

Documentation

This prototype was described in a methodological paper [available here](#).

If you just want to know how to play with the interface, [skip ahead to the prototype interface documentation](#), or [the video guides](#).

The objective of this page is to briefly overview the data and methods used and explain how to interpret the results illustrated in the interface. No expert knowledge is required beyond a basic grasp of statistics (specifically, [boxplots](#)).

Overview and objectives

The objective of this prototype is to **explore census data**. We employ data clustering to simplify the information, which is fairly commonplace for this type of analysis. In this scenario, a cluster is a set of geographic regions that are more similar to themselves than to the other regions.

The objectives of this project are:

- Make evident what clusters are present in a geographic region (city, metro, etc).
- Understand how these clusters evolved over time, both spatially and in content.

Data

We used data from [NHGIS](#), for the USA, using census years 1970, 1980, 1990, 2000, and 2010, and from [CHASS](#) for Canada, census years 1971, 1981, 1991, 2001, and 2011. Neither allows redistribution of the data, which is why our prototype does not export the results in a downloadable format. In both cases, we used **census tracts** as the basic geographic unit.

When dealing with census data, the main challenge is the lack of uniformity on the measurements. Everything changes, from the shapes of the census tracts, to what variables are measured and what each measured variable actually means. **Here the spatial changes do not matter, because the method was developed specifically to deal with different geographic regions**, but we still had to find the correspondences between the variables across different years. Since the objective of this work is to demonstrate that the prototype works, not do a deep analysis, we only matched a subset of the available variables, just enough for it to be useful, we hope.

More importantly, we did not consider each measurement as a separate entity, but we grouped them into aspects (for lack of a better term). Each aspect is then characterised by a probability distribution of its values. For instance, the aspect "Race" in the USA includes White; Black; Indian, eskimo and aleut; Asian, Hawaiian, other pacific islander; and other.

Since each aspect is a probability distribution (a histogram, to be precise), we can compare any two regions without further concerns. There are many ways to compare two probability distributions, but we adopted a simple euclidean distance between the vectors, because it gave reasonable results and it is fast to compute.

ATTENTION - INCOME: While some aspects match perfectly between census years (age, for one), income does not. We adjusted the values using publicly available inflation data and considered rather large ranges, but the results are not absolutely reliable. They are not wrong either, as far as we know, but should be **interpreted with extreme caution**. For instance, there is a "baseline" oscillation in all cities of both countries, most likely caused by a slight mismatch in the values, that should not be confused with actual change in the data. Additionally, accurately adjusting historical monetary values is a challenging problem by itself, and out of the scope here.

For the details of which census field went where, [click here](#).

For the USA, we used: Education level, Family Income, Marital status, Occupation, and Race.

For Canada: Age, Education level, Home language, Household income, Marital status, Occupation, Place of birth, and Religion

Clustering method

Data representation and geographic relationships

One of the distinctive features of this work is that we use a [graph](#) to represent the census tracts. Each CT is represented as node, with probability distributions of each aspect, and we place edges between the nodes of CTs that share geographic boundaries. Since we also have temporal evolution, we connect the graphs of each year by linking regions that spatially overlap, leading to a single graph (with possibly several connected components). That graph is the data used for the whole method, and its inherent flexibility is what allows us to compare different geographic regions over time (no need for harmonized regions into a single CT division). In other words, **the data is not a matrix**.

Watershed cuts and hierarchies

The actual clustering method is based on the [Watershed cuts](#) algorithm for graphs, coupled with a [greedy edge matching method](#), leading to a result that is not a single clustering, but a hierarchy. This hierarchical result is what allows the easy change in the numbers of clusters, without the need to redo the computational process. We limit the number of clusters to eight because we cannot visually display more than that using colours.

Aspect Weights

The user can assign different weights to each aspect. By default, their values are equal to one. Use this parameter to increase/decrease the importance of the *differences* in each aspect. For instance, using "Race" with weight one and "Income" with weight two means that a difference in race needs to have twice the value as one in income to be equally distinctive, leading to a clustering that mostly considers differences in Income, using Race as a "tie breaker" of sorts.

