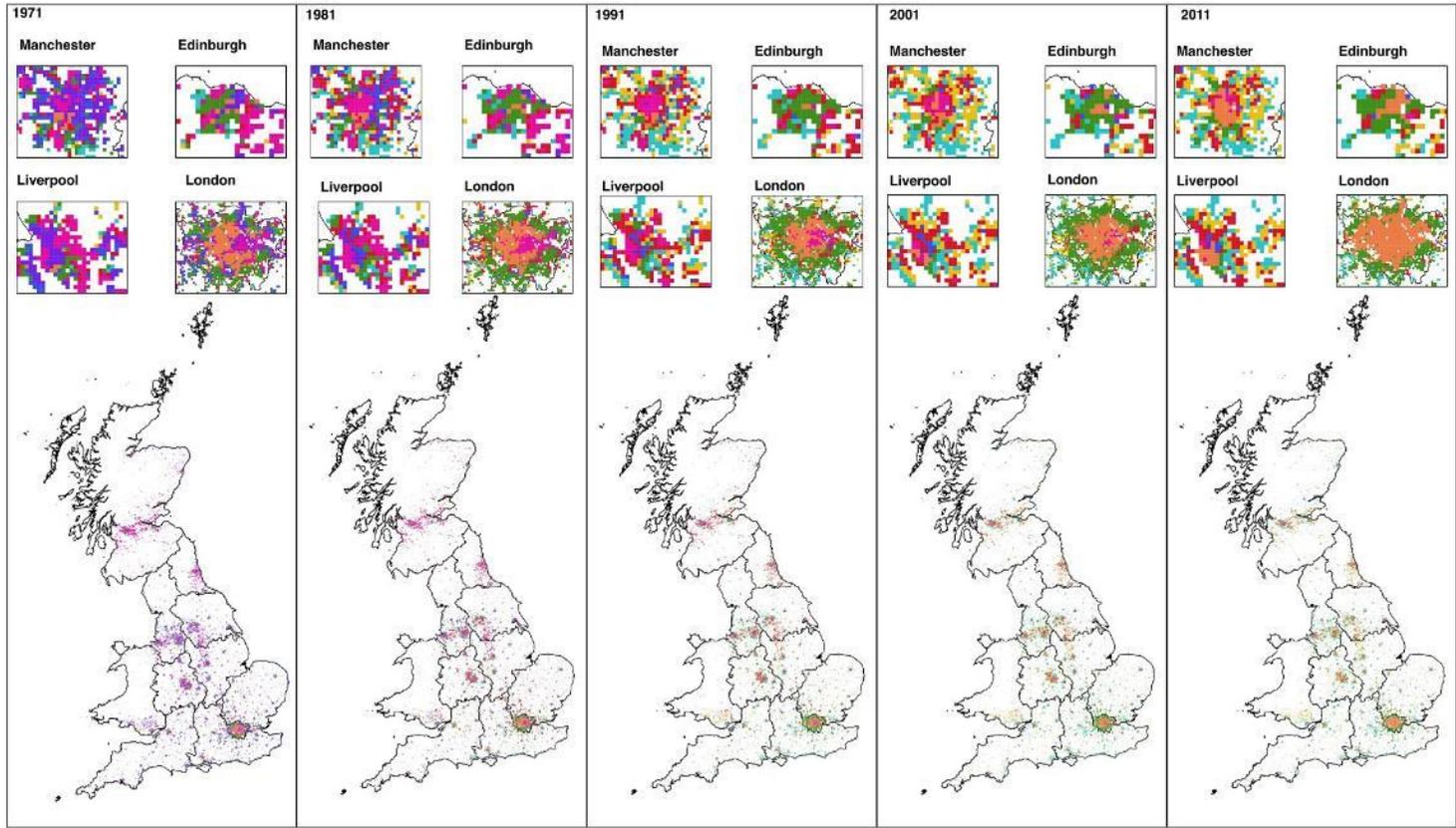


Figure 2 Representative variables across neighborhood type

This classification can be used to analyze spatio-temporal changes of neighborhood types. For example, a marked decrease in the number of blue collar families and families in council rent can be observed across Great Britain over the 40 -year period. Liverpool emerges a prominent example changing from predominantly pink and purple in 1971 to red and yellow in 2011 in Figure 3. The number of multicultural urban neighborhoods have significantly increased from 1971, especially between 2001 and 2011. These changes at the neighborhood level reflect structural

shifts in the population and economy. Key structural changes emerging from the observed patterns are:

- The shrinkage of manual jobs in Great Britain after 1970s;
- The ethnic diversification of urban centers in the 2001 and 2011.



