Clustering age-sex structures to monitor their development over time: Latin America and the Caribbean sub-national areas 1960-2011

Simona Korenjak-Černe, University of Ljubljana, Faculty of Economics, Slovenia, and Ludi Simpson, University of Manchester, UK

Una contribución al pre-evento del Congreso de ALAP: *Demografía subnacional de América Latina y el Caribe: Proyecto s-ALyC*. **30 September 2016**

1. Introduction

As governments attempt to develop their nation's infrastructure, subnational demographic projections play a part in assessing both the future demand for services and the impact of new investments on population change. In Latin America the development of subnational demographic projections is a priority for public policy (Jannuzzi 2012) but has uneven experience, with few countries providing regular updates (Gonzalez and Torres 2012).

The aims of this paper are to examine sub-national time series of age-sex-structures for many countries in Latin America and the Caribbean, summarize the diversity and the socio-demographic associates of changing age-sex structures, and to identify and characterize the development of those age-sex structures over time, in ways useful to the practice of demographic projections. In particular, we are interested in the similarity of sub-national areas across national boundaries. The work is part of a wider project on sub-national demography in Latin America and the Caribbean (University of Manchester 2016).

Many countries have no robust estimates of sub-national population between the decennial censuses that do take place in almost all countries. The censuses are supported by the United Nations and its regional demographic office CELADE, which prepares a common set of national population estimates and projections from 1950, but not sub-national equivalents. The investment in national censuses is the basis for this paper, as many of them have been archived as sampled micro-data by the University of Minnesota (IPUMS, 2015).

2. Methodology and short explanations of the obtained results

The data set contains age-sex distributions of 1444 census samples representing sub-national areas of Latin America and the Caribbean from 1960 to 2011, downloaded from IPUMS. We included in analysis only 1396 sub-national areas without missing values (denoted with NA) where all 'blank' values were considered as missing data. The frequencies of them by country and by years are presented in the appendix (Appendix, A1 and A2). All of the excluded areas date from before 1996. One quarter of them (12) are from Paraguay mostly from the sample from 1962 which has erroneous entries for its women aged 65 and older (IPUMS 2016). Another quarter of them are from Colombia and the other half of the excluded areas are from nine other countries, among them also both Saint Lucia data. These data may be real zeros (no person in an age group), or the result of top-coding of age. All the included samples have non-zero data for males and females in quinary age groups 0-4 to 80-84, and 85+.

We examine sub-national time series of age-sex-structures for many countries in Latin America and the Carribean with a clustering approach, where we want to identify the main

shapes of the structures. Further we describe the association of the obtained clusters with socio-demographic/economic indicators to observe the relationship between structure shape and selected indicators, and to identify and characterize the development of the structures over time.

We considered two different clustering approaches: one based only on the structures relative to the whole population in the area, and the second one which weights by population size of each sex.

The first one is based on a classical unit's representation where each sub-national area (DAM) is represented by a single vector of 36 components (one for each of 18 age groups and each sex) representing the age-sex structure of population relative to the whole population in this sub-national area. Dissimilarity between two DAM is measured with a squared Euclidean distance between the two vectors to be able to discuss the variation of age structures. Main shapes (clusters) are detected with the Ward agglomerative hierarchical clustering method. Graphical presentation of the obtained hierarchy enables us to decide upon the possible number of clusters. For analysis we used procedure hclustSO from the R program `clamix` (Kejžar and Batagelj, 2010).

In the second approach we represented the age-sex distributions of sub-national areas with two vectors – separate distributions of men and women over age groups and clustering with agglomerative hierarchical clustering weighted by the DAM population. Two vectors represent distributions of men and women over 18 five-year age groups. A weighted agglomerative clustering method (Korenjak-Černe et al., 2014) is used. In this approach, the weighting ensures that each cluster's average remains the age-sex distribution of the aggregate population of the cluster and has as such meaningful interpretation by itself. The second difference from the first approach is that because of the two vectors, relative distributions by each sex are recognized in the clustering, not relative to the whole population in the area. Imbalances between population of men and women in the area are included by weighting. Also here we used procedure hclustSO from the R program `clamix` on appropriate units and clusters' representations with the included weights. The main disadvantage of weighting by population size is that areas with relatively large populations are considered distant from each other and from areas with small populations, even if their relative age-sex structures are not very different.

We focused detailed study on the results of the first approach which are provided and discussed in section 3. Main advantages and disadvantages of both clustering approaches are explained in the last section, where we give short summary of the analytical strategy with the results.

We linked the clusters to some socio-demographic indicators of classical development characteristics. Some of them are directly related to the age-sex distribution (e.g. % of children in the population, % of elderly in the population), while others detect the development stage of the sub-national areas in the clusters (agricultural activity, urban population, women's economic activity, and primary education). The strong connection between cluster representatives and socio-demographic indicators, and the movement of each DAM over time between clusters, help to establish an optimum ordering of the clusters that best coincides with progress of the demographic transition and economic development. We identify DAMs that move over time in ways that do not conform to a notion of gradual progress, and will gather local knowledge to explain these unusual cases.

With additional descriptive statistics we also examined the presence of each country in each cluster (Appendix, Table A3). Large differences among sub-national age-sex distributions (relative to the sub-national area population) are detected in Costa Rica, and to a less extreme extent in Brazil, and Panama. On the other hand, there are countries in which shapes of sub-national distributions are very similar (Uruguay, Jamaica, Cuba).

Since we don't have data for all sub-national areas for the same years, we made comparisons for each pair of contiguous decades. We calculated average dissimilarity between pairs of sub-national areas for which we have data in both decades. The average increased only from 1960 to 1970. In all later sequential pairs of decades the average dissimilarity decreased, suggesting a slight convergence of age-sex structures during the forty years 1970-2010.

3. Analysis of age-sex distributions relative to the whole area population (first approach)

3.1 Cluster analysis

Figures 1 and 2 and Table 1 identify clusters of sub-national areas from Latin America and the Caribbean with similar shapes of the population age-sex distribution using Ward's agglomerative hierarchical clustering, indicating 4 main and 11 more detailed clusters.

