

Visualizing demographic evolution using geographically inconsistent census data

Fabio Dias and Daniel Silver

Abstract—We propose a visual analytics system that enables the exploration evolutionary patterns in geographically inconsistent data, removing the need for harmonization of the geographical regions into a common geometry, a time consuming and error-prone process that is currently used in virtually all longitudinal analysis of geographical data. This work also includes incremental developments in the representation, clustering, and visual exploration of region-based geographical data. While we leverage the well known context of census data, our proposal is suitable for any region-based data. The method enables an easier identification and understanding of the demographic groups present in a city and their evolution over time. We present the feedback of experts in urban sciences and sociology, along with illustrative scenarios in the USA and Canada on the decennial censuses between 1970 and 2010.

1 INTRODUCTION

URBAN sciences are blooming thanks to a renewed interest in understanding and improving the urban environment. Visual analytics is following this trend, fueled by new public datasets that encompass progressively more of our daily lives [1]. There is no shortage of methods to explore mobility patterns [2], social media [1], traffic [3], and so on, providing experts, planners, policy makers, and the general population with deeper insights about their cities.

These new datasets usually contain GPS coordinates for the records, leading to *point-based* data. Combined with the corresponding timestamps, this data is easily suitable for longitudinal analysis. But most demographic datasets are *region-based*, where the measurements are associated with pre-defined regions, not only for an additional level of privacy protection, but because some measurements only make sense over a defined area. Census data is a classic example of this format, with datasets available from 1790 onward for the US [4]. Despite this unmatched temporal availability, longitudinal analyses of census data are often restricted in time, especially when smaller tabulation areas are considered, such as census tracts (CT) or dissemination areas, which evolve to reflect changes in population density, leading to geographic inconsistencies across time, and the traditional time-series based approach is no longer viable.

However, these analyses are necessary to understand the urban environment. Indeed, two different regions can have similar average income for a given year, while one is experiencing a process of economic improvement and the other one impoverishment. A single snapshot cannot be used to identify gentrification, migration, education changes, or any of the relevant processes that happen over time.

To overcome these inconsistencies, the traditional approach is the *geographical harmonization* of the data, the interpolation of the measurements into a common set of

regions [5], [6], [7], so that each variable can be represented using time-series. This is laborious work that inevitably introduces some amount of error [8], even when additional data is provided [9]. Nevertheless, this step is considered mandatory in the current literature: “(...) *tract-by-tract comparison is not possible unless data from 2000 is interpolated to 2010 boundaries (...)*” [10].

The main contribution of this application paper is an visualization-based alternative to the geographical harmonization, a combination of established graph based processing and information visualization techniques allowing tract-by-tract comparison, the identification and visualization of patterns of demographic evolution without geographic harmonization, effectively removing one of the most challenging problems in longitudinal demographic analysis. We also include illustrative scenarios and our prototype is available at <http://uoft.me/piccard>, including more than forty regions in the US and Canada. The source code is publicly available at <https://github.com/fabioasdias/piccard>.

2 RELATED WORK

Since our problem encompasses several fields, we divided this section into specific sub problems: *longitudinal demographic studies*, describing the traditional approach to perform longitudinal studies; *data representation*, exploring how evolving geographic data can be represented for processing; *Data clustering*, briefly reviewing existing clustering methods; and *cluster characterization*, exploring how the clusters can be visually summarized.

2.1 Longitudinal demographic studies

Census data is used not only to discover demographic patterns [11], but to correlate demographic characteristics to other measurements [12]. However, longitudinal studies are rare: “(...) One of the most challenging and fascinating areas in spatial statistics is the synthesis of spatial data collected at different spatial scales(...)” [13].

While CT level data is readily available for the US since 1910 [4], most studies consider the period between 1970

• F. Dias is with the Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, ON, M5S 3G8.
E-mail: fabio.dias@utoronto.ca

• D. Silver is with the Department of Sociology, University of Toronto, Toronto, ON, M5S 2J4.
E-mail: dsilver@utsc.utoronto.ca

Manuscript received MMMM DD, YYYY; revised MMMM DD, YYYY.

73 and 2010, using pre-harmonized data [4], [5]. Despite the
 74 inherent errors [6], [8], this dataset became the standard
 75 source for longitudinal demographic data, with similar ef-
 76 forts appearing in other countries [7], [14], [15]. This result
 77 was significant for the field, but it also restricts the usable
 78 data, since new datasets need to be similarly processed.

79 Another option considers the use of grid data [10], [16],
 80 where small rectangular areas are used, in an approach
 81 similar to satellite imagery. Beyond the increased spatial ac-
 82 curacy, this approach does not require complex harmonization
 83 when new data is considered. However, demographic
 84 data is usually not available in this format, especially from
 85 older sources, and the conversion from tabulation areas can
 86 introduce significant errors.

87 In the proposed methodology, we avoid the harmo-
 88 nization by considering each measurement using its actual
 89 geographic region. It does not require the regions to be
 90 consistent across time because they are already represented
 91 as different entities.

92 **2.2 Data representation**

93 Most data is represented in tabular form, where the rows
 94 and columns have coherent definitions. For example, con-
 95 sider a table with rain measurements over time, with the
 96 rows representing different locations and the columns dif-
 97 ferent times. This representation can also be interpreted as
 98 a collection of time-series, one for each location. Geographic
 99 data followed this format, only including an additional field
 100 that describes the associated geographic area. Following the
 101 example, the data would now represent the amount of rain
 102 for a given region and time. As long each region remains
 103 the same, the data is coherent and can be interpreted again
 104 as a collection of time-series.

105 In the proposed method, we remove the requirement
 106 for consistency in the measurement regions by leveraging a
 107 graph-based representation, where each region in time cor-
 108 responds to a different node. Instead of a collection of time-
 109 series, the data is represented as a dynamic graph. Graph
 110 based representation of geographic information is fairly well
 111 explored in the literature, as a basis for topological methods
 112 for event detection [17], leveraging signal processing on
 113 graphs [18], [19] to find patterns and outliers [20], [21],
 114 [22]. Graphs are well suited to represent trajectories as
 115 well [2], [3], [23], allowing the use of graph visualization
 116 methods [24], [25].

117 Graphs were used to represent census data for clustering
 118 purposes before [21], [26], but these works did not explore
 119 temporal evolution, where graphs are particular powerful as
 120 they allow a natural representation of inconsistent regions,
 121 with both spatial and temporal connections. Note that there
 122 are other possible representations that have similar proper-
 123 ties, but we adopted graphs to allow the use of the existing
 124 literature and methods.

125 **2.3 Data clustering**

126 Data clustering is one of the elementary processes for data
 127 analysis, simplifying the data into a smaller number of
 128 homogeneous sets that can be interpreted in the same way.

While there is no shortage of contributions for this prob-
 129 lem [27], most applications still rely on k-means [28], [29]
 130 and, to a lesser extent, Self Organizing Maps [30], [31].

131 However, a method for geographic data analysis should
 132 not ignore the geographic component of the data. One
 133 straightforward option, for agglomerative methods [32], is
 134 to consider only nearby clusters for merging [33], which
 135 can also be done for k-means [34]. Alternatively, the spatial
 136 distance could be directly added to the inter-cluster met-
 137 ric [33] via a mixing parameter, which adds flexibility to the
 138 method, but introduces the problem of finding the correct
 139 application-dependent values.

140 Indeed, one crucial step in most clustering algorithms
 141 is the definition of the number of clusters. We sidestep
 142 this problem by considering hierarchical methods [35],
 143 where the result is not a partition of the data, but a tree
 144 of partitions. This approach is interesting for interactive
 145 methods, because it allows the user to change the number
 146 of displayed clusters with minimal processing. Since our
 147 data is represented as a graph, one option would be the
 148 watershed cuts algorithm [36], inspired by the well known
 149 image processing segmentation and equally prone to over
 150 segmentation. Considering that the processing time is also
 151 a relevant factor, we opted for an heuristic variation of
 152 the maximum weighted matching algorithm called *sorted*
 153 *maximal matching* [37], which merges clusters based on the
 154 weights of the edges between pairs of clusters.