Neighborhood types

- Affluent
- Blue collar families
- Families in council rent
- Mixed workers suburban
- Multicultural urban
- Older striving
- Struggling
- Thriving suburban

Figure 3 Temporal neighborhood clusters in Great Britain

3.2 Sequence analysis

Using sequence analysis, a more comprehensive understanding of spatiotemporal process can be achieved by examining the occurrence, timing, duration and order of transitions between neighborhood types. Figure 4 displays the year-to-year substitution cost matrix to define the sequence of a neighborhood. It shows that lower substitution costs for earlier years, reflecting the higher degree of neighborhood transformation from 2001 onwards (full description of the substitution costs provided in the Appendix). Neighborhoods during the 1970s were more likely to transition between mixed workers suburban to thriving suburban.

In addition to this, the results show that some neighborhood transitions between particular types are more common than others. Thus, the probability of transitioning between affluent and thriving suburban or blue collar families and struggling is higher compared to the probability of transitioning between affluent and blue collar families through all the decades. Yet from 2001 onwards these probabilities have been decreased as mentioned above.

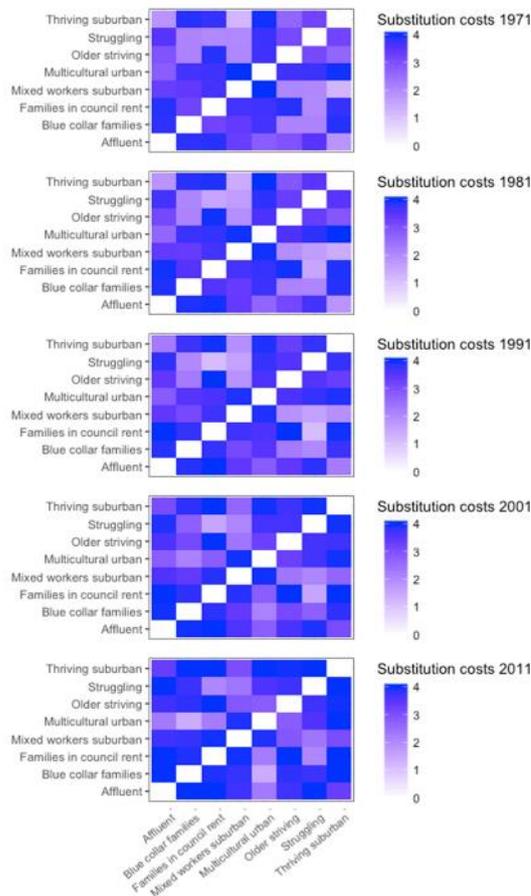


Figure 4 Substitution costs from 1971 to 2011

3.3 Sequences clustering

The substitution costs matrices were then used, to calculate a dissimilarity matrix between individual sequences and derive a typology of neighborhood sequences using the dynamic Hamming method and weighted PAM clustering. Figure 5 displays the resulting typology of seven neighborhoods representing pathways of stability, improvement and decline across the national socio-economic hierarchy. Three panels of graphs are shown in Figure 5. The top panel shows individual sequences. Each line in this graph represents a neighborhood. Each color denotes a

particular type of neighborhood and the x-axis represents each census year. So, horizontally each line shows the transition of a neighborhood between neighborhood type over time. The middle panel displays the year-specific distribution of each neighborhood type. The bottom panel shows the mean time remaining in each neighborhood type.

