Sidestepping geographic harmonisation: a new method for handling data over evolving spatial regions

August 30, 2019

Abstract

This paper proposes a novel method for data-driven identification of spatio-temporal homogeneous regions and their dynamics, enabling the exploration of their composition and extents. Using a simple network representation, the method enables temporal regionalisation *without the need for geographical harmonisation.* 1.1/2.1

To allow for a transparent corroboration of our method, we use it as a basis for an interactive and intuitive interface for the progressive exploration of the results. The interface guides the user through the original data, enabling both experts and non-experts to characterise broad patterns of stability and change and identify detailed local processes.

The proposed methodology is suitable for any region-based data, and we validate our method with illustrative scenarios from Chicago and Toronto, with results that match the established literature. The system is publicly available, with demographic data for over forty regions in the USA and Canada between 1970 and 2010.

1 Introduction

Neighbourhoods have increasingly become a central concept in social research and targets for social policy (Sampson, 2012; Galster, 2019; Stone et al., 2015; Looker, 2015). To be sure, a focus on neighbourhoods extends to the formative period of the modern social sciences (Abbott, 1997). Recent interest has at least partly been rekindled through newly available longitudinal demographic datasets (Logan et al., 2014; Manson et al., 2017), convenient computational tools (Rey et al., 2018), and new sources of data (Poorthuis, 2018).

Yet new challenges have also emerged, especially at the convergence of research on neighbourhood effects and neighbourhood dynamics. Neighbourhood effects research assumes knowledge about the nature and scope of "the neighbourhood" that presumably shapes individual outcomes (Kwan, 2018; Shelton and Poorthuis, 2019). Concurrently, researchers note that neighbourhoods are not necessarily fixed containers in which other processes occur, but themselves dynamically evolve (Delmelle, 2017; Reades et al., 2019; Li and Xie, 2018). The result is to open up key assumptions about neighbourhoods for theoretical and empirical examination: how do we appropriately define and compare neighbourhoods at a given time?; how do we appropriately define and compare the temporal trajectories of neighbourhoods?; and can we do both at once, "fully interactionally" (Abbott, 1997): classify neighbourhoods now based on where they came from and where they are going?

In principle, much of the recent research is committed to the proposition that neighbourhoods are open and evolving entities. Ironically, its empirical practice tends to rely on methods that require fixed geographical regions. This requirement is not trivial to satisfy in general, as most longitudinal datasets are based on pre-defined tabulation areas that are routinely modified by data collection agencies, usually to follow population changes.

The standard approach then is to *geographically harmonise* the data. This involves interpolating existing measurements into a common set of regions (Logan et al., 2014; Hallisey et al., 2017; Allen and Taylor, 2018). Recent computational tools have somewhat simplified this process (Rey et al., 2018), but it still involves non-trivial questions: which geometry to use as target, how to apportion 1.18 the variables, or how to combine data from different sources. Further, these question do not necessarily have optimal answers. Indeed, regardless of how well this process is performed, it still introduces errors (Logan et al., 2016), even when additional data is provided (Eicher and Brewer, 2001). Essentially, harmonisation generates *artificial data points* that can potentially lead to inaccurate results, even though they are seldom interpreted as such. Nevertheless, because there has been no viable alternative, and the results often appear plausible, these concerns are overlooked. The result is that the harmonisation approach is virtually mandatory in the current literature: *”(...) tract-by-tract comparison is not possible unless data from 2000 is interpolated to 2010 boundaries (...)”* (Dmowska et al., 2017), *”(...) This limits cross-year comparison since data are not representative of the same spatial units. (...)”* (Allen and Taylor, 2018).

In some cases, harmonisation might seem methodologically adequate. If a study is focused on 1.1 a small subregion that is fixed in time, limiting and modifying the data to fit that specific region of interest sounds not only reasonable, but recommended, even if it comes at the cost of producing artificial data points. However, this view is limited in that it ignores interactions beyond the fixed borders: each different portion of the area of interest may belong to larger homogeneous processes beyond these arbitrarily defined borders. This effect, along with the creation of artificial points, is caused by the geometry mismatch between the data and the region of interest. Instead of forcing one set of borders into another, our method sidesteps this issue by changing how the data is represented and interpreted. It therefore allows researchers to examine changes in both the composition and extent of local areas.

The main contribution of this paper is a method for longitudinal geographical data representation and processing that works with the original data by leveraging a network based representation. It enables direct comparison and temporal regionalisation without the need to match arbitrarily defined borders. Even in the case of a small, temporally fixed, region of interest, our approach leads to a better understanding of its dynamics by exploring the related data-driven homogeneous regions.

To allow a proper examination of our method and its results, we built a prototype online interactive system using this representation. It enables users to visualise, interpret, and explore trajectories of neighbourhood change. This prototype helps validate our method, by allowing it to be compared to existing and future methods. Further, it is a significant contribution to the research community: it provides a vehicle for quickly and easily grasping complex long-term changes, experimenting with different parameters to interactively learn from data, and making neighbourhood change research publicly transparent. The interface thus responds to increasing concerns about reproducibility and transparency, as well as ongoing attention to the value of visualisation in scientific research and communication.

We start by presenting an intuitive example of our representation in Section 1, then we review the relevant literature on longitudinal studies, data representation, clustering and regionalisation, and spatio-temporal visualisation in Section 3. Our methodology is introduced in detail in Section 4, along with the included interface. Illustrative scenarios for Chicago and Toronto are presented in Section 5 and the feedback of five field experts are summarised in Section 6. Our prototype system is available at , including more than forty regions in the US and Canada. The source code is publicly available at .[1]

2 Intuition

While utterly simple, the network model breaks from the deeply rooted traditional tabular paradigm in a significant way. The traditional method requires the data to be treated as a collection of fixed

[1] The editors are considering, at our request, an exception to the double-blind requirement to allow access to the system. We provided them with the URLs of the system, code, and documentation separately.

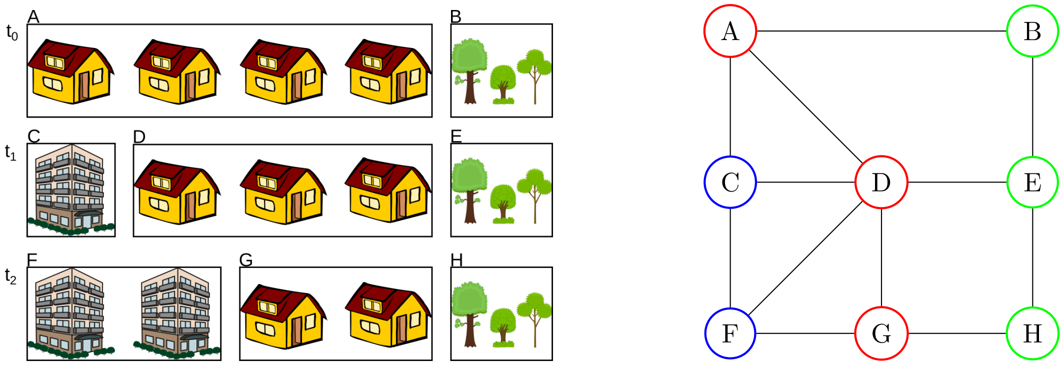


Figure 1: Network based spatio-temporal data representation. **Left**: Three temporal stages of the evolution of a fictitious urban area, with aggregation areas A to H. **Right**: Network representation of the aggregation areas where the colours identify similar regions.

entities with properties that evolve over time – rows in a table with temporal values as columns. By contrast, our method represents each measurement as a separate entity and encodes the evolution of these entities over time.

To ease the cognitive transition to this paradigm, we start with an intuitive example of how the method works, using a small portion of a fictitious urban region illustrated on the left part of Figure 1. This example includes three different times (t_0, t_1, t_2), with different aggregation areas identified as letters from A to H. For t_0, the initial time, we have areas A and B, with small houses and a park, respectively. The park remains stable (B, E, and H), but the houses are partially replaced by larger buildings (C and F).

The aggregation areas of none of these years is clearly suitable as an interpolation target. Adopting any of them would require merging heterogeneous regions and/or dividing homogeneous regions. For instance, by choosing the regions of t_1, A would be split to match regions C and D, which appears to be a rather reasonable approximation in this homogeneous artificial example (even if it commits the fallacy of division). Region F would be similarly split to match C, but F would be split and merged with G to match D, potentially leading to statistical measurements that do not properly represent either region.