Weighted agglomerative hierarchical clustering on 1396 areas from 1960 to 2011

(4 main clusters – red, 11 main clusters – blue)

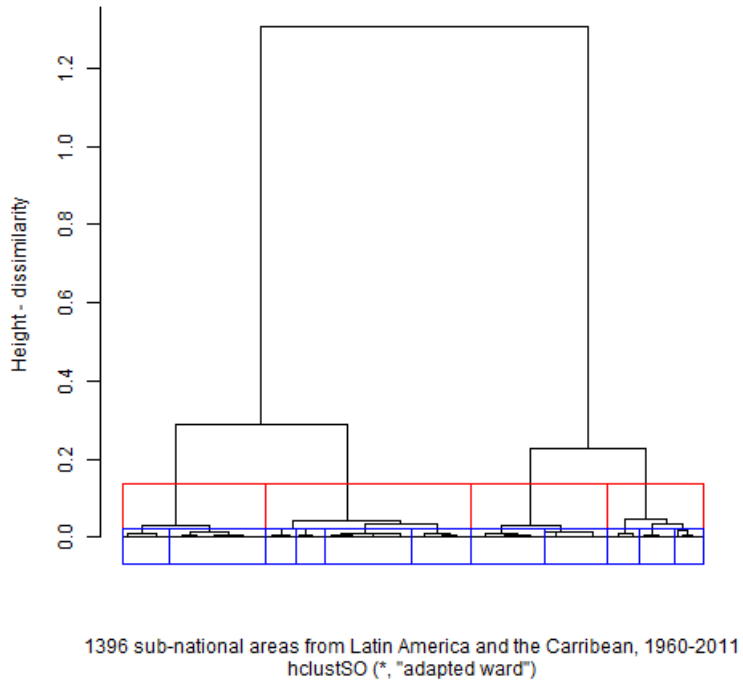


Figure 1: Dendrogram obtained with the Ward's agglomerative hierarchical clustering on 1396 areas from 1960 to 2011 with the partitions into 4, and 11 clusters.

The ordering of the four main clusters matches the ordering in the hierarchy in Figure 1 from left to the right. Among possible reorderings of the 11 more detailed clusters in the obtained hierarchy (clusters must remain under their main cluster in the hierarchy, but the left-right ordering in the level below each main cluster can be changed) we selected the one that creates the most monotonic ordering of the socio-demographic/economic indicators, and we also considered counts of time movements (subsection 3.3).

Count

		4 clusters				
		C1_4	C2_4	C3_4	C4_4	Total
11 clusters	C01_11	114	0	0	0	114
	C02_11	230	0	0	0	230
	C03_11	0	73	0	0	73
	C04_11	0	69	0	0	69
	C05_11	0	141	0	0	141
	C06_11	0	210	0	0	210
	C07_11	0	0	152	0	152
	C08_11	0	0	176	0	176
	C09_11	0	0	0	86	86
	C10_11	0	0	0	69	69
	C11_11	0	0	0	76	76
Total		344	493	328	231	1396

Table 1: Overlapping of the 4 main and 11 more detailed clusters obtained with the Ward hierarchical clustering of the age-sex distributions (relative to the whole population in the sub-national area) of the sub-national areas in Latin America and the Caribbean from 1960 to 2011.

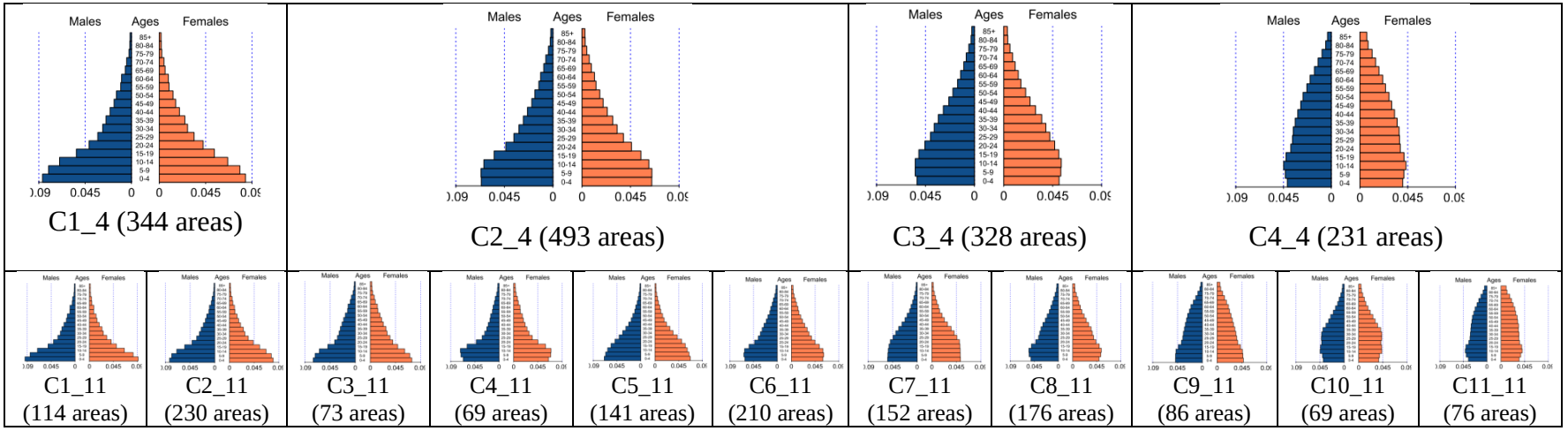
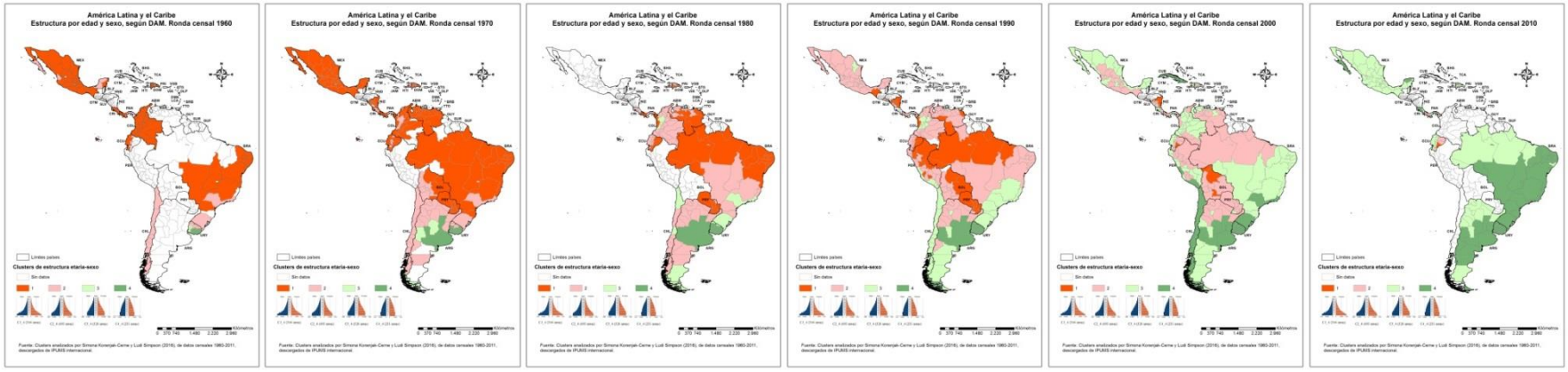