Augmented edges

One slight caveat of Watershed cuts is that each cluster is composed of a connected subgraph of the original graph. In other words, if we have two separate sets of census tracts, with identical features (aspects), they will form two distinct clusters, which is not ideal for this specific application.

To overcome this, we augment the original graph with two edges from a [K-NN graph](#), generated without considering the geographic information. In this example, these edges will connect the two previously disconnected regions, assigning both regions to the same cluster.

Prototype interface

Important

This tool is still in the *prototype* phase. That means it should work, using Chrome (Safari might crash due to heavy memory usage), but it will crash sometimes. We tried to mitigate this, by breaking NYC into boroughs for instance, but problems are still possible, especially when analysing larger regions (LA, Greater Toronto, Vancouver, San Francisco).

The results are computed on request *on the server*, then saved and reused if the same configuration is requested again. The default configuration is probably already saved for every region, so the results are generally fast to load. However, more "exotic" parameters (custom weights, removing variables) will need to be processed, and that might take a while (around 5 minutes for Chicago/LA/etc). When that happens, the page might crash, or just get stuck with the "loading" message forever. Just refresh the page after a few minutes and try again, the server will have processed the request and it should load quickly. I do apologise for this.

Trajectories

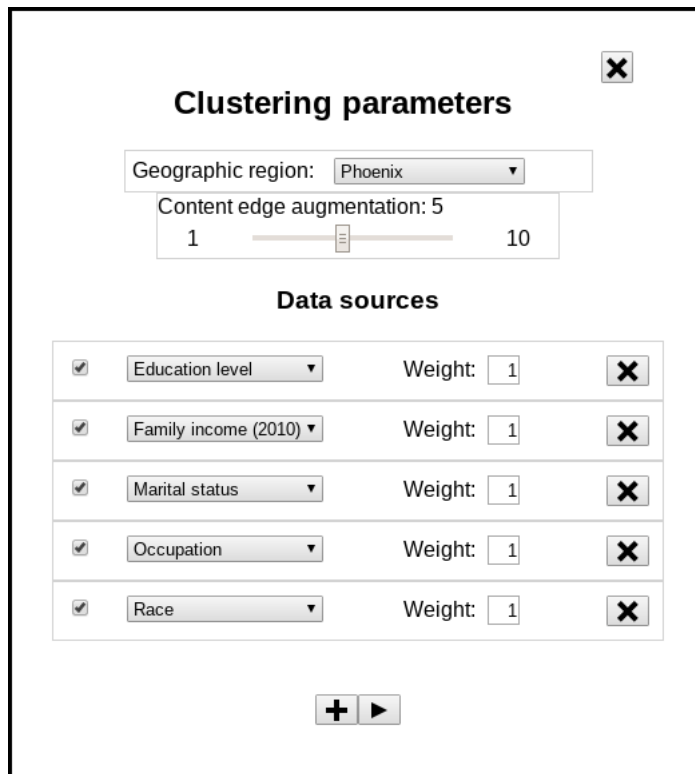
For each little piece of the region under analysis we can associate the corresponding clusters over time. We call this sequence a *trajectory*.

The interface

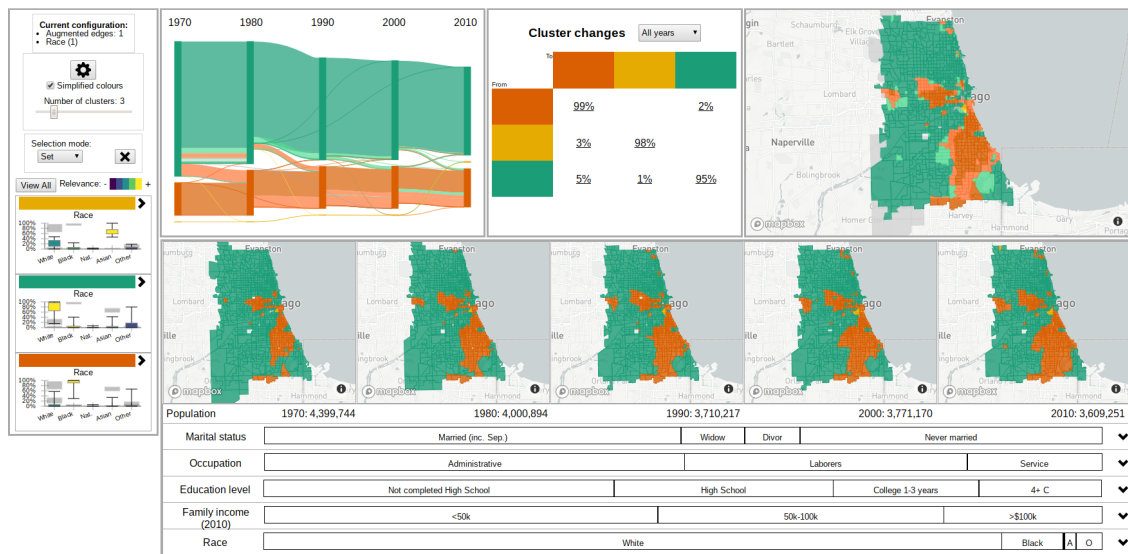
When loaded, the interface will show a default configuration for an arbitrary region.