156 **2.4 Cluster characterization**

157 Visually representing evolving spatial data is a challenging
 158 old problem [38], [39], [40], [41]. Most geographic data
 159 is naturally bi dimensional and maps work well in this
 160 case [41], [42], but the temporal dimension cannot be so nat-
 161 urally represented. One straightforward option is to lever-
 162 age tridimensional plots [43], [44], but this can lead to visual
 163 obstructions or scaling problems unless a tridimensional
 164 display device is used. Animation can also be explored in
 165 some specific cases [45], but it is not a general approach.
 166 Glyphs can also be used [46], [47], but this may lead to
 167 cluttering when many small regions are present. A simpler,
 168 well adopted, option is to display a map that corresponds
 169 to a subset of the temporal information, allowing the user
 170 to change the time with an associated control [1], [17], [20],
 171 [22]. Small multiples can be used [2], but only when there
 172 are few temporal snapshots. However, none of these options
 173 is suitable to represent many variables at the same time.

174 Using data clustering, we can represent the region's
 175 cluster instead of all the its variables [2], [20], [22]. While
 176 this simplifies the geographic portion of the visualization, it
 177 introduces the problem of how to summarize the contents of
 178 each cluster. One traditional approach is to use parallel coor-
 179 dinates plot [48], [49], [50], [51], but these they can get clut-
 180 tered representing similar clusters over several variables.
 181 Further, for demographic applications, the clusters are usu-
 182 ally strongly characterized by a small subset of values [29],
 183 [30]. Therefore, in the proposed method, we identify the
 184 variables that are most relevant to the characterization of
 185 each cluster. The distribution of values on that variable is
 186 then represented using a boxplot, a well known statistical
 187 plot displaying basic properties of the distributions.

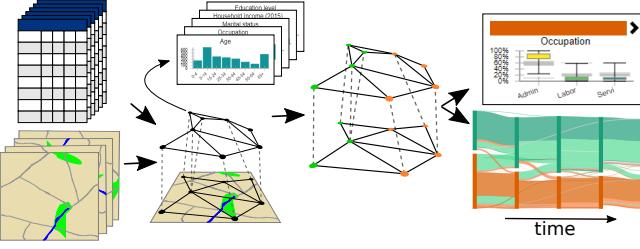


Fig. 1. Overview of the proposed method. A graph is generated combining the original census data, encoding the changing geographical information. The graph is partitioned into an hierarchy [37]. The characteristics and evolution of the clusters are then visually represented.

188 3 VISUALIZING THE DEMOGRAPHIC SPATIO- 189 TEMPORAL EVOLUTION

190 Beyond the objective of allowing the study of inconsistent
191 data, our method includes incremental developments in
192 most steps of the analysis, from data representation to the
193 visualization method for the clusters. Figure 1 presents an
194 overview of the processing steps of the proposed method.

195 3.1 Census methodology and data representation

196 Census data is disseminated in a tabulated form for aggre-
197 gation areas: whole country, state/province, metropolitan
198 region, and so on. To provide as much detail as possible, we
199 focus on the smallest region with available data: *census tracts*
200 (CT). They are usually defined to maintain the anonymity
201 of the population, leading to a population count in the
202 order of thousands in densely populated areas. Physical
203 barriers are usually adopted as borders, so these regions can
204 change because of new roads, construction or removal of
205 high density buildings, and so on. Some census entities also
206 consider demographic characteristics, aiming to establish
207 the CTs as a cohesive unit. Therefore, CTs are the least
208 geographically stable tabulation area.

209 Each CT is associated with a series of variables, with
210 counts derived from the census questionnaires, covering
211 several aspects of the demographic characteristics of the
212 population. Some questions allow for multiple choices
213 or open answers, that are then tabulated into the most
214 frequent categories. Since the census is often used to
215 direct government initiatives, which variables are mea-
216 sured/disseminated is dependent on administrative inter-
217 ests, the general understanding of the population, and cur-
218 rent customs. For instance, income is disseminated with a
219 finer tabulation in the lower portion than on the higher.

220 To match these variables over time and allow for direct
221 comparison across different census years, we aggregated
222 similar ones (e.g. White, Black, Asian, Other) into *aspects*
223 (e.g. Race), encoding the distribution of that facet of the
224 population. In this convention, we refer to the composing parts
225 of an aspect as a *part* or the traditional *variable*. Internally,
226 the aspects are represented using normalized histograms.
227 This normalized representation is crucial for the comparison
228 between inconsistent regions.

229 In our graph based representation, each CT of each
230 census year is represented as a node, and edges are placed
231 between nodes if the corresponding CTs share geographic

232 borders in the same year. Further, edges are placed be-
233 tween nodes if the corresponding CTs belong to sequential
234 years and there is geographical overlap between them. This
235 approach leads to a single graph representing the whole
236 spatio-temporal space of the data. Our objective then be-
237 comes to identify partitions of this graph such that the
238 nodes of each partition are more similar between themselves
239 than to the other nodes. This representation is not the only
240 option, nor unique, but it allows the use of existing graph-
241 based methods for the other steps.

242 3.2 Geographic content clustering

243 To partition the graph we must first establish a distance
244 function between the nodes, measuring the data similarity.
245 This similarity is then associated with the edges, leading
246 to a weighted dynamic graph. Every node has a collection
247 of histograms, each representing the distribution of certain
248 aspect in the population.

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

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

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

263 Once the distances are associated to the edges, we use
264 watershed cuts [36] to create an initial clustering, which
265 is then refined into a hierarchy using the Sorted Maximal
266 Matching (SMM) [37] with median linkage. The initial wa-
267 tershed step is performed to create an initial clustering and
268 reduce the running time of the SMM. For completeness, we
269 briefly review this method, but we refer the reader to the
270 original paper [37] for more details, including a complete
271 performance evaluation using several metrics.

272 We included two application-specific parameters: the
273 maximum number of clusters to be shown and a distance
274 threshold. Contrarily to the original SMM, which merges all
275 clusters in all steps, we only merge two clusters where the
276 distance is above the threshold after we reach the maximum
277 number of displayed clusters. Without this restriction, sig-
278 nificantly different clusters would be merged early, leading
279 to increased intra cluster variance and the disappearance of
280 small outlier regions. Further, after the maximum number
281 of clusters is reached, we create one step of the hierarchy
282 for each merge, leading to a binary partition tree. In this
283 structure we can directly access a result with an arbitrary
284 number of clusters.

285 Each resulting cluster is contiguous in the graph. This
286 means that two similar, but non-contiguous, sets of CTs
287 will be classified into two different clusters, which can

be counter-intuitive. To overcome this issue, we *augment* the graph with two new edges per node from a nearest neighbors graph [52] 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. Theoretically, adding more of these content based edges could be used to decrease the impact of the spatio-temporal edges, controlling the balance between content and topology in the result. In practice, the effect is dependent on the data itself, and the results are not consistent, or predictable, across different cities. We fixed it at two edges because it was the lowest number that empirically led to consistent clusters, but we believe that this idea warrants further investigation, as an alternative to mixing parameters in the distance metric [33].

3.3 Cluster characterization and variable relevance

The composition of each cluster is determined by simple statistical measures, considering each aspect separately. We compute the minimum, maximum, median, 25%, and 75% quantiles for each part 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 characterizes each cluster, that is, what makes it different from the others. We define the *relevance* of a part of an aspect based on the distance between the interquartile ranges (IQR) of the clusters in the same hierarchical level. If the IQRs overlap for all clusters, that variable is not relevant to the characterization 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.

3.4 Clusters and trajectories

While the partition of the data into different clusters helps the user to understand what demographic groups exist and where they are, we are also interested in the changes in those groups. To this end, we introduce the concept of *trajectories*, which are composed by regions that are classified into the same sequence of clusters over the considered period. This enables direct access to regions that evolved in the same manner. While the interface provides access to data by census tract, the trajectories are the main unit of exploration in this work.

3.5 Colors

As illustrated by Figure 3 and further explored in the next subsection, our proposed interface heavily relies on color to express cluster-related information. We adopted this convention because colors can be used in all our visual tools in a coherent manner. However, this also introduced significant challenges. The first is the limit on the number of clusters that can be visually represented. We limited the number of clusters to eight because this was the largest number of colors that we could reliably use, derived from the 8-class Dark2 set from ColorBrewer [53].