Figure 2: Representatives of 4 and 11 clusters, respectively, obtained with the Ward agglomerative hierarchical clustering on 1396 age-sex distributions of the population (relative to the whole population of the sub-national area) sub-national areas in Latin America and the Caribbean from 1960 to 2011, with the corresponding number of areas in them.

One of the most informative pictures about the time changes of the observed structures with the inclusion of the sub-national areas into obtained clusters can be provided with maps that show distributions of clusters for each decade.



1960s 1970s 1980s 1990s 2000s 2010s

Figure 3: Maps with four main clusters over time 1960 - 2011

3.2 Cluster descriptions using socio-demographic indicators

For additional cluster descriptions we used some socio-demographic indicators to be able to link the clusters with classical development characteristics. Some of them are directly related with the age-sex distribution (e.g. % of children in the population, % of elderly in the population), while others are relevant to the development stage of the sub-national areas in the clusters.

Ninos Niños 0-14, % de poblacion - Children 0-14, % of population
Ancianos Ancianos 60+, % de poblacion - Elderly 60+ % of population
Ninos/Anc Ninos por anciano (Children per elderly person)
Agric Agricultura, % de 15-59 trabajando - Agriculture, % of 15-59 working
Urbano Residencia urbana, % de poblacion - Urban residence, % of population
MujEcAc Mujeres % económicamente activa (15-59) - Women, % economically active (15-59)
EdPrim Educacion primaria (o mas), % de 15-59 - Achieved at least primary education, % of 15-59

4 main clusters	# of areas	Ninos	Ancianos	Ninos_Anc	Agric	Urbano	MujAcEc	EdPrim
C1_4	344	45%	5%	8.7	52%	45%	19%	26%
C2_4	493	39%	7%	5.9	28%	65%	30%	55%
C3_4	328	32%	8%	3.9	15%	79%	38%	70%
C4_4	231	25%	12%	2.0	10%	89%	54%	79%

Table 2: Descriptions of 4 main clusters of sub-national areas from Latin America and the Caribbean with socio-demographic indicators.

11 clusters	# of areas	Ninos	Ancianos	Ninos_Anc	Agric	Urbano	MujAcEc	EdPrim
C01_11	114	48%	5%	9.7	55%	41%	15%	21%
C02_11	230	45%	5%	8.6	51%	46%	20%	21%
C03_11	73	42%	7%	5.8	41%	48%	31%	34%
C04_11	69	42%	7%	6.3	36%	57%	23%	52%
C05_11	141	39%	5%	7.1.	25%	69%	31%	52%
C06_11	210	37%	7%	5.1.	28%	65%	31%	52%
C07_11	152	32%	7%	4.3.	13%	83%	40%	67%
C08_11	176	31%	9%	3.4.	20%	71%	34%	74%
C09_11	86	27%	13%	2.0	9%	89%	48%	82%
C10_11	69	24%	11%	2.2	10%	89%	56%	77%
C11_11	76	21%	19%	1.1	2%	98%	55%	92%

Table 3: Descriptions of 11 clusters of sub-national areas from Latin America and the Caribbean with socio-demographic indicators.

From Table 2 and Table 3 we can clearly see connections between the obtained clusters and the values of development indicators. The ordering rather well corresponds to the monotonic change for most of the indicators. But since the process of demographic transition is not associated with these variables according to a single relationship, a precise relationship between age-sex structure and socio-demographic variables should not be expected, especially when considering single DAMs. For example, a large number of children does not necessarily indicate a less developed stage – it might be caused by the migration of the working population into the area, especially when smaller areas are observed which are most sensitive to migration influences.

3.3 Movements over time among clusters for each sub-national area

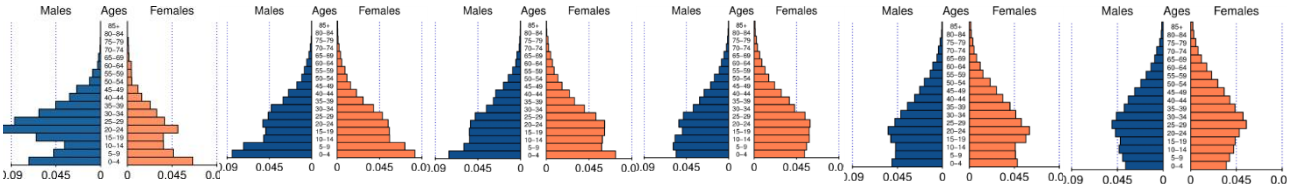
We observed changes of age-sex distributions by counting transitions from one cluster to another for each pair of contiguous censuses. For example the City of Buenos Aires' results for 1970, 1980, 1991, 2001 and 2010 provide data for four transitions. Cuba, with results in the IPUMS database only for 2002, does not contribute to this analysis.