The bottom part of the left panel contains an overview of what composes each cluster, where you can click on the arrow to see a detailed composition of each cluster or click on the 'view all' button to see all the features that composed all of the clusters at the same time. You can also click on the coloured bar besides the arrow to have an overview of that specific cluster.

The region and the clustering parameters can be altered by clicking on the gear located on the top left side. This panels allows the user to add or remove other aspects, and set the desired weights for each feature. To run the new configuration, just click on the run button.



After the page is loaded, you can see the current configuration shown on the top left. Below you can choose if you want "simplified colours", as shown in the section [Colours, simplified?](#), and adjust the number of clusters.



Clicking on any coloured region on the left and top panels will select the corresponding trajectories and update the bottom panel to show details about these regions. How these selections are made can be configured using the selection mode combo box.

The default 'set' option will replace any current selection with whatever trajectories are chosen next. 'Add' will add the new trajectories to the currently selected, and 'remove' will remove from it. The 'x' button on this panel clear the current selection. The bottom part of the interface will update the details accordingly.

For example, if you click on the green bar of the year 2010, you will be able to see all the trajectories that belong to the green cluster in that year. By clicking on the individual edges, you can select a single trajectory. Alternatively, trajectories can be manipulated by clicking on regions on the top right map or on the cluster changes matrix.

The other panels on the top displays how the clusters evolved over time (the different trajectories). Here, you can also click on different parts of said trajectories to further analyze them. For example, choosing Chicago, Race with weight 1, we can clearly see three main colours and trajectories that go from the year 1970 to the year 2010. Now, this is the first general overview you get, but you may be interested in something more specific, like what regions remained the same over the years by clicking on the coloured lines between the bars representing the years.

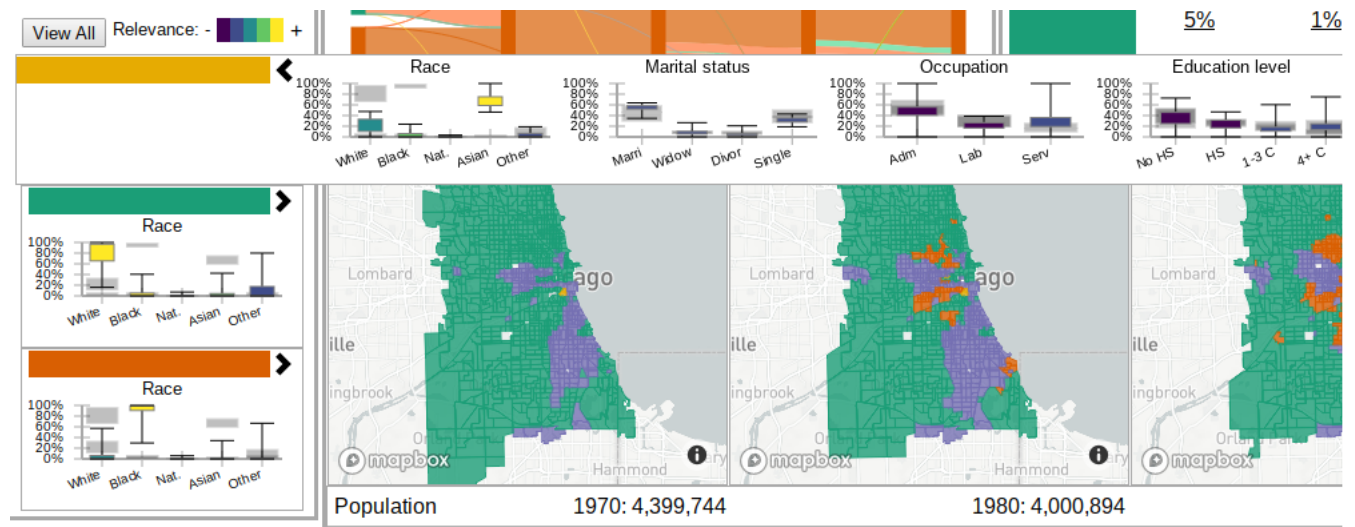
On the middle top panel, the percentage of changes between them is shown, and an overview map, illustrating the cluster evolution for the whole considered period.

