Exploratory analysis of clustering Swiss population using Self-Organizing Map

Axelle Bersier, 2022, Université of Lausanne, Faculty of Geosciences.
Keywords: Machine Learning, Self-Organizing Map, cluster, population dynamic

Abstract

This paper describes a method to cluster socio-economic data using the Self-Organizing Map algorithm. It defines the relevant data and the number of clusters and compare the results with a K-means. The purpose of this study is to specify the characterization of the clusters defined by the census data of Switzerland in 2020. As a result, the analysis shows that dense city centers have their own dynamic and that clusters are spatialy distributed around them.

1. Introduction

The results of the Swiss votes often illustrate the cleavage between rural and urban municipalities. Those outcomes are mainly due to different lifestyles and socio-economics factors. However, how could we define those patterns and what are their characterizations? Using unsupervised machine learning clustering such as self-organizing map (SOM) helps to understand high dimensional data behaviors. SOMs utilizes an artificial neural network algorithm developed by Kohonen (1982). It contributes to understand socio-economic distribution and patterns by computing a high dimensional dataset onto lower dimensional output space. These algorithm has two interesting properties, one is to output a two-dimensional grid representing each feature as a heat map, and the other result is to cluster data by grouping similar information together. In this study, we analyzed census data of the Swiss population in 2020, that focus mainly on the household's properties and is grouped by municipalities (OFS, 2020). This paper is focusing on describing a method that could be used to cluster socio-economics information and the way of characterizing different groups.

2. Methodology

In this paper the focus is in on providing an approach of clustering some socio-economics information and assess assumptions on the results of the clusters of the Swiss population census data in 2020. To deal with that, at first the data have been extracted, then standardized. This standardization as been done with the Z-score method calculated as : $z = (x - mean)/variance$. After the data selection has been refined with a correlation matrix analysis and then the number of cluster has been chosen trough an elbow analysis and a silhouette score. The next step is to fill the data into the unsupervised algorithm of SOM and then to compare it with the K-means algorithm.K-means clustering and Self-Organizing Maps (SOMs) are both unsupervised machine learning algorithms that are used for clustering data points. K-means clustering is an iterative algorithm, that defines centroids by minimizing the Euclidean distance between the closest centroid and the data point. On the other hand, SOMs are neural network model trained by adjusting the weights of the network's neuron in a way that preserves the topological structure of the data. The final step is to visualize the resulting clusters with a box plot, an heat map for the SOM algorithm and a projection of the data into the Swiss map (Whelan et al., 2010),(figure 1).

