Clustering of demographic structures: Observations on US counties for 2000 and 2010 and on Brazilian municipalities for 2010

Simona Korenjak-Černe*
University of Ljubljana, Faculty of Economics, Ljubljana, Slovenia - simona.cerne@ef.uni-lj.si

Vladimir Batagelj
IMFM - Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia - vladimir.batagelj@fmf.uni-lj.si

Nataša Kejžar
University of Ljubljana, Faculty of Medicine, Ljubljana, Slovenia - natasa.kejzar@mf.uni-lj.si

Abstract

Demographic structures offer important information for demographers, sociologists, politicians, and other experts. Clustering methods allow us to systematically arrange and observe a large number of such structures. We present clustering results of age-sex distributions on US counties for the years 2000 and 2010. The clustering results are obtained by weighted agglomerative clustering method, adapted for modal valued symbolic data (i.e. demographic structures). Systematic arrangement of changes of demographic structures from 2000 to 2010 can also be observed with clustering method. We expect to detect clusters with similar structural changes from 2000 to 2010 and view these changes further in light of the ethnicity component. Urban and rural areas are not separated only by location, they also describe geographic area with demographic and many other characteristics connected with other fields (like agriculture, natural resources etc.). Finding clusters of Brazilian municipalities with similar population distributions by geographic area and age groups can offer additional insight into this field. We will discuss obtained clustering results with additional demographical characteristics for Brazilian municipalities data from 2010.

Keywords: weighted agglomerative hierarchical clustering; demographic structures; population pyramid; symbolic object.

1. Introduction

Socioeconomic data are often presented in an aggregated way. Demographic structures, i.e. distributions of population by age and sex or by some other variables, are very important for studying not only demographic characteristics, but also for observing other influences like social, and political situations in time on the observed region. When dealing with many observed regions, clustering results can provide better (more concise and systematic) view on all that. In this presentation, we want to observe clustering results on US counties and on Brazilian municipalities, obtained with such a clustering method which would preserve as much information about the structures as possible and at the same time provide meaningful clusters' representatives.

Although there are some approaches based on representations with distributions (like compositional data, symbolic objects), most of the analytical methods are still based on classical data representations, where each observed variable is represented only with a single value. In the case of a demographic structure, a lot of information is lost when the structure is represented in a classical way. Such information is also the population size, which therefore can not be properly preserved or included in the analyses. Due to this limitations in our analysis the clustering method based on data, described with discrete distributions is used. Such data representations are a special kind of so called modal multi valued data in symbolic data analysis (Billard, L., & Diday, E. (2006); Noirhomme-Fraiture, M., & Brito, P. (2011); Brito, P. (2014)). They preserve structures with included information on the population size of considered variables. Appropriate adaptation with included weights in the clustering methods enables to preserve real structures in the clusters with population sizes. One of the most important advantages of this clustering approach are meaningful

clusters' representatives with real demographic structures, which offer more useful and easier interpretable results.
The above described weighted clustering method has already been used to study clusters of the world's countries and their changes over 5-year periods from 1996 to 2006 (Korenjak-Černe, S., Kejžar,N., & Batagelj, V. (2014)). We presented some results on US counties also at IFCS conference in Tilburg in 2013. For this contribution we extended our study on US counties from 2000 and 2010, where we considered also the distributions by ethnicity. Additionally, some preliminary results on data of Brazilian municipalities from 2010 were included, where we were especially interested in observing results based on structures by geographic areas, i.e. distribution of the population of urban and rural area by age groups, which might be interesting and also useful for experts, such as agronomists and other experts in agriculture. Due to the space limitation we present here only some of the obtained results.

2. Data

The datasets of US counties consist of population structures by age and sex for the years 2000 and 2010. The data for 2000 with 3 219 US counties in which the population is divided into 18 five-years age groups were obtained from US Census 2000: `http://factfinder.census.gov/`, described in Census 2000 Summary File 1. The data for 2010 with 3 221 US counties in which the population is divided into 23 age groups, were obtained from US Census 2010 `http://www.census.gov/2010census/data/`.
For each gender, the population is distributed over age groups, therefore, each county is represented with two symbolic variables (one for each sex), described with (absolute) frequency distributions over age groups. This representation is a special kind of more complex data representation named *symbolic object* — SO (Billard, L., & Diday, E. (2006)).
For an illustration of the SO representation of a county, let us split the population of the county "New York County, New York" with $n_1 = n_{men} = 729\,534$ men and $n_2 = n_{women} = 807\,661$ women into 4 twenty-years age groups 0-19, 20-39, 40-59, 60+. The corresponding frequency distributions $\mathbf{f}_{NYC,i}$, $i \in \{\text{men, women}\}$, are

$$\mathbf{f}_{NYC,men} = [148\,355, \underset{\substack{\uparrow \\ \text{men}}}{288\,029}, 191\,199, 101\,951] \quad \text{and} \quad \mathbf{f}_{NYC,women} = [145\,525, \underset{\substack{\uparrow \\ \text{women}}}{302\,589}, 211\,035, 148\,512]$$