While we can reasonably limit the number of clusters, there are far more possible trajectories. And the color associated with each trajectory should bear some resemblance

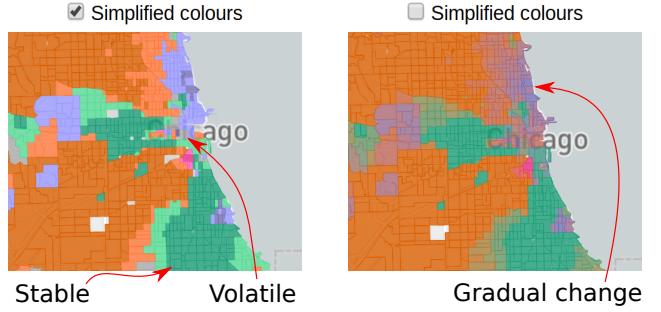


Fig. 2. Different color schemes for Chicago with four clusters. Left: simplified, right: average color.

to the clusters included in it. Therefore, we were left with a conundrum: *Should we associate each trajectory with a unique color, which the user probably cannot distinguish, or should we use a reduced set of colors and associate the same color to different trajectories?* Since there are advantages and disadvantages for each of those options, we adopted both. The user can control which color policy is used via a checkbox in the configuration panel, on the top left of the interface.

By default, the interface adopts a simplified color scheme, where a trajectory is painted in the same color of a cluster if the regions were associated to that cluster for *all* times; in a slightly less saturated version of that color if the regions were associated to that cluster for *the simple majority* of the time, and gray otherwise. In this mode, the colors will mostly represent stability, immediately identifying the regions that were consistently associated with each cluster. It also easily identifies volatile regions, painted gray.

When this simplified color scheme is disabled, each trajectory will be painted using the average of the colors of the involved clusters, in the LAB color space. In this mode, the map becomes more similar to a heatmap, where stronger presence of a color indicates more temporal affinity to the cluster. Volatile regions will also tend to be displayed in gray, as the average of three or more colors.

While both approaches will use more than eight colors, in practice this is not as significant because most cities can be explained using less than eight clusters. In fact, articles in the literature usually employ from two to five, which are fairly stable across time. For the more dynamic scenarios, user interaction can be used to alleviate the shortcomings of both approaches.

3.6 User interface

The initial interface is illustrated in Figure 3. Since demographic data can be nuanced, with intricate interconnections, we decided against validating the interface using a synthetic dataset, considering instead data from the Chicago region between 1970 and 2010, using previous published studies as corroboration. This region is known for its entrenched racial divide and the emergence of a '*young urban*' population with a higher education level [29], [30]. More details about this dataset are presented in Section 4, with an workflow example in Figure 5.

The configuration panel, on top left in Figure 3, displays which aspects were used and their weights (following Equation 1). It also includes other configuration options that

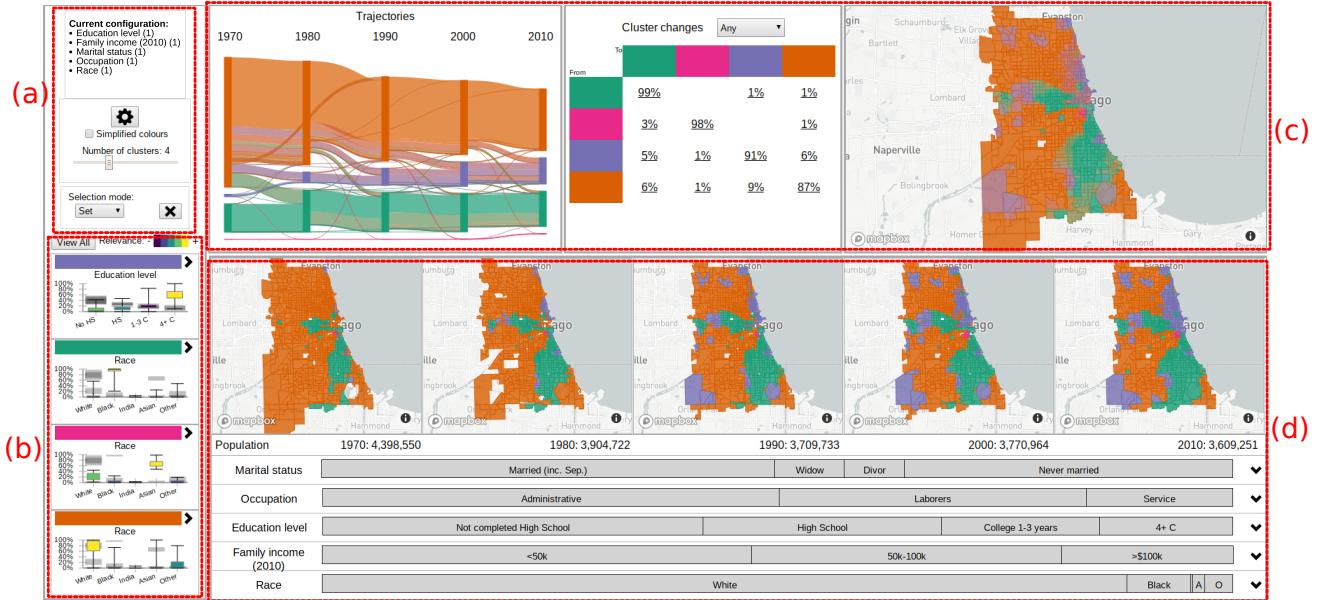


Fig. 3. Initial interface of our method showing the demographic evolution of Chicago. (a): Configuration panel with the current clustering parameters and controls. (b): Cluster overview illustrating the most relevant aspect for each cluster. (c): Trajectories overview and the general evolution of the population, geographical information, and how it changed. (d): Details of the selected trajectories, including precise geographic locations, population numbers, and the composition of the aspects.

can be altered without re-processing the data, such as the number of clusters and the color option. The gear button allows access to the other configuration options that do require further processing, such as changing location, aspects, and weights. This panel also includes the configuration of the selection mode for the trajectories, which allows the user to set, add, or remove the next selected trajectories to the current selection. This feature enables the analysis of complex sets of trajectories.

The cluster overview panel, on the bottom left in Figure 3, displays a brief summary of each cluster, based on the distance between the IQRs, as detailed in Section 3.3. The *View all* button opens a new panel where all aspects are represented, while the chevron at the side of the color lets the user expand each cluster separately. While the standard approach to represent cluster characteristics is to use parallel coordinates [50], [51], this representation occupies screen space proportional to the number of variables and can get cluttered with a higher number of clusters, or when the clusters are not well defined for multiple variables. To save space and leverage the familiarity scientists have with statistical tools, we opted to use boxplots to properly convey the distribution of each variable in the current cluster. However, a simple boxplot would not include information about the other clusters, forcing the user to mentally compare them to find what is relevant.

We adopt an *enhanced* version of the traditional boxplot, which includes the minimum, maximum, 25% and 75% quantiles for the current cluster, but also the IQRs for the other clusters, in slightly larger and faded black rectangles. We also color the current IQR according to its relevance. While there might be some degree of similarity between the color schema for relevance and for trajectory identification, none of the experts consulted reported confusion. Indeed, one expert reported confusion regarding the grey rectangles

that represent the IQRs for other distributions when they are not colored; when that variable is not relevant to the characterization of the cluster. These simple changes allow the user to easily understand the composition of the cluster and how it relates to the others. Violin plots [54] could also be used, providing more information about the shape of the distribution, but with increased potential for obstructing the representation of the other clusters.

For instance, the boxplot that summarizes the purple cluster illustrated in Figure 3, detailed on Figure 4, illustrates that 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 colored 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. By clicking on the colored bar above the boxplot, the user can select all trajectories that contain this cluster at any point in time.

The other clusters identified on Figure 3 correspond to higher concentration of people that identify as Black in the green cluster, people that identify as "Asian, Hawaiian, other pacific islander" in the pink cluster, and people that identify as White in the orange cluster. From these plots, it is clear that the city is indeed racially divided [29], with several CTs that are almost exclusively occupied by people of the same racial category.