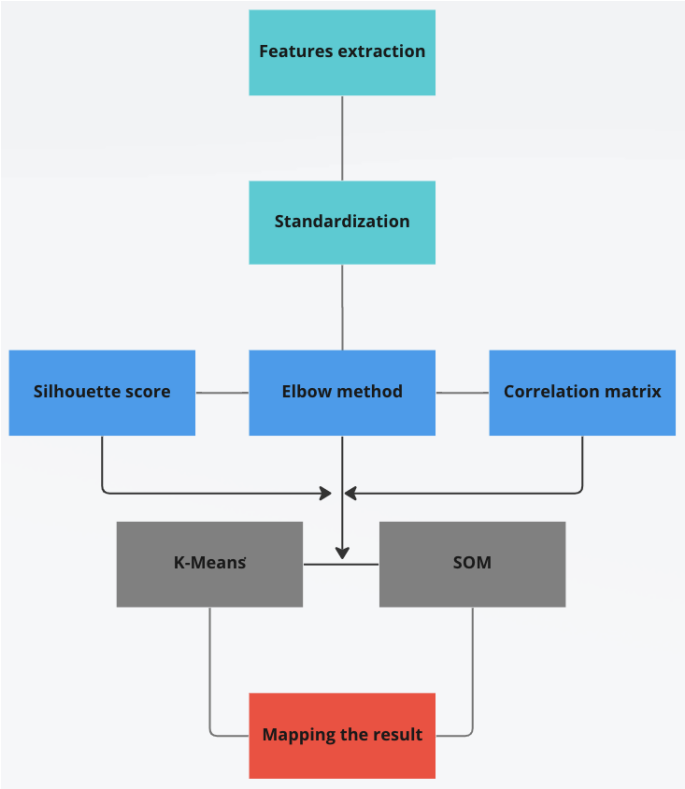

Figure 1. Workflow description

Name	Description
D_pop	Density of the population
N_communes	Name of the municipalities in 2020
M_home	Mean size of people in the household
Little_house	Percentage of housing with 3-4
N_foreigners	Number of people with a foreigner nationality
Natural_growth	Difference between birth and death
N_cinema	Number of movie theatre
N_wedding	Number of weddings
N_indivhome	Percentage of individual houses
N_nlog	New housing built
N_death	Number of deaths
Prop_fam	Proportion of family of five or more people

Figure 2. data description

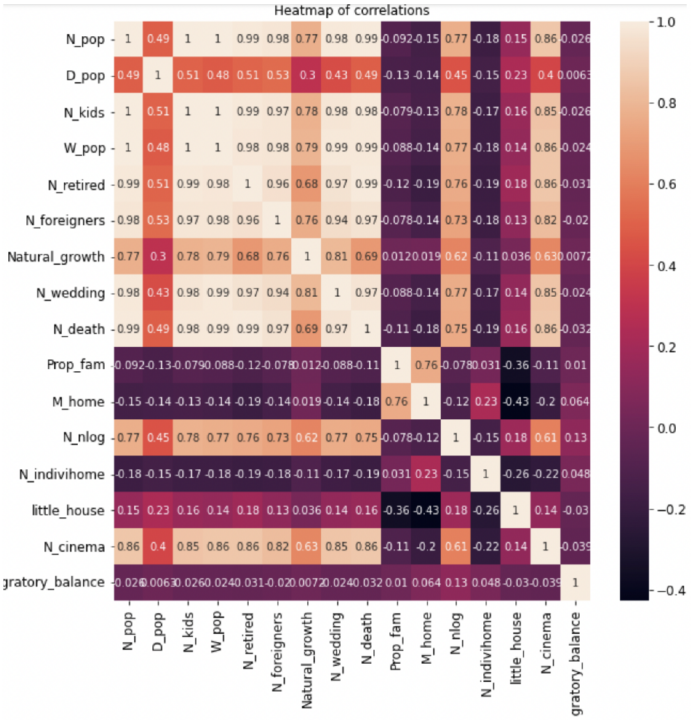

Figure 3. correlation matrix

3. Data and variables

11 features divided by municipalities have been chosen for the analysis. Those are the following described in figure 2. The main themes of those variables are households' characterization and type of houses. Those are data from the year 2020, because it was the most recent year with available data. The matrix feed into the algorithms is a 2198 (municipalities) * 10 features. The municipalities names are not included in the quantitative analysis.

4. Results

4.1. Statistical description of the features

During the data processing step, the curse of dimensionality is analyzed.The curse of dimensionality refers to problems that arise when working with a too high-dimensional dataset. The volume of the space increases, the variable is more sparce and therefore the accuracy of the results can be biased. This is the reason why it is interesting to evaluate a correlation matrix in this case, or a princical composant analysis, to have an idea of the dependencies of the variables and to define which information is redundant. The figure 3 illustrates the correlation of the extracted data of this analysis. The lighter colored cells represent the more correlated data. Hence the variables describing the number of the population, the number of kids, the percentage of the working population, the number of retired, the number of foreigners, the number of wedding and the number of deaths seem to be very correlated. They all are information about the age of the population living in the municipality. In order to avoid to have an overrepresentation of structure of the population, some of those variables have not be fitted into the clustering algorithms (figure 2).

4.2. Parameter estimation

In cluster analysis, the elbow method and the silhouette coefficient are heuristic analysis that help define the best number of clusters for an analysis. The elbow method consists of finding the best compromise between the number of clusters, in the x axis, anf the Within Cluster Sum of Squares. In this case, figure 4, around 5, but this choice of 5 clusters could be discussed.