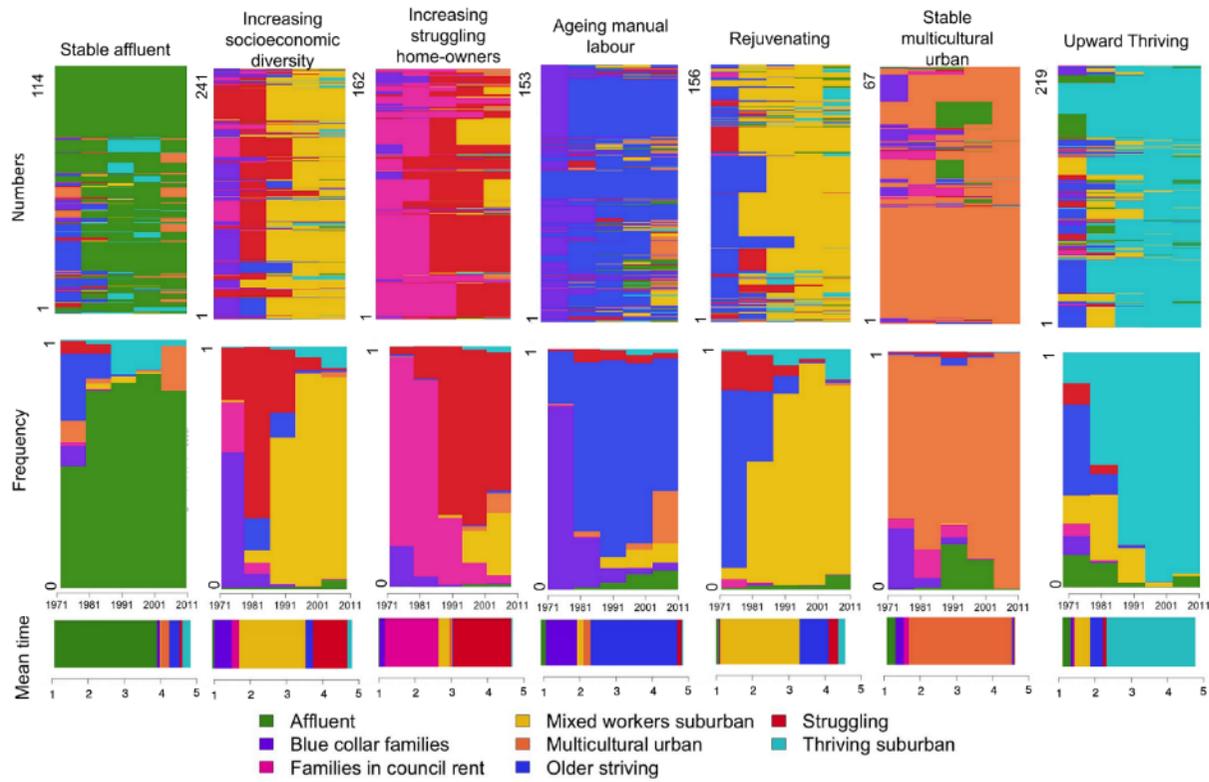


Figure 5 Neighborhood trajectories clusters

The name and key features of the seven main neighborhood transition patterns are described below and displayed in Figure 6:

- Stable affluent neighborhoods: Areas remaining persistently affluent over 1971 and 2011.
- Ageing manual labor neighborhoods: Areas transitioning from being dominated by blue collar families to an older striving neighborhood type.
- Increasingly socio-economically diverse neighborhoods: Areas transitioning from a struggling or blue collar families type to a mixed workers suburban type.
- Increasingly struggling home-owners neighborhoods: Areas transitioning from a families in council rent type to a struggling type.
- Stable multicultural urban neighborhoods: Areas remaining multicultural in urban locations.
- Rejuvenating neighborhoods: Areas transitioning from an older striving type to a mixed workers suburban type.
- Up-warding thriving neighborhoods: Areas transitioning from an older striving type to, or remaining in, a thriving suburban type.

The spatial distribution of these neighborhood trajectories varies between and within areas. There are areas such as Edinburgh and London suburbs that are dominated by stable affluent neighborhoods, while others such as Liverpool and Newcastle have more increasingly struggling home-owners neighborhoods. Regarding the distribution of neighborhood trajectories within areas there are few interesting patterns. One example is the rejuvenating neighborhoods that are characterized by younger people with various socioeconomic backgrounds ‘replacing’ the older population in suburban areas. Another example is the upward trajectories from struggling neighborhoods to more socio-economic diverse and the massive increase of thriving neighborhoods in suburban areas too. Lastly, stable multicultural urban areas appear in all big urban conurbations city or close to city centers.

Finally, ageing of British population is clearly reflected in the results. Suburban and rural areas are largely occupied by retired and older people. Interestingly, this pattern has changed slightly in the last decade, reflecting more inclusive communities both socio-economically and ethnically.

