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

Nikos Patias*[1], Francisco Rowe[1] and Stefano Cavazzi[2],

[1]Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool
[2]Ordnance Survey Limited

Summary

Spatio-temporal changes reflect the complexity of real-life events. 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. We use gridded cell Census data for Great Britain from 1971 to 2011 with 10-year intervals, creating neighbourhood typologies for each Census year. These typologies are then used to analyse transitions of grid cells between different types of neighbourhoods and define representative trajectories of neighbourhood change. The results reveal seven prevalent trajectories of neighbourhood change across Great Britain, identifying neighbourhoods which have experienced stable, upward and downward pathways through the national socioeconomic hierarchy over the last four decades.

KEYWORDS: Neighbourhood change, Sequence analysis, Spatio-temporal data analysis, Classification, Population dynamics

1. Introduction

Changes over space and time reflect the complexity of real-life events (Miller, 2015). Yet measuring the actual magnitude, location and temporal frequency of change has been challenging. Using traditional forms of data (i.e. longitudinal data), change can only be captured infrequently, namely every month, year or decade. Using new forms of data (i.e. social media, mobile phone data), change can be captured in near-real time, every few seconds or minutes. Spatial attributes can also vary, particularly the geographical level of analysis. An example can be seen in the inconsistency of spatial frameworks (i.e. administrative boundaries) over time (Goodchild, 2013). The spatial reference of new forms of data, however, is typically at a fine level of detail, referring to individual locational points (An *et al.*, 2015). Despite this, data aggregation is often used to extract conclusions, regardless the nature of the data (i.e. individual or aggregations).

Different levels of spatial aggregation can, however, produce different representations of socio-economic processes as a result of the Modifiable Areal Unit Problem (MAUP). The MAUP refers to the statistical sensitivity and variability of results related to the spatial framework of analysis (Openshaw, 1983). The most appropriate spatial framework of analysis may thus differ according to the selection of variables (Prouse *et al.*, 2014). MAUP can create 'artificial' 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, helps to tackle this issue by offering the possibility to analyze temporal data, regardless of changes at administrative geographical boundaries by creating appropriately customised spatial frameworks.

While there is an increasing number of methods for spatio-temporal data processing, they often focus on exploring observations rather than explicitly capture changes between their spatial and temporal 'states'. The most important part of spatio-temporal data processing is knowledge generation and how we can draw useful generalisable conclusions. Clustering techniques are usually employed on space-

* n.patias@liverpool.ac.uk

time data to identify systematic prevalent patterns (Warren Liao, 2005). While clustering approaches on their own can provide useful information about movements and points of interest. Integration of different approaches can provide context 'aware' data and overcome the limitation of comparing points in time but provide a more complete picture of evolution by analysing sequences.

Creating sequences of events or 'states' can help gaining a better understanding of spatio-temporal patterns. A useful approach is sequence analysis. Sequence analysis was originally developed for analysing DNA sequences and measuring their similarities (Sanger and Nicklen, 1977). In the 1980s, it was introduced in the social sciences (Abbott, 1983), offering the opportunity to be applied to longitudinal data to measure change across space and over time (Studer and Ritschard, 2016).

Usefully, sequence analysis can be applied to spatio-temporal data to better understand the 'life-cycle' of places. According to Chicago School scholars, neighbourhoods do not exist in isolation but they interact, socially and spatially (Park and Burgess, 1921). These interactions have, however, predominantly been perceived in a static cross-sectional framework, neglecting the diversity and sequence of transitions of change through neighbourhoods progress (Hoover & Vernon 1959). Conceptually, neighbourhoods are expected to undergo the same rigid experience of socio-demographic change from emergence, increasing affluence, affluence to decline. Arguably the limited development in the conceptualisation of neighbourhood change is partly due to the lack of comprehensive longitudinal data based on temporally consistent geographical boundaries at a meaningful detailed spatial scale and purposely-built methodological tools to analyse trajectories of neighbourhood change in multi-dimensional data sets.

In this study, we aim to provide a scalable analytical framework for spatio-temporal data analysis contributing in two ways to the current literature:

1. Taking explicitly into account the timing, frequency and sequence of the transitions between different states;
2. Provide a robust scalable analytical framework that can be applied to large-scale datasets.

