Clustering of Population Pyramids

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

Nataša Kejžar
University of Ljubljana, Faculty of Social Sciences, Department of Informatics and Methodology
natasa.kejzar@fdv.uni-lj.si

Vladimir Batagelj
University of Ljubljana, Faculty of Mathematics and Physics, Department of Mathematics
vladimir.batagelj@fmf.uni-lj.si

Keywords: clustering, population pyramid, Ward hierarchical method, hierarchical clustering with relational constraint

Received: March 6, 2008

Population pyramid is a very popular presentation of the age-sex distribution of the human population of a particular region. The shape of the pyramid shows many demographic, social, and political characteristics of the time and the region. In the paper results of hierarchical clustering of the world countries based on population pyramids are presented. Special attention is given to the shapes of the pyramids. The changes of the pyramids' shapes, and also changes of the countries inside main clusters are examined for the years 1996, 2001, and 2006.
Also smaller territorial units of a country can be observed through clusters. To illustrate this, clusters of 3111 mainland US counties in the year 2000 obtained using the hierarchical clustering with relational constraint of counties' population pyramids are examined. In the paper, the results for clustering into nine main clusters are presented.

Povzetek: Prikazano je grupiranje demografskih piramid.

1 Introduction

Population pyramid is a very popular presentation of the age-sex distribution of the human population of a particular region. It gives picture of a population's age-sex structure, and can also be used for displaying historical and future trends.

Generally, there are three main pyramids' shapes: expansive, constrictive, and stationary (Figure 1). The *expansive* shape is typical for fast-growing populations where each birth cohort (a group of people born in the same year or years period) is larger than the previous one (Latin America, Africa).

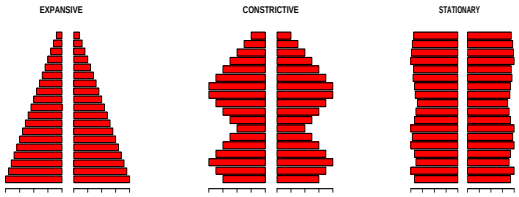


Figure 1: Three main shapes of population pyramids

Constrictive shape displays lower percentages of younger population (United States).

Stationary shape present somehow similar percentages for almost all age groups. The population pyramids of the Scandinavian countries tend to fall in this group.

Since the biggest influence on the pyramid's shape have fertility and mortality, the explanation of the pyramids' shapes is often related to the "Demographic Transition Model" (DTM) that describes the population changes over time (Figure 2). It is based on an interpretation that begun in 1929 by the American demographer Warren Thompson, of the observed changes, or transitions, in birth and death rates in industrialized societies over the past two hundred years.

Besides births and mortality, also other processes, depending on social or/and political policy and events (migrations, birth control policy, war, life-style etc.) have strong influence on age-sex structure of the population, that reflect also on the shape of the population pyramid.

Population pyramids are very easily understandable to almost everyone. In the combination with professional knowledge, they offer also many additional explanations about different processes to experts (e.g. demographers, sociologists, politicians, economists, geographers).

Due to these facts we decided to observe clusters of the world countries based on the shapes of their population pyramids. We observed how countries and clusters of them fit with the main pyramids' shapes and how stable are clusters that we got with hierarchical clustering. Selected clustering procedure to determine clusters of the world coun-

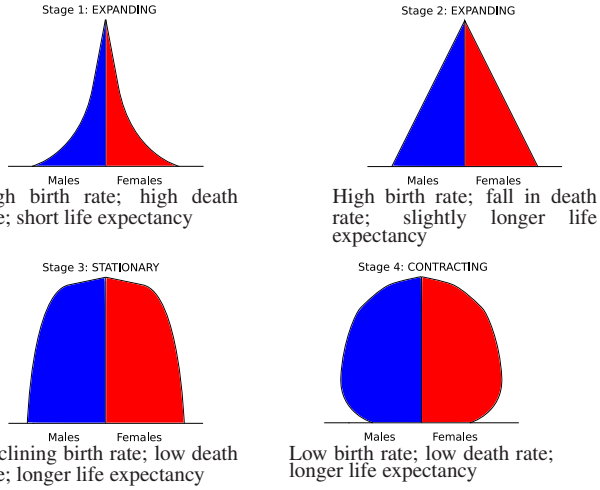


Figure 2: Shapes of population pyramids for 4 stages of the demographic transition model

tries is based on the Ward's hierarchical clustering method.

For each country a clustering can be applied also on smaller regions. In smaller regions the influence of other processes besides births and mortality (for example local migrations caused by schooling, religion, work or health reasons etc.) is even more emphasized, which reflect in differences of pyramids' shapes. Therefore we further examined clusters of 3111 mainland US counties. To determine clusters, hierarchical clustering with relational constraint was used.

Since we are not demographers we focus on the presentation of the methods and results they produce, and we limited the explanation of them to the 'technical' characteristics (results that can be directly seen from the obtained graphs and clusters), hoping that the proposed approaches will catch the attention of researches using pyramids' analysis in their work.