While didactic, this example is not realistic because each building is visible in Figure 1, so we 2.6.4 know each individual of the tabulation area, which is never the case. Further, real measurements are seldom as homogeneous and noise-free as this artificial example. In this example, one possible harmonisation would split region A in two to provide different ancestors to C (A_C) and D (A_D). Then A_D would not be considered when analysing the evolution of C, nor A_C for D. This is inaccurate, because A was the original state for both evolutions, and we should not assume that these processes are disconnected. Either way, by splitting and merging the data to fit arbitrary borders, the harmonisation process increases the distance between the data and reality by creating artificial data points.

Instead, we propose a network-based representation. A *network* (also called a *graph*) is a collection of entities (nodes) that are related to each other (edges). In this case, each different aggregation area is represented as a node and we connect nodes that have overlapping geographical areas in different times or are neighbours in the same time, leading to the network illustrated on the right of 1.2 Figure 1. By partitioning the network into connected nodes that are similar, we are effectively identifying clusters in the spatio-temporal data, as illustrated by the colours of the nodes on the right side of Figure 1. Further, all the possible paths of change can be obtained by computing sequences of nodes over time, in this case: (A, C, F), (A, D, F), (A, D, G), and (B, E, H). This representation is also suited for geographically consistent regions, as illustrated by the stable park in this example, and is therefore a generalisation of the traditional paradigm.

Note that the edges of this network merely encode that two regions are related. This is binary information, there is no apportionment, no areal measurements, no population percentages, or weights of any kind associated with the edge. Indeed, our method also connects regions of the same time that share borders, representing exactly that they are neighbouring areas.

3 Related Work

Since our problem encompasses several fields, we divide this section into specific sub problems: *longitudinal demographic studies*, describing the traditional tabular approach to longitudinal studies; *data representation*, elaborating how evolving geographic data can be represented for processing; *data clustering and regionalisation*, briefly reviewing existing data clustering methods, geographic constraints and regionalisation; and *cluster characterisation*, articulating how clusters can be visually 2.1 summarised.

3.1 Longitudinal demographic studies

Census data is used not only to discover demographic patterns (Firebaugh and Farrell, 2016), but to correlate demographic characteristics to other measurements (Diez-Roux et al., 1997). However, longitudinal studies are rare, because they are difficult : *"(...) One of the most challenging and fascinating areas in spatial statistics is the synthesis of spatial data collected at different spatial scales(...)"* (Gotway and Young, 2002). While census tract level data is readily available for the US since at least 1910 (Manson et al., 2017), most studies consider the period between 1970 and 2010, using pre-harmonised data from the Longitudinal Tract Data Base (Logan et al., 2014) or the 1.3 Neighborhood Change Database (GeoLytics et al., 2010), albeit their inherent errors (Logan et al., 2016; Hallisey et al., 2017). Similar harmonisation efforts appear in other countries (Liu et al., 2015; Lee and Rinner, 2015; Allen and Taylor, 2018). These datasets have been highly significant for the field. Yet they also limit the universe of data that can be used to study neighbourhood change, since any new datasets would need to be similarly processed in order to be rendered compatible with these sources.

Another option considers the use of grid data (Dmowska et al., 2017; Dmowska and Stepinski, 2018; Stepinski and Dmowska, 2019). Beyond the potentially increased spatial precision, this 1.4 approach does not require complex harmonisation when new data is considered, if the grids are compatible. However, demographic data is usually not available in this format, especially from older sources. Additionally, the conversion from tabulation areas can introduce significant errors.

Given these challenges, it is worth considering new alternatives. In this work, we propose a novel methodology that entirely avoids the problems of geographical harmonisation, considering each measurement using its actual geographic region. It does not require regions to be consistent across time because they are naturally represented as different entities.

3.2 Data representation

Network based representation of geographic information is fairly well explored in the literature, as a basis for topological methods for event detection (Doraiswamy et al., 2014), leveraging signal processing on networks (Shuman et al., 2013; Sandryhaila and Moura, 2013) to find patterns and outliers (Valdivia et al., 2015; Dias and Nonato, 2015; Dal Col et al., 2018). Networks are well suited to represent trajectories as well (Von Landesberger et al., 2016; Huang et al., 2016; Chen et al., 2015), allowing the use of network visualisation methods (Vehlow et al., 2015; Beck et al., 2014). Our proposed method builds upon this literature. We leverage a network-based representation that removes the rigidity in the measurement regions. Each region in time corresponds to a different node. Instead of a collection of time-series, the data is represented as a dynamic network.

Networks have been used to represent census data for clustering purposes (Dias and Nonato, 2015; Setiadi et al., 2017), but these works did not explore temporal evolution, where they are particularly powerful. Networks allow a natural representation of these inconsistent regions, with both spatial and temporal connections. There are other possible representations that have similar properties, but we adopted networks to allow the use of the vast existing literature and methods.

3.3 Data clustering and regionalisation

Data clustering is one of the elementary processes for data analysis, simplifying the data into a smaller number of homogeneous sets that can be interpreted in the same way. There is no shortage of contributions for this problem (Fahad et al., 2014), since variations of it appear in almost all scientific fields.

2.1 In geography, this problem is known as *regionalisation* (Montello, 2003), a rather old problem that has been thoroughly explored, leveraging different mathematical tools, including discrete topology (Brantingham and Brantingham, 1978) and discrete geometry (Assunção et al., 2006). Indeed, network-based methods are among the current state-of-the-art (Guo, 2008; Duque et al., 2012). However, *temporal* regionalisation is significantly less explored, especially in a demographic context, arguably due to the difficulties in dealing with unharmonised longitudinal data. Recent neighbourhood related applications rely on k-means (Jain, 2010; Delmelle, 2016), the Louvain method for community detection (Blondel et al., 2008; Thomas et al., 2012), or, to a lesser extent, Self Organising Maps (Delmelle, 2017; Ling and Delmelle, 2016; Arribas-Bel and Schmidt, 2013).

1.5/2.2 Since we adopted a network-based data representation and our objectives include an interactive interface, we opted for an heuristic variation of the maximum weighted matching algorithm called *sorted maximal matching* (Dias et al., 2017), because of its simplicity, customisability, and fast computation times. This algorithm operates on weighted networks, where a distance metric between the nodes is associated with the edges. We adopted a distance based on the data associated with each node. The algorithm merges clusters based on these distances, creating a hierarchy instead of a fixed number of clusters. This hierarchical result is rendered by the interface, allowing the user to change the number of clusters without reprocessing. Changing the clustering algorithm would lead to different results, but any hierarchical network clustering method, and distance metric, can be used in our framework.

3.4 Cluster characterisation

While visualisation has gained prominence as a crucial component of scientific discovery, justification, and communication (Tufte et al., 1998), visually representing evolving spatial data is a challenging old problem (Monmonier, 1990; Andrienko et al., 2003; Ferreira, 2015; Zheng et al., 2016).

Most geographic data is naturally bidimensional and maps work well in this case (Zheng et al., 2016; Ward et al., 2015), but the additional temporal dimension cannot be so naturally represented. One straightforward option is to leverage tridimensional plots (Andrienko et al., 2014; Tominski and Schulz, 2012), but this can lead to visual obstructions or scaling problems unless a tridimensional display device is used. A simpler, well adopted, option is to display a map that corresponds to a subset of the temporal information, allowing the user to change the time with an associated control (Chen et al., 2017; Valdivia et al., 2015; Dal Col et al., 2018; Doraiswamy et al., 2014). Small multiples can be used (Von Landesberger et al., 2016), but only when there are few temporal snapshots. However, none of these options is suitable to represent many variables at the same time.