2. Data and Methods

2.1. Data

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

Popchange, is a visualisation platform that uses 1km^2 grid as a reference geography and calculates the correspondence between those and low-level Census administrative geographies (Lloyd *et al.*, 2017). Considering this geographical inconsistency of boundaries over time, we have used the *Popchange* project code which uses as input the Census data at the lowest geographical level (i.e. enumeration districts for 1971, 1981 and 1991 and output areas for 2001 and 2011) and outputs the data in 1km^2 grid. While the grid data values are estimations of variables based on Census data, they offer two advantages. Firstly, it makes possible to directly compare grid level data over a period of time (i.e. 50 years in our case) but also helps to tackle the MAUP. Selecting a spatial unit that is based on geographical coordinates (i.e. 1km^2 grid) rather than aggregation of Census statistical units should provide comparable and scalable results over the 40 years period investigated.

2.2. Methods

The methodological framework developed in this study is split into five steps as presented in **Figure 1**. The first stage involves the creation of gridded population data. This data is then used in the second stage to create neighbourhood typologies using k-means clustering. The neighbourhood typologies are used in the third stage to analyse the transition of individual grid cells across neighbourhood types and calculate transition rates which is the cost of transitioning from one neighbourhood type to another. These rates vary over time as they are based on probabilities calculated for each year. Using the transition rates, a dissimilarity matrix between each neighbourhood sequence (i.e. trajectory) pair is created to measure their similarity using optimal matching method. Finally, the measure of similarity is used in the fourth stage to define typologies of neighbourhood trajectories using weighted k-medoids clustering approach.

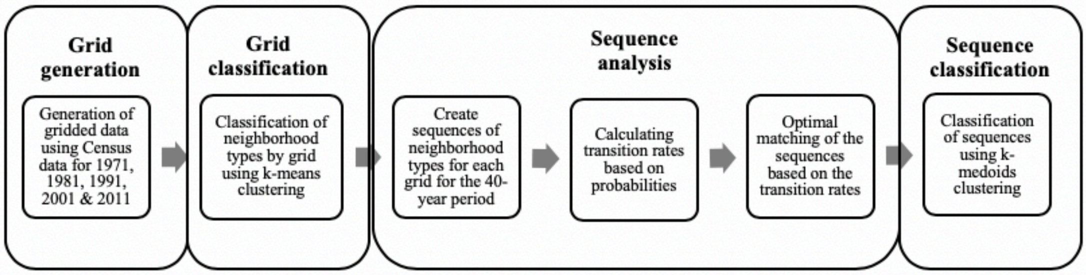

Figure 1 Methodological framework workflow.

3. Results and discussion

3.1. Temporal clustering

Geodemographic classification provides an effective approach to represent the internal socio-economic structure of neighbourhoods and it is used in this paper to identify eight neighbourhood signatures based on the gridded data from the British 1971 to 2011 censuses. These signatures were defined using k-means clustering applied to data capturing three key dimensions: demographic (the percentage of population by age band, ethnicity and student status), socio-economic (the percentage of population by socio-economic group, mode of travel to work, and unemployment status) and housing (the percentage of population by home ownership status, and vacancy rate) dimensions. The eight neighbourhood typologies are described in more detail in the following paragraphs and shown spatially in **Figure 3**:

1. Affluent neighbourhoods: large shares of managerial and professional non-manual occupations and owned houses.
2. Mixed workers suburban neighbourhoods: high shares of manual and non-manual workers of British nationality in suburban areas.
3. Families in council rent neighbourhoods: high shares of households with children living in council rented housing, in manual occupations and of British nationality.
4. Blue collar families neighbourhoods: high shares of manual workers and children using active modes of commuting (i.e. walking or cycling).
5. Thriving suburban neighbourhoods: high shares of middle-age and older adults, living in owner occupied housing and working in non-manual occupations.
6. Older striving neighbourhoods: high shares of retirees and vacancy rates.
7. Struggling neighbourhoods: high shares of British born and unemployed population.
8. Multicultural urban neighbourhoods: high shares of young, middle aged and student populations from ethnically diverse background, living in private rented housing and heavy users of public transport.