	4 main clusters	**11more detailed clusters**
same	**644**	**422**
higher	**418**	**622**
(+1)	408	154
(+2)	10	223
(+3)		181
(+4)		57
(+5)		5
(+6)		2
lower	**13**	**31**
(-1)	13	20
(-2)		6
(-3)		3
(-4)		0
(-5)		2

Table 4: Movements of sub-national areas from Latin America and the Caribbean from 1960 to 2011 within 4 main and 11 more detailed clusters over time. The numbers in the brackets indicate for how many clusters (+ to the right and – to the left) they moved.

Counting of movements shows that most of the areas stayed in the same cluster over time or moved to the right to a more developed stage, although some rare areas moved also to the left to the clusters with socio-demographic characteristics that describe a less developed stage.

Besides observation of how sub-national areas move through clusters in general, we are sometimes interested to observe changes of the population pyramid of an individual sub-national area and connect it with the clusters. Here, we demonstrate this with the Brasilian sub-national area Distrito Federal, the Brasilian capital.



1960 (C3_4, C07_11) | 1970 (C2_4, C05_11) | 1980 (C2_4, C05_11) | 1991 (C3_4, C07_11) | 2000 (C3_4, C07_11) | 2010 (C4_4, C10_11)

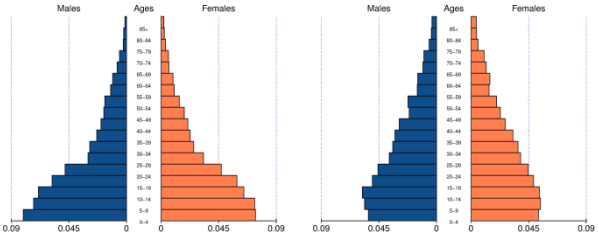
Figure 4: Population pyramids for the sub-national area **Distrito Federal (Brazil)** 1960 – 2010 with the clusters to which they belong.

In Figure 3 we can see very asymmetrical shape of the population pyramid in the first year (1960), strongly influenced by the population of the included Brasilian capital Brasilia, founded in 1960. In the following two decades (1970 and 1980) there is still a large number of children in the first age group, but that number decreased noticeably from 1991 on. We presume that the 1960 population includes many male construction workers, some temporarily in the area, while by 1970 and 1980 the shape is influenced by incoming government workers, their partners and their newly born children. Due to its shapes, this area was included in the third (out of four) main clusters in the year 1960, than moved and remained in the second out of four clusters in the next two decades, moved to the right, i.e. to the third out of four clusters in 1991, where stayed also in 2000, and finaly ended in the last out of four clusters. From clustering results we detected the unusual shape of this area in 1960 which can be explained with the additional local knowledge – the foundation of Brasilia.

Observing movements among 4 main clusters there are 418 areas that moved over time to the right, among them 10 moved for 2 clusters. On the other hand, there are all together only 13 areas that moved to the left and all of them moved for only one cluster (one of them being the Distrito Federal of Brazil between 1960 and 1970).

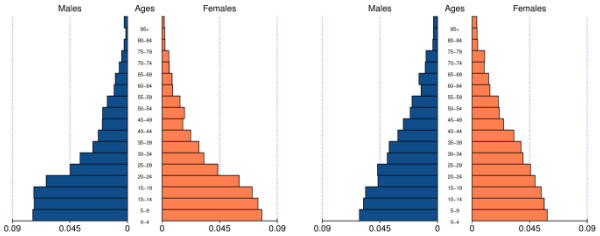
Observing more detailed clusters, even more areas moved to the right (622 movements). The largest "jumps", i.e. six clusters to the right, are noticed in 2 sub-national areas, both moved (from C02_11 in 1981 to C08_11 in 2002): Duarte [Province: Dominican Republic] and Peravia and San José de Ocoa [Province: Dominican Republic].

Duarte
[Province: Dominican Republic]



1981 (C02_11) | 2002 (C08_11)

Peravia and San José de Ocoa
[Province: Dominican Republic]



1981 (C02_11) | 2002 (C08_11)

Figure 5: Population pyramids for the sub-national areas with the biggest movements to the right (+6): Duarte and Peravia and San José de Ocoa [Province: Dominican Republic] in 1981 and 2002.

Both jumps are from 1981 to 2002 and they can be explained with the large time lap between data points (21 years) and the reduction of fertility at the time of the earlier census.

The largest to the left were detected in 2 areas: Aguascalientes [State: Mexico] and Baja California Sur [State: Meico]. Both moved for 5 clusters to the left from C06_11 in 1960 into C01_11 in 1970.

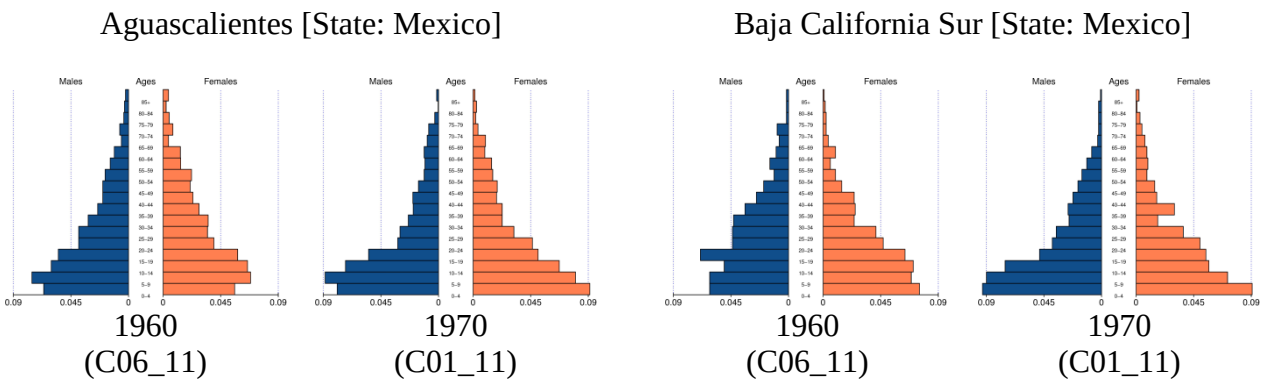


Figure 6: Population pyramids for the sub-national areas with the biggest movements into the lower numbered clusters (-5): Aguascalientes and Baja California Sur [State: Mexico] in 1960 and 1970.