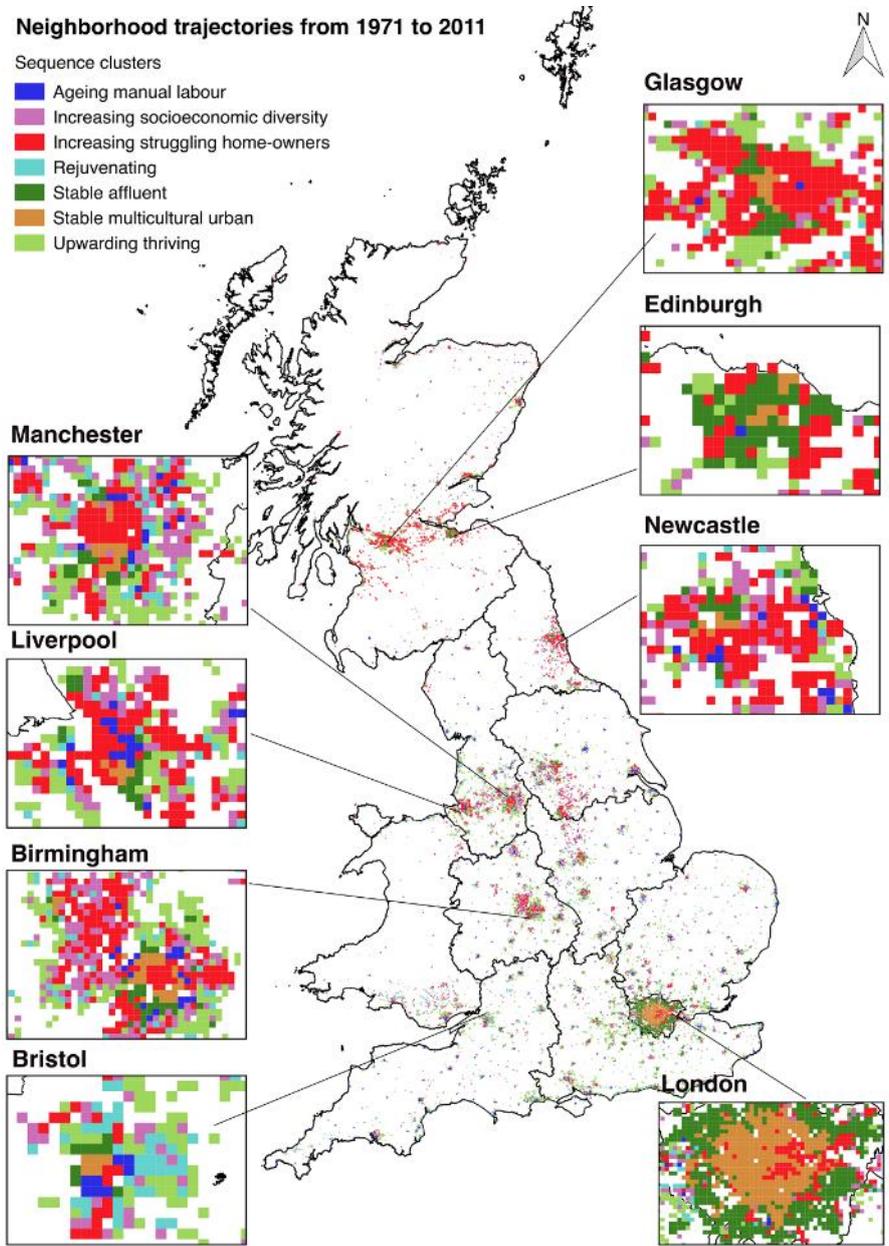


Figure 6 Neighborhood trajectories map

4 Conclusion

This study proposed a novel scalable analytical framework for spatiotemporal data analysis. It does so by (1) producing a temporally consistent spatial framework and geodemographic classification based on 1km² grids; (2) offering the potential to perform analysis at particular administrative, functional or purpose-built areas; (3) implementing a weighted approach to measure dissimilarity between individual neighborhood trajectories; and, (4) integrating multiple approaches (population grid surface estimation, clustering analysis and optimal matching) to analyze long-term change.