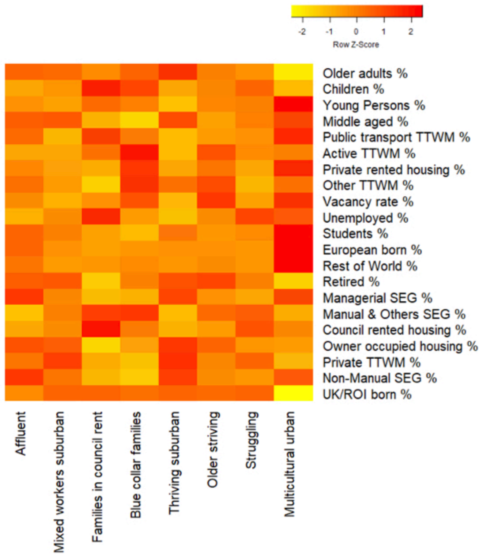

Figure 2 Representative variables across neighbourhood types.

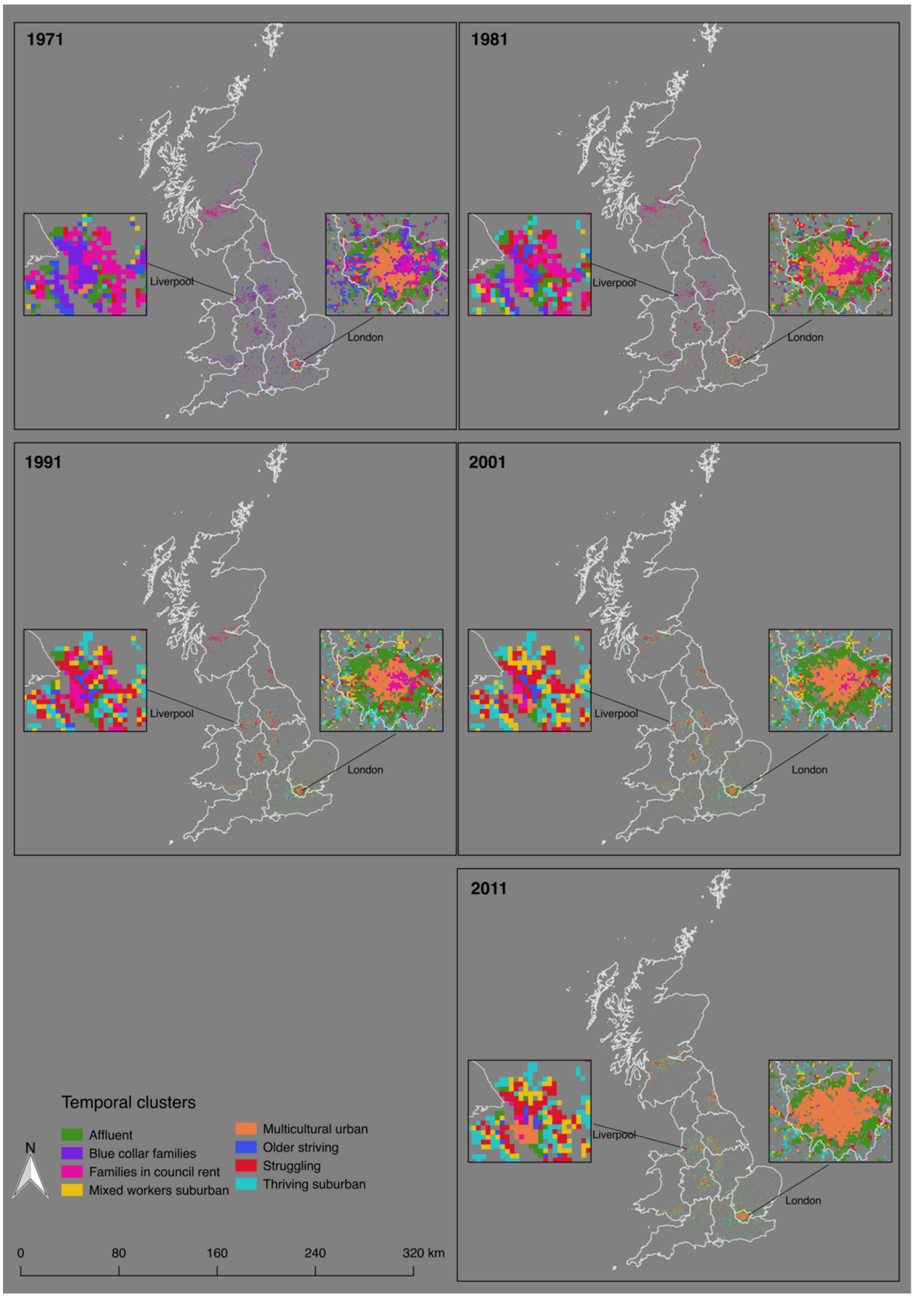

Figure 3 Temporal neighbourhood clusters in Great Britain.

3.2. Sequence analysis

We understand that urban response to neighbourhood transition can vary across space and time but also across different geographical levels. Sequence analysis provides a comprehensive way to analyze spatiotemporal data considering the number of times that a data point was in a particular state (i.e. neighbourhood type), the duration, the time when the transitions happen and finally the actual chronological trajectory of transitions. In this study, we consider the timing as an important factor influencing transitions. As such, we explicitly consider the variations in the probability of transitions by calculating year-to-year substitution cost matrices to define the sequence of a neighbourhood.

