# Latin American age structure 1960-2011

## Estructura etaria en América Latina, 1960-2010

*Simona Korenjak-Černe*[1]
*Ludi Simpson*[2]

1 Assistant Professor at the University of Ljubljana, Faculty of Economics, Slovenia. <simona.cerne@ef.uni-lj.si>

2 Professor of Population Studies, University of Manchester, UK. <ludi.simpson@manchester.ac.uk>

*Abstract*

This chapter examines a sub-national time series of age-sex-structures for twenty countries in Latin America and the Caribbean, summarises the diversity and the socio-demographic associates of changing age-sex structures, and characterises the development of those age-sex structures over time. Ward's agglomerative hierarchical method clearly detects four main clusters of areas and more detailed clusters within them. Socio-demographic indicators show the main stream of improvements in the sense of a starting or deepening of the demographic transition of lower fertility and mortality, alongside a reduction in agricultural employment and increases in female participation in the labour force and participation in education. The clustering method also reveals some areas with unusual trajectories or pyramid shapes, which deserve to be studied separately with additional local knowledge. Comparison of average dissimilarity among age-sex structures between contiguous decades indicates that age structures have become more similar over time.

**Key words:** Age-sex structure. Sub-national. Clustering. Time series

*Resumen*

Este capítulo examina una serie temporal de estructuras por sexo y edad para veinte países de América Latina y el Caribe, a nivel subnacional. Esta serie resume la diversidad y los asociados sociodemográficos de las cambiantes estructuras por sexo y edad, y caracteriza el desarrollo de estas estructuras a través del tiempo. El método de aglomeración jerárquico de Ward detecta claramente cuatro grupos principales de áreas y clústeres más detallados dentro de ellos. Los indicadores sociodemográficos muestran la principal corriente de mejoras en el sentido de iniciar o profundizar la transición demográfica de menores tasas de fecundidad y mortalidad, junto con la reducción del empleo agrícola y el aumento de la participación femenina en la fuerza de trabajo y la participación en la educación. El método de aglomeración también revela algunas áreas con trayectorias o formas piramidales atípicas, que merecen ser estudiadas por separado con conocimiento local adicional. La comparación de la disimilitud promedio entre las estructuras por sexo y edad entre las décadas contiguas indica que las estructuras por edad se han vuelto más similares en el tiempo.

**Palabras claves:** Estructura por sexo y edad. Sub-nacional. Aglomerados. Serie temporal



## Introduction

As governments attempt to develop their nation's infrastructure, subnational demographic trends 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 (González and Torres, 2012).

The aims of this chapter are to examine sub-national time series of age-sex-structures for Latin America and the Caribbean, to summarise the diversity and the socio-demographic associates of changing age-sex structures, and to identify and characterise the development of those age-sex structures over time. 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 (Asociación Latinoamericana de Población (ALAP), 2017).

The paper first describes our data sources and our analytical approach. Analysis is then presented which identifies clear differences in age structure between areas and across time, which group into four main clusters of age-sex composition, with eleven sub-clusters. In the next section, the relationship of each cluster to socio-demographic variables shows an ordering of clusters from young to old which is also an ordering of social progress, away from agricultural dependence and accompanied by increased labour market participation by women, increased achievement of primary education, and increased residence in urban areas. In line with the demographic transition to lower fertility and lower mortality, the clusters ordered by age and social progress moves from a young age-sex pyramid that is triangular with a broad base, to older pyramids with a more rectangular profile. The following section examines areas' moves between clusters over time, and confirms that the ordering of social progress is also an ordering of time from the 1960s to the 2010s. Some specific areas follow a trajectory opposite to these general trends, that can only be explained by local knowledge. The final analytical section examines the average dissimilarity of age-sex structures. It allows us to conclude a divergence of age-sex structures during the 1960s, followed by slow but steady convergence since the 1970s. It also shows that the degree of homogeneity of age-sex structures differs between countries. The chapter ends with a brief discussion of the findings, focusing on the powerful illustration of general trends that it provides.

## Data and methodology

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 The Latin American and Caribbean Demographic Centre (Centro Latinoamericano y Caribeño de Demografía, 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 chapter, because many of them have been archived as sampled micro-data by the University of Minnesota (Integrated Public Use Microdata Series (IPUMS), 2015), which we use in this study.