Using data clustering, we can represent the region's cluster instead of all the its variables (Dal Col et al., 2018; Valdivia et al., 2015; Von Landesberger et al., 2016). While this simplifies the geographic portion of the visualisation, it introduces the problem of how to summarise the contents of each

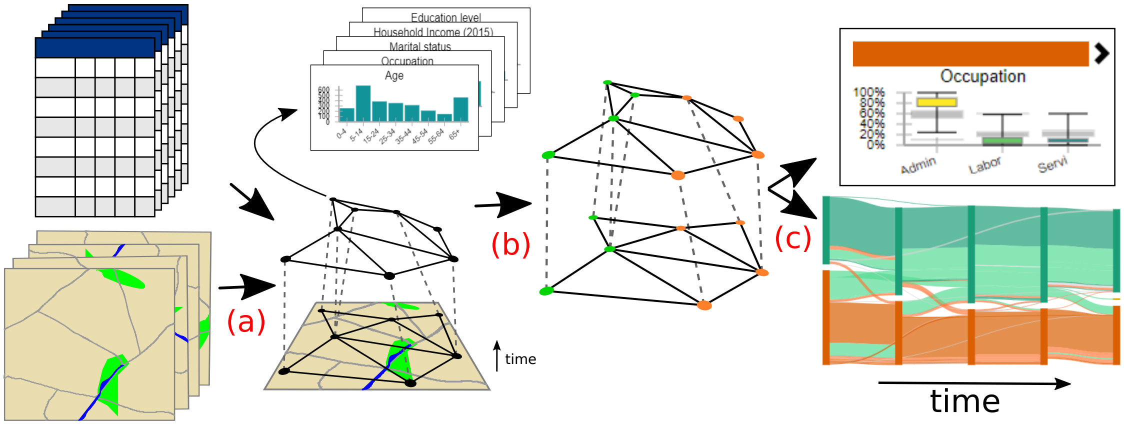


Figure 2: Overview of the proposed method. **(a)** A network is generated from the original data, encoding the changing geographical information. **(b)** The network is partitioned into an hierarchy. **(c)** The characteristics and evolution of the clusters are then visually represented.

cluster. One traditional approach is to use parallel coordinates plot (Ferreira et al., 2015), but they can get cluttered when representing similar clusters over several variables. Further, for demographic applications, the clusters are usually strongly characterised by a small subset of values (Delmelle, 2016, 2017). Therefore, in the proposed method, we identify the variables that are most relevant to the characterisation of each cluster. The distribution of values on that variable is then represented using a boxplot, a well known statistical plot displaying basic properties of the distributions.

4 Visualising the demographic spatio-temporal evolution

Figure 2 presents an overview of the processing steps of the proposed method, illustrating how the nodes of the network are used to represent the regions. The following sections elaborate on this figure and explain the main features of the interface we built to visualise and explore the evolution of neighbourhoods on the basis of our proposed method.

4.1 Census methodology and data representation

Census data is disseminated in a tabulated form for aggregation areas: whole country, state/province, metropolitan region, and so on. To allow for a more meaningful comparison of the data, we aggregated related variables (e.g. White, Black, Asian, Other) into what we called an *aspect* (e.g. Race). The aspects are represented using normalised histograms. This normalisation is crucial for direct comparison. In essence, it is a generalisation of the standard method of comparing percentages, since each aggregation area has a different total population.

Each area of each census year is represented as a node, and edges are placed between nodes if the corresponding regions share geographic borders in the same year. Further, edges are placed between nodes if the corresponding regions belong to sequential years and their geometries intersect. 1.6/2.6.3 To avoid spurious connections caused by small fluctuations, one of the geometries is slightly shrunk before the intersection, using a buffer of -1e6. While the result would not change considerably by adding more connections between nearby regions, these extra connections would greatly increase the computational cost. More importantly, while weights will be associated with these edges before they are processed, they are not derived from the geometry, but from the data. The actual intersection area is not considered in this representation. This approach leads to a single network representing the whole spatio-temporal space of the data.

4.2 Geographic content clustering

Having tied the regions together into a network, we can now partition it to identify similar sets of regions. In essence, we are performing temporal regionalisation by applying a clustering method over this spatio-temporal network. 2.3

We start by adopting a distance function between the nodes, measuring the difference between the data of the regions. This value is then associated with the edges, leading to a weighted dynamic network. Every node has a collection of histograms, each representing the distribution of certain aspect in the population.

Let $G = (V, E)$ be a network, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes and $E = \{(v_i, v_j), i \neq j \text{ and } i, j \in [1, n]\}$ is the set of edges. A function H associates each node to a set of K histograms. We define the distance D between two nodes v_i and v_j as:

$$D(v_i, v_j) = \sum_{k \in [1,K]} w_k \, d(H_k(v_i), H_k(v_j)) \tag{1}$$

where d is a distance metric between histograms and w is a sequence of non-negative weights associated with each aspect, $\sum_{k\in[1,K]} w_k = 1$. While any histogram metric can be used, we adopted a euclidean distance between the vectors, because it led to reasonable results with reduced computational cost. Therefore the distance between two nodes is defined as the weighted average distance between its associated histograms, where the weights can be adjusted by the user.

Once the distances are associated to the edges, we use watershed cuts (Cousty et al., 2009) to create an initial clustering, which is then refined into a hierarchy using the Sorted Maximal Matching (SMM) (Dias et al., 2017). The initial watershed step is performed to create an initial clustering and reduce the running time of the SMM. We introduced one new parameter to this method: a maximum distance threshold for the merges, to avoid the early merging of outliers. We refer the reader to the original paper (Dias et al., 2017) for more details, including a complete performance evaluation using several metrics. We chose this algorithm because it is fast, simple, and easily customisable, but our methodology should work with any hierarchical network clustering algorithm. 1.5/2.2 However, we caution against the use of single-linkage methods due to their tendency to form chain-like clusters with elements that are similar pairwise, but significantly different on the ends. That is also the reason we did not use a Watershed hierarchy directly (Najman, 2011).

Each resulting cluster is contiguous in the network. This means that two similar, but non-contiguous, sets of areas will be classified into two different clusters, which can be counter-intuitive. To overcome this issue, we *augment* the network with two new edges per node from a nearest neighbours graph (Pedregosa et al., 2011) using only the distances between the histograms. These edges connect nodes with similar content, if they are not already connected, providing a path for the algorithm to group similar nodes. The regions connected by those edges will be merged on the first stages of the clustering, since the nodes they connect are, by definition, as similar as possible[2], 2.4 leaving the remaining steps of the hierarchy to be determined only by the geographical edges. We experimented with different numbers of augmentation edges, but the results were not consistent, since the distribution of the edges is data dependent. Adding two edges per node was the smallest number of edges that led to stable and consistent results in the scenarios available in our prototype. Since the problem of balancing the data space with the geographical space is relevant for geographical data analysis, this idea potentially warrants further exploration, beyond the scope of this work.

4.3 Cluster characterisation and variable relevance

A crucial step in understanding neighbourhood change is to characterise the evolving clusters. The composition of each cluster is represented here by simple statistical measures, considering each aspect separately. We compute the minimum, maximum, median, 25%, and 75% quantiles for each variable

[2] This assertion relies on the euclidean distance and its relationship with the space where k-nn operates.

of each aspect for all clusters in the hierarchy. While interpreting these values is more complex than interpreting just the average, they provide far more information about the underlying distribution.

We also use these statistical measurements to discover what characterises each cluster, that is, what makes it different from the others. We define the *relevance* of a variable of an aspect based on the distance between the interquartile ranges (IQR) of the clusters in the same hierarchical 1.8 level. If the IQRs overlap for all clusters, that variable is not relevant to the characterisation of the cluster, but if the IQRs are distant, it means that this specific range of values is something that only occurs in this cluster. Examining IQRs therefore provides users a straightforward visual method for determining what variables most clearly define a given cluster.

To allow for an easier visualisation of these ideas, we represent it using an *enhanced* version of the traditional boxplot, which includes the IQRs for the other clusters, in slightly larger and faded black rectangles. We also colour the current IQR according to its relevance, as illustrated in Figure 4. In this example, this cluster is best defined by the proportion of the population with four or more years of college. The user can quickly see that this is relevant because the corresponding IQR is coloured with the highest relevance present in the legend. It is also clear that, while this cluster includes CTs that have between 10% to 90% of people in this variable, approximately, half of them have about 60% of the population with four or more years of college. Since all the other IQRs are well separated, this is a defining characteristic of this cluster. Conversely, the proportion of the population with one to three years of college is not relevant, as indicated by black fill in the rectangle representing the IQR of this cluster, in overlapping position with the rectangles of the other clusters.