3.3. Sequence 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 so to group more similar sequences. We call these sequences neighbourhood trajectories, as they show the ‘route’ each area followed over a 40-year period. **Figure 4** illustrates the sequence clusters making apparent that sequences can be stable or could follow upward/downward trends in the socioeconomic hierarchy over time. The three main elements of the graph show all the sequences in the first row, the distribution of these sequences in the second row and the mean time spent in each state in the last (i.e. neighbourhood type).

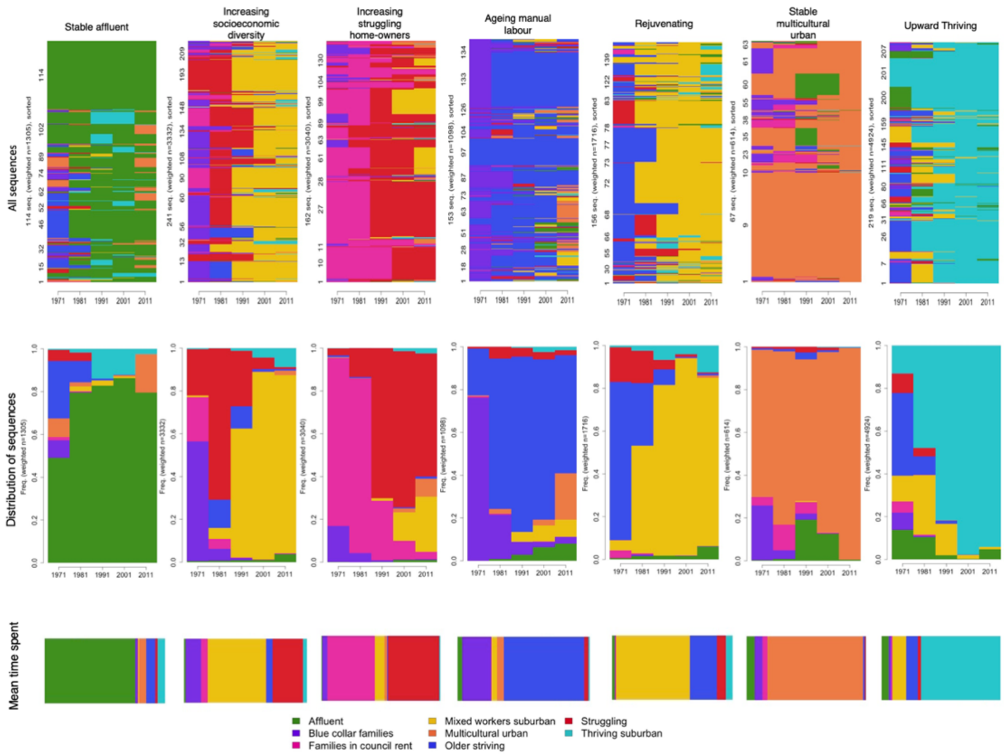

Figure 4 Neighbourhood trajectories clusters

These three elements of **Figure 4** along with the map showing the spatial structure of these neighbourhood trajectories (see **Figure 5**) can help us to gain some useful insights. The seven main sequence patterns identified are:

1. Stable affluent neighbourhoods: areas remaining persistently affluent over 1971 to 2011.
2. Ageing manual labour neighbourhoods: areas transitioning from being dominated by blue collar families to an older striving neighbourhood type.
3. Increasingly socio-economically diverse neighbourhoods: areas transitioning from a struggling or blue collar families type to a mixed workers suburban type.
4. Increasingly struggling home-owners neighbourhoods: areas transitioning from a families in council rent type to a struggling type.
5. Stable multicultural urban neighbourhoods: areas remaining multicultural in urban locations.
6. Rejuvenating neighbourhoods: areas transitioning from an older striving type to a mixed workers suburban type.
7. Up-warding thriving neighbourhoods: areas transitioning from an older striving type to, or remaining in, a thriving suburban type.

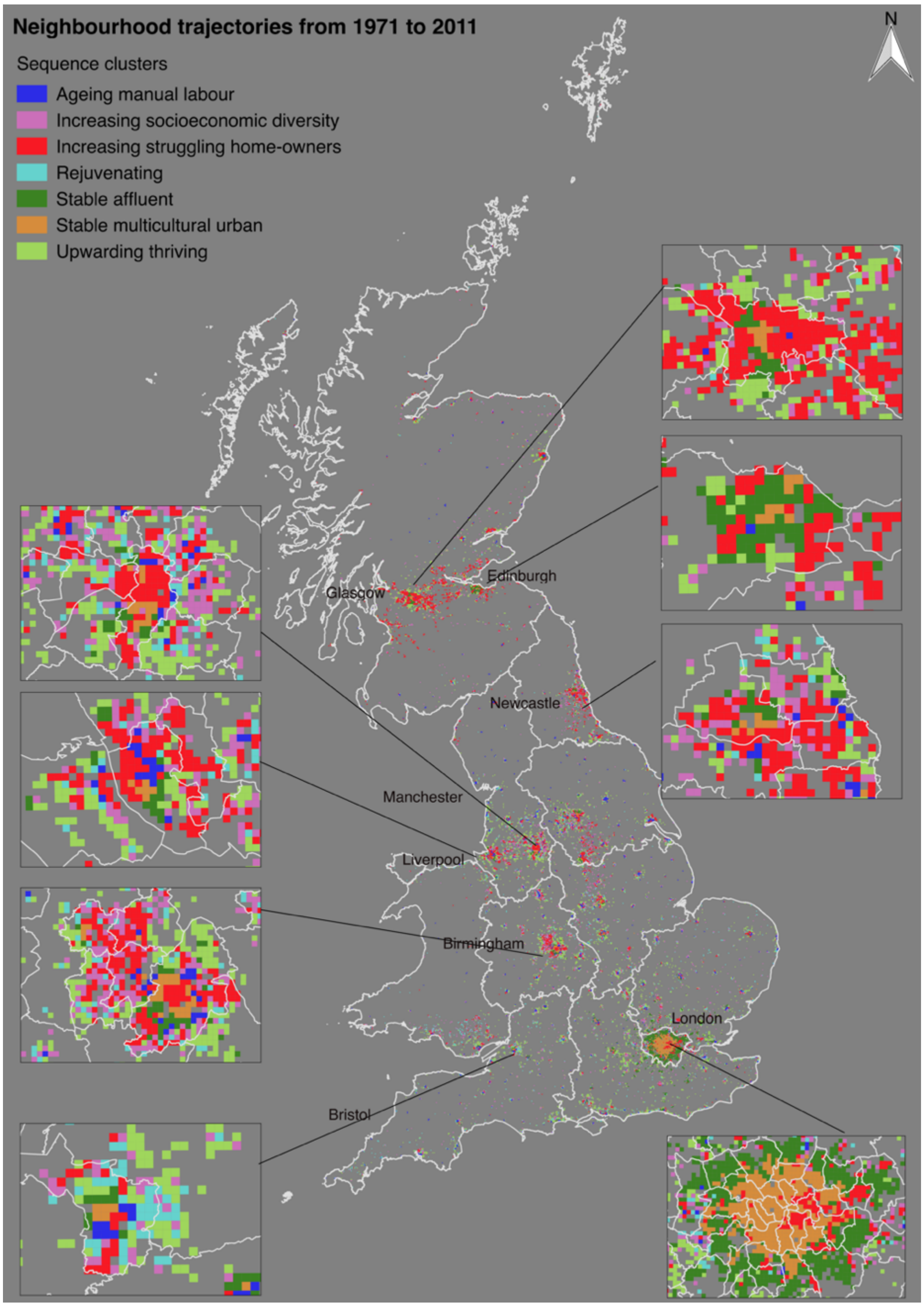

Figure 5 Neighbourhood trajectories map

4. Conclusion

This study provided a novel scalable analytical framework for spatiotemporal data analysis. The originality of this framework can be summarised in three points:

1. The proposed framework explicitly incorporates the timing of the transitions between 'states' and captures the prevalence or short temporality of neighbourhood states over time;
2. The framework supports large multi-dimensional data sets by using a weighted clustering method to optimise computer memory and time;
3. The framework is built through the integration of sophisticated analytical techniques, namely conversion of administrative boundary data to a gridded format, clustering methods and optimal matching, providing a more comprehensive analysis and presentation of the results.

5. Acknowledgements

This abstract is a part of a four-year PhD project assessing the extent, sequence, pace and spatial pattern of neighbourhood change in Britain over a 40-year period from 1971 to 2011. The project is funded by the Economic and Social Research Council (ESRC) and Ordnance Survey through the Data Analytics and Society Centre for Doctoral Training.

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.

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

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.

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.

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.

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.

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.

Park, R. E. and Burgess, E. W. (1921) *Introduction to the Science of Sociology*. University of Chicago Press.

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.

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

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.

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.

Biographies