The bottom part of the interface contains the details for the selected regions, with maps illustrating which CTs are used, since they can be different over time, and plot comparing the selected region against the whole.

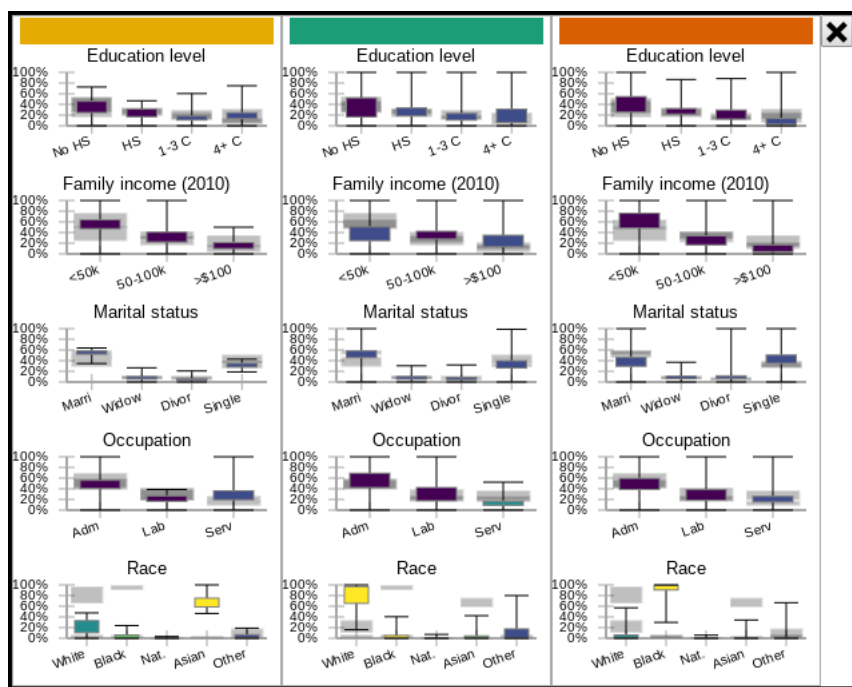
Overview of cluster composition

The left panel contains one boxplot per cluster, illustrating the aspect that better characterises that cluster. **It was not necessarily used for the clustering**, but it is the one with less overlap between the 25% and 75% quantiles. We define relevance as the distance between the interquartile ranges, meaning that an aspect is relevant if those specific values are not present in the other classes. The relevance of an aspect is the maximum relevance of its composing variables.

The relevance is represented by the colour of the interquartile box. The whiskers represent the maximum and minimum. To allow for an easier comparison, the quantiles of the other clusters are represented in gray. The user can expand that panel and see all the aspects of that cluster, or all of them at the same time.