To support the argument of choosing 5 clusters, the Silhouette coefficient is calculated. This method calculates the mean of intra-cluster distance (a) over the mean nearest-cluster distance (b) such as $(b-a)/max(a,b)$. The plot of

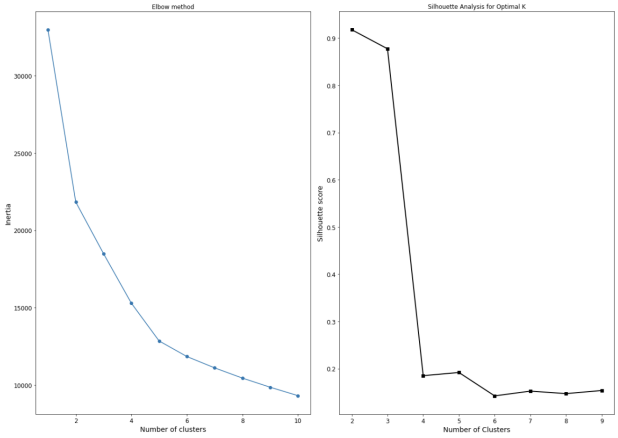

Figure 4. estimation of the numeber of clusters

this result over the number of clusters shows us that using 5 clusters, is a better than 4 or 6 in term of the distance between a sample and the nearest cluster.

Even though those graphics justify choosing 5 clusters, over 4 other 6, it could be also arguable to define 3 clusters. Another point to notice is that those information was done over the K-means results and generalizes over the SOMs algorithm by assuming that it would give the same result.

As indication, the main difference between K-means and SOM clustering is that for K-means the clusters are formed through centroids, that are independent, whereas with SOM the clusters are formed geometrically.

4.3. Visualization of the results with SOM

The mapping of the clusters (figure 5) illustrates that the center of big cities, such as Geneva, Zurich, Bern have very different characteristic from other municipalities (cluster 5). Those cities are followed by a group of cities with the same tendency, but less extreme. Those are Winterthur, Lausanne, Basel and Luzern (cluster 4). Those are also followed then, by more little cities such as Bellinzona, Sion, Neuchâtel. Those clusters, with 3 different levels are represented by more foreigners, more birth than death (positive natural growth), more weddings, more culture, more new built housing than in the other surroundings areas. Logically living in the city center means having less chance to have an individual house, it doesn't mean living in a small house. It is also characterized by a small number of big families and also a small number of people by housing. In the clusters 1 and 2, there is less variance and many municipalities. The main difference between those two clusters are the proportion of family with 5 or more children that also have a tendency to live in bigger houses(figure 6,7,8,9,10).