4.4 Clusters and trajectories

While the partition of the data into different clusters helps the user to understand what groups exist and where they are, we are also interested in the evolution of these groups. To examine this process of evolution directly, we introduce the concepts of *temporal paths* and *trajectories*.

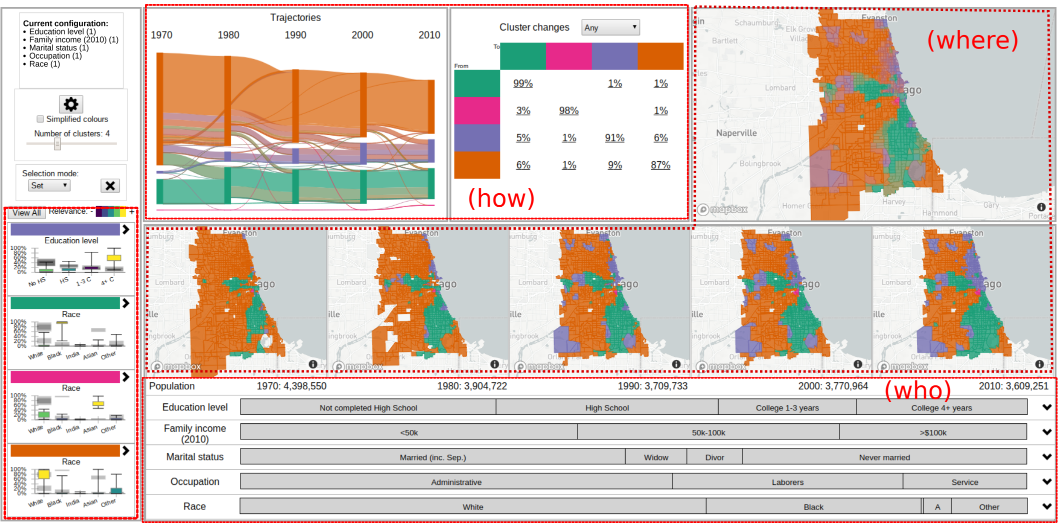
We call a temporal path any sequence of nodes in our representation network such that the 2.6.4 temporal information associated with the nodes only increases. For instance, in Figure 1, the sequences ACF, ADF, ADG, and BEH are temporal paths. With harmonised data, as illustrated by the path BEF, the time-series of each region would form a temporal path, each node would be connected only to its older and newer versions, and would belong to only one temporal path. Since our data is not harmonised, more connections are allowed and each node can belong to an arbitrary number of paths.

Semantically, this is a generalisation of the idea of geographical time-series, because each temporal path is one possible option for the data to change over time. Returning to Figure 1, the paths ACF, ADF, and ADG all start on the same homogeneous region, but evolve differently over time. In other words, this network encodes the information that portions of the region A changed to form regions C and D, but we do not know specifically which parts. Nor do we need to, since the same region can belong to several temporal paths. Interestingly, when interpreted in this framework, geographical harmonisation is a method to split and/or merge nodes so each belongs to a single temporal path.

Since each node in the sequence that forms the temporal paths has an associated cluster, we can classify the paths based on the sequence of clusters. We call each unique sequence of clusters present in this result a *trajectory*. Regions on the same trajectory had the same sequence of clusters, therefore had similar temporal evolution.

4.5 Overview of the prototype interface

To validate and explore the results of our methodology, we built a user interface, illustrated in Figure 3, considering census tract (CT) level data from the Chicago region between 1970 and 2010. This region is known for its entrenched racial divide and the emergence of a *'young urban'* population with a higher education level (Delmelle, 2016, 2017). More details are presented in Section 5.



Current configuration:
- Education level (1)
- Family income (2010) (1)
- Marital status (1)
- Occupation (1)
- Race (1)

☐ Simplified colours

Number of clusters: 4

Selection mode: Set

Cluster changes: Any

From \ To	🟩	🟥	🟪	🟧
🟩	99%		1%	1%
🟥	3%	98%		1%
🟪	5%	1%	91%	6%
🟧	6%	1%	9%	87%

Population	1970: 4,398,550	1980: 3,904,722	1990: 3,709,733	2000: 3,770,964	2010: 3,609,251

Education level	Not completed High School	High School	College 1-3 years	College 4+ years	
Family income (2010)	<50k	50k-100k	>$100k		
Marital status	Married (inc. Sep.)	Widow	Divor	Never married	
Occupation	Administrative	Laborers	Service		
Race	White	Black	A	Other	

Figure 3: Initial interface of our method showing the demographic evolution of Chicago and identifying the objectives of each plot. **(what)**: Cluster overview illustrating the most relevant aspect for each cluster. **(where)**: Maps illustrating the geographical location, both as an overall summary and for each year. **(how)**: Tabular and visual summary of how the regions classification changed over time. **(who)**: Demographic details for the currently selected data.

As illustrated by Figure 3, our proposed interface heavily relies on colour to express cluster-related information. We adopted this convention because colours can be used in all our visual tools in a coherent manner. However, there is a limit on the number of distinct colours that can be used. We limited the number of clusters to eight because this was the largest number of colours that we could reliably and accessibly use, derived from the 8-class Dark2 set from ColorBrewer (Harrower and Brewer, 2003).

The configuration panel, on top left in Figure 3, displays the processing configuration of the visualised results: which aspects were used and their weights (following Equation 1). It also includes other configuration options that can be altered without re-processing the data, such as the number of clusters and the colour option. The gear button allows access to the other configuration options that do require further processing, such as changing location, aspects, and weights.

The main functionalities of the interface allow users to examine four inter-related questions central to understanding local area dynamics. 1) *What* do the clusters consist in?; 2) *Where* are they located; 3) *How* do they change?; and 4) *Who* occupy them? We discuss each in turn.

The cluster overview panel, marked as *what* in Figure 3, displays the most relevant aspects for each cluster, based on the distance between the IQRs, as detailed in Section 4.3.This feature allows the user to quickly understand what exists in this region and which aspect makes each cluster unique, without going into excessive detail. In Chicago's example, we expect race to be relevant in general, and indeed, it is recognised as the most relevant aspect for three of the four most distinct clusters, with the fourth being education level. As we increase the number of clusters, they get more nuanced, requiring more than one aspect for their proper characterisation, as each cluster is divided into its composing sub-clusters. The *View all* button opens a new panel where all aspects are included, while the chevron at the side lets the user expand each cluster separately, allowing the exploration of more details on demand. 2.5

With the knowledge of what clusters are present, we can explore where they are through the maps, marked as *where* in Figure 3. While the maps on the lower part of the interface show the location of the clusters at each time, the larger map shows a summarised version, based on the

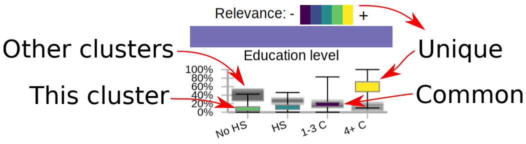


Figure 4: Enhanced boxplot of the clusters' characteristics allows a quick comparison to the other clusters.

trajectories of each region. In simplified mode the colour corresponds to the cluster present in that trajectory for more than half the temporal samples, and is grey otherwise. 1.9 With simplified colours disabled, the colours are the average of the colours of the clusters in that trajectory in RGB colour space [3].

2.5 Each trajectory represents a different temporal path through the clusters. Therefore the trajectories encode how the demographic characteristics changed over time. These changes are summarised by the two sections identified as *how* in Figure 3. In the left panel, there is a Sankey diagram where each ribbon represents a different trajectory. The widths are proportional to the population involved. In our example in Figure 3, the orange and green clusters contain most of the population and are fairly stable over time. The pink cluster is small and mostly stable. The purple cluster is increasing, mostly by incorporating areas that were previously orange. Since the purple group corresponds to the emergent 'young urban' group, this corroborates the findings of Delmelle (Delmelle, 2016, 2017), showing that our network-based method can recover results from the traditional data processing approach.

2.5 In the next panel, illustrated in the top middle of Figure 3, is a transition matrix between the clusters. It indicates a rounded percentage of regions whose area changed between each pair of clusters. In this example, we can see that all clusters are fairly stable, with more than ninety percent of identity transitions. This kind of table can be found in the related literature (Delmelle, 2016), so it is familiar to the advanced users. It not only informs the proportional changes, but allows the selection of specific changes for further analysis.