Technically, we store each distribution $\mathbf{f}_{xi}$ as a couple $(n_{xi}, \mathbf{p}_{xi})$, where n_{xi} is the total number of all values in the distribution of a variable divided into k_i categories and $\mathbf{p}_{xi}$ is the corresponding relative frequency (probability) distribution:

$$n_{xi} = \sum_{j=1}^{k_i} f_{xij}, \quad \mathbf{p}_{xi} = \frac{1}{n_{xi}} \mathbf{f}_{xi},$$

and the corresponding SO is

$$\begin{aligned} X_{\text{NYC,NY}} &= [\mathbf{f}_{NYC,men}, \mathbf{f}_{NYC,women}] \\ &\quad \Big[\big(729\,534, [0.2034, 0.3948, 0.2621, 0.1397]\big), \big(807\,661, [0.1802, 0.3746, 0.2613, 0.1839]\big)\Big]. \end{aligned}$$

In general, the number of categories in the distributions are not necessary equal for all variables and we could add some other variables with a different number of categories (education, marital status etc.) under the same general description, which allows us to use the same clustering method.
Since this representation preserves the same description of a structure as known from descriptive statistics, and it is preserved through clustering process, it enables the usage of the same graphical presentation with a population pyramid for the clusters representatives.

The other dataset consists of age-sex distributions of the population of 5 570 Brazilian municipalities for 2010, obtained from Brazilian Institute of Geography and Statistics (IBGE), Census 2010: `http://www.cidades.ibge.gov.br/xtras/home.php`.

Besides the age-sex distributions data, also data on ethnicity are available for all US counties in 2010 and some additional data for Brazilian municipalities, such as the share of rural/urban areas. These data (vari-

ables) help us putting the clustering results into a broader demo-geographical context.

3. Weighted agglomerative clustering method

For the analysis we use clustering procedure, implemented in R within the `clamix` package available at the R-forge (Batagelj, V., & Kejžar, N. (2010)), adapted with weights. This weighted clustering method was already used to study clusters of the world's countries and their changes over 5-year periods from 1996 to 2006 (Korenjak-Černe, S., Kejžar, N., & Batagelj, V. (2014)). It is based on the definition of the clustering as an optimization problem, where the value of the criterion function is minimized. The criterion function is defined as a sum of the dissimilarities between each unit in the cluster (i.e. in our case county or municipality) and the cluster's representative. An agglomerative hierarchical method (Anderberg,M. R. (1973)) is adapted for the data representation with discrete distributions (Batagelj et al. (2015)). In this case, cluster's representative is determined with a weighted average of relative distributions over all units inside the cluster and so it represents the age-sex distribution of the whole population in the cluster. This agglomerative hierarchical clustering method is a generalization of the well known Ward's agglomerative hierarchical method (Ward, J. H. (1963); Späth, H. (1977)).

4. Some of the results

4. 1. Clustering US counties

Here we present some of the clustering results, obtained on US counties from 2000 where the population is divided into 18 five-years age groups. Weighted clustering of age-sex distributions of 3 219 US counties from 2000 revealed four main clusters, presented in Figure 1.