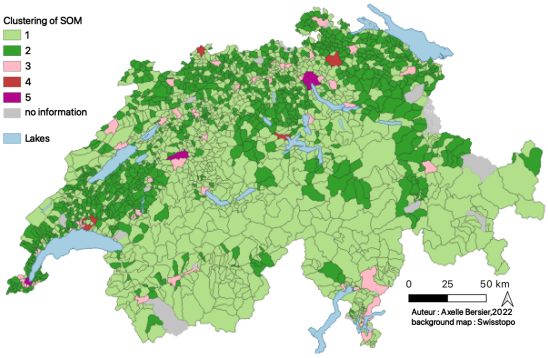

Figure 5. Map of the clusters with SOM

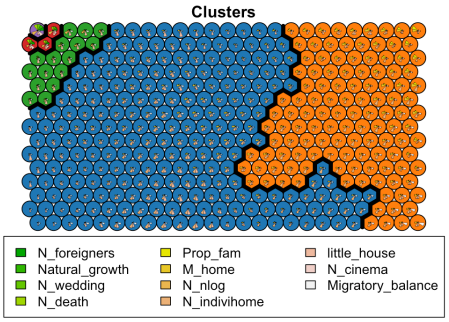

Figure 6. Spatial representation of the data in the output space

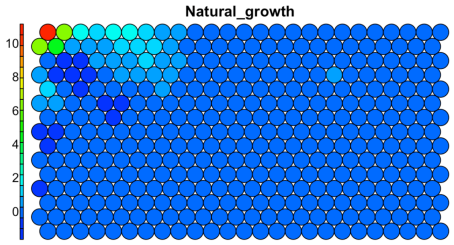

Figure 7. heat map of the natural growth

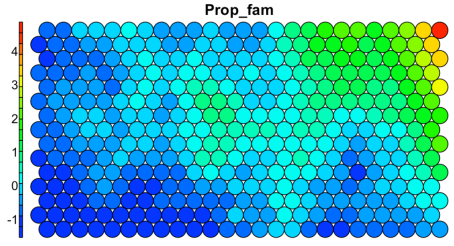

Figure 8. heat map of the percentage of families with 5 or more people

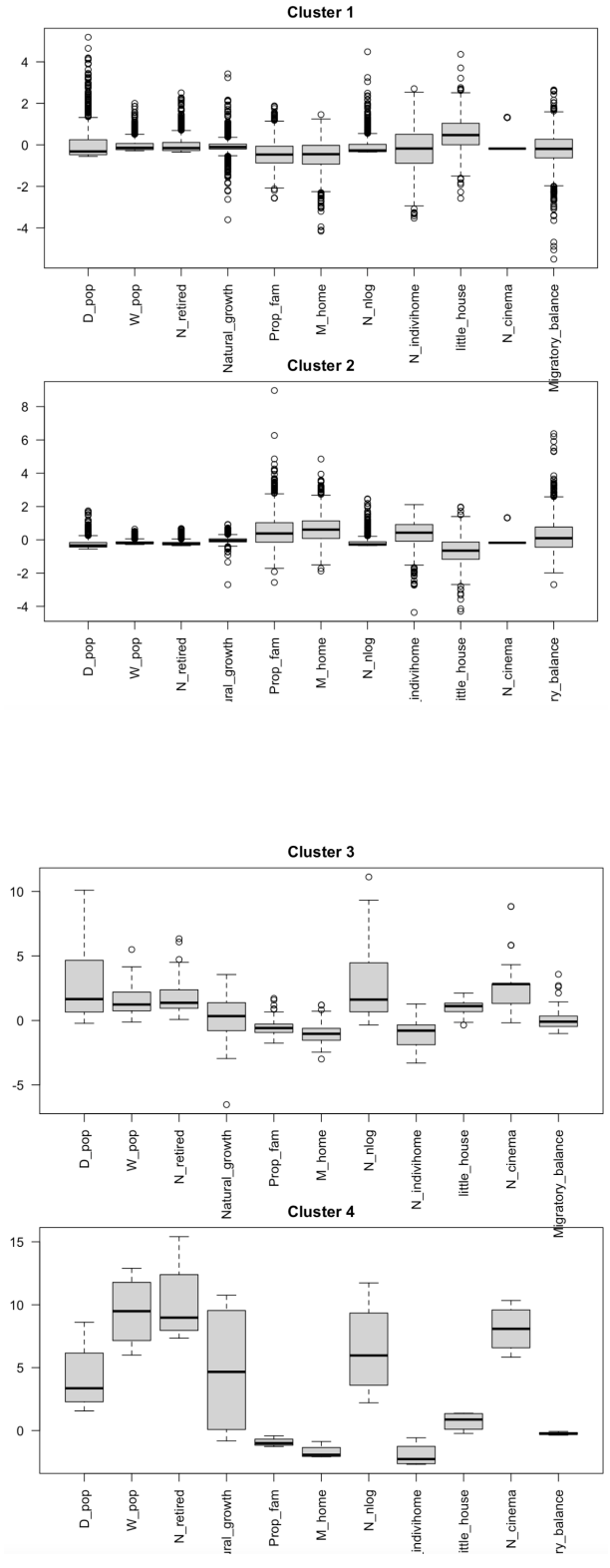

Figure 9. Information of the clusters

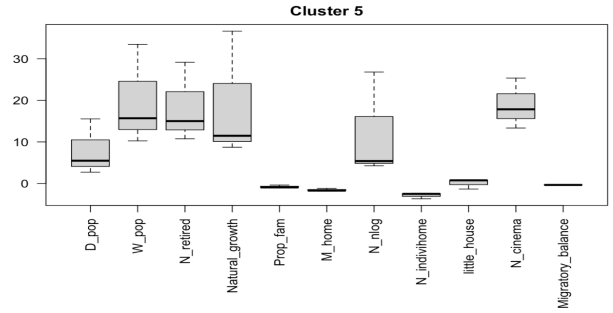

Figure 10. Information of the clusters

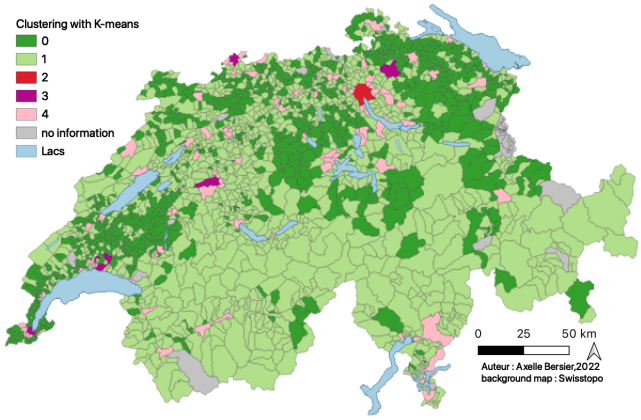

Figure 11. Map of the clusters with K-means

4.4. Comparison of the results with K-means

In this case, the dataset was not devide into a training set, a test set and a validation set, because only one year was studied and all the municipalities needed to be taken into account in the same algorith. This is the reason why a K-means clustering is also presented to check if the cluster are very different and how could it be analyzed in another way. The conclusion of this validation method via the K-means is that it seems to be similar (figure 10).

5. Discussion and conclusion

The neural network algorithm using SOM is interesting to describe clusters because it enables a better understanding of the patterns through heat map analysis. The reduction of variables was judicious to analyze the difference between the clusters 1 and 2, that had fewer extreme differences. The points that could be discussed and improved in this analysis are the number of clusters that could also be 3, also, the method used to standardize the variables, because there are not only numbers between -1 and 1 after the standardization.