The proposed spatiotemporal analytical approach offers a framework within which the evolution of complex demographic and socio-economic processes can be effectively captured and enables understanding of the ways in which past conditions influence present and future transitional changes. Unlike commonly used longitudinal approaches such as event history analysis, which focuses on a single transition, the proposed sequence analysis provides a more comprehensive representation of present and future changes by examining the chronological sequence of events. Such approach enables to unravel key dimensions of changing socio-economic processes in terms of their incidence, prevalence, duration, timing and sequencing – which can serve as useful guidance for policy development.

The proposed approach offers the potential to expand understanding on key demographic and socio-economic processes. A key area for future research is the analysis of trajectories of socio-inequality examining at various levels of spatial aggregation and determining the extent of intra-regional and inter-regional inequalities. Such analysis can guide policy intervention by identifying spatial concentration of poverty and areas undergoing continuous economic decline. Another area of future investigation is the analysis of population change by identifying areas experiencing rapid and continuous population loss or population ageing in the light of sustained low patterns of fertility and signs of declining life expectancy (Green *et al.*, 2017).

References

- Abbott, A. (1983) ‘Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes’ Abbott, Andrew *Historical Methods*; Fall 1983; 16, 4; Periodicals Archive Online pg. 129’.
- Aghabozorgi, S., Seyed Shirkhorshidi, A. and Ying Wah, T. (2015) ‘Time-series clustering - A decade review’, *Information Systems*, 53, pp. 16–38. doi: 10.1016/j.is.2015.04.007.
- An, L. *et al.* (2015) ‘Space–Time Analysis: Concepts, Quantitative Methods, and Future Directions’, *Annals of the Association of American Geographers*, 105(5), pp. 891–914. doi: 10.1080/00045608.2015.1064510.
- Arribas-Bel, D. and Tranos, E. (2018) ‘Characterizing the Spatial Structure(s) of

Cities “on the fly”: The Space-Time Calendar’, *Geographical Analysis*, 50(2), pp. 162–181. doi: 10.1111/gean.12137.

Backman, M., Lopez, E. and Rowe, F. (2018) ‘Career trajectories and outcomes of forced migrants in Sweden: Self-employment, employment or persistent inactivity?’, *Small Business Economics*.

Bailey, T. L. (2017) *Bioinformatics*. doi: 10.1007/978-1-4939-6622-6.

Casado-Díaz, J. M., Martínez-Bernabéu, L. and Rowe, F. (2017) ‘An evolutionary approach to the delimitation of labour market areas: an empirical application for Chile’, *Spatial Economic Analysis*, 12(4), pp. 379–403. doi: 10.1080/17421772.2017.1273541.

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(1), pp. 36–56. doi: 10.1080/00045608.2015.1096188.

Delmelle, E. C. (2017) ‘Differentiating pathways of neighborhood change in 50 U.S. metropolitan areas’, *Environment and Planning A*, 49(10), pp. 2402–2424. doi: 10.1177/0308518X17722564.

Fotheringham, A. S. and Wong, D. W. S. (1991) ‘The Modifiable Areal Unit Problem in Multivariate Statistical Analysis’, *Environment and Planning A*, 23(7), pp. 1025–1044. doi: 10.1068/a231025.

Gabardinho, A. *et al.* (2009) ‘Mining sequence data in R with the TraMineR package: A user’s guide for version 1.1.1’, 1, pp. 1–129.

Gentle, J. E., Kaufman, L. and Rousseeuw, P. J. (1991) *Finding Groups in Data: An Introduction to Cluster Analysis*, *Biometrics*. doi: 10.2307/2532178.

Goodchild, M. F. (2013) ‘Prospects for a Space-Time GIS’, *Annals of the Association of American Geographers*, 103(5), pp. 1072–1077. doi: 10.1080/00045608.2013.792175.

Green, M. A. *et al.* (2017) ‘Could the rise in mortality rates since 2015 be explained by changes in the number of delayed discharges of NHS patients?’, *Journal of Epidemiology and Community Health*, 71(11), pp. 1068–1071. doi: 10.1136/jech-2017-209403.

Hayward, P. and Parent, J. (2009) ‘Modeling the influence of the modifiable areal unit problem (MAUP) on poverty in Pennsylvania’, *The Pennsylvania Geographer*, 47(1), pp. 120–135.