The trajectories overview aims to convey basic informa-

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

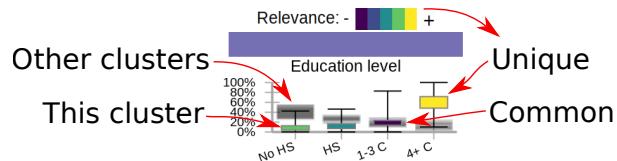


Fig. 4. Enhanced boxplot of the clusters' characteristics allows a quick comparison to the other clusters.

tion about the trajectories present in the data, where they are, and what changes are involved. This is done using three sub panels. The first, on the left, contains a Sankey diagram illustrating the evolution of the clusters over time. The widths are proportional to the population involved, the colors follow a policy detailed in Section 3.5. A stacked graph could also be used to represent the proportions of each cluster [20] with less clutter, since the transitions between clusters would not be represented. However, this is only viable if more temporal steps are available, making the plot smoother. Another option to remove clutter is to remove portions of this plot when trajectories are selected, but this would change the layout and compromise the user's mental map.

In our example in Figure 3, we can see that the total population of Chicago is decreasing. Additionally, 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 [29], [30]. This diagram can also be used to select specific trajectories, by clicking on the bands, or all trajectories that contain a specific cluster at a specific time, by clicking on the rectangles.

The next sub panel, in the top middle of Figure 3, is a transition matrix between the clusters. It indicates the percentage of the population whose area changed in each pattern. This kind of table can be found in the related literature [29], so it is familiar to the advanced users, and it not only informs the proportional changes, but allows the selection of the corresponding trajectories for further analysis.

Contrary to the trajectories plot, this representation is more Markovian, where only the current and next state are considered. This panel also enables easier access to trajectories with specific changes, by clicking on the corresponding percentage values. The combo box allows the user to refine the transitions, from 'Any', which includes all transitions between years, to specific transitions, to changes from the first year to the last year. In this example, approximately 99% of the population in areas classified as green were also in areas classified as green in the next year, while 1% changed to purple at some point and another 1% to orange. The total is over 100% due to rounding errors. Regions changed from the orange to the green cluster for 6% of its population, 1% to pink, and 9% to purple. This further corroborates the fact that most of the growth of the purple cluster came from the orange cluster. Additionally, the lack of transitions is also relevant, for instance, no CT changed from majority of Black

population (green) to Asian population (pink), and no CT with significant Asian population had significant increase in education levels (purple).

The panel in the top right of Figure 3 is a map of the region under analysis, summarizing the geographical evolution of the clusters over time. The colors are derived from the clusters involved in each trajectory as detailed in Section 3.5, which are consistent across the linked views.

The bottom part of the interface contains the details for the selected trajectories, or for the whole city if nothing is selected, as illustrated in Figure 3. This panel contains two main regions: the small multiple maps, depicting the clusters at each year, and the stacked bar plots that summarize the overall composition of these regions. Some finer localization information is lost using small multiples, such as small border changes, but that information is available at the larger map. All the maps are linked with synchronized navigation, and the use of small multiples allows the exploration of each temporal census individually, and its comparison to the others, with minimal interaction.

In this example, the maps show the transition from orange to green and purple in several regions over time. Clicking on a region in these maps will bring up a new panel with the original census data of this specific region. The actual population numbers are below the maps, and they confirm the notion provided by the Sankey diagram that the total population is indeed decreasing.

Each aspect is represented by a stacked bar plot, where the width of each rectangle corresponds to the average percentage of that variable over the considered period. We chose stacked bar plots to represent the composition of the regions because they can accurately and succinctly inform the proportions of each aspect, without any interaction. In this case, about half of the people in Chicago in the considered period 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 half have gross family income below 50,000USD per year. The vast majority identify as white. Placing the mouse over one of the bars will open a small panel with the temporal evolution of that specific variable, and clicking on the chevron on the right side expands the corresponding aspect, showing details of the temporal evolution of each variable and also the corresponding IQRs for the whole city.

4 ILLUSTRATIVE SCENARIOS

We used decennial census data from the United States [4] and Canada¹, tabulated by CTs, from 1970 to 2010. The prototype allows access to 40 regions, 28 in the US and 12 in Canada. Due to the high number of CTs, New York City was split into its boroughs.

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. While our method does not require geographic harmonization, it requires matching variables

1. <http://datacentre.chass.utoronto.ca/census/>

568 over time. The supplementary material contains the details
 569 of which census columns were used for each aspect. Income
 570 is slightly inaccurate, even though we did correct for official
 571 inflation. We grouped the original ranges into three larger
 572 ranges, but they do not match precisely.

573 4.1 Chicago

574 We selected a region loosely following the administrative
 575 borders. The demographic composition is well explored in
 576 the literature, with reports of racial divide and gentrification
 577 [29], [30], [55]. While the definition of gentrification is
 578 still unclear and out of the scope of this paper, we associate
 579 gentrification with higher education and income levels.

580 The initial state of the prototype is illustrated in Figure 3,
 581 and its findings are explained in Section 3.6, where the
 582 racial divide is clear. Starting from this initial state, the
 583 specific workflow used to identify the existence and details
 584 of the gentrification process are illustrated in Figure 5. For
 585 the users, the first step is to identify the compositions of
 586 each cluster from the boxplots, so orange is associated with
 587 majority of White population, green with majority Black,
 588 and purple with higher proportion of four years of college
 589 or more (high education level). The expanded version of the
 590 boxplots for the purple cluster shows a higher income level
 591 and majority of occupations in administrative jobs, therefore
 592 the purple cluster identifies gentrified regions.

593 The trajectories plot illustrates the process of gentri-
 594 fication, progressively absorbing regions from the orange
 595 cluster (White). This corroborates results from the literature
 596 reporting that Black neighborhoods are less likely to gentri-
 597 fy [55]. Moreover, this process is unlikely to be reversed,
 598 as indicated by the limited number of trajectories leaving the
 599 purple stream. Next, we select the region that is gentrified
 600 in 2010, by clicking on the corresponding rectangle in the
 601 trajectories plot, updating the information on the maps and
 602 the details portion of the interface.

603 The corresponding CTs are highlighted in the maps,
 604 where the spatial pattern is clear, corresponding exactly
 605 to previous findings in the literature [55]. Further, we can
 606 also identify the regions that gentrified earlier, with an
 607 stronger purple hue, compared to newer regions, where the
 608 orange and green colors are still present. This also indicates
 609 from which clusters they belonged before gentrifying. In the
 610 details portion of the interface, the order of the aspects was
 611 updated to reflect the order of relevance considering only
 612 the selected region. The most relevant aspect is the Educa-
 613 tion, specifically "Four or more years of college", illustrated
 614 in the rightmost portion of Figure 5, which is increasing for
 615 the whole city (grey band), but faster and to a higher level
 616 in this region (black band). Indeed, while the IQR for the
 617 city goes from 11% to 45%, the IQR for this region goes
 618 from 55% to 77%. However, there are portions of this region
 619 with significantly lower or higher proportions, as indicated
 620 by the dotted black lines representing the minimum and
 621 maximum for the selected region.

622 4.2 Toronto

623 We considered a region that is approximately the adminis-
 624 trative border of the current city of Toronto and all seven

available aspects with equal weights. The results are sum-
 625 marized in Figure 6, considering eight clusters.

626 The population with low percentage of University de-
 627 grees is represented in orange, mostly anglophone popula-
 628 tion in green, Asian immigrants in yellow, high percentage
 629 of income in the highest bracket in purple, high percentage
 630 of Jewish people in light green and brown, high percentage
 631 of Eastern Non-Christian religion in pink, and high concen-
 632 tration of single people in dark gray. From the trajectories
 633 plot, we can see that Toronto is more dynamic than Chicago,
 634 with one cluster constantly shrinking. In the 1970s, the city
 635 was divided into four clusters: low number of university
 636 degrees, Jewish population, majority anglophones, and high
 637 income. Interestingly, the more recent clusters that absorbed
 638 regions from the orange cluster have similar education
 639 profiles and are differentiated by other aspects. In this sense,
 640 the city is growing diverse, changing from a common low
 641 education profile to a higher level of education with more
 642 diversity in religion (pink) and immigration (yellow).