Furthermore, if the goal is to study one specific question, the data chosen are important. Having nice clusters but no representative or too corelated data won't give good results.

In conclusion this analysis describes a method to cluster socio-economics information that could be expand to different problems such as the difference of life quality between the urban and the rural areas, dynamic of the population in different areas or urban sprawl. Those problematics could also have a time axis and those cluster could be analyze at different years.

Code link : click there or copy this url: https://github.com/axellebersier/2022_ML_EES/blob/main/project/datadescrp.ipynb

Data link : click there or copy this url: https://github.com/axellebersier/2022_ML_EES/blob/main/project/data.csv

Source of the data : Office Fédéral de la Statistique. (2020, mise à jour en 2022). *Atlas statistique de la Suisse*. https://www.atlas.bfs.admin.ch/

6. References

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1), 59-69.

Office Fédéral de la Statistique. (2020, mise à jour en 2022). *Atlas statistique de la Suisse*. https://www.atlas.bfs.admin.ch/

Tonini, M., Rabelo, M., Silvestri, N. (2022). Change dynamics of agricultural land systems in Europe: an innovative approach based on an unsupervised clustering technique. *Conference proceeding*, iEMSs, Bruxelles

Whelan, C. T., Lucchini, M., Pisati, M., Maître, B. (2010). Understanding the socio-economic distribution of multiple deprivation: An application of self-organising maps. *Research in Social Stratification and Mobility*, 28(3), 325-342.