Hoover, E. M. and Vernon, R. (1959) *Anatomy of a metropolis; the changing distribution of people and jobs within the New York metropolitan region, New York metropolitan region study*. Cambridge (Mass.): Harvard University Press, 1959. (New York metropolitan region. Study: no. 1). Available at: <http://mirlyn.lib.umich.edu/Record/004478493>.

Huang, B. (2017) *Comprehensive Geographic Information Systems*. Elsevier.

Janssen, H. J. and Ham, M. Van (2019) ‘Resituating the Local in Cohesion and Territorial Development Report on multi-scalar patterns of inequalities’, (January). doi: 10.13140/RG.2.2.23832.65287.

Kyriakidis, P. C. and Journel, A. G. (1999) ‘Geostatistical space-time models: A review’, *Mathematical Geology*, 31(6), pp. 651–684. doi:

10.1023/A:1007528426688.

Lesnard, L. (2009) *Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns*, *Sociological Methods and Research*. doi: 10.1177/0049124110362526.

Lloyd, C. D. *et al.* (2017) 'Exploring the utility of grids for analysing long term population change', *Computers, Environment and Urban Systems*. The Authors, 66, pp. 1–12. doi: 10.1016/j.compenvurbsys.2017.07.003.

Van Der Maaten, L. J. P. and Hinton, G. E. (2008) 'Visualizing high-dimensional data using t-sne', *Journal of Machine Learning Research*, 9, pp. 2579–2605. doi: 10.1007/s10479-011-0841-3.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) 'Multivariate statistics'. Academic Press,.

Miller, H. J. (2015) 'Space-time Data Science for a Speedy World', *I/S: A Journal of Law and Policy for the Information Society*, 10(3), pp. 705–720. doi: 10.1525/sp.2007.54.1.23.

Openshaw, S. (1983) 'The modifiable area unit problem', *Concepts and Techniques in Modern Geography*, 38, pp. 1–41. doi: 10.1177/1077558707312501.

Prouse, V. *et al.* (2014) 'How and when Scale Matters: The Modifiable Areal Unit Problem and Income Inequality in Halifax', *Canadian Journal of Urban Research*, 23(1), pp. 61–82.

Rowe, F. (2017) 'The CHilean Internal Migration (CHIM) database: Temporally consistent spatial data for the analysis of human mobility', *Region*, 4(3), p. 1. doi: 10.18335/region.v4i3.198.

Rowe, F., Casado-Díaz, J. M. and Martínez-Bernabéu, L. (2017) 'Functional Labour Market Areas for Chile', *Region*, 4(3), p. 7. doi: 10.18335/region.v4i3.199.

Rowe, F., Corcoran, J. and Bell, M. (2017) 'The returns to migration and human capital accumulation pathways: non-metropolitan youth in the school-to-work transition', *Annals of Regional Science*. Springer Berlin Heidelberg, 59(3), pp. 819–845. doi: 10.1007/s00168-016-0771-8.

Sanger, F. and Nicklen, S. (1977) 'DNA sequencing with chain-terminating', 74(12), pp. 5463–5467.

Studer, M. (2013) 'WeightedCluster Library Manual', pp. 1–34. doi: 10.12682/lives.2296-1658.2013.24.

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(2), pp. 481–511. doi: 10.1111/rssa.12125.

Teernstra, A. B. and Van Gent, W. P. C. (2012) 'Puzzling Patterns in Neighborhood Change: Upgrading and Downgrading in Highly Regulated Urban Housing Markets', *Urban Geography*, 33(1), pp. 91–119. doi: 10.2747/0272-3638.33.1.91.

Warren Liao, T. (2005) 'Clustering of time series data - A survey', *Pattern Recognition*, 38(11), pp. 1857–1874. doi: 10.1016/j.patcog.2005.01.025.

Appendix

Substitution cost matrices

1971 substitution costs

	Affluent	Blue col- lar fami- lies	Families in council rent	Mixed workers suburban	Multicul- tural ur- ban	Older striving	Struggling	Thriving suburban
Affluent	0.00	3.96	3.97	3.81	3.65	3.70	3.87	3.24
Blue collar families	3.96	0.00	3.78	3.83	3.92	3.39	3.38	3.97
Families in council rent	3.97	3.78	0.00	3.92	3.93	3.97	3.33	3.95
Mixed workers suburban	3.81	3.83	3.92	0.00	4.00	3.36	3.34	2.87
Multicul- tural urban	3.65	3.92	3.93	4.00	0.00	3.93	3.94	3.99
Older striv- ing	3.70	3.39	3.97	3.36	3.93	0.00	3.75	3.60
Struggling	3.87	3.38	3.33	3.34	3.94	3.75	0.00	3.78
Thriving suburban	3.24	3.97	3.95	2.87	3.99	3.60	3.78	0.00