2 Clustering procedure

Data on the population pyramids of the world countries used in our analysis were taken from the web page of the International Data Base (IDB). Age is divided into 17 five-years groups (0-4 years, 5-9 years, 10-14 years, ..., 75-79 years, 80+). In our model, for each country each age group is considered as a separate variable for each sex, so each country is presented with 34 variables: 17 variables for 5-years age groups for men, and 17 variables for 5-years age groups for women. Values are normalized so that they present percentages of the country's population in each age group. Euclidean distance between corresponding vectors is used. Although some objections against the usage of this difference measure can be found (Andreev, 2004), in our opinion in observation of the shapes of the population pyramids, each age-group can be considered as a separate variable.

For clustering procedure, the algorithm for the Ward's hierarchical method, implemented in a package 'cluster' in the statistical environment R was used. This clustering procedure is implemented based on the description of the method in Kaufman and Rousseeuw, 1990.

Since each country is represented with normalized vector (relative age-sex distribution), the 'centroid' pyramids of the clusters of countries are not real population pyramids describing the whole population in the clusters, but are based on the new vector got with the clustering procedure. So they can be interpreted only in terms of shapes, not as a population pyramids of countries' clusters, because the population size of the whole cluster is not taken into account. But on the other hand such approach enables us to detect even small countries with very different pyramids' shapes based on special countries' characteristics.

3 Clusters of the world's countries over time

We observed clusters of the world countries obtained by the Ward's clustering procedure for each year from 1996 to 2007. Although the time period is rather short for the human life, substantial changes can be seen. In Figure 9, Figure 10, and Figure 11 respectively, hierarchical trees for the years 1996, 2001, and 2006 with the appropriate pyramids' shapes for some of the main clusters are presented.

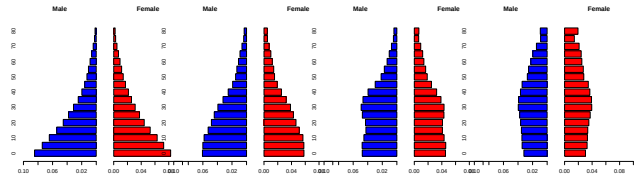
3.1 Pyramids' shapes of the clusters of world countries

Our first examination is concentrated on the shapes of the population pyramids of the clusters in the hierarchies. For each of the years 1996, 2001, and 2006, some of the interpretations are given. Much more detailed information can be obtained depending on the cutting level in the hierarchy and on the interest of the observer.

3.1.1 Year 1996

For the year 1996, the dataset contains 215 countries. Four main pyramids's shapes are presented in Figure 3. The first cluster has typical expansive pyramid's shape. It includes 77 countries (most of African countries etc.). When comparing these pyramids' shapes with pyramids' shapes of the DTM shown in Figure 2, we can say that the first one corresponds to the Stage 1, the second and the third have characteristics of the Stage 2 and 3, and the last looks between stages 3 and 4. For easier explanation of later observations we will denote them with letters A, B, C, and D respectively.

Clusters at the bottom levels of the hierarchy are more and more similar. When we cut at the level with eight clusters, clusters A and C are each divided into two smaller clusters presented in Figure 4 for cluster A and in Figure 5 for cluster C. Cluster B remains the same also when cutting into eight clusters. First shape from Figure 4 corresponds



A B C D

Figure 3: Pyramids's shapes of four main clusters of the countries for the year 1996

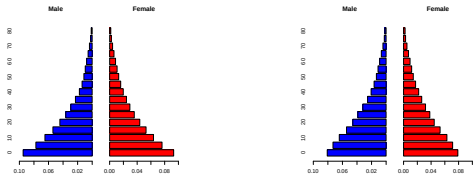


Figure 4: Pyramids's shapes of the first two (from the left) of eight clusters of the countries for the year 1996

to the Stage 1, but the second one looks closer to Stage 2 of the DTM from Figure 2.

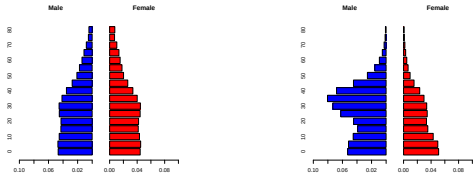


Figure 5: Pyramids's shapes of the fourth and fifth (from the left) of eight clusters of the countries for the year 1996

The last cluster in Figure 5 includes countries with very big differences between gender's distributions: Bahrain, Kuwait, Qatar, and United Arab Emirates. Since in all these countries the population of men in the data is much bigger (specially in the middle ages) than of women, this has effect also on the pyramid's shape of the cluster. We also calculated dissimilarities between gender's distributions for all countries in 1996, and five of them with the largest differences are: United Arab Emirates, Qatar, Kuwait, Oman, and Bahrain. Four of them are included in the described cluster, detected in the hierarchy.