From Figure 4 we see rather strange shapes. We do not yet know the reason for these two changes.

With the clustering method we were able to detect main stream of the changes of age-sex structures. It can be seen that in most of the areas changes over time showed mainly improvements. On the other hand, the clustering method revealed also some areas with the most deviating shapes which deserve to be studied separately with additional local knowledge.

3.4 Average dissimilarities in each decade

To observe average dissimilarities between age-sex structures (relative to the area population) over time, we divided the censuses by decades. Decade 1960s includes censuses dated from 1960 to 1969, decade 1970s includes censuses dated from 1970 to 1979 and so on. For Mexico there are data for two years in 1990s and in 2000s (Appendix, table A2). We excluded from the calculation data from Mexico for 1995 and 2005. For the same reasons we also excluded from decade 2000s data from Puerto Rico for 2005. For the rest of the sub-national areas we calculated average dissimilarity for each decade, where we also here (as in the clustering approach) used squared Euclidean distance to be able to connect the values with the component variances.

The average dissimilarities in each decade are the following (all average dissimilarities are multiplied by 10,000 for easier comparisons, as if the age-sex distributions were expressed as percentages of the population): 24.8 in 1960s, 26.2 in 1970s, 27.2 in 1980s, 24.3 in 1990s, 2.42 in 2000, and 18.5 in 2010. These values cannot be compared since we don't have the same sub-national areas in each decade. Therefore, we made comparisons for each pair of contiguous decades separately, for each pair using only the DAMs that are recorded in both decades. We then calculated the average dissimilarity between all the sub-national areas in the first decade and then the second decade:
between 1960 and 1970 (data 135 common sub-national areas) from 24.8 to 25.9,
between 1970 and 1980 (data 181 common sub-national areas) from 29.8 to 26.4,
between 1980 and 1990 (data 172 common sub-national areas) from 28.6 to 24.9,
between 1990 and 2000 (data 267 common sub-national areas) from 24.3 to 21.1,
between 2000 and 2010 (data 159 common sub-national areas) from 22.9 to 18.5.

As we can see the only increase of the average dissimilarity between sub-national areas is from 1960 to 1970. In all other pairs of decades (from 1970 to 1980, from 1980 to 1990, from 1990 to 2000, and also from 2000 to 2010) the averages decreased. We can say that in each decade except from 1960 to 1970 the observed age-sex structures had become more similar. The biggest reduction in average difference between areas is detected from 2000 to 2010.

If we compare sub-national areas from 1960 with the same areas in 2010, there are 111 common areas in our data set and the dissimilarity between their relative age-sex distributions (measured with squared Euclidean distance, multiplied with 1000 for comparison reasons) decreased from 27.8 in 1960 on 14.9 in 2010.

3.5. Dissimilarities between sub-national areas inside countries and inside clusters

Another important issue for our research is to study the similarity within each country and between areas in the same cluster which are in different countries. To measure dissimilarities between clusters and between the age-sex distributions of each DAM we considered the average dissimilarity of age-sex structures:

- between areas in a country,

- between all 1396 areas,

- between a country's areas and their country's average age-sex structure,

- between a country's areas and their cluster's average age-sex structure (for 4 and 11 clusters).

Formulae for the calculations are presented in Appendix B.

Squared Euclidean distance is simply the sum of the 36 squared differences between age-sex proportions of two areas. It is also worthwhile measuring the Euclidean distance by taking the square root, to return to the scale of the original measurements. This has similarities to the root mean square error (RMSE) in other statistical evaluations. The values for squared Euclidean distance are multiplied by 10,000 and the Euclidean distance by 100, for easier comparisons.

These are the results as if we had used the percentage age-sex distributions rather than proportions.

When measuring with squared Euclidean distance, the average of the dissimilarities among all pairs of sub-national areas, multiplied by 10,000, is 37.24; and 5.446 when measured with Euclidean distance. The average dissimilarities between areas inside each cluster are presented in Table 5 and Table 6 which are helpful for understanding how compact the clusters are.

4 main clusters	# of areas	Dissimilarity between areas in a country – squared Euclidean distance	Relative to dissimilarity for the whole continent (37.2)	Dissimilarity between areas in a country – Euclidean distance	Relative to dissimilarity for the whole continent (5.4)
C1_4	344	8.72	23%	2.81	52%
C2_4	493	11.09	30%	3.16	58%
C3_4	328	9.90	27%	2.86	52%
C4_4	231	16.88	45%	3.88	71%

Table 5: The average dissimilarities between age-sex structures of the sub-national areas in four main clusters, measured with squared Euclidean distance and with Euclidean distance.

From Table 5 can be seen that the biggest differences among areas are in the fourth cluster, which can be detected also from the hierarchy in Figure 1. As also expected the measures confirmed that more detailed clusters are also more homogenous (Table 6).

11 clusters	# of areas	Dissimilarity between areas in a country – squared Euclidean distance	Relative to dissimilarity for the whole continent (37.2)	Dissimilarity between areas in a country – Euclidean distance	Relative to dissimilarity for the whole continent (5.4)
C01_11	114	7.48	20%	2.62	48%
C02_11	230	6.59	18%	2.45	45%
C03_11	73	6.61	18%	2.47	45%
C04_11	69	6.56	18%	2.40	44%
C05_11	141	7.19	19%	2.52	46%
C06_11	210	7.13	19%	2.55	47%
C07_11	152	10.23	27%	2.78	51%
C08_11	176	6.27	17%	2.38	44%
C09_11	86	7.60	20%	2.65	49%
C10_11	69	10.39	28%	2.91	53%
C11_11	76	11.62	31%	3.23	59%

Table 6: The average dissimilarities between age-sex structures of the sub-national areas in eleven more detailed clusters, measured with squared Euclidean distance and with Euclidean distance.