643 Indeed, the influx of Asian population is visible starting
 644 in the 1980s and building thereafter, leading to the yellow
 645 and pink clusters. While both include a high percentage of
 646 people born in Asia, the pink is more defined by religion,
 647 with low percentage of university degrees, and contains the
 648 lowest percentage of people in the highest income bracket
 649 for these clusters; the yellow is less defined by religion,
 650 and has higher education and income, geographically cor-
 651 responding to the Markham region, known for its Chinese
 652 population. A similar division also happens for the two
 653 Jewish clusters, where the light green cluster has lower
 654 education and income levels than the brown cluster. The
 655 purple cluster of high income is somewhat stable. This
 656 cluster includes the Bridle Path neighborhood, known for
 657 its wealthy population, until 2011. In 2011 it was classified
 658 into the yellow cluster of Asian immigration, since about
 659 35% of the population for this CT were then born in Asia.
 660 The income distribution did not change, with 85% of the
 661 population with an income of 90k CAD or more.

662 The most significant indicator of Toronto's dynamism is
 663 the presence of grey regions on the simplified color map;
 664 representing regions associated to three or more clusters
 665 over this five census period. Using the 'Add' mode for
 666 the trajectory selection, we select their trajectories, and a
 667 subset of the details is illustrated in Figure 7. These regions
 668 account for about 5% of Toronto's population. The whole
 669 region was classified into the orange cluster in 1971 (low
 670 level of university degrees). By 1991, most of the region was
 671 classified into the green cluster, representing anglophone
 672 population, mostly Canadian born, with a higher level of
 673 education. As the corresponding plot indicates, this trend in
 674 increasing education is city-wide, but this region has people
 675 with better education than most.

676 In 2001, the purple cluster of high income annexes neigh-
 677 bor parts of the volatile region, and the Asian born pop-
 678 ulation increases sharply, as illustrated by the appearance
 679 of the yellow cluster. This cluster indicates well educated,
 680 higher income, and about 30%-50% Asian born population.
 681 By 2011, the yellow cluster increased considerably, annex-
 682 ing parts of the high income purple cluster, including the
 683 neighboring Bridle Path area.

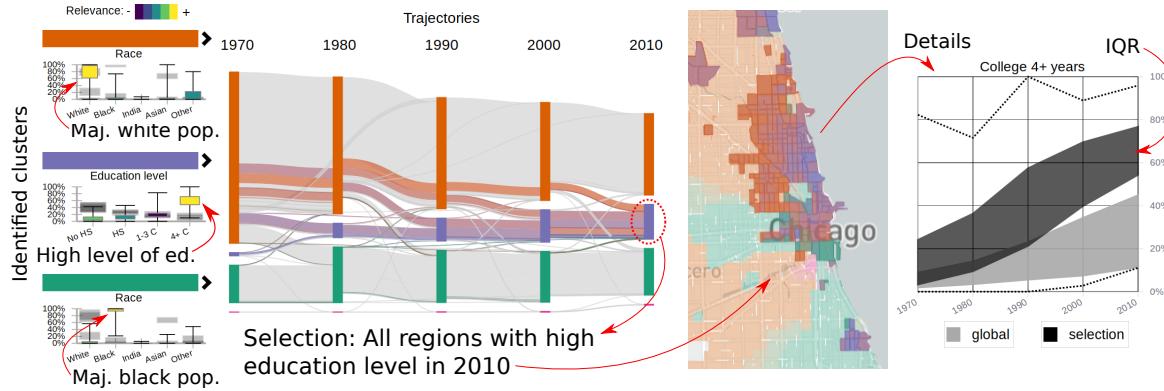


Fig. 5. Workflow to discover gentrification in Chicago: the purple cluster corresponds to high education / income. Its population is increasing over time, absorbing from the majority White cluster (orange). By selecting the purple cluster in 2010, the region is highlighted in the maps. The proportion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

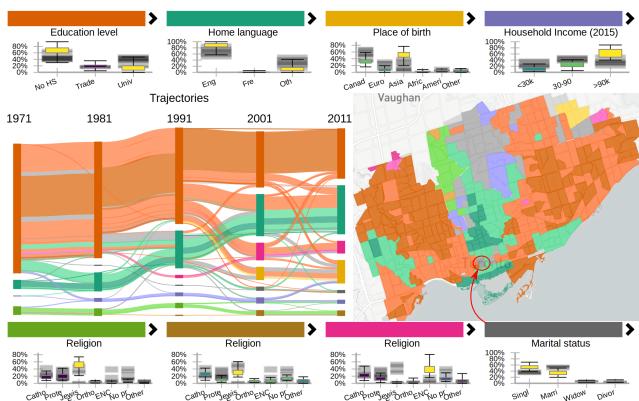


Fig. 6. Clustering results for Toronto, with eight clusters, including clusters representing Jewish population, high and low income, low education, and Asian immigration.

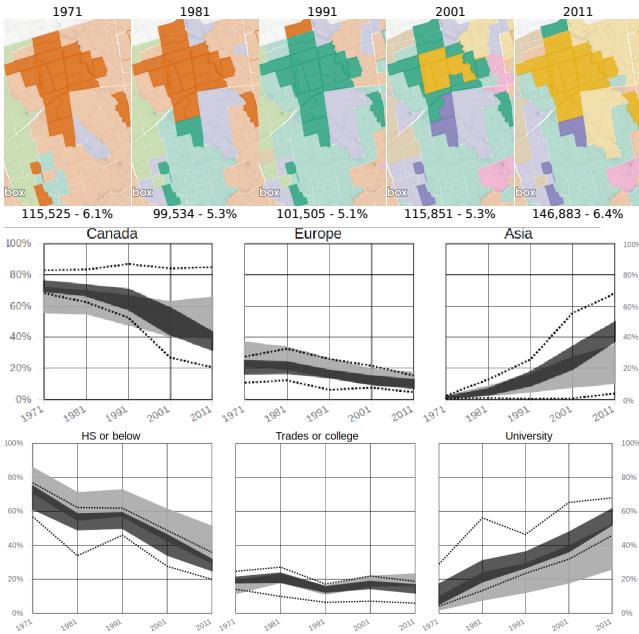


Fig. 7. Details for some regions of Toronto that were classified into 3 or more clusters over time.

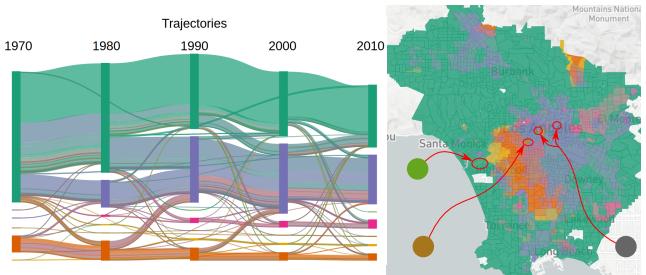


Fig. 8. Result for Los Angeles with 8 clusters, including three small and ephemeral clusters. Cluster characterization is displayed in Figure 9.

4.3 Los Angeles

We selected a region around the metropolitan area of Los Angeles (LA), following urban density. The summary of the results using all aspects with equal weights and eight clusters is illustrated in Figure 8. The full statistical description of the clusters is illustrated in Figure 9, where the most relevant aspect of each cluster is highlighted. From the trajectories plot, we can see that there is a large but shrinking cluster, depicted in green, one increasing cluster in purple, an almost constant orange cluster, a smaller but increasing pink cluster, and three other small clusters. The corresponding map illustrates where these clusters are located, and that they are somewhat geographically stable, with some movement between the green, orange, and purple clusters.

From Figure 9, we can see that the green cluster is characterized by a high percentage of White population, low percentage of population in the lowest income bracket, mostly administrative occupations, and about 30% of the population with four or more years of college. The orange cluster is characterized by a high percentage of Black population, with few people in the highest range of income and education. The purple cluster corresponds to a high concentration of "Other" in race, which includes Hispanic for this dataset, high concentration of Laborers, and low education and income. The pink cluster contains a high percentage of Asian population and about 30% of the population with four or more years of college. The light green cluster contains very few people in the lower income bracket, mostly White population, with the highest percentage of population with

686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713

714 four or more years of college, working administrative jobs,
 715 and a high concentration of singles. The yellow cluster
 716 represents Black population, with higher level of education
 717 and income, mostly working administrative jobs. The brown
 718 cluster represent a majority of single population, working
 719 administrative jobs with mostly low income. The dark gray
 720 cluster is characterized by all its population in the lowest
 721 income bracket, low education level, with a majority of
 722 White population. Since the extremes in the boxplots of
 723 the grey cluster are not significantly different, we can also
 724 surmise that this cluster is either small or homogeneous.