Cluster D is at lower level divided into three smaller clusters. Their pyramids's shapes are presented in Figure 6. Slovenia is in the last cluster from the right together with 34 other countries. The most similar country to Slovenia is Croatia, and after it also Belgium, France, Finland, and Gibraltar what can be seen on the dendrogram in Figure 9. The age distributions of both genders in the pyramid's shape are quite similar, although there are slightly more women, specially those that are older than 70.

Similar and even more detailed descriptions at lower levels can be obtained for each group in the hierarchy. With additional knowledge about countries more detailed explanations of the shapes can be given which also offer many

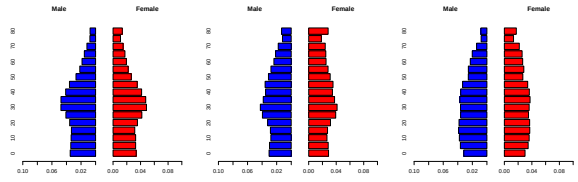


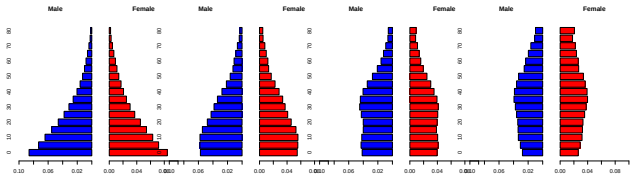
Figure 6: Pyramids' shapes of the last three (from the right) of eight clusters of the countries for the year 1996

interesting points for discussion about similarities and differences among countries and/or clusters.

3.1.2 Year 2001

Hierarchy on the 222 world countries for the year 2001 is presented in Figure 10. At the upper level of the dendrogram two main shapes of pyramids can be seen. The first pyramid's shape approximately corresponds to the Stage 2 of the DTM shown in Figure 2, and the second one approximately corresponds to the Stage 4.

One level lower each of two main clusters is divided into two additional clusters. Their pyramids' shapes are presented in Figure 7. Their shapes are similar to those in the year 1996, therefore we denoted appropriate clusters with the same letters A, B, C and D.



A B C D

Figure 7: Pyramids's shapes of four main clusters of the countries for the year 2001

Inspecting lower level of the hierarchy, eight clusters can be detected. First cluster is the same as the first one (A) presented in Figure 7. It includes 60 countries and has typical expanding shape of the population pyramids.

Cluster B is at the lower level divided into two smaller clusters. The shapes of the pyramids are presented in Figure 8. They approximately fit to Stage 2 and Stage 3 of the DTM. That shows some progress in countries' development comparing with the year 1996.

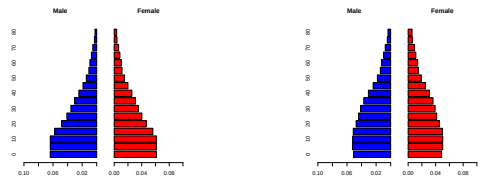


Figure 8: Pyramids's shapes of the second and third (from left) of eight clusters of the countries for the year 2001