Comparison of the averages of the dissimilarities among all pairs of sub-national areas from each country with the overall average for the continent gives us the information about the "closeness" of the units inside each country (Table 7). The country with least variation in age-structure between its DAM is Cuba, where the average dissimilarity is only 9% of the overall continental average. The biggest ratio is in Costa Rica, where differences between its DAM's age structures are 10% more than the overall continental variation. From the table A3 in the Appendix can be seen that all 15 areas from Cuba are in the same cluster out of 11 obtained clusters (in C10_11), while 35 areas from Costa Rica can be found in 8 out of 11 clusters. Using this measure, quite big differences among sub-national areas can be found also in Brazil, Dominican Republic, and Panama. While it is certain that Cuba's age-structure is more homogeneous than other countries, some caution must be applied, as all the Cuban data are from one year. In our next version of this paper we will compare each area's age-sex structure with the country's average for the same year, so that our comparisons between countries do not depend on the number of years involved.

As the second measure we used average dissimilarity between sub-national areas and their country average. The values are multiplied by 10,000 for easier comparison. This measure is simply half the first measure because we measure dissimilarity with squared Euclidean distance as a dissimilarity. However, it is useful to then see whether a country's areas are closer to their country average or to their cluster. The largest values are in the same countries as before, i.e. in Costa Rica, Brazil, Dominican Republic, and Panama.

Further we want to observe if the units, i.e. age-sex structures of sub-national areas, are closer to the country average or to the nearest cluster average. The largest value for average distance from cluster is in Cuba. As we noticed before, there are the smallest differences among areas in Cuba and all areas are in the same cluster. But in the same cluster are many other areas which influence the cluster's average so much that all Cuba's areas become rather "far away" from it. This also suggests that Cuba's area age-sex compositions have special characteristics, and may become a cluster on their own if more detailed analysis were undertaken.

Of course each sub-national area is closer to its nearest cluster out of 11 than out of 4 clusters, and so the country distances to each DAM's cluster out of 11 is smaller than to the distances to each DAM's cluster out of 4.

For each country we also compared the average dissimilarity of its areas from their cluster to the average dissimilarity from their country. Again Cuba stands out in Table 7: the average dissimilarity between areas in Cuba to their cluster (in the 4-cluster typology) is 946% of their distance to Cuba's average, i.e. more than nine times as large. But the ratios for all other countries are much smaller and near to or less than 100%. El Salvador, Jamaica and Uruguay have average dissimilarities to their clusters approximately the same as the dissimilarity to their country's averages. On the other hand, Brazil, Costa Rica and Mexico confirm large dissimilarities inside the countries, because the areas average distance to their cluster was about 30% or less than their distance to their country. However, some caution must be exercised as these three are among the countries that are represented in five or more decades, so their 'country average' is taken over a wide span of time.

When we choose more detailed clustering in 11 clusters, we can see that only in two countries the average dissimilarity of their sub-national areas to the nearest cluster are bigger than to the

country's average (Cuba with 677% and Haiti with 107%). In all other countries areas are on average closer to the clusters' average than to the country 's average. In Brazil, Mexico, and Panama, this ratio is the smallest – even less than one quarter.

Comparisons between age-sex distributions based on Euclidean distance produce very similar conclusions to those with squared Euclidean distance. Although this is not necessarily always so – it is in this case and so we have not reproduced tables 5 and 6 using Euclidean distance.

Country	Country code	N of cases	Average dissimilarity between areas in a country	Relative to dissimilarity for the whole continent (37.2)	Between a country's areas and their country age-sex structure (a)	Between a country's areas and their 4-cluster age-sex structure (b)	Ratio (b)/(a)	Between a country's areas and their 11-cluster age-sex structure (c)	Ratio (c)/(a)
Argentina	32	117	22.35	60%	11.18	7.12	64%	4.94	44%
Bolivia	68	25	8.77	24%	4.39	4.72	108%	3.21	73%
Brazil	76	138	34.26	92%	17.13	5.28	31%	3.69	22%
Chile	152	39	25.16	68%	12.58	5.79	46%	3.88	31%
Colombia	170	112	25.45	68%	12.72	4.55	36%	3.31	26%
Costa Rica	188	35	40.80	110%	20.40	5.93	29%	3.65	18%
Cuba	192	15	3.41	9%	1.70	16.12	946%	11.54	677%
Dominican Republic	214	122	34.16	92%	17.08	5.78	34%	4.08	24%
Ecuador	218	80	19.62	53%	9.81	5.32	54%	3.25	33%
El Salvador	222	28	10.80	29%	5.40	6.07	112%	3.75	69%
Haiti	332	12	11.29	30%	5.64	8.04	142%	6.03	107%
Jamaica	388	42	14.21	38%	7.10	7.20	101%	5.25	74%
Mexico	484	222	24.47	66%	12.24	3.55	29%	2.73	22%
Nicaragua	558	45	20.00	54%	10.00	5.09	51%	3.83	38%
Panama	591	41	32.73	88%	16.37	6.00	37%	3.78	23%
Paraguay	600	40	22.29	60%	11.14	6.00	54%	5.01	45%
Peru	604	50	19.53	52%	9.77	4.77	49%	3.22	33%
Puerto Rico	630	31	22.99	62%	11.49	11.00	96%	6.72	58%
Uruguay	858	114	14.66	39%	7.33	7.38	101%	4.33	59%
Venezuela	862	88	19.46	52%	9.73	3.64	37%	2.50	26%

Table 7: Some dissimilarity measures for the countries (the calculations are described in Appendix B) based on squared Euclidean distance.