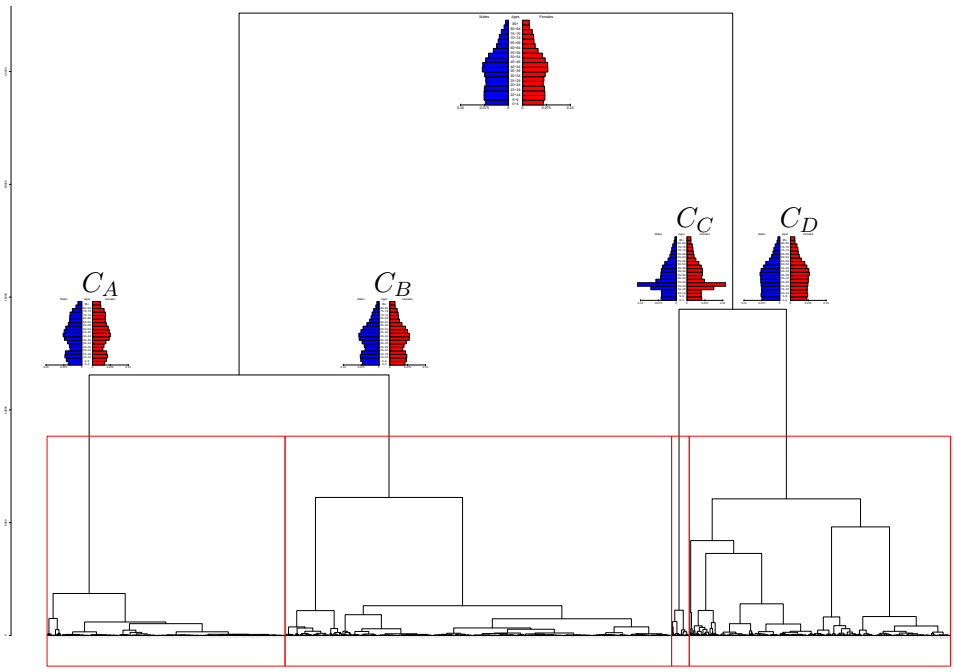


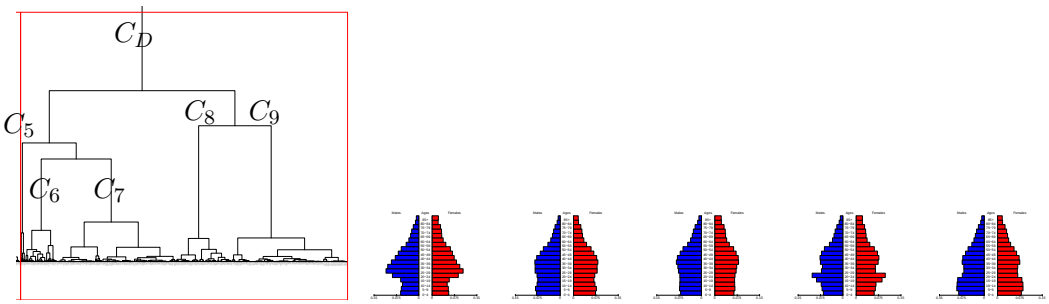
Figure 1: Dendrogram of 3 219 US counties with population pyramids of four main clusters for the year 2000

The level of the last aggregation inside the four main clusters in the dendrogram is the lowest in the third cluster C_C. It includes 63 counties, where the young population highly dominates. The counties in this cluster are the ones with very large universities that have strong influence on the shape of the population pyramid. As can be seen from Table 1, more than half of the USA population is included in the last cluster C_D with 931 counties. Its pyramid's shape is the most similar to the all-American pyramid's shape. But

Table 1: The number of counties and the population sizes included in four main clusters in USA for the year 2000

	C_A	C_B	C_C	C_D
number of counties	848	1377	63	931
population size	23 802 850	106 511 453	4 756 639	150 159 574
% of all population	8.3	37.3	1.7	52.6

from dendrogram we can very easily see five sub-clusters in only a little lower cutting level of the hierarchy. Their pyramid's shapes with the number of included counties are presented separately in Figure 2. The pyramid's shape of the cluster C_5 shows the largest percentage of people around their thirties, less youth population under 20 years old and declining number of people in older age groups. This cluster includes fourteen very densely populated counties with the largest cities in the United States. In this 0.4 per cent among the observed 3 219 counties live 10.5 per cent of the population from the cluster C_D, which represents 5.3 per cent of all population in USA for the year 2000.



	C_5	C_6	C_7	C_8	C_9
# of counties:	14	92	334	162	329

Figure 2: Population pyramids of five sub-clusters of the cluster C_D

4. 2. Clustering of Brazilian municipalities

Population of the 5 570 Brazilian municipalities for the year 2010, obtained from IBGE, is divided into 11 age groups: 6 five-year groups (0-4 years, 5-9 years, ..., 25-29 years), 4 ten-year groups (30-39 years, ..., 60-69 years), and the last year group (70 and older). Age groups have different ranges, but this is not the problem for the used adapted agglomerative hierarchical clustering procedure since in the symbolic representation the age-groups are considered as categories.

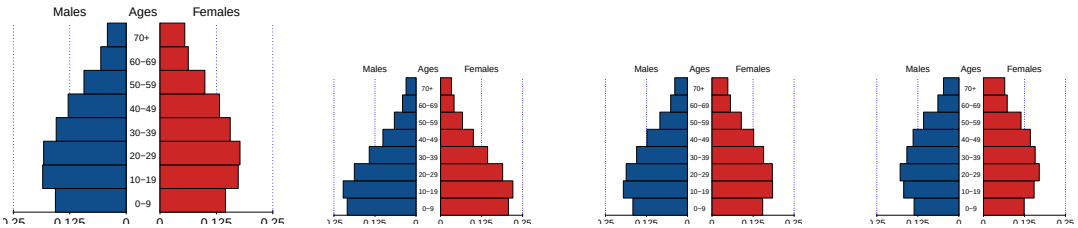
The data include also several other information and more detailed structures like additional division of the population on urban and rural areas. This enables us to produce clustering results based on several different symbolic descriptions of the municipalities, where symbolic representation of the municipality (SO) consists of:

1. two (one for each sex) discrete distributions of length 11 (based on the original age groups);
2. two (one for each sex) discrete distributions of length 8 (based on 7 ten-year age groups and the last 70+ age group);
3. two (one for each area: urban/rural) discrete distributions of length 11 (based on the original age groups);
4. two (one for each area: urban/rural) discrete distributions of length 8 (based on 7 ten-year age groups and 70+ age group);

5. four (for each sex and for each area: urban/rural) discrete distributions of length 8 (based on 7 ten-year age groups and 70+ age group).

Municipalities with incomplete data (NAs) were excluded from clustering process.

In clustering age-sex structures of the Brazilian municipalities, five municipalities with incomplete data were excluded from clustering. On 5 565 Brazilian municipalities for the year 2010, described with age-sex structures of the population with ten-year age groups, three main clusters were detected: one with rather typical pyramidal shape and the other two more similar to the shape of the structure for the whole population, although with some noticeable differences between them. The population pyramids of the whole population and of the obtained three main clusters are presented in Figure 3. Their order from left to right rather well coincides with the location of the municipalities, included in the clusters, on a map from northwest to southeast of the country.



840 mun. 12.7% of all Braz.popul. | 2 124 mun. 36.1% of all Braz.popul. | 2 601 mun. 51.2 % of all Braz.popul.

Figure 3: Population pyramids of three main clusters of Brazilian municipalities described with the age-sex structures of the population for the year 2010, obtained with the weighted agglomerative hierarchical method

Similarly as with UC counties, we can study more detailed descriptions of the clusters and differences inside them by observing smaller clusters on the lower levels of the dendrogram.

5. Conclusions

Demographic structures offer important and interesting information for experts from many different sciences. When we want to study a large number of structures, clustering method can be very useful. It enables us arrange the different types of structures in a systematic way. It is important for any analytical tool to have a proper theoretical background and on the other hand to provide experts with meaningful interpretations. Therefore we chose weighted agglomerative clustering method for the observation of demographic structures of more than 3 000 US counties and more than 5 500 Brazilian municipalities. The method is based on data representations with modal multi valued data, which enables us to preserve the original structures in the most natural way. With such a representation demographic structures are preserved and also the clusters' representatives are demographic structures by themselves. Selected clustering method detected clusters of US counties with similar age-sex structures. With the observation of structural changes from 2000 to 2010 we detect clusters of counties with the most similar and also with the biggest differences in structural changes (where also the ethnicity component play a role). The same method can be used to identify Brazilian municipalities with similar urban and rural structures by age groups, which might be interesting for demographers, sociologists and other experts, for example experts in agriculture.

Acknowledgements

This work was supported in part by the Slovenian Research Agency (research programs P1-0294 and P3-0154 and research projects J5-5537 and J1-5433).

References

Anderberg, M. R. (1973). Cluster Analysis for Applications. Academic Press, New York.

Batagelj, V. (1988): Generalized Ward and Related Clustering Problems. In: H.H. Bock (Ed.): Classification and Related Methods of Data Analysis. North-Holland, Amsterdam, pp. 67-74.

Batagelj, V., & Kejžar, N. (2010). `clamix` Clustering symbolic objects. [Computer software R program]. Vienna, Austria: R Foundation for Statistical Computing. Available at: `https://r-forge.r-project.org/projects/clamix/`

Batagelj, V., Kejžar, N., & Korenjak-Černe, S. (2015). Clustering of Modal Valued Symbolic Data. ArXiv e-prints, Available at: `http://adsabs.harvard.edu/abs/2015arXiv150706683B`.

Billard, L., & Diday, E. (2006). Symbolic data analysis. Conceptual statistics and data mining. Wiley.

Bock, H. H., & Diday, E. (editors and coauthors) (2000). Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg.

Brito, P. (2014). Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Vol. 4, Issue 4, pp. 281-295.

Noirhomme-Fraiture, M., & Brito, P. (2011). Far beyond the classical data models: symbolic data analysis. Statistical Analysis and Data Mining. Vol. 4, Issue 2, pp. 157-170.

Korenjak-Černe, S., Kejžar, N., & Batagelj, V. (2015, Epub 2014 Oct 13). A weighted clustering of population pyramids for the world's countries, 1996, 2001, 2006. Population Studies: A Journal of Demography. Vol. 69, No. 1, pp. 105-120.

Späth, H. (1977). Cluster-Analyse-Algorithmen. R. Oldenbourg, München.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association. Vol. 58, pp. 236-244.