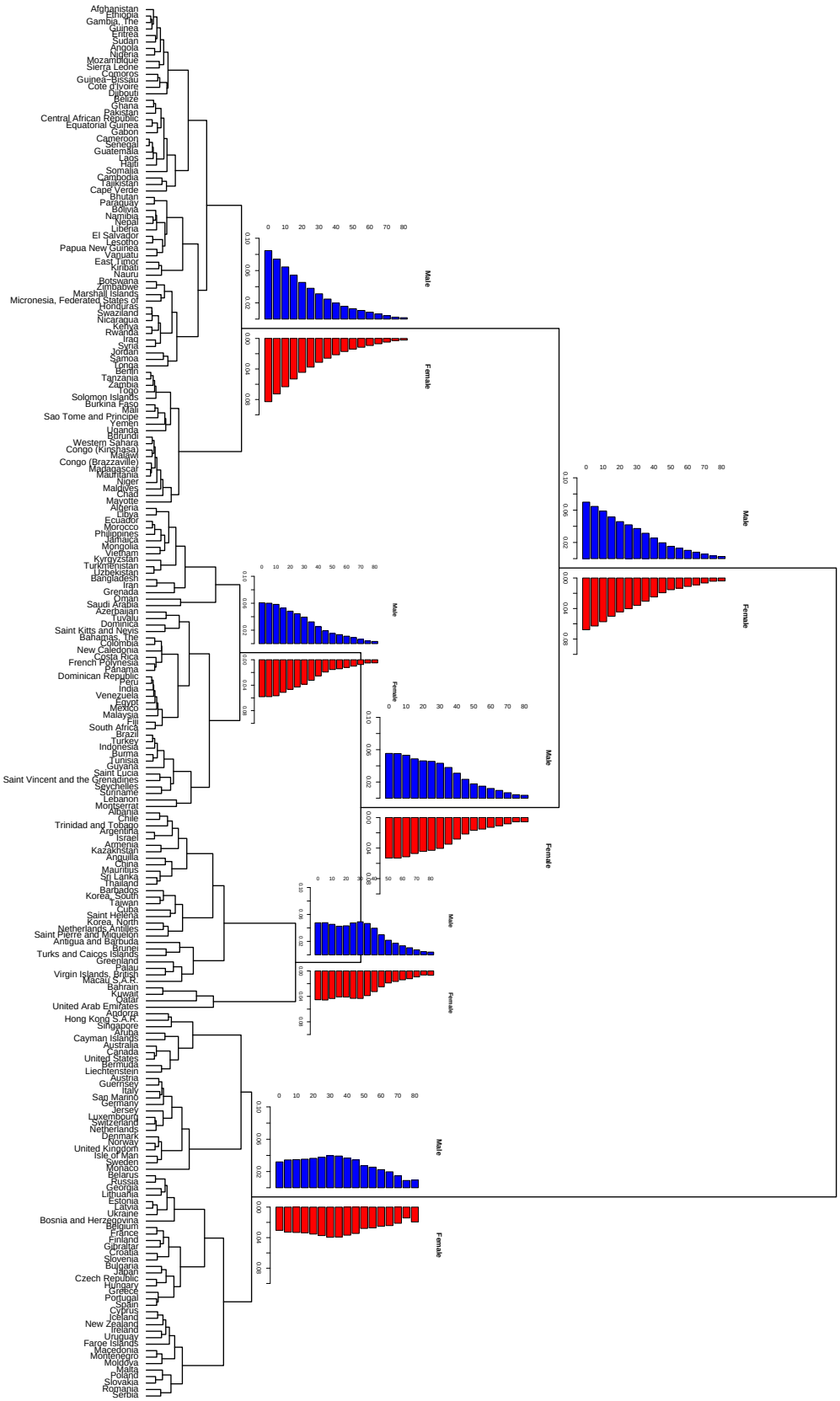


Figure 9: Clusters of the countries and main pyramids' shapes for the year 1996

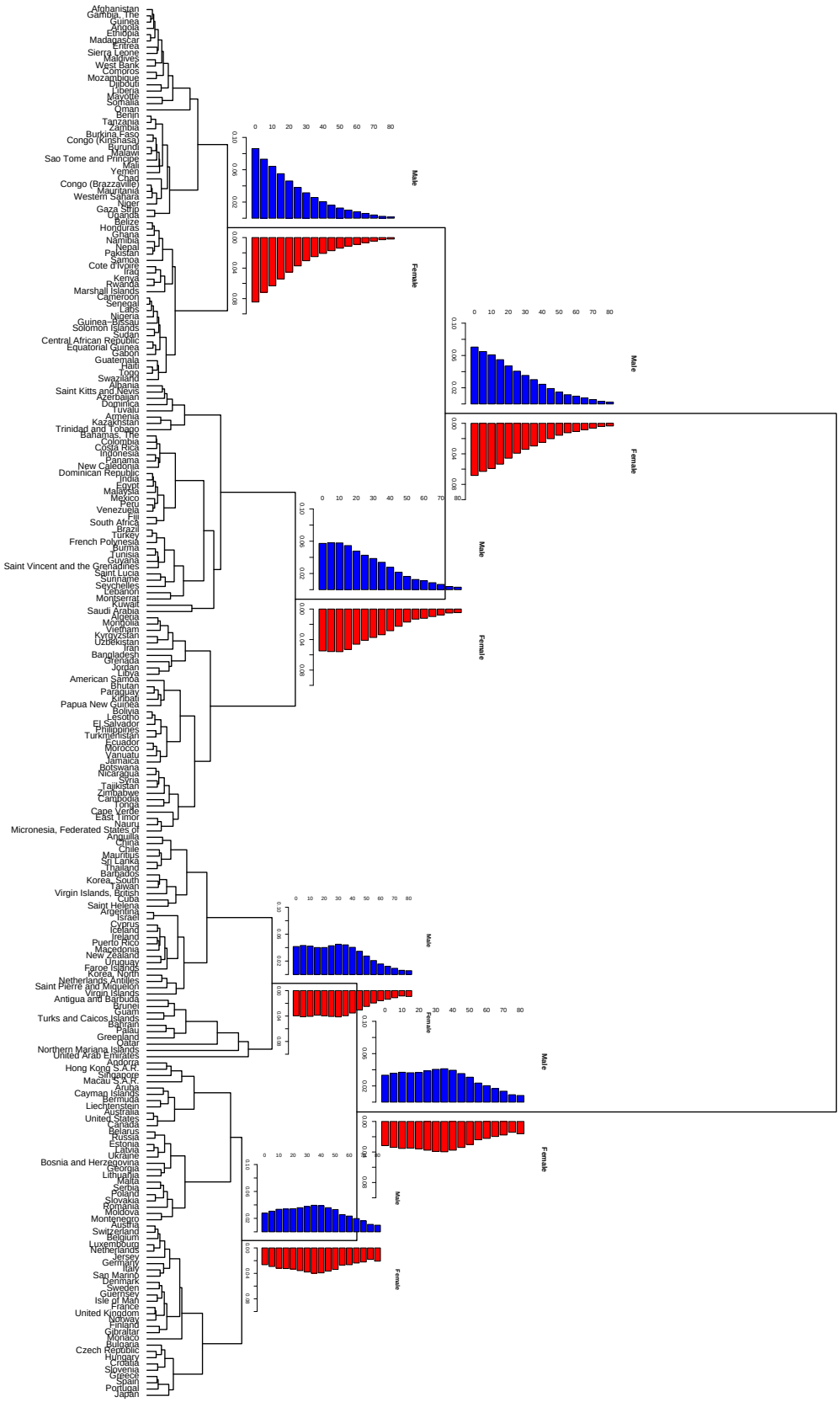


Figure 10: Clusters of the countries and main pyramids' shapes for the year 2001

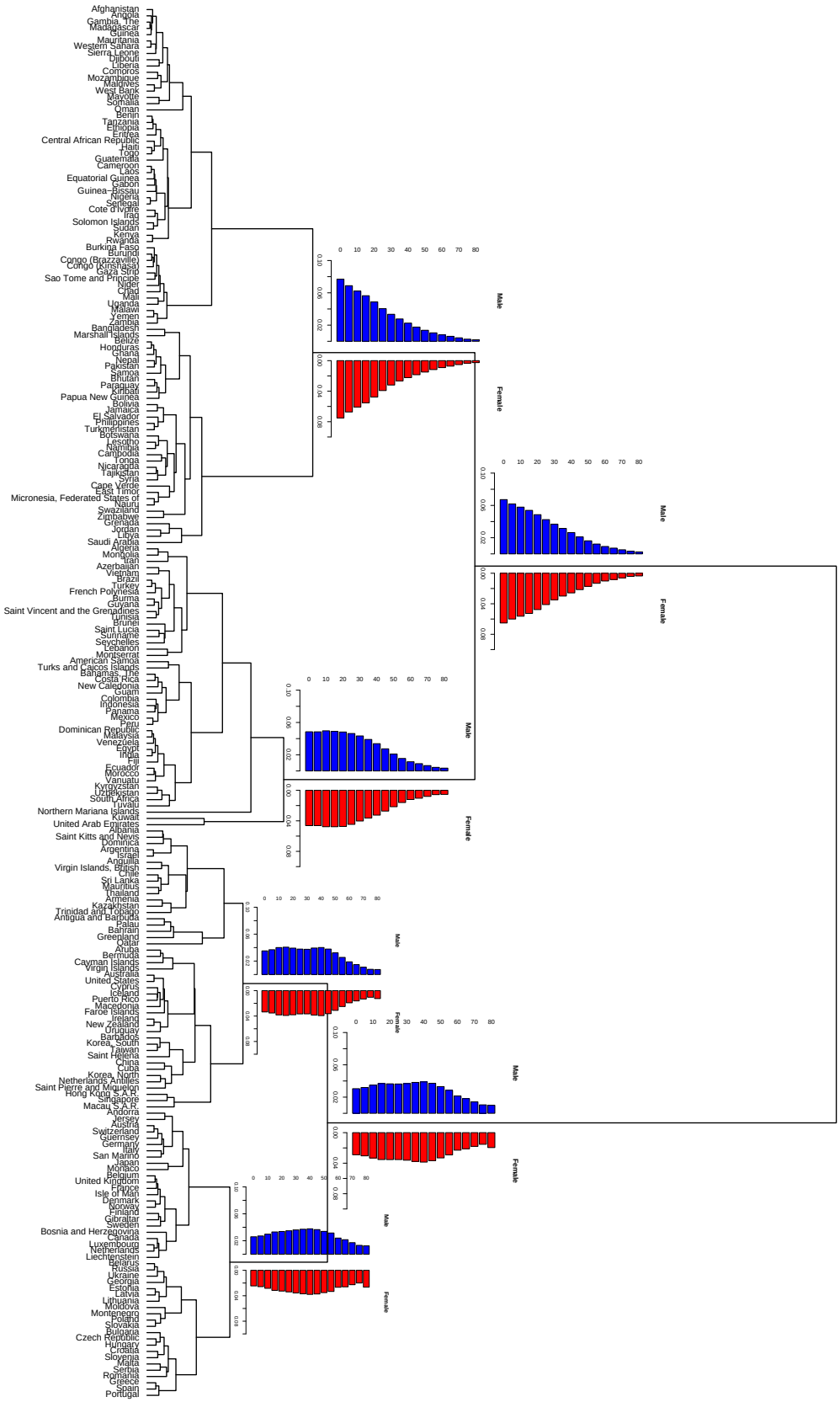


Figure 11: Clusters of the countries and main pyramids' shapes for the year 2006

Cluster C is at the lower level divided into three additional smaller clusters. Their pyramids' shapes are presented in Figure 12. Among eight clusters United Arab Emirates forms separate cluster by itself (the right pyramid in Figure 12).