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 within twenty countries without missing values, where all 'blank' values were considered as missing data. The number of areas in each country and by years are presented in Tables A1 and A2 in Appendix A. 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[3]. Another quarter of them are from Colombia and the other half of the excluded areas are from nine other countries, among them also both samples from Saint Lucia. These excluded 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+.

One advantage of the IPUMS datasets is their consistent sub-national time series. Where there have been boundary changes, areas have been aggregated until consistent boundaries are attained. We use the data from twenty countries for every year in which they are present in the IPUMS datasets, including nine from the 2010 round of national censuses. At the time of writing, sample data from a further four countries in the region for the 2010 round were about to be released via IPUMS. Other data are available direct from the countries' national statistical institutes but have not been used in this study.

We examine sub-national time series of age-sex-structures for twenty countries in Latin America and the Caribbean 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 indicators to observe the relationship between structure and selected indicators, and to identify and characterise the development of the structures over time.

Since we are interested in the diversity of the areas we chose a clustering method that looks at a clustering as an analysis of variance problem: Ward's agglomerative hierarchical clustering method based on squared Euclidean distance. 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 approach which is reported in this chapter, is based on a classical representation where each sub-national area (DAM, from its Spanish description: *Division Administrativa Mayor*) is represented by a single vector of 36 components (a percentage for each of 18 age groups and each sex adding to 100) 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-sex 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 (Batagelj and Kejžar, 2010).

In an alternative approach we represented the age-sex distributions of sub-national areas with two vectors – separate distributions of men and women over age groups. Agglomerative hierarchical clustering is weighted by the DAM population for each sex. Two vectors represent distributions of men and women over 18 five-year age groups. A weighted agglomerative clustering method (Korenjak-Černe *et al.*, 2015; Batagelj *et al.*, 2015) is used. The advantage of this approach would be that the weighting ensures that each cluster's average remains the age 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 within each sex are recognised in the clustering, not relative to the whole population in the area. Imbalances between the 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. When weighting by population size, areas with relatively large populations are considered distant from each other and from areas with small populations,

3 IPUMS (2016), personal communication fom Joe Grover to Ludi Simpson, 15 August 2016.

even if their relative age-sex structures are not very different. For this reason we focused this report on the results of the first approach.

We use simple socio-demographic indicators to link the clusters with classical development characteristics. Three indicators directly summarise the age-sex distribution, while the remaining four describe the socio-economic development of the sub-national areas. Other indicators could have been chosen but their close relation to the clusters is post-hoc justification for their use here.

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 DAM that move over time in ways that do not conform to a notion of gradual progress, where local knowledge is needed 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).

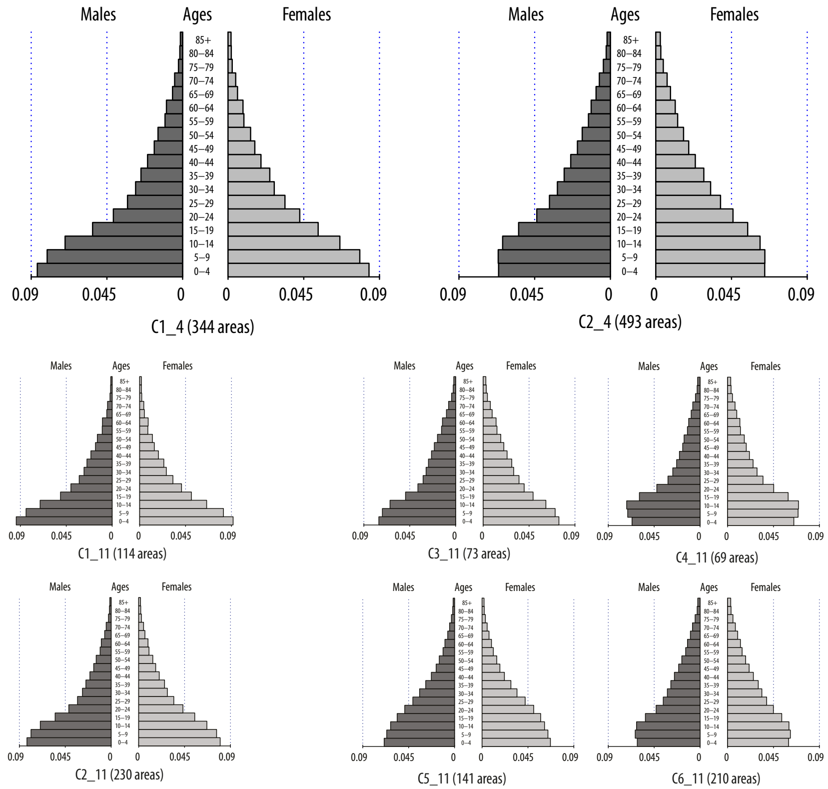
Since we do not have data for all sub-national areas for the same years, we made comparisons for each pair of contiguous decades to measure the variation of age structures across time. We calculated average dissimilarity between pairs of sub-national areas for which we have data in both decades. The 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.