The bottom part of the interface contains the details for the selected trajectories, or for the whole city if nothing is selected, marked as *who* in Figure 3. The stacked bar plots summarise the overall demographic composition of these regions. In this example, the maps show the transition from orange to green and purple in several regions over time. Each aspect is represented by a stacked bar plot, where the width of each rectangle corresponds to the percentage of population in that variable 1.10 over the considered period. In this case, a little less than half of the population are married, and the percentage that are Widowers or Divorced is roughly similar. About half of the population work in Administrative jobs, a third never completed high-school, approximately forty percent have gross family income below 50,000USD per year. About sixty percent identify as white. Placing the mouse over one of the bars will open a small panel with the temporal evolution of the population percentage of that specific variable, and clicking on the chevron on the right side expands the corresponding aspect, showing census tract level statistics, with details of the temporal evolution of each variable and also the corresponding IQRs for the whole city. Those values differ from the total population percentage. While the overall population of this region is about sixty percent White, the average percentage of White population over the CTs is a little over forty percent.

More importantly, this interface is fully interactive, allowing for a progressive exploration of the data. Clicking on a cluster in the cluster overview panel and the other panels will filter the results to consider only that cluster, highlighting the geographic regions and involved changes, and updating population count and demographic composition. If the user selects a region from the larger map, all identical trajectories are considered, and the whole interface changes to allow for further study of that specific region. Clicking on a region in small maps will bring up a new panel with the

[3] `https://gka.github.io/chroma.js/#chroma-average`

original census data of this specific region. Further, these selectors can be combined in different ways, enabling complex queries with instantaneous results.

5 Illustrative scenarios

In this section we present two illustrative scenarios, using census data from the United States (Manson et al., 2017) and Canada[4], tabulated by CTs, from 1970 to 2010.

The prototype interface allows access to 41 regions, 29 in the US and 12 in Canada. New York City was split into its boroughs to avoid memory crashes on the client browser due to the high number of CTs. We used five aspects for the USA: Education level, Family income, Marital status, Occupation, and Race; and seven for Canada: Age, Education level, Home language, Household Income, Marital status, Occupation, Place of birth, and Religion. The supplementary material contains the details of which census columns were used for each aspect.

1.12.(1-5) These results are meant to demonstrate the utility of the interface for understanding the evolutionary dynamics of urban neighbourhoods. They also show the face validity of the results generated by our novel network-based approach. While our method does not require geographic harmonisation, it does however require matching the variables over time. This is (currently) a painstaking, manual process, and the current version of the interface remains somewhat limited. For example, we selected a minimal set of aspects for each country, which are not necessarily similar to each other, or similar to studies in the current literature. Income moreover is slightly inaccurate, even though we did correct for official inflation. We grouped the original ranges into three larger ranges, but they do not match precisely. In addition, the US data labelled ”2010” is not from the decennial census, but from the ACS 2006-2010, somewhat compromising the temporal stability of the data. Further, the regions selected do not correspond to any pre-defined regions (metros, cities, census areas), but to somewhat arbitrary regions defined around a location of interest, making direct comparisons harder. Taken together, these factors would no doubt require caveats for a substantive study using the current prototype version of the interface. Nevertheless, in the present context we find that the interface results are well-aligned with several other studies, illustrating the robustness of our approach to imperfect conditions.

5.1 Chicago

Our first scenario examines Chicago, using an arbitrary region larger than the administrative borders. Its demographic composition is well explored in the literature, with reports of racial divide and gentrification (Delmelle, 2016, 2017; Hwang and Sampson, 2014), so we expect our results to contain stable regions where the Race aspect is relevant, and some degree of population change, with increasing income and education levels.

The initial state of the prototype is illustrated in Figure 3. The first step is to identify *what* the clusters consist in. To do so, we examine the compositions of each cluster from the boxplots. Orange is associated with majority of White population, green with majority Black, and purple with higher proportion of four years of college or more (high education level). The expanded version of the boxplots for the purple cluster shows a higher income level and majority of occupations in administrative jobs. The purple cluster therefore could be described as gentrified regions.

The trajectories plot shows *how* clusters have changed over time. For example, it illustrates the process of gentrification, also illustrated in Figure 5, with the purple cluster progressively absorbing regions from the orange cluster (White). This corroborates results from the literature reporting that Black neighbourhoods are less likely to gentrify (Hwang and Sampson, 2014). The transition matrix provides more details, showing that around one percent of all regions that became purple at any time transitioned from the green cluster (Black population), while nine percent were previously

[4] http://datacentre.chass.utoronto.ca/census/

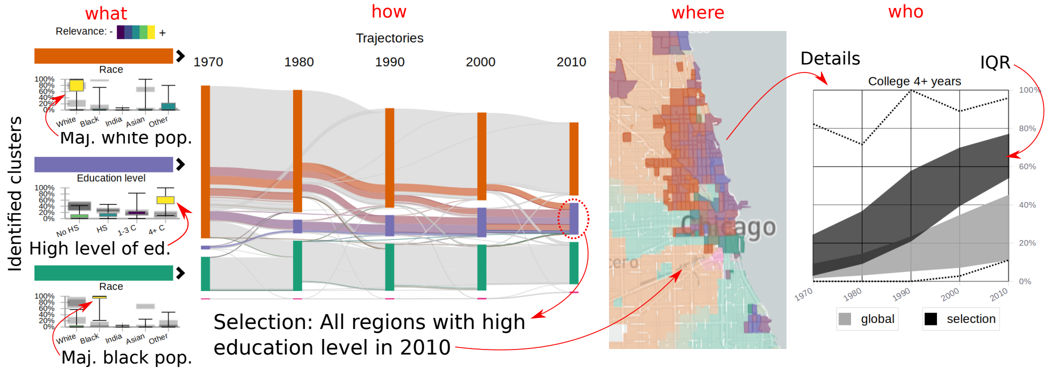


Figure 5: Workflow to discover gentrification in Chicago: the purple cluster corresponds to high education / income. Its population is increasing over time, absorbing from the majority White cluster (orange). By selecting the purple cluster in 2010, the region is highlighted in the maps. The proportion of people with 4+ years of college is increasing in the whole city (grey IQRs), but significantly more in this region (black).

associated with the orange cluster (White). Next, we select the region that is gentrified in 2010, by clicking on the corresponding rectangle in the trajectories plot, updating the information on the maps and the details portion of the interface.

The maps show *where* these processes are occurring, by highlighting the corresponding regions. The spatial pattern is clear, corresponding exactly to previous findings in the literature based upon harmonisation (Hwang and Sampson, 2014). Further, we can also identify the regions that gentrified earlier on the small maps that depict the involved regions over time. Finally, we can examine *who* is driving these changes. Since the most relevant aspect is Education, specifically ”Four or more years of college”, we can expand the details of this aspect, as illustrated in the rightmost portion of Figure 5, which is increasing for the whole city (grey band), but faster and to a higher level in this region (black band).

5.2 Toronto

We consider a region that corresponds approximately to the administrative border of the current city of Toronto, using all seven available aspects with equal weights. While Chicago was fairly stable, Toronto is known to be a more dynamic and diverse city, with significant and increasing immigrant population (Hulchanski, 2007; Fong and Chan, 2011), especially Asian (Fong and Wilkes, 2003). Toronto is also known for a stable and well defined Jewish community (Harold and Fong, 2018; Fong and Chan, 2011). Therefore, we expect a combination of stable and dynamic regions on the results, with Place of Birth, Home Language, and Religion identified as relevant aspects. The results are summarised in Figure 6, considering eight clusters. 1.14

We again are guided by the four questions embedded in the interface: what, how, where, and who. The boxplots address what clusters defined Toronto over this period. The population with low percentage of University degrees is represented in orange, mostly anglophone population in green, Asian immigrants in yellow, high percentage of income in the highest bracket in purple, high percentage of Jewish people in light green and brown, high percentage of Eastern Non-Christian religion in pink, and high concentration of single people in dark grey.