725 The green, orange, and purple clusters present a signif-
 726 icant intra-cluster variance in most variables, as indicated
 727 by extreme whiskers of the boxplots. While fifty percent of
 728 the CTs in the green cluster have between 20% and 40%
 729 of people in the lowest income bracket, that cluster also
 730 includes CTs where none and all the population belongs to
 731 that bracket. This might indicate that this cluster represents
 732 different groups of people that are not different enough to be
 733 separated at this level of the hierarchy. Conversely, the light
 734 green, brown, and dark gray clusters are different enough
 735 to be separated into their own clusters at this level, despite
 736 being small and ephemeral, including only a few CTs.

737 The orange area in the map in Figure 8 presents move-
 738 ment, indicated by the presence of green and purple tones
 739 mixed with the orange, which may warrant further explo-
 740 ration. By clicking on the orange bar above the boxplot,
 741 we select all trajectories that contain the orange cluster.
 742 The corresponding details are illustrated in Figure 10. This
 743 shows a location change, where the orange cluster is pro-
 744 gressively replaced by the purple cluster on its east side, and
 745 in turn expanding to the west. Interestingly, the population
 746 increased, the racial profile changed, but the distribution of
 747 income was reasonably stable, with a higher amount of the
 748 population in the lowest income range and very few people
 749 in the highest income range. Indeed, the income difference
 750 is significant when compared to the city-wide distribution.

751 A portion of this region is classified into the green cluster
 752 in 2010, indicating a majority white population. To further
 753 understand that change, we clear the current selection,
 754 and select all regions that changed from orange in 1970
 755 to green in 2010, using the transition matrix. A portion of
 756 the resulting region, near the Florence-Graham region, is
 757 depicted in Figure 11, along with the temporal evolution of
 758 Race. Despite this difference, the other aspects are similar
 759 to the ones from the region in Figure 10, with slightly
 760 lower income and education profiles. While the racial aspect
 761 changed considerably, the economic and educational aspects
 762 stayed the same.

763 While Toronto is more dynamic than Los Angeles, pos-
 764 sibly due to size differences, the volatile regions shown in
 765 Figure 7 did not change as quickly or dramatically as the
 766 ones shown in Figure 11, which involved twice as many
 767 people. We found this trend to be related to the countries
 768 themselves, Canadian cities have larger areas undergoing
 769 slow, gradual changes, whereas American cities have more
 770 general stability, but quicker changes in smaller scales. The
 771 supplementary material contains brief summaries of all the
 772 regions accessible in this prototype.

5 EXPERT FEEDBACK

To assess the proposed method, we contacted academic
 774 and industry experts in sociology and urban sciences. The
 775 supplementary material contains the communication with
 776 five experts, verbatim, identified from A to E.

Their overall response was positive, mentioning that the
 778 prototype allows them to analyze census data without the
 779 additional work of obtaining and cleaning the data (A, B,
 780 E), and it allows the inclusion of geographic visual analysis
 781 tools in their research process (D). It enables the users to
 782 tell different stories about neighborhoods/cities and their
 783 changes (A), visualize the relationship between key urban
 784 variables over time (D), offering a quick way to identify
 785 particular neighborhoods that one may be interested in
 786 studying more in depth around a particular issue or effi-
 787 ciently understanding the context of an area (E). Indeed,
 788 the experts identified gentrification processes in Manhattan
 789 (B) and Dallas (E), reinforced a hypothesis for occupational
 790 clustering (D), and highlighted how the method can be used
 791 to compare neighborhoods and cities (A).

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

While the map of trajectories was mentioned as a "good
 803 summary map", how it related to the clustering method
 804 was unclear (C). While we provide different options on
 805 how the colors are used, both are sub optimal since reliably
 806 representing several distinct entities using colors is humanly
 807 unfeasible. Indeed, the number of distinguishable colors
 808 was a significant constraint, we found indications that more
 809 clusters should be used in some cases, even if eight clusters
 810 is more than what is traditionally considered in these anal-
 811 yses. Conversely, increasing the number of clusters would
 812 also complicate the interpretation of the results.

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

Most of the experts demonstrated interest in using our
 824 method in their research (A, B, D, E), aiming to use the
 825 census data as a backdrop for other datasets, providing
 826 demographic context. They also mentioned the need to
 827 export subsets of data, plots, and maps to be used in reports
 828 and publications (C, D, E).

More importantly, while these experts were aware that
 830 our method does not perform geographic normalization,
 831 none of them mention it. We did not specifically ask if this

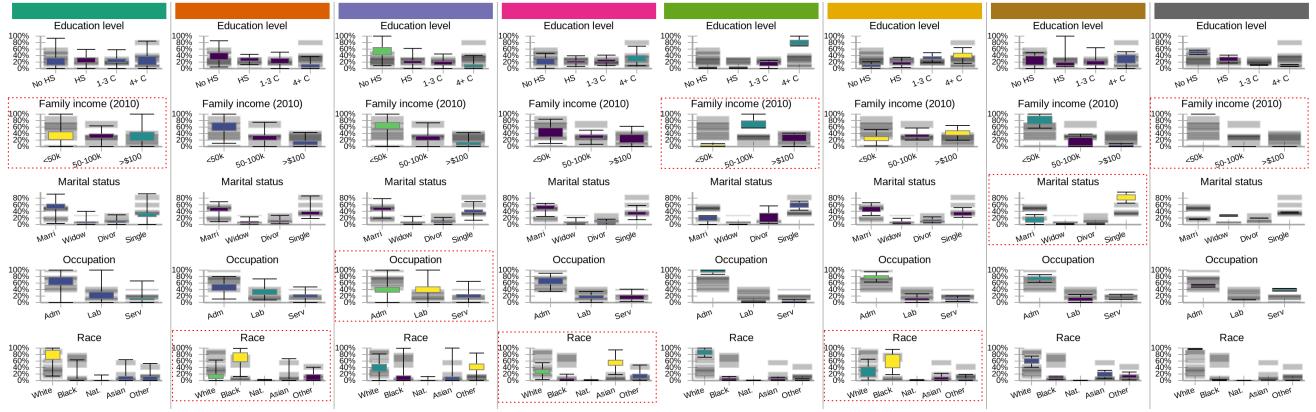


Fig. 9. Full characterization of the eight clusters found for LA. The red rectangles indicate the most relevant aspects for each cluster.

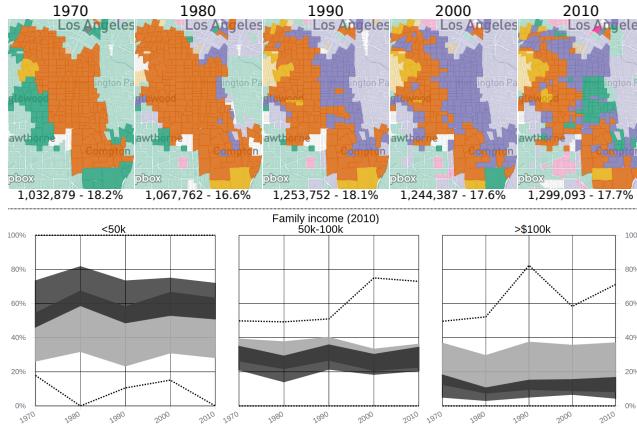


Fig. 10. Top: Geographic changes in the majority Black population cluster (orange) and Laborers cluster (green). Bottom: Income evolution for this region (black) and the whole city (gray).