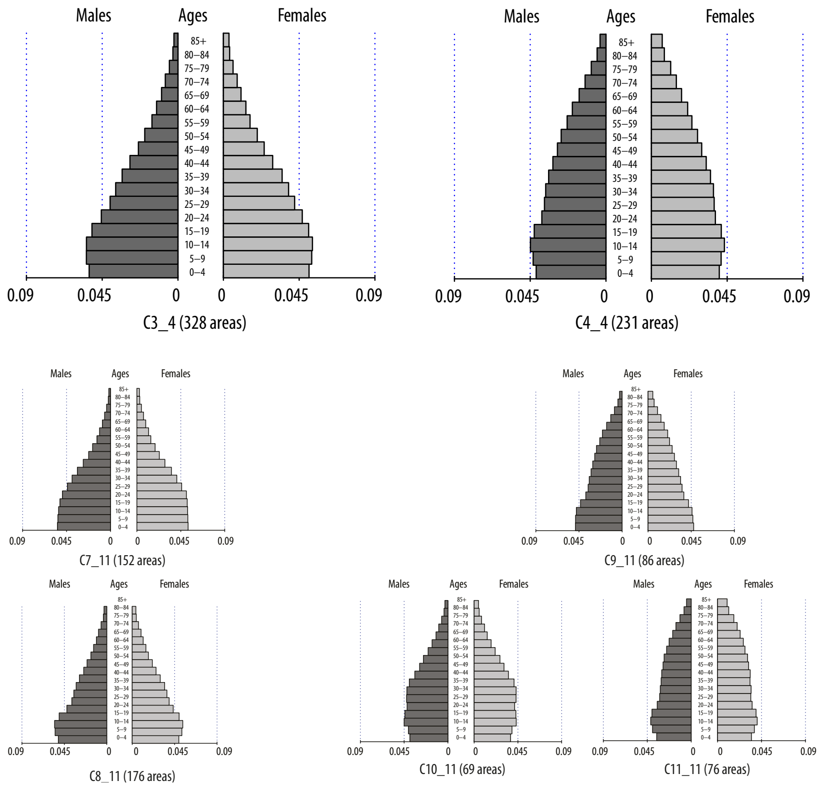
## Cluster analysis of age-sex distributions

Table 1 and Figure 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.



Figure 1
Latin America and the Caribbean from 1960 to 2011:
Unweighted average age-sex structure of 4 and 11 clusters






Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.
Note: 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.

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 (see below).

The four main clusters represented visually in Figure 2 are clearly representative of the stages of the demographic transition acknowledged in demographic literature. Cluster 1 includes areas without fertility or mortality reductions in which the base of the pyramid is relatively very wide and the reduction in cohort size with age is steady. Cluster 2 includes areas with a



reduction in fertility and child mortality in recent years, such that the base of the pyramid between ages 0 and 14 is vertical, while the steady reduction in older cohorts continues. In cluster 3 the reduction in adult mortality is evident from the cohorts at ages above 20 do not reduce as rapidly as in clusters 1 and 2. Further reduction in fertility reduction has meant that for the first time the youngest cohort is not the largest. Finally, cluster 4 represents an advanced stage of the demographic transition where the adult cohorts reduce more slowly and have a convex rather than concave shape, emphasising a more aged structure.