From the trajectories plot, we can see how the city changed. It shows that Toronto is more dynamic than Chicago, with one cluster constantly shrinking. In the 1970s, the city was divided into four clusters: low number of university degrees, Jewish population, majority anglophones, and high income. Interestingly, the more recent clusters that absorbed regions from the orange cluster

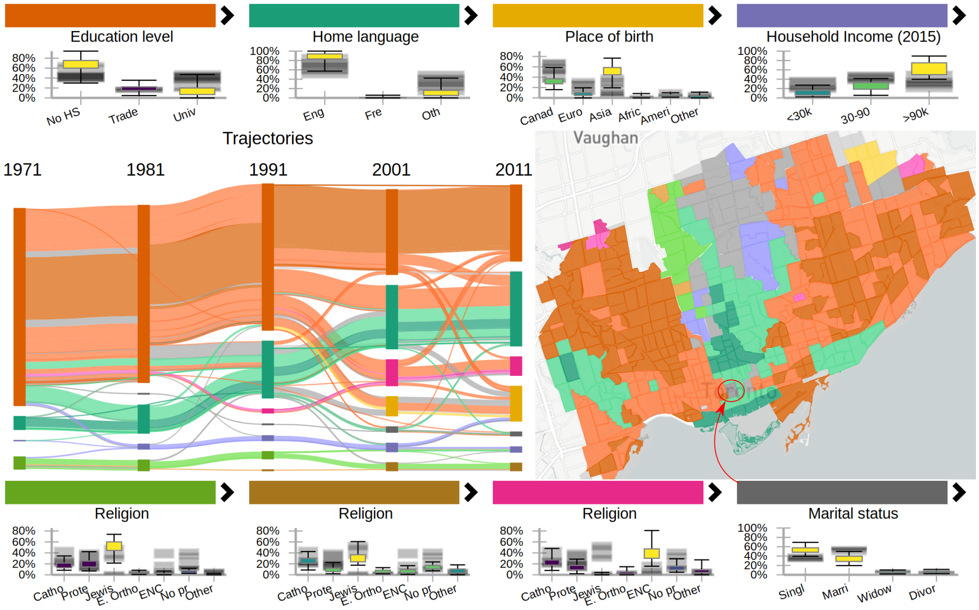


Figure 6: Clustering results for Toronto, with eight clusters, including clusters representing Jewish population, high and low income, low education, and Asian immigration.

have similar education profiles and are differentiated by other aspects. In this sense, the city is growing diverse, changing from a common low education profile to a higher level of education with more diversity in religion (pink) and immigration (yellow). Indeed, the growing Asian population is visible starting in the 1980s and building thereafter, leading to the yellow and pink clusters. While both include a high percentage of people born in Asia, the pink is more defined by religion, with low percentage of university degrees, and contains the lowest percentage of people in the highest income bracket for these clusters; the yellow is less defined by religion, and has higher education and income, geographically corresponding to the Markham region, know for its Chinese population. A similar division also happens for the two Jewish clusters, where the light green cluster has lower education and income levels than the brown cluster. The purple cluster of high income is somewhat stable. Until 2011 the cluster included the Bridle Path neighbourhood, known for its wealthy population. In 2011 it was classified into the yellow cluster of Asian immigration, since about 40% of the population for this CT were then born in Asia. The income distribution did not change, with 85% of the population with an income of 90k CAD or more.

Examining where and by whom these changes occurred reveals additional information about the city's underlying dynamics. The most significant indicator of Toronto's dynamism is the presence of grey regions on the map. These represent regions associated to three or more clusters over this five census period. Using the 'Add' mode for the trajectory selection, we select their trajectories, and a subset of the details is illustrated in Figure 7. These regions account for about 5% of Toronto's population. The whole region was classified into the orange cluster in 1971 (low level of university degrees). By 1991, most of the region was classified into the green cluster, representing anglophone population, mostly Canadian born, with a higher level of education. As the corresponding plot indicates, this trend in increasing education is city-wide, but this region has people with better education than most. In 2001, the purple cluster of high income annexes neighbouring parts of the volatile region, and the Asian born population increases sharply, as illustrated by the appearance of the yellow cluster. This cluster indicates well educated, higher income, and about 30%-50% Asian born population. By 2011, the yellow cluster increased considerably, annexing parts of the high

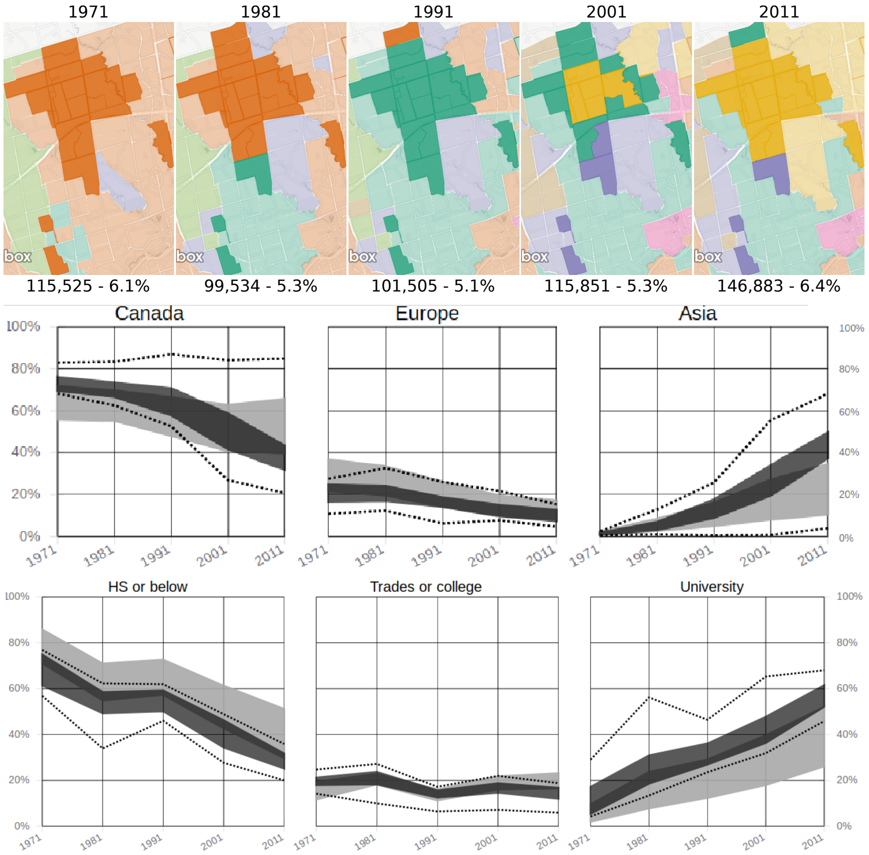


Figure 7: Details for some regions of Toronto that were classified into 3 or more clusters over time.

income purple cluster, including the neighbouring Bridle Path area.

The geographical borders of the clusters obtained using our method are similar to the regions presented by previous studies considering Toronto (Hulchanski, 2007). However, our interface provides a deeper insight into their demographic composition, since we consider more data than solely Average Income, which appears to be a good proxy variable nonetheless. This scenario showcases the ability of our method and interface to capture and understand the sources of urban volatility.

6 Expert feedback

As our method and tool are novel to the field, and somewhat exotic, we subjected them to the critical scrutiny of experts. We contacted academic and industry experts in sociology and urban sciences to solicit their evaluation of our methodology. They had access to the prototype tool, a descriptive documentation of the features (included in the supplementary material), and a sequence of documentation videos illustrating how to perform specific tasks. The documentation explains which datasets are used and how the data is represented and processed, noting explicitly that there is no geographic harmonisation. We focused our inquires on the results obtained, asking if they found anything interesting in the data. The message sent and their full response is included in the supplementary material. Each of the five experts is identified by a letter, from A to E.

The overall overall response of the experts was positive, mentioning that the prototype allows them to analyse census data without the additional work of obtaining and cleaning the data (A, B, E), and it allows the inclusion of geographic visual analysis tools in their research process (D). It enables the users to tell different stories about neighbourhoods/cities and their changes (A), visualise the relationship between key urban variables over time (D), offering a quick way to identify particular neighbourhoods that one may be interested in studying more in depth around a particular issue or efficiently understanding the context of an area (E). Indeed, the experts identified gentrification processes in Manhattan (B) and Dallas (E), reinforced a hypothesis for occupational clustering (D), and highlighted how the method can be used to compare neighbourhoods and cities (A). In summary, their view was that the proposed methodology can be a viable alternative for the visual analytics of evolving demographic data.