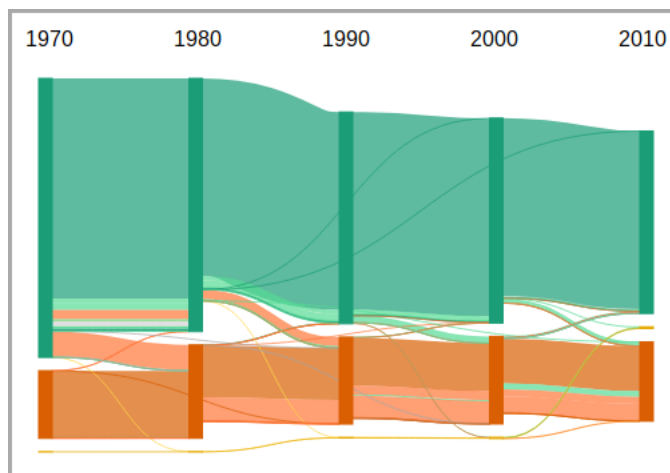
In this image, we can see that the yellow cluster corresponds a concentration (~70%) of people that identify as "Asian, Hawaiian, other pacific islander" and that this is unique to this cluster. The other boxplots are ordered by relevance, left to right, but there are no more relevant aspects, with all "boxes" almost in the same position. This means that, while Race is a relevant aspect for this clustering, the other variables are not, and have similar values across the different clusters. This can be confirmed by clicking on "view all":



Trajectories

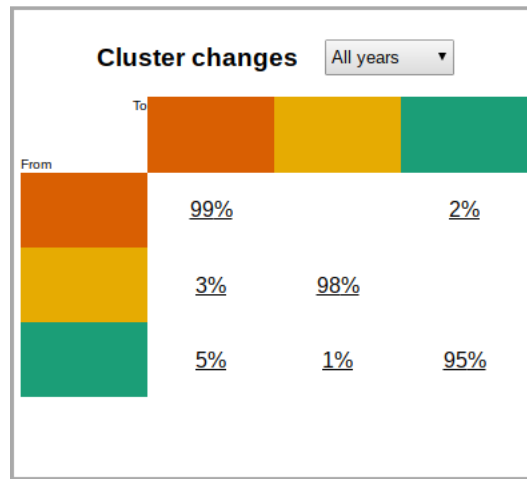
We defined a *trajectory* as the sequence of clusters assigned to a specific region. This allows for the study of the **changes** in the demographic groups.

The trajectories are depicted using a [Sankey diagram](#), where the width is proportional to the population involved.



We can see that the green cluster contains more people, followed by the orange, and that the yellow cluster is small. The clusters are mostly stable, with a large trajectory composed by the green cluster on all times. A better explanation on the colour choices is [here](#).

Another way of identifying change is using the cluster changes matrix, which indicates the percentage of the population of a given cluster that was later classified into another cluster.



In this image, we can see that, considering all years, 95% of the population in the green cluster stayed green, while 5% became orange and 1% yellow. The total is not 100% due to rounding. The panel on the top right depicts the geographic positions of the trajectories, considering all years (the other maps show the clustering results for specific years).

More importantly, **the user can select (multiple) trajectories**. The "Selection mode" control on the left panel allows the user to Set the selection, Add to the selection, or remove from it. The button next to it clears the selection. The bottom part of the interface will then show details for the selected trajectories (or for the whole, if nothing is selected).

Colours, simplified?

In this work, we use colours to represent the clusters and the trajectories. That is the reason the number of clusters is limited to 8, we cannot really tell colours apart beyond that. However, we have considerably more trajectories, so we are left with a conundrum: Do we accurately represent each different trajectory with a different colour or do we use small set of colours and paint different, but similar, trajectories in the same colour? Well, instead of "solving" this problem, we made both options available, let the users decide.

Simplified colours: When this is active, the colours on the interface are determined by the rule:

- If a region belongs to the same cluster on all years, use the colour of that cluster,
- If a region belongs to a given cluster for most of the years (simple majority), use a "lighter" version of the colour of the cluster,
- Gray otherwise.

Non-simplified colours: When this is **not** active, the colours on the interface are determined as the average of the colours of the involved clusters.

Characterization of the selected trajectories

The bottom part of the interface corresponds to the *details*, either for the whole region under analysis or the selected trajectories.

Following the Chicago example, this is what happens when the user selects all trajectories involving the yellow cluster (by clicking on the yellow bar over the boxplot)