Its population pyramid's shape shows big differences between gender's distributions in the country. At the upper levels, Northern Mariana Islands, Qatar etc. join United Arab Emirates.

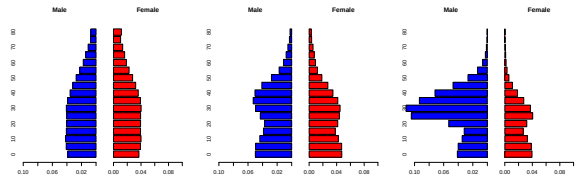


Figure 12: Pyramids's shapes of three of eight clusters of the countries for the year 2001

The last cluster D among four main clusters is at lower level divided into two smaller and more similar clusters of countries. Their pyramids' shapes are presented in Figure 13. Slovenia is in the right cluster in Figure 13 together with 27 other countries. Among them were twelve countries (Belgium, Bulgaria, Croatia, Czech Republic, Finland, France, Gibraltar, Greece, Hungary, Japan, Portugal, and Spain) also in the selected cluster with Slovenia among eight main clusters for the year 1996. The shape of the pyramid in this cluster is rather symmetric, although there are larger values in the women side (specially for older than 75).

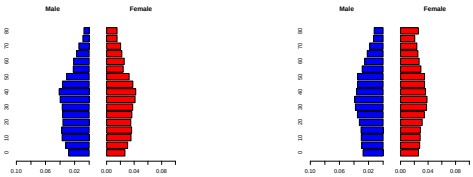


Figure 13: Pyramids's shapes of the last two among eight clusters of the countries for the year 2001

3.1.3 Year 2006

The last hierarchy we present is the hierarchy of 222 the countries for the year 2006. It is presented in Figure 11. The shapes of the pyramids of the four main clusters are presented separately in Figure 14. Also these shapes are similar as in the years 1996 and 2001 therefore appropriate clusters are denoted with the same letters. As for the previous two years also for the year 2006 more detailed explanations for each cluster of the hierarchy could be found.

3.2 Stability of the clusters in the hierarchies

In the following section we observe how stable are the main clusters over time. For each of the years 1996, 2001, and

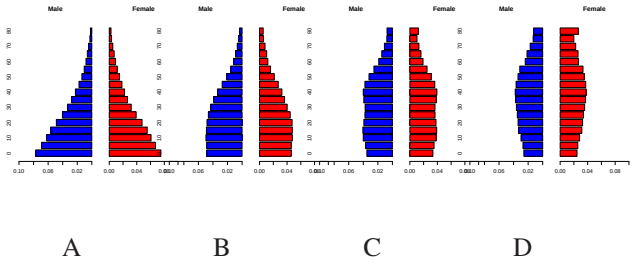


Figure 14: Population pyramids of the main clusters of the countries for the year 2006

2006 we present the results of observation of four main clusters. From Figure 3, Figure 7, and Figure 14 we can see four main rather similar pyramids' shapes, although even these are slightly changing over time.

Observing changes of countries inside each of the clusters A, B, C and D for each of the years 1996, 2001 and 2006, we can conclude the following:

Countries from the cluster A in the year 1996 (Figure 3) are in the year 2001 in clusters A and B (Figure 7). All these countries except one (Vanuatu) are included in the cluster A in the year 2006 (Figure 14).

Cluster B in the year 1996 is included in cluster B in 2001 (the only exception is Oman), and also mostly corresponds to the cluster B in the year 2006.

Cluster C in the year 1996 is mainly presented in clusters B and C in the year 2001, and all except three of the countries from it are included in the cluster C in the year 2006. The remaining three countries (Kuwait, Turks and Caicos Islands, and United Arab Emirate) are in the year 2006 included in the cluster B. This is not surprising because of the differences of gender's distributions in these countries.

The fourth cluster D of four main clusters in the year 1996 is mainly included in cluster D in the year 2001, and most of the countries, precisely 46 from 60 countries from it, are also included in cluster D in the year 2006. The remaining 14 countries from it are in cluster C in 2006.

Similar comparisons were made for the years 2001 and 2006:

The first cluster A of four main clusters in the year 2001 (Figure 7) is included in cluster A in 2006 (Figure 14).

Most of the countries from cluster B in 2001 are included in clusters A and B in 2006, except six countries (Albania, Armenia, Dominica, Kazakhstan, Saint Kitts and Nevis, and Trinidad and Tobago), that are in 2006 included in cluster C.

Five of the countries (Brunei, Guam, Northern Mariana Islands, Turks and Caicos Islands, and United Arab Emirates) from cluster C in 2001 are moved to cluster B in 2006, all the remaining countries are in 2006 included in cluster C.

46 of 54 countries (including Slovenia) from cluster D in 2001 create cluster D in 2006, the remaining eight countries are in 2006 included in cluster C.

In the Figure 15 we present these movements among four

main clusters for the years 1996, 2001 and 2006 with the number of countries.