1981 substitution costs

	Affluent	Blue col- lar fami- lies	Families in council rent	Mixed workers suburban	Multicul- tural ur- ban	Older striving	Struggling	Thriving suburban
Affluent	0.00	3.97	3.98	3.82	3.58	3.75	3.91	3.23
Blue collar families	3.97	0.00	3.87	3.81	3.93	3.37	3.36	3.96
Families in council rent	3.98	3.87	0.00	3.91	3.94	3.98	3.06	3.97
Mixed workers suburban	3.82	3.81	3.91	0.00	3.98	3.29	3.15	2.99
Multicul- tural urban	3.58	3.93	3.94	3.98	0.00	3.89	3.95	3.99
Older striv- ing	3.75	3.37	3.98	3.29	3.89	0.00	3.80	3.67
Struggling	3.91	3.36	3.06	3.15	3.95	3.80	0.00	3.85
Thriving suburban	3.23	3.96	3.97	2.99	3.99	3.67	3.85	0.00

1991 substitution costs

	Affluent	Blue col- lar fami- lies	Families in council rent	Mixed workers suburban	Multicul- tural ur- ban	Older striving	Struggling	Thriving suburban
Affluent	0.00	3.97	4.00	3.84	3.66	3.84	3.96	3.45
Blue collar families	3.97	0.00	3.95	3.75	3.87	3.45	3.33	3.94
Families in council rent	4.00	3.95	0.00	3.93	3.90	4.00	2.79	3.99
Mixed workers suburban	3.84	3.75	3.93	0.00	3.98	3.26	3.11	3.28
Multicul- tural urban	3.66	3.87	3.90	3.98	0.00	3.90	3.95	3.98
Older striv- ing	3.84	3.45	4.00	3.26	3.90	0.00	3.88	3.81
Struggling	3.96	3.33	2.79	3.11	3.95	3.88	0.00	3.95
Thriving suburban	3.45	3.94	3.99	3.28	3.98	3.81	3.95	0.00

2001 substitution costs

	Affluent	Blue col- lar fami- lies	Families in council rent	Mixed workers suburban	Multicul- tural ur- ban	Older striving	Struggling	Thriving suburban
Affluent	0.00	3.99	4.00	3.89	3.64	3.90	3.98	3.74
Blue collar families	3.99	0.00	3.96	3.83	3.40	3.75	3.63	3.96
Families in council rent	4.00	3.96	0.00	3.97	3.64	4.00	3.08	4.00
Mixed workers suburban	3.89	3.83	3.97	0.00	3.99	3.49	3.37	3.59
Multicul- tural urban	3.64	3.40	3.64	3.99	0.00	3.80	3.92	3.99
Older striv- ing	3.90	3.75	4.00	3.49	3.80	0.00	3.92	3.93
Struggling	3.98	3.63	3.08	3.37	3.92	3.92	0.00	3.99
Thriving suburban	3.74	3.96	4.00	3.59	3.99	3.93	3.99	0.00

2011 substitution costs

	Affluent	Blue col- lar fami- lies	Families in council rent	Mixed workers suburban	Multicul- tural ur- ban	Older striving	Struggling	Thriving suburban
Affluent	0.00	4.00	4.00	3.93	3.46	3.93	4.00	3.81
Blue collar families	4.00	0.00	3.98	3.95	3.00	3.96	3.94	4.00
Families in council rent	4.00	3.98	0.00	3.99	3.44	4.00	3.36	4.00
Mixed workers suburban	3.93	3.95	3.99	0.00	3.98	3.68	3.49	3.73
Multicul- tural urban	3.46	3.00	3.44	3.98	0.00	3.65	3.89	4.00
Older striv- ing	3.93	3.96	4.00	3.68	3.65	0.00	3.93	3.99
Struggling	4.00	3.94	3.36	3.49	3.89	3.93	0.00	4.00
Thriving suburban	3.81	4.00	4.00	3.73	4.00	3.99	4.00	0.00