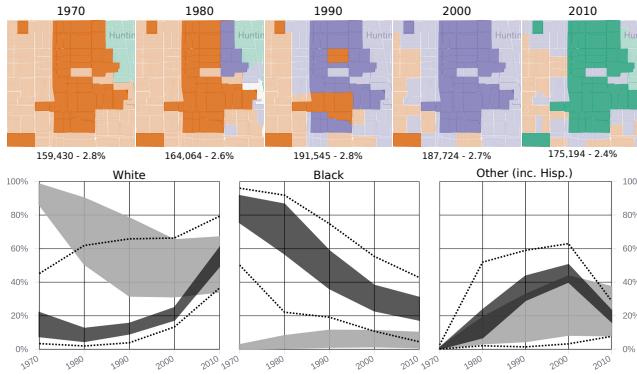


Fig. 11. Details for a volatile region contained in the area of Figure 10. This region went from Black to Hispanic to White.

difference led to unexpected results, but rather inquired if they found interesting insights, leading them to consider the interface, its options, and results as a whole. Most experts found phenomena corroborated by the specialized literature, indicating that our methodology produces equivalent results, with a fraction of the effort. We interpret the fact that most of them were interested in the next steps as an affirmation of the accuracy of the method.

6 DISCUSSION AND LIMITATIONS

The objective of this work was to leverage a graph based data representation with visual tools to allow for the exploration of geographically inconsistent census data. While we successfully replicated and corroborated results from the literature, this method still has significant limitations.

Removing the geographical normalization/interpolation step greatly reduces the amount of work necessary, but the method still requires consistent variables across the years. Matching the fields of the public census can be trivial for some aspects (Age), but challenging for others (Income). The divulged income ranges vary over time, the actual value changes due to inflation and other factors, and so on. Moreover, some fields were not considered in earlier censuses, such as Race in Canada, or Hispanic population in the USA, hampering its use when they are available. Since this is only a prototype, we matched few aspects, but a deeper demographic analysis would greatly benefit from all available information.

The limitation on the number of displayed clusters because of the limited number of distinguishable colors was significant. While increasing the number of clusters would further complicate an already complex analysis, it might be warranted for some regions. Color 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 colors, there was overlap between some clusters, the relevance gradient, and the color combination.

Another limitation is the lack of control on how much the geographical information will impact the clustering result. While the adopted method met our needs for this work, a configurable control would add another dimension to the exploration, allowing for more intra-cluster variance to obtain more 'compact' clusters. We explored changing the number of content based augmented edges, but this proved to be unreliable and hard to interpret. The *ClustGeo* method [33] can be a viable option for this, allowing a graph based input and a hierarchical output, combined using a single mixing parameter. Alternatively, one could cluster the changes [56] instead of the stable states.

There are also technological limitations, such as memory use on the visualization client. To allow for changes on the

884 CTs over the years, we use a geographic file that contains
 885 all possible intersections, which can grow rather large if the
 886 original city was expansive and contained several CTs, like
 887 NYC or LA. However, the most significant technological
 888 limitation relates to parameters that are not immediately
 889 interactive, such as the clustering configuration. Since the
 890 clustering is computationally expensive and performed on
 891 the server, which allows for cached results, some changes
 892 can take a few minutes to be considered, removing any
 893 possibility of a continuous exploration.

894 Indeed, the cognitive load on the user is already sig-
 895 nificant, as we compromised simplicity for accuracy. While
 896 other works labelled the clusters, as 'young urban', 'strug-
 897 gling', and so on [29], [30], we show the statistical char-
 898 acteristics of the clusters, which are harder to interpret, as the
 899 data may have subtle nuances that labels would otherwise
 900 hide. This also led to a crowded interface, mitigated some-
 901 what the use of pop-up panels and collapsible sections. For
 902 some cities, especially if they are small and stable, the panels
 903 can appear redundant, but each provide a different way to
 904 interact with the information that can ease the exploration
 905 process for larger and dynamic cities.

906 7 CONCLUSION

907 Our objective was to allow for the exploration of census data
 908 without geographical harmonization, an approach that is
 909 usually not considered. Our method was able to corroborate
 910 previous findings from the specialized literature, with an
 911 increased level of detail due to our data representation
 912 and visualization choices. The feedback from experts was
 913 remarkably positive, most of them were able to extract
 914 insight from the prototype and demonstrated interest in
 915 using it on their research efforts.

916 While we focused on census data in this work, the
 917 proposed method can be used with most similarly area-
 918 based datasets. Further, while there is interest in the use
 919 of census data, it is mostly in combination with other, larger
 920 and more dynamic, datasets. We postulate that, while census
 921 data is interesting, the amount of effort required to obtain
 922 and prepare it often discourages its use as complementary
 923 information, reinforcing our hypothesis that there is need
 924 for methods that are able to deal with geographically incon-
 925 sistent data.

926 "(...) tract-by-tract comparison is not possible unless data from
 927 2000 is interpolated to 2010 boundaries (...)" [10]

928 ACKNOWLEDGMENTS

929 This research was supported by a University of Toronto
 930 Connaught Global Challenge grant and is part of the Urban
 931 Genome Project. The authors thank Cary Wu, Ethan A.
 932 Fosse, Fernando Caldern Figueroa, Patrick Adler, and James
 933 Murdoch for their expert opinions; Mark S. Fox, Robert M.
 934 Wright, Ultan Byrne, Matti Siemiatycki, Shauna Brail, and
 935 Richard Florida for general guidance and support; and the
 936 anonymous reviewers for their constructive comments.

937 REFERENCES

- [1] W. Chen, Z. Huang, F. Wu, M. Zhu, H. Guan, and R. Maciejewski, "Vaud: A visual analysis approach for exploring spatio-temporal urban data," *IEEE Transactions on Visualization & Computer Graphics*, no. 1, pp. 1–1, 2017.
- [2] T. Von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren, "Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 11–20, 2016.
- [3] W. Chen, F. Guo, and F.-Y. Wang, "A survey of traffic data visualization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 2970–2984, 2015.
- [4] S. Manson, J. Schroeder, D. V. Riper, and S. Ruggles. (2017) Ipums national historical geographic information system: Version 12.0 [database]. Minneapolis: University of Minnesota. [Online]. Available: <http://doi.org/10.18128/D050.V12.0>
- [5] J. R. Logan, Z. Xu, and B. J. Stults, "Interpolating us decennial census tract data from as early as 1970 to 2010: A longitudinal tract database," *The Professional Geographer*, vol. 66, no. 3, pp. 412–420, 2014.
- [6] E. Hallisey, E. Tai, A. Berens, G. Wilt, L. Peipins, B. Lewis, S. Graham, B. Flanagan, and N. B. Lunsford, "Transforming geographic scale: a comparison of combined population and areal weighting to other interpolation methods," *International Journal of Health Geographics*, vol. 16, no. 1, p. 29, Aug 2017.
- [7] J. Allen and Z. Taylor, "A new tool for neighbourhood change research: The canadian longitudinal census tract database, 19712016," *The Canadian Geographer / Le Géographe canadien*, vol. 0, no. 0, 2018. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cag.12467>
- [8] J. R. Logan, B. J. Stults, and Z. Xu, "Validating population estimates for harmonized census tract data, 2000–2010," *Annals of the American Association of Geographers*, vol. 106, no. 5, pp. 1013–1029, 2016.
- [9] C. L. Eicher and C. A. Brewer, "Dasymetric mapping and areal interpolation: Implementation and evaluation," *Cartography and Geographic Information Science*, vol. 28, no. 2, pp. 125–138, 2001.
- [10] A. Dmowska, T. F. Stepinski, and P. Netzel, "Comprehensive framework for visualizing and analyzing spatio-temporal dynamics of racial diversity in the entire united states," *PLOS ONE*, vol. 12, no. 3, pp. 1–20, 03 2017.
- [11] G. Firebaugh and C. R. Farrell, "Still large, but narrowing: The sizable decline in racial neighborhood inequality in metropolitan america, 1980–2010," *Demography*, vol. 53, no. 1, pp. 139–164, Feb 2016. [Online]. Available: <https://doi.org/10.1007/s13524-015-0447-5>
- [12] A. V. Diez-Roux, F. J. Nieto, C. Muntaner, H. A. Tyroler, G. W. Comstock, E. Shahar, L. S. Cooper, R. L. Watson, and M. Szkló, "Neighborhood environments and coronary heart disease: a multilevel analysis," *American journal of epidemiology*, vol. 146, no. 1, pp. 48–63, 1997.
- [13] C. A. Gotway and L. J. Young, "Combining incompatible spatial data," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 632–648, 2002.
- [14] X. Liu, Y. Song, K. Wu, J. Wang, D. Li, and Y. Long, "Understanding urban china with open data," *Cities*, vol. 47, pp. 53 – 61, 2015, current Research on Cities (CRoC).
- [15] A. C.-D. Lee and C. Rinner, "Visualizing urban social change with self-organizing maps: Toronto neighbourhoods, 1996–2006," *Habitat International*, vol. 45, pp. 92–98, 2015.
- [16] A. Dmowska and T. F. Stepinski, "Spatial approach to analyzing dynamics of racial diversity in large u.s. cities: 199020002010," *Computers, Environment and Urban Systems*, vol. 68, pp. 89 – 96, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S019897151730371X>
- [17] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. T. Silva, "Using topological analysis to support event-guided exploration in urban data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2634–2643, Dec 2014.
- [18] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.
- [19] A. Sandryhaila and J. M. Moura, "Discrete signal processing on graphs," *IEEE transactions on signal processing*, vol. 61, no. 7, pp. 1644–1656, 2013.
- [20] P. Valdivia, F. Dias, F. Petronetto, C. T. Silva, and L. G. Nonato, "Wavelet-based visualization of time-varying data on graphs," in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct 2015, pp. 1–8.