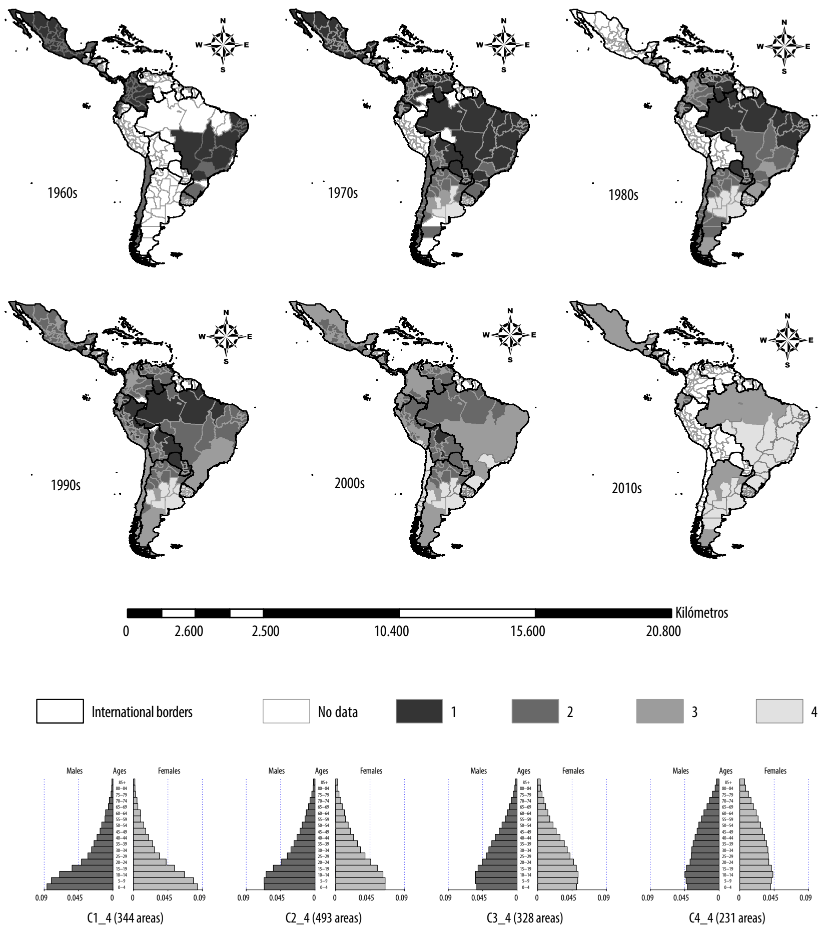
Table 1
Latin America and the Caribbean, from 1960 to 2011: Overlapping of the 4 main and 11 more detailed clusters obtained with the Ward hierarchical clustering of the age-sex distributions (relative to the whole population in the sub-national area) of the sub-national areas

| Count | | 4 clusters | | | | Total |
|---|---|---|---|---|---|---|
| | | C1_4 | C2_4 | C3_4 | C4_4 | |
| 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 |

Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.

The eleven clusters are subdivisions of the four main clusters, and are generally intermediate positions in the demographic transition. However, they also indicate that there is not a single smooth route through the demographic transition, because the pace of fertility and mortality reduction is not the same in each area, producing slightly different trajectories of age-sex structure. For example, cluster 4 in the 11-cluster solution, a subset of the main cluster 2, seems to have sharply lower fertility, seen from the reduction in child cohorts that is more reminiscent of main clusters 3 and 4, but not the improvement in adult mortality, seen in the very concave adult pyramid that is more like main cluster 1. The ordering of the sub-clusters is discussed in the next section, according to their association with socio-demographic variables.

Figure 2
Latin America and the Caribbean from 1960 to 2011:
Maps with four main clusters of age-sex pyramids




Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.



One of the most informative pictures about the change over time of sub-national age-sex structures is provided by maps that show distributions of clusters for each decade (Figure 2). The progress of demographic transition in the continent from generally young pre-transition structures in 1960 (main cluster 1) to generally advanced (main clusters 3 and 4) is clear, as is the early progress of sub-national areas in southern Uruguay and northern of Argentina. By the decade of 2000, areas with pre-transition structures are very limited, and seem to be associated with areas of substantial indigenous populations in Nicaragua, Ecuador and Bolivia, although we do not have the data to make this association explicitly. The maps also indicate the areas with and without data in ipums in each decade. Where there were two censuses in a single decade (Mexico, Puerto Rico), the earlier census has been used in the maps.