The interface was "easy to navigate" (B), but it was also considered "overwhelming" (A), "intimidating" (E), and "tricky to interpret" (C), possible side-effects of our effort to increase representational accuracy, where we avoided using simplified representation or labels. Identifying clusters by their most relevant variables was welcome, but the overlap of information from different clusters in the boxplot was "a bit confusing" (C) when colour was not present. Further, most clusters can be sufficiently characterised using only the most relevant aspect, but this is not generally true.

While the map of trajectories was mentioned as a "good summary map", how it related to the clustering method was unclear (C). The methods include different options on how the colours are used, but both are works in progress since reliably representing several distinct entities using colours is humanly unfeasible. Indeed, the number of distinguishable colours was a significant constraint, we found indications that more clusters should be used in some cases, even if eight clusters is more than what is traditionally considered in these analyses. Conversely, increasing the number of clusters would also complicate the interpretation of the results.

The experts also mentioned the poor responsiveness of the method when changes in the clustering parameters required server-side processing (B,D). Indeed, the current implementation can take a few minutes to cluster regions with high number of CTs, like Los Angeles or Brooklyn. Server-side processing reduced the amount of data transferred to client, but it might increase the response time under load. We implemented a caching policy, but fully pre-processing the results is not practical due to size of the parameter space.

Most of the experts demonstrated interest in using our method in their research (A, B, D, E), aiming to use the census data as a backdrop for other datasets, providing demographic context. They also mentioned the need to export subsets of data, plots, and maps to be used in reports and publications (C, D, E). More importantly, while these experts were aware that our method does not perform geographic harmonisation, none of them mention it. We did not specifically ask if this difference led to unexpected results, but rather if they found interesting insights.

7 Discussion and limitations

Our objective was to leverage a network based data representation and visualisation methods for the exploration of geographically inconsistent region-based data. While we successfully replicated and corroborated results from the literature, this method still has significant limitations.

We removed the need for geographical harmonisation, but the method still requires consistent variables across the years. Matching the variables can be trivial for some aspects (Age), but challenging for others (Income). The divulged income ranges vary over time and the actual values change due to inflation. Since this is only a prototype, we matched few aspects, but a proper demographic analysis would benefit from all available information.

The limitation on the number of displayed clusters because of the limited number of distinguishable colours was significant. While increasing the number of clusters would further complicate an already complex analysis, it might be warranted for some regions. Colour is a fundamental and intuitive tool for information representation that can be coherently used across different plots, so we opted to use it, even if in a limited way. With eight colours, there was overlap between some clusters, the relevance gradient, and the colour combination.

The cognitive load on the user is significant, as we compromised simplicity for accuracy. While other works labelled the clusters, as 'young urban', 'struggling', and so on (Delmelle, 2016, 2017), we show the statistical characteristics of the clusters, which are harder to interpret, as the data may have subtle nuances that labels would otherwise hide. This also led to a crowded interface, mitigated somewhat the use of pop-up panels and collapsible sections. For some cities, especially if they are small and stable, the panels can appear redundant, but each provide a different way to interact with the information that can ease the exploration process for larger and dynamic cities.

8 Conclusion

The objective of this work was to demonstrate that temporal regionalisation can be performed without geographical harmonisation. 2.1 We proposed an alternative methodology that robustly considers the data in its original geography, without the creation of arbitrary artificial data points.

This methodology was then used to create a publicly accessible system, with an interactive and intuitive interface, allowing a transparent evaluation and replication of our results. We used this interface to corroborate results from the literature and we hope that it will be used to corroborate future results as well.

The feedback from experts was positive and most of them were able to extract insight from the prototype while indicating interest in using it for their research efforts. Since our interface can be used by non-experts as well, we also contributed to scientific dissemination and stakeholder transparency in urban sciences. Building on the comments of the experts and other colleagues exposed to our prototype, a future iteration of this prototype would include a simplified interface 1.12.6/2.6.5 and faster computation times, but more importantly the option to define the region of study and add more data. This would also allow for a direct comparison between original and harmonised data, to further validate the prototype, and explore if our approach offers any new insights. Methodologically, the variable matching step is the current manual bottleneck of the system, and the next target for removal.

More importantly, we introduced a new idea that, apparently, was never considered in the literature. While significant resources have been invested to improve geographical harmonisation, rarely, if at all, has anybody doubted that it was really necessary in the first place. We proved that it is not necessary for regionalisation, especially for neighbourhood effects and neighbourhood dynamics. 2.1

References

Abbott, A. (1997, 06). Of Time and Space: The Contemporary Relevance of the Chicago School. *Social Forces 75*(4), 1149–1182.

Allen, J. and Z. Taylor (2018). A new tool for neighbourhood change research: The canadian longitudinal census tract database, 1971–2016. *The Canadian Geographer / Le Géographe canadien*.

Andrienko, G., N. Andrienko, H. Schumann, and C. Tominski (2014). Visualization of trajectory attributes in space–time cube and trajectory wall. In *Cartography from Pole to Pole*, pp. 157–163. Springer.

Andrienko, N., G. Andrienko, and P. Gatalsky (2003). Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing 14*(6), 503–541.

Arribas-Bel, D. and C. R. Schmidt (2013). Self-organizing maps and the us urban spatial structure. *Environment and Planning B: Planning and Design 40*(2), 362–371.

Assunção, R. M., M. C. Neves, G. Câmara, and C. da Costa Freitas (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science 20*(7), 797–811.

Beck, F., M. Burch, S. Diehl, and D. Weiskopf (2014). The State of the Art in Visualizing Dynamic Graphs. In R. Borgo, R. Maciejewski, and I. Viola (Eds.), *EuroVis - STARs*. The Eurographics Association.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment 2008*(10), P10008.

Brantingham, P. L. and P. J. Brantingham (1978). A topological technique for regionalization. *Environment and Behavior 10*(3), 335–353.

Chen, W., F. Guo, and F.-Y. Wang (2015). A survey of traffic data visualization. *IEEE Transactions on Intelligent Transportation Systems 16*(6), 2970–2984.

Chen, W., Z. Huang, F. Wu, M. Zhu, H. Guan, and R. Maciejewski (2017). Vaud: A visual analysis approach for exploring spatio-temporal urban data. *IEEE Transactions on Visualization & Computer Graphics*.

Cousty, J., G. Bertrand, L. Najman, and M. Couprie (2009, Aug). Watershed cuts: Minimum spanning forests and the drop of water principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(8), 1362–1374.

Dal Col, A., P. Valdivia, F. Petronetto, F. Dias, C. T. Silva, and L. G. Nonato (2018). Wavelet-based visual analysis of dynamic networks. *IEEE Transactions on Visualization and Computer Graphics PP*(99), 1–1.

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), 36–56.

Delmelle, E. C. (2017). Differentiating pathways of neighborhood change in 50 u.s. metropolitan areas. *Environment and Planning A: Economy and Space 49*(10), 2402–2424.

Dias, F. and L. G. Nonato (2015). Some operators from mathematical morphology for the visual analysis of georeferenced data. In *Workshop on Visual Analytics, Information Visualization and Scientific Visualization - SIBGRAPI.*

Dias, M. D., M. R. Mansour, F. Dias, F. Petronetto, C. T. Silva, and L. G. Nonato (2017, Oct). A hierarchical network simplification via non-negative matrix factorization. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 119–126.

Diez-Roux, A. V., F. J. Nieto, C. Muntaner, H. A. Tyroler, G. W. Comstock, E. Shahar, L. S. Cooper, R. L. Watson, and M. Szklo (1997). Neighborhood environments and coronary heart disease: a multilevel analysis. *American journal of epidemiology 146*(1), 48–63.

Dmowska, A. and T. F. Stepinski (2018). Spatial approach to analyzing dynamics of racial diversity in large u.s. cities: 1990–2000–2010. *Computers, Environment and Urban Systems 68*, 89 – 96.

Dmowska, A., T. F. Stepinski, and P. Netzel (2017, 03). Comprehensive framework for visualizing and analyzing spatio-temporal dynamics of racial diversity in the entire united states. *PLOS ONE 12*(3), 1–20.