- [21] F. Dias and L. G. Nonato, "Some operators from mathematical morphology for the visual analysis of georeferenced data," in *Workshop on Visual Analytics, Information Visualization and Scientific Visualization - SIBGRAPI*, 2015.
- [22] A. Dal Col, P. Valdivia, F. Petronetto, F. Dias, C. T. Silva, and L. G. Nonato, "Wavelet-based visual analysis of dynamic networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2018.
- [23] X. Huang, Y. Zhao, C. Ma, J. Yang, X. Ye, and C. Zhang, "Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 160–169, Jan 2016.
- [24] C. Vehlow, F. Beck, and D. Weiskopf, "The State of the Art in Visualizing Group Structures in Graphs," in *Eurographics Conference on Visualization (EuroVis) - STARs*, R. Borgo, F. Ganovelli, and I. Viola, Eds. The Eurographics Association, 2015.
- [25] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, "The State of the Art in Visualizing Dynamic Graphs," in *EuroVis - STARs*, R. Borgo, R. Maciejewski, and I. Viola, Eds. The Eurographics Association, 2014.
- [26] T. Setiadi, A. Pranolo, M. Aziz, S. Mardiyanto, B. Hendrajaya, and Munir, "A model of geographic information system using graph clustering methods," in *2017 3rd International Conference on Science in Information Technology (ICSTech)*, Oct 2017, pp. 727–731.
- [27] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 267–279, sep 2014. [Online]. Available: <https://doi.org/10.1109/tetc.2014.2330519>
- [28] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [29] E. C. Delmelle, "Mapping the dna of urban neighborhoods: Clustering longitudinal sequences of neighborhood socioeconomic change," *Annals of the American Association of Geographers*, vol. 106, no. 1, pp. 36–56, 2016.
- [30] ——, "Differentiating pathways of neighborhood change in 50 u.s. metropolitan areas," *Environment and Planning A: Economy and Space*, vol. 49, no. 10, pp. 2402–2424, 2017.
- [31] C. Ling and E. C. Delmelle, "Classifying multidimensional trajectories of neighbourhood change: a self-organizing map and k-means approach," *Annals of GIS*, vol. 22, no. 3, pp. 173–186, 2016.
- [32] J. Han, M. Kamber, and A. K. Tung, "Spatial clustering methods in data mining," *Geographic data mining and knowledge discovery*, pp. 188–217, 2001.
- [33] M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco, "Clustgeo: an r package for hierarchical clustering with spatial constraints," *Computational Statistics*, pp. 1–24, 2017.
- [34] S. Soor, A. Challa, S. Danda, B. D. Sagar, and L. Najman, "Extending k-means to preserve spatial connectivity," 2018.
- [35] P. Soille and L. Najman, "On morphological hierarchical representations for image processing and spatial data clustering," in *Applications of Discrete Geometry and Mathematical Morphology*. Springer, 2012, pp. 43–67.
- [36] J. Cousty, G. Bertrand, L. Najman, and M. Couprise, "Watershed cuts: Minimum spanning forests and the drop of water principle," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1362–1374, Aug 2009.
- [37] M. D. Dias, M. R. Mansour, F. Dias, F. Petronetto, C. T. Silva, and L. G. Nonato, "A hierarchical network simplification via non-negative matrix factorization," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Oct 2017, pp. 119–126.
- [38] M. Monmonier, "Strategies for the visualization of geographic time-series data," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 27, no. 1, pp. 30–45, 1990.
- [39] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization: an analytical review," *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 503–541, 2003.
- [40] N. Ferreira, "Visual analytics techniques for exploration of spatiotemporal data," Ph.D. dissertation, Polytechnic Institute of New York University, 2015.
- [41] Y. Zheng, W. Wu, Y. Chen, H. Qu, and L. M. Ni, "Visual analytics in urban computing: An overview," *IEEE Transactions on Big Data*, vol. 2, no. 3, pp. 276–296, Sept 2016.
- [42] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: foundations, techniques, and applications*. AK Peters/CRC Press, 2015.
- [43] G. Andrienko, N. Andrienko, H. Schumann, and C. Tominski, "Visualization of trajectory attributes in space-time cube and trajectory wall," in *Cartography from Pole to Pole*. Springer, 2014, pp. 157–163.
- [44] C. Tominski and H.-J. Schulz, "The Great Wall of Space-Time," in *Vision, Modeling and Visualization*, M. Goesele, T. Grosch, H. Theisel, K. Toennies, and B. Preim, Eds. The Eurographics Association, 2012.
- [45] S. Buschmann, M. Trapp, and J. Döllner, "Real-time animated visualization of massive air-traffic trajectories," in *Cyberworlds (CW), 2014 International Conference on*. IEEE, 2014, pp. 174–181.
- [46] D. Seebacher, J. Häußler, M. Hundt, M. Stein, H. Müller, U. Engelke, and D. Keim, "Visual analysis of spatio-temporal event predictions: Investigating the spread dynamics of invasive species," in *2017 IEEE Visualization Conference (VIS)*, 2017.
- [47] G. Andrienko, N. Andrienko, G. Fuchs, and J. Wood, "Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 9, pp. 2120–2136, Sept 2017.
- [48] N. Ferreira, M. Lage, H. Doraiswamy, H. Vo, L. Wilson, H. Werner, M. Park, and C. Silva, "Urbane: A 3d framework to support data driven decision making in urban development," in *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. IEEE, 2015, pp. 97–104.
- [49] M. Li, Z. Bao, T. Sellis, S. Yan, and R. Zhang, "Homeseeker: A visual analytics system of real estate data," *Journal of Visual Languages & Computing*, vol. 45, pp. 1 – 16, 2018.
- [50] J. Johansson, P. Ljung, M. Jern, and M. Cooper, "Revealing structure within clustered parallel coordinates displays," in *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. IEEE, 2005, pp. 125–132.
- [51] D. Guo, J. Chen, A. M. MacEachren, and K. Liao, "A visualization system for space-time and multivariate patterns (vis-stamp)," *IEEE transactions on visualization and computer graphics*, vol. 12, no. 6, pp. 1461–1474, 2006.
- [52] F. Pedregosa, 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, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [53] M. Harrower and C. A. Brewer, "Colorbrewer.org: An online tool for selecting colour schemes for maps," *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, 2003.
- [54] J. L. Hintze and R. D. Nelson, "Violin plots: a box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.
- [55] J. Hwang and R. J. Sampson, "Divergent pathways of gentrification: Racial inequality and the social order of renewal in chicago neighborhoods," *American Sociological Review*, vol. 79, no. 4, pp. 726–751, 2014.
- [56] J. Bian, D. Tian, Y. Tang, and D. Tao, "A survey on trajectory clustering analysis," *arXiv preprint arXiv:1802.06971*, 2018.