4. Discussion

This analysis has taken a novel approach to understanding variation and demographic trends by grouping area age-sex pyramids using cluster analysis. Latin American and Caribbean areas' membership of clusters of similar age-sex composition reveals some countries which are very homogenous (Cuba, Puerto Rico), and others whose areas are more like those of different countries than the average of their own country (Costa Rica, Brazil, Mexico).

The clusters are distinguished by their percentage of young and of elderly and by socio-economic variables including the percentage that work in agriculture, the percentage of women working, and the percentage achieving primary education. They clearly reflect different levels of progress both in the demographic transition and in economic development. This idea of progress is confirmed by the movement of individual areas across time which is generally along an order of the clusters from youthful to older and from low to higher levels of socio-economic development.

Exceptions to progressive change with time are also instructive, sometimes involving shocks of migration or catastrophe to particular areas.

The regularity of changing age-sex structures suggests that their variation fairly measures social inequalities. It is less clear, as yet, how the regularity could be of use in demographic projections. A projection of an area's population total might be given the age-structure indicated by the latest census age-structure and the progression over time suggested in this report.

References:

Batagelj, V., Kejžar, N., and Korenjak-Černe, S. (2015) Clustering of Modal Valued Symbolic Data. ArXiv e-prints, 1507.06683, July 2015.

Batagelj, V., and Kejžar, N. (2010) `clamix` - Clustering symbolic objects R package. https://r-forge.r-project.org/projects/clamix/, 2010.

Gonzalez L and Torres E. (2012) Estimaciones de Población en Áreas Menores en América Latina: revisión de métodos utilizados. In: Cavenaghi S (ed) Estimaciones y proyecciones de población en América Latina: desafíos de una agenda pendiente. Rio de Janeiro: ALAP; Serie e-Investigaciones N.º 2, 105-138.
http://www.alapop.org/alap/index.php?option=com_content&view=article&id=1210&Itemid=573

IPUMS: Minnesota Population Center. *Integrated Public Use Microdata Series, International: Version 6.4* [Machine-readable database]. Minneapolis: University of Minnesota, 2015.
https://international.ipums.org/international/

IPUMS: personal communication fom Joe Grover to Ludi Simpson, 15 August 2016.

Jannuzzi PdM. (2012) Proyecciones de Población y Políticas Públicas: Importancia y Desafíos de las Nuevas Agendas. In: Cavenaghi S (ed) Estimaciones y proyecciones de población en América Latina: desafíos de una agenda pendiente. Rio de Janeiro: ALAP; Serie e-Investigaciones N.º 2,

87-104.
http://www.alapop.org/alap/index.php?option=com_content&view=article&id=1210&Itemid=573

Korenjak-Černe S., Kejžar, N., and Batagelj, V. (2014) A weighted clustering of population pyramids for the world's countries, 1996, 2001, 2006. Population Studies: A Journal of Demography.

University of Manchester (2016) Comparative subnational demographic development in Latin America and the Caribbean (website), http://www.cmist.manchester.ac.uk/research/projects/s-alyc/

Appendix A

A1: 48 excluded IPUMS sub-national areas with missing data (NA)

DAM_code+Year:

"32058 1970" "32078 1970" "32094 1970" "32099 1970" "68009 1976"
"68009 1992" "76011 1970" "76014 1970" "152012 1960" "152099 1960"
"152099 1970" "170018 1964" "170081 1964" "170086 1964" "170088 1964"
"170091 1964" "170095 1964" "170081 1973" "170091 1973" "170095 1973"
"170091 1985" "170095 1985" "170091 1993" "214010 1960" "214015 1960"
"214016 1970" "218014 1962" "218016 1962" "218019 1962" "218021 1962"
"218099 1962" "484023 1960" "484023 1970" "591005 1960" "591005 1962"
"600001 1962" "600002 1962" "600007 1962" "600008 1962" "600009 1962"
"600010 1962" "600011 1962" "600013 1962" "600015 1962" "600099 1962"
"600099 1972" "662 1980" "662 1991"

Country * nearYear Crosstabulation

Count

		nearYear								
		1960-1962	1963-1964	1970-1972	1973-1976	1980-1982	1984-1985	1990-1992	1993-1996	Total
Country	Argentina	0	0	4	0	0	0	0	0	4
	Bolivia	0	0	0	1	0	0	1	0	2
	Brazil	0	0	2	0	0	0	0	0	2
	Chile	2	0	1	0	0	0	0	0	3
	Colombia	0	6	0	3	0	2	0	1	12
	Dominican Republic	2	0	1	0	0	0	0	0	3
	Ecuador	5	0	0	0	0	0	0	0	5
	Mexico	1	0	1	0	0	0	0	0	2
	Panama	1	0	0	0	0	0	0	0	1
	Paraguay	11	0	1	0	0	0	0	0	12
	Saint Lucia	0	0	0	0	1	0	1	0	2
Total		22	6	10	4	1	2	2	1	48

Clustering was done on the remaining 1396 sub-national areas (DAMs) - noNA.

A2: Frequencies (number of sub-national areas – DAM) by nearYear for 1396 areas (these without NAs)

Country * nearYear Crosstabulation

		nearYear											Total
		1960-1962	1963-1964	1970-1972	1073-1976	1980-1982	1984-1985	1990-1992	1993-1996	2000-2002	2003-2007	2010-2011	
Country	Argentina	0	0	21	0	24	0	24	0	24	0	24	117
	Bolivia	0	0	0	8	0	0	8	0	9	0	0	25
	Brazil	15	0	23	0	25	0	25	0	25	0	25	138
	Chile	7	0	8	0	8	0	8	0	8	0	0	39
	Colombia	0	18	0	22	0	23	0	24	0	25	0	112
	Costa Rica	0	7	0	7	0	7	0	0	7	0	7	35
	Cuba	0	0	0	0	0	0	0	0	15	0	0	15
	Dominican Republic	23	0	24	0	25	0	0	0	25	0	25	122
	Ecuador	10	0	0	14	14	0	14	0	14	0	14	80
	El Salvador	0	0	0	0	0	0	14	0	0	14	0	28
	Haiti	0	0	4	0	4	0	0	0	0	4	0	12
	Jamaica	0	0	0	0	14	0	14	0	14	0	0	42
	Mexico	31	0	31	0	0	0	32	32	32	32	32	222
	Nicaragua	0	0	15	0	0	0	0	15	0	15	0	45
	Panama	6	0	7	0	7	0	7	0	7	0	7	41
	Paraguay	0	0	10	0	10	0	10	0	10	0	0	40
	Peru	0	0	0	0	0	0	0	25	0	25	0	50
	Puerto Rico	0	0	1	0	6	0	6	0	6	6	6	31
	Uruguay	0	19	0	19	0	19	0	19	0	19	19	114
	Venezuela	0	0	22	0	22	0	22	0	22	0	0	88
Total		92	44	166	70	159	49	184	115	218	140	159	**1396**