Nikos Patias is a PhD student at the Department of Geography and Planning in the University of Liverpool and member of Geographic Data Science Lab. His research focuses on assessing the extent, sequence, pace and spatial pattern of neighbourhood change in Britain over a 40-year period from 1971 to 2011. He holds a master’s degree in ‘Real estate and planning’ from Heriot Watt University. Prior to this, he completed his undergraduate degree on ‘Spatial planning and development’ at the Aristotle University of Thessaloniki.

Francisco Rowe is a Senior Lecturer in Human Quantitative Geography at the University of Liverpool, member of the Geographic Data Science Lab and a Project Associate of the international project IMAGE: Comparing Internal Migration Around the Globe. Francisco joined as a Lecturer at Liverpool in 2016. Prior to this, he was a Research Postdoctoral Fellow at the Queensland Centre for Population Research at the University of Queensland in Australia (2013-2016). During this time, he was involved in a longitudinal study exploring the migration, educational and employment outcomes and pathways of young people in Australia, working with the State Government of Victoria. Francisco completed his PhD in Economic Geography at the University of Queensland (2009-2013) under the supervision of Martin Bell and Robert Stimson. His PhD explored the evolution of internal migration and long-distance commuting patterns of the workforce in Chile during the transition to economic liberalisation. He also holds a MSc in Regional Science (2007-2008) and a First Class BA in Business Management with specialisation in Economics (2002-2007) from the Universidad Catolica del Norte, Chile. His commitment to continued professional development was recognised with the award of Fellow of the Higher Education Academy status (2018). Francisco is editor of REGION, the journal of the European Regional Science Association (2018-present) and a member of the Regional Studies Association. His research has been published in leading international journals, such as Transport Research Part C, Applied Geography and Population Studies. Francisco received an award for the best paper published in Spatial Economic Analysis in 2018 and works closely with government organisations, the United Nations Economic Commission for Latin America and the Caribbean, Ordnance Survey, and the ONS Data Campus.

Dr Stefano Cavazzi is a Principal Innovation and Research Scientist at Ordnance Survey where he leads the development of GIS and geomatics research programmes. Ordnance Survey (OS) is the national mapping agency for Great Britain, and a world-leading geospatial data and technology organisation.
Stefano has fifteen years of experience in the geospatial sector in both academia and industry specialising in geospatial data science. He holds a PhD from the University of Cranfield where he investigated spatial scale analysis to model earth's terrain attributes used as primary inputs in machine learning methods.