## Cluster descriptions using socio-demographic indicators

We used simple socio-demographic indicators to link the clusters with classical development characteristics. Three indicators directly summarise the age-sex distribution, while the remaining four describe the development stage of the sub-national areas. Other indicators could have been chosen but their close relation to the clusters is justification for their use here.

| | |
|---|---|
| Children | Children 0-14, % of population |
| Elderly | Elderly 60+ % of population |
| Chi/Eld | Children per elderly person |
| Agric | Agriculture, % of 15-59 working |
| Urban | Urban residence, % of population |
| FemEcAc | Females, % economically active 15-59 |
| EdPrim | Achieved at least primary education, % of 15-59 |

From Table 2 and Table 3 we can clearly see connections between the obtained clusters and the values of development indicators. The ordering of the main clusters corresponds to monotonic change for all of the indicators. The four main clusters progress through demographic transition described above also progresses through steadily lower proportions of children and of adults working in agriculture, and higher proportions of urban residence, women economically active and primary education achieved. The eleven clusters also achieve a near monotonic relationship with each of the socio-demographic indicators. But since the process of demographic transition is not associated with these variables according to a single or exact relationship, a precise relationship between age-sex structure and socio-demographic variables should not be expected, especially when considering single DAM. For example, a large number of children does not necessarily indicate a less developed stage – it might be caused by migration of the working population into the area, especially when smaller areas are observed which are most sensitive to migration influences.

Table 2
Latin America and the Caribbean from 1960 to 2011: Descriptions of 4 main clusters of sub-national areas with socio-demographic indicators

| 4 main clusters | # of areas | Children | Elderly | Chi/Eld | Agric | Urban | FemEcAc | 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% |

Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.

Table 3
Latin America and the Caribbean from 1960 to 2011: Descriptions of 11 clusters of sub-national areas with socio-demographic indicators

| 11 clusters | # of areas | Children | Elderly | Chi/Eld | Agric | Urban | FemEcAc | 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% |

Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.



## 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.

Table 4
Latin America and the Caribbean from 1960 to 2011: Movements of sub-national areas within 4 main and 11 more detailed clusters over time

| | 4 main clusters | 11 more 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 |

Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.
Note: 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 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 in the changes of the population pyramid of an individual sub-national area and its connection with the clusters. First we demonstrate this with the Brazilian sub-national area Distrito Federal, the Brazilian capital.

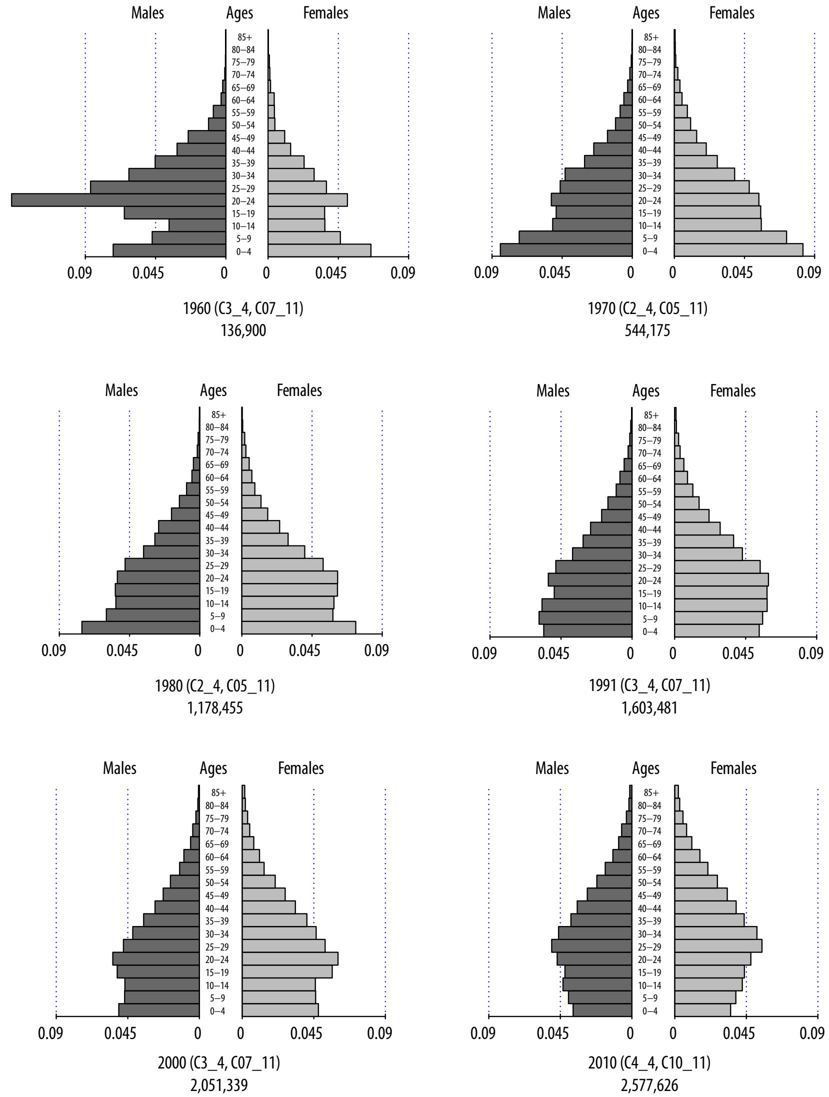
In Figure 3 we can see the very asymmetrical shape of the population pyramid in the first year (1960), strongly influenced by the population of the included Brazilian capital Brasilia, founded in 1960. In the following two decades (1970 and 1980) the population has grown rapidly. There is still a large proportion of children in the first age groups, but that proportion 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, then moved and remained in the second out of four clusters in the next two decades; it moved to the right, i.e. to the third out of four clusters in 1991, where it stayed also in 2000, and finally ended in the last out of four clusters. From our clustering results we detected the unusual trajectory and then 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 in Table 4, there are 418 areas that moved over time to the right, among them 10 moved for 2 clusters. On the other hand, there are 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 and therefore smaller clusters, there are naturally fewer that remain in the same cluster from one census to the next. Many 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 (Figure 4), both moved from C02_11 in 1981 to C08_11 in 2002: Duarte and Peravia and San José de Ocoa, both Provinces of the Dominican Republic.