A3: Frequencies (number of sub-national areas – DAM) in 11 clusters of 1396 areas (these without NAs), detected with Ward agglomerative hierarchical clustering (rel1var)

		C11_rel1var											Total
		C01_11	C02_11	C03_11	C04_11	C05_11	C06_11	C07_11	C08_11	C09_11	C10_11	C11_11	
Country	Argentina	0	0	18	0	1	30	4	27	29	3	5	117
	Bolivia	1	7	6	0	6	5	0	0	0	0	0	25
	Brazil	6	44	1	1	20	10	29	6	0	21	0	138
	Chile	0	0	9	0	1	9	9	2	0	9	0	39
	Colombia	12	24	0	4	22	17	13	19	0	1	0	112
	Costa Rica	5	5	0	4	7	2	2	4	0	6	0	35
	Cuba	0	0	0	0	0	0	0	0	0	15	0	15
	Dominican Republic	35	22	4	9	6	13	6	24	2	1	0	122
	Ecuador	9	14	19	3	6	15	1	13	0	0	0	80
	El Salvador	0	1	0	15	0	8	2	2	0	0	0	28
	Haiti	0	0	4	3	1	3	1	0	0	0	0	12
	Jamaica	0	0	0	11	4	10	7	10	0	0	0	42
	Mexico	16	40	3	12	22	30	52	41	0	6	0	222
	Nicaragua	11	20	0	1	4	7	2	0	0	0	0	45
	Panama	6	3	6	1	3	10	2	6	1	1	2	41
	Paraguay	3	19	1	3	2	8	3	1	0	0	0	40
	Peru	0	7	1	0	7	13	10	7	0	5	0	50
	Puerto Rico	0	0	0	0	0	3	1	4	2	0	21	31
	Uruguay	0	0	1	0	0	8	0	5	52	0	48	114
	Venezuela	10	24	0	2	29	9	8	5	0	1	0	88
Total		114	230	73	69	141	210	152	176	86	69	76	1396

APPENDIX B

CALCULATIONS FOR THE FIRST APPROACH (STRUCTURE REPRESENTATION WITH 1 VARIABLE)

The age-sex distribution of a population in sub-national area X at a specific census is represented with one 36-component vector (18 age groups for each sex), where the sum of all components equals 1:

$$\underset{\rho}{p}_X = [p_{X1}, p_{X2}, ..., p_{X36}]^T, \sum_{j=1}^{36} p_{Xj} = 1.$$

Dissimilarity between sub-national areas X and Y measured with

- squared Euclidean distance[1] $d(X,Y) = \left\| \underset{\rho}{p}_X - \underset{\rho}{p}_Y \right\|^2 = \sum_{j=1}^{36} (p_{Xj} - p_{Yj})^2$

- Euclidean distance $d(X,Y) = \left\| \underset{r}{p}_X - \underset{r}{p}_Y \right\| = \sqrt{\sum_{j=1}^{36} (p_{Xj} - p_{Yj})^2}$

(Average) Dissimilarity between areas in a country = $\frac{1}{nCo \cdot nCo} sum\_Co$

$$sum\_Co = \sum_{X \in Co} \sum_{Y \in Co} d(X,Y)$$

$nCo =$ number of sub-national areas X in the country Co

Dissimilarity between sub-national area X and representative (average) of the cluster Cl measured with

- squared Euclidean distance $d(X,Cl) = \left\| \underset{\rho}{p}_X - \underset{\rho}{p}_{Cl} \right\|^2 = \sum_{j=1}^{36} (p_{Xj} - p_{Clj})^2$

- Euclidean distance $d(X,Cl) = \left\| \underset{r}{p}_X - \underset{r}{p}_{Cl} \right\| = \sqrt{\sum_{j=1}^{36} (p_{Xj} - p_{Clj})^2}$

Dissimilarity between sub-national area X and average of the country Co measured with

[1] For the squared Euclidean distance, the average dissimilarity between areas represents the sum of components' variance, multiplied by 2.

- squared Euclidean distance $d(X,Co) = \left\| \underset{\rho}{p}_X - \underset{\rho}{p}_{Co} \right\|^2 = \sum_{j=1}^{36} (p_{Xj} - p_{Coj})^2$

- Euclidean distance $d(X,Co) = \left\| \underset{\mathrm{r}}{p}_X - \underset{\mathrm{r}}{p}_{Co} \right\| = \sqrt{\sum_{j=1}^{36} (p_{Xj} - p_{Coj})^2}$

(Average) Dissimilarity between a country's areas and their country age-sex structure (a) =

$$\frac{1}{nCo} sum\_d(X,Co)$$

$$sum\_d(X,Co) = \sum_{X \in Co} d(X,Co)$$

(Average) Dissimilarity between a country's areas and their 4-cluster age-sex structure (b) = $\frac{1}{nCo} sum\_d(X,4Cl)$

$$sum\_d(X,4Cl) = \sum_{X \in Co} d(X,4Cl)$$

(Average) Dissimilarity between a country's areas and their 11-cluster age-sex structure (c) = $\frac{1}{nCo} sum\_d(X,11Cl)$

$$sum\_d(X,11Cl) = \sum_{X \in Co} d(X,11Cl)$$