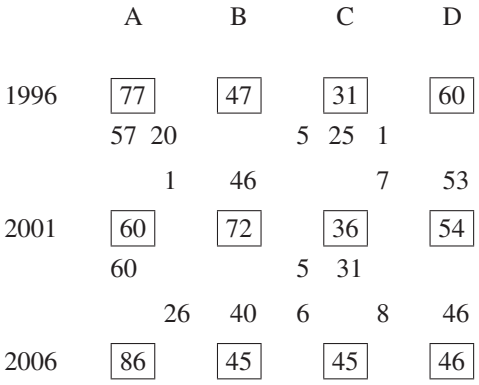


Figure 15: Movements presented with the number of countries among four main clusters for the years 1996, 2001 and 2006

Differences among clusters can be observed in greater detail considering the hierarchies in each of the clusters.

4 Hierarchical clustering with relational constraint of US Counties

For US counties age in population pyramids is divided into 18 five-years groups (0-4 years, 5-9 years, 10-14 years, ..., 75-79 years, 80-84 years, 85+). In our model, for each US county each age group is considered as a separate variable for each gender, so each county is presented with 36 variables: 18 variables for 5-years age groups for men, and 18 variables for 5-years age groups for women. Values are normalized so that they present percentages of the county's population in each age group among the whole county population. Euclidean distance between the corresponding vectors is used. NA values in data for 6 counties were replaced with 0.

For clustering procedure, the algorithm for hierarchical clustering with relational constraints based on the maximum hierarchical method was used. It is implemented in Pajek, the program for analysis and visualization of large networks. The agglomeration of two counties was restricted with the relational constraint based on neighboring counties (Ferligoj, Batagelj, 1983). The maximal method to calculate new dissimilarity between clusters was used. The neighboring relation is symmetric. Therefore the tolerant strategy to determine the relation between the new cluster and other clusters is used (Batagelj, Ferligoj, Mrvar, 2008).

As in the case with the world's countries, also here the 'centroid' pyramids of the clusters of counties are not real population pyramids describing the whole population in the clusters, but are based on the new vector produced by the clustering procedure. So they can be interpreted only in terms of shapes, not as a population pyramids of counties' clusters, because the population size of the whole cluster is not taken into account.

The cut of the dendrogram was done at height 0.06, which divided counties in 36 clusters with 155 isolates (counties that are very different from all their neighbors). Out of these 36 clusters were only 14 clusters with more than 10 vertices, therefore we decided to increase the height.

At height 0.1 we obtained 9 clusters and 54 isolated counties. There are 6 clusters of more than 6 counties (precisely with 7, 15, 69, 402, 1152 and 1406 counties) and 3 of them with 2 counties. Clusters (groups of counties) are presented in Figure 16 with different shapes and colors. Group 1 is the largest group situated at eastern part of the USA (light gray circles). Darker gray group with triangles that borders group 1 is group 2, the second largest group of counties. Dark gray circular vertices at the Florida peninsula belong to group 5. 7 dark gray vertices in the middle of group 2 (in the center of the USA) represent group 7. The large white squares in the north and middle of the USA represent group 4, while area with lighter gray circles inside the bottom of group 2 belongs to group 3. Groups 6, 8 and 9 with 2 counties each are in Figure 16 represented with dark gray diamonds (group 6 is in the north-west of the USA, group 8 at the south east of group 3, and group 9 in the middle of the largest group 1).

Further inspection of the pyramids' shapes of the clusters of counties shows that all three 2-vertex clusters (groups 6, 8 and 9) have average population pyramid with mostly young people in their 20s (Figure 17). We conjecture these are counties with mainly student population (surroundings of larger universities and colleges). More precisely: cluster 6 includes counties with University of Idaho and Washington State University. In cluster 8 are Madison and Walker County, Texas, with median ages 33 and 31 and with male population for more than 50% larger than female population. Cluster 9 includes Montgomery and Radford Counties, Virginia, with West Virginia University Institute of Technology, popularly called WVU Tech, and Radford University, which have strong influence on the age distribution.

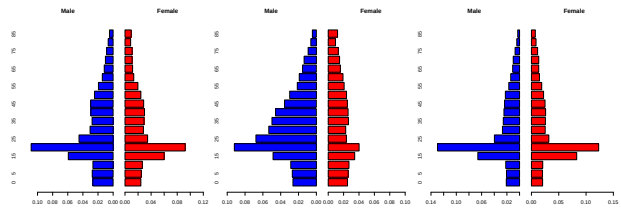


Figure 17: Pyramids' shapes of three clusters with two counties

Pyramids' shapes of the two largest clusters show typical all-American population pyramid (Figure 18) with rather typical constrictive shape (Figure 1).