Doraiswamy, H., N. Ferreira, T. Damoulas, J. Freire, and C. T. Silva (2014, Dec). Using topological analysis to support event-guided exploration in urban data. *IEEE Transactions on Visualization and Computer Graphics 20*(12), 2634–2643.

Duque, J. C., L. Anselin, and S. J. Rey (2012). The max-p-regions problem*. *Journal of Regional Science 52*(3), 397–419.

Eicher, C. L. and C. A. Brewer (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science 28*(2), 125–138.

Fahad, A., N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras (2014, sep). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing 2*(3), 267–279.

Ferreira, N. (2015). *Visual analytics techniques for exploration of spatiotemporal data.* Ph. D. thesis, Polytechnic Institute of New York University.

Ferreira, N., M. Lage, H. Doraiswamy, H. Vo, L. Wilson, H. Werner, M. Park, and C. Silva (2015). Urbane: A 3d framework to support data driven decision making in urban development. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*, pp. 97–104. IEEE.

Firebaugh, G. and C. R. Farrell (2016, Feb). Still large, but narrowing: The sizable decline in racial neighborhood inequality in metropolitan america, 1980–2010. *Demography 53*(1), 139–164.

Fong, E. and E. Chan (2011). Residential patterns among religious groups in canadian cities. *City & Community 10*(4), 393–413.

Fong, E. and R. Wilkes (2003, Dec). Racial and ethnic residential patterns in canada. *Sociological Forum 18*(4), 577–602.

Galster, G. C. (2019). *Making our neighborhoods, making our selves.* University of Chicago Press.

GeoLytics, I. et al. (2010). Census neighborhood change database 1970—2010 census tract data.

Gotway, C. A. and L. J. Young (2002). Combining incompatible spatial data. *Journal of the American Statistical Association 97*(458), 632–648.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). *International Journal of Geographical Information Science 22*(7), 801–823.

Hallisey, E., E. Tai, A. Berens, G. Wilt, L. Peipins, B. Lewis, S. Graham, B. Flanagan, and N. B. Lunsford (2017, Aug). Transforming geographic scale: a comparison of combined population and areal weighting to other interpolation methods. *International Journal of Health Geographics 16*(1), 29.

Harold, J. and E. Fong (2018). Mnemonic institutions and residential clustering: Jewish residential patterns in toronto. *Canadian Review of Sociology/Revue canadienne de sociologie 55*(2), 257–277.

Harrower, M. and C. A. Brewer (2003). Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal 40*(1), 27–37.

Huang, X., Y. Zhao, C. Ma, J. Yang, X. Ye, and C. Zhang (2016, Jan). Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. *IEEE Transactions on Visualization and Computer Graphics 22*(1), 160–169.

Hulchanski, D. J. (2007). The three cities within toronto: Income polarization among toronto's neighbourhoods, 1970-2005.

Hwang, J. and R. J. Sampson (2014). Divergent pathways of gentrification: Racial inequality and the social order of renewal in chicago neighborhoods. *American Sociological Review 79*(4), 726–751.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters 31*(8), 651–666.

Kwan, M.-P. (2018). The limits of the neighborhood effect: Contextual uncertainties in geographic, environmental health, and social science research. *Annals of the American Association of Geographers 108*(6), 1482–1490.

Lee, A. C.-D. and C. Rinner (2015). Visualizing urban social change with self-organizing maps: Toronto neighbourhoods, 1996–2006. *Habitat International 45*, 92–98.

Li, Y. and Y. Xie (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*(5), 1313–1337.

Ling, C. and E. C. Delmelle (2016). Classifying multidimensional trajectories of neighbourhood change: a self-organizing map and k-means approach. *Annals of GIS 22*(3), 173–186.

Liu, X., Y. Song, K. Wu, J. Wang, D. Li, and Y. Long (2015). Understanding urban china with open data. *Cities 47*, 53 – 61. Current Research on Cities (CRoC).

Logan, J. R., B. J. Stults, and Z. Xu (2016). Validating population estimates for harmonized census tract data, 2000–2010. *Annals of the American Association of Geographers 106*(5), 1013–1029.

Logan, J. R., Z. Xu, and B. J. Stults (2014). Interpolating us decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer 66*(3), 412–420.

Looker, B. (2015). *A nation of neighborhoods: imagining cities, communities, and democracy in postwar America*. University of Chicago Press.

Manson, S., J. Schroeder, D. V. Riper, and S. Ruggles (2017). Ipums national historical geographic information system: Version 12.0 [database].

Monmonier, M. (1990). Strategies for the visualization of geographic time-series data. *Cartographica: The International Journal for Geographic Information and Geovisualization 27*(1), 30–45.

Montello, D. R. (2003). Regions in geography: Process and content. *Foundations of geographic information science*, 173–189.

Najman, L. (2011). On the equivalence between hierarchical segmentations and ultrametric watersheds. *Journal of Mathematical Imaging and Vision 40*(3), 231–247.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Poorthuis, A. (2018). How to draw a neighborhood? the potential of big data, regionalization, and community detection for understanding the heterogeneous nature of urban neighborhoods. *Geographical Analysis 50*(2), 182–203.

Reades, J., J. D. Souza, and P. Hubbard (2019). Understanding urban gentrification through machine learning. *Urban Studies 56*(5), 922–942.

Rey, S., E. Knaap, S. Han, L. Wolf, and W. Kang (2018). Spatio-temporal analysis of socioeconomic neighborhoods: The open source longitudinal neighborhood analysis package (oslnap). In *Proceedings of the 17th Python in Science Conference (SciPy 2018)*, pp. 121–128.

Sampson, R. J. (2012). *Great American city: Chicago and the enduring neighborhood effect.* University of Chicago Press.

Sandryhaila, A. and J. M. Moura (2013). Discrete signal processing on graphs. *IEEE transactions on signal processing 61*(7), 1644–1656.

Setiadi, T., A. Pranolo, M. Aziz, S. Mardiyanto, B. Hendrajaya, and Munir (2017, Oct). A model of geographic information system using graph clustering methods. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*, pp. 727–731.

Shelton, T. and A. Poorthuis (2019). The nature of neighborhoods: Using big data to rethink the geographies of atlanta's neighborhood planning unit system. *Annals of the American Association of Geographers 0*(0), 1–21.

Shuman, D. I., S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine 30*(3), 83–98.

Stepinski, T. F. and A. Dmowska (2019). Imperfect melting pot–analysis of changes in diversity and segregation of us urban census tracts in the period of 1990–2010. *Computers, Environment and Urban Systems 76*, 101–109.

Stone, C. N., R. P. Stoker, J. Betancur, S. E. Clarke, M. Dantico, M. Horak, K. Mossberger, J. Musso, J. M. Sellers, E. Shiau, et al. (2015). *Urban neighborhoods in a new era: Revitalization politics in the postindustrial city.* University of Chicago Press.

Thomas, I., C. Cotteels, J. Jones, and D. Peeters (2012, March). Revisiting the extension of the brussels urban agglomeration : new methods, new data ... new results ? *Belgeo* (1-2).

Tominski, C. and H.-J. Schulz (2012). The Great Wall of Space-Time. In M. Goesele, T. Grosch, H. Theisel, K. Toennies, and B. Preim (Eds.), *Vision, Modeling and Visualization.* The Eurographics Association.

Tufte, E. R., S. R. McKay, W. Christian, and J. R. Matey (1998). Visual explanations: images and quantities, evidence and narrative.

Valdivia, P., F. Dias, F. Petronetto, C. T. Silva, and L. G. Nonato (2015, Oct). Wavelet-based visualization of time-varying data on graphs. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 1–8.

Vehlow, C., F. Beck, and D. Weiskopf (2015). The State of the Art in Visualizing Group Structures in Graphs. In R. Borgo, F. Ganovelli, and I. Viola (Eds.), *Eurographics Conference on Visualization (EuroVis) - STARs.* The Eurographics Association.

Von Landesberger, T., F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren (2016). Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE transactions on visualization and computer graphics 22*(1), 11–20.

Ward, M. O., G. Grinstein, and D. Keim (2015). *Interactive data visualization: foundations, techniques, and applications.* AK Peters/CRC Press.

Zheng, Y., W. Wu, Y. Chen, H. Qu, and L. M. Ni (2016, Sept). Visual analytics in urban computing: An overview. *IEEE Transactions on Big Data 2*(3), 276–296.