In two of other four clusters pyramids' shapes older

Figure 16: Clustering of US counties in the year 2000 with relational constraints

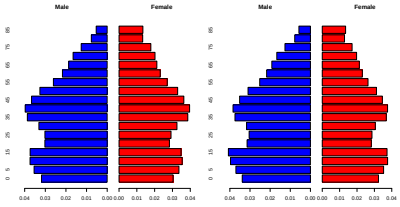


Figure 18: Pyramids' shapes of two largest clusters

population is more pronounced (looking bottom-up the pyramid bars start shrinking later than the overall American population pyramid). The groups are concentrated in Florida (first in Figure 19) and in Missouri (the second one in Figure 19).

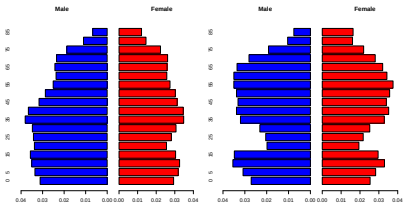


Figure 19: Pyramids' shapes of clusters with older population

First of the last two among nine clusters shows relatively less people older than 30 than the overall American population, while the second one (North and middle of the USA) indicate less people in the 20s (they might be away for study).

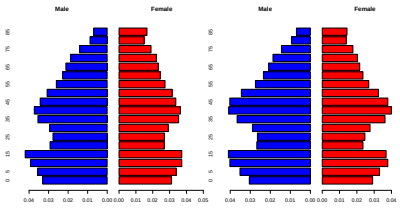


Figure 20: Pyramids' shapes of the two remaining clusters

Because of the relational constraint (regional neighborhood) we got 54 isolated counties at the cutting level 0.1. In Figure 16 they are represented with black circles. They remain isolated because they have different pyramids' shapes than the neighboring counties.

Most of the isolates (29 of them) are due to the proximity of a university. The shape of their population pyramids looks very much like those in Figure 17. 8 other isolates are more gender specific (have mostly more men than women). There are 5 isolates in the state of New York that have slightly different population distribution as the shape of group around them. The other isolates are of two types: they have either considerably less youngsters (or older people) than the surrounding counties or their pyramids look very random due to the small number of inhabitants.

5 Conclusion

Population pyramid is a very popular graphical presentation of the age-sex distribution of the human population of a particular region. Its shape is influenced besides fertility and mortality (usually presented with demographical indicators as birth rates, death rates and growth rates) also by many other social and political policies and events, such as migrations, birth control policy, wars, life-style etc. Population pyramid offers insight into different phenomena in many fields interested in population observations, such as demography, geography, sociology, economy, politics etc.

The aim of the paper was to observe how population pyramids of the world's countries corresponds to the main pyramids' shapes, which are usually related to the "Demographic Transition Model". Although the observation period of 10 years was short for the human life, substantial changes can be seen. Roughly speaking we can conclude from our observations, that the pyramid's shapes of the main clusters correspond to the Stages 1 and Stage between 3 and 4 of the "Demographic Transition Model" in the year 1996, and later are moved to the Stages 2 and even closer to 4 in the 2001 and 2006. The divide between the undeveloped and developed countries is increasing.

Most of the main four clusters are quite stable through observed years. We are aware that for some observers differences are more important than generalization and they can be observed in detail with the separate inspection of the smaller parts of the hierarchy that belong to each cluster.

In the second part, we examine pyramids' shapes of clusters of US counties, because in smaller territorial units the influence of local characteristics is even more emphasized, which reflects also on pyramids' shapes. For clustering 3111 mainland US counties, hierarchical agglomerative clustering procedure with neighborhood constraint was used. The results confirm strong influences of local characteristics (for example universities) on the pyramids' shapes of smaller populations. The clustering procedure exposed some groups of counties with pyramids' shapes which strongly differs from all-American constrictive population pyramid's shape.

Acknowledgement

This work was partially supported by the Slovenian Research Agency, Project J1-6062-0101.

References

[1] Andreev, L. and Andreev, M. (2004) Analysis of Population Pyramids by a New Method for Intelligent Pattern Recognition, *Matrixreasoning*, Equicom, Inc.

[2] Andreev, L. (2004) Fusion of Socio-Demographic Variables, *Matrixreasoning*, Equicom, Inc.

http://www.matrixreasoning.com/publications/1.html

[3] Batagelj V., Ferligoj A. and Mrvar A. (2008): Hierarchical clustering in large networks. Presented at Sunbelt XXVIII, 22-27. January 2008, St. Pete Beach, Florida, USA.

[4] Ferligoj A. and Batagelj V. (1983): Some types of clustering with relational constraints. *Psychometrika*, 48(4), p. 541-552.

[5] Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

[6] Pressat, R. (1978) *Statistical Demography* (Translated and adapted by Damien A. Courtney), Methuen, University Press, Cambridge.

[7] Pressat, R. (1988) *The Dictionary of Demography* (Edited by Wilson, C.), Basil Blackwell.

[8] International Data Base.
http://www.census.gov/ipc/www/idbnew.html

[9] Mrvar, A. and Batagelj, V. (1996-2008) The Pajek program – home page.
http://pajek.imfm.si/

[10] R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
http://www.R-project.org.