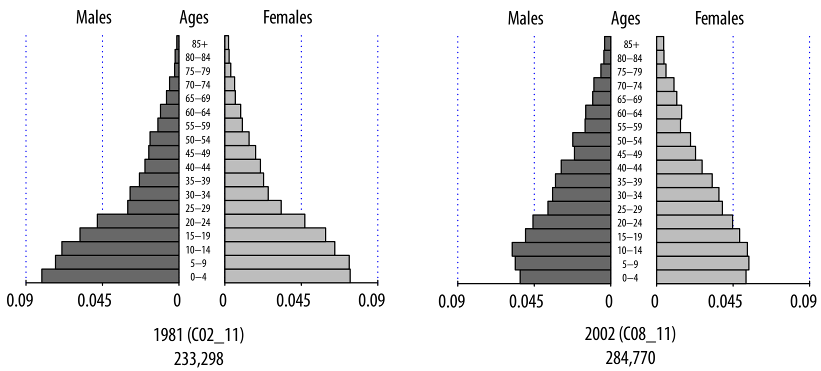
Figure 3
Distrito Federal (Brazil) 1960 – 2010: Population pyramids for the sub-national area with the clusters to which they belong, and the population size



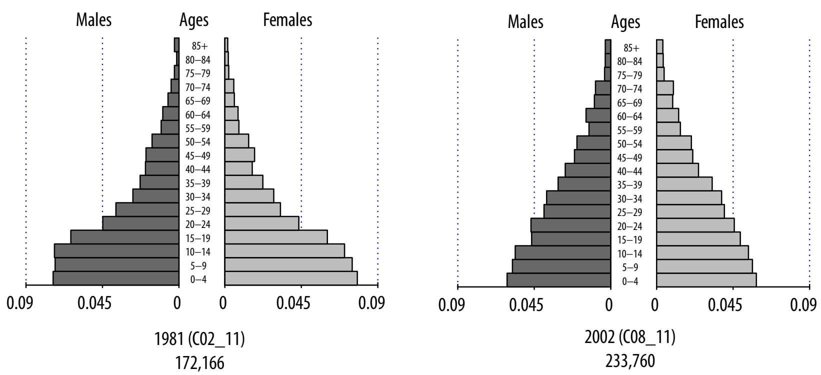

Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.

Figure 4
Duarte and Peravia and San José de Ocoa [Province: Dominican Republic] in 1981 and 2002: Population pyramids for the sub-national areas with the biggest movements to the right (+6) with the cluster to which they belong, and population size

Duarte [Province: Dominican Republic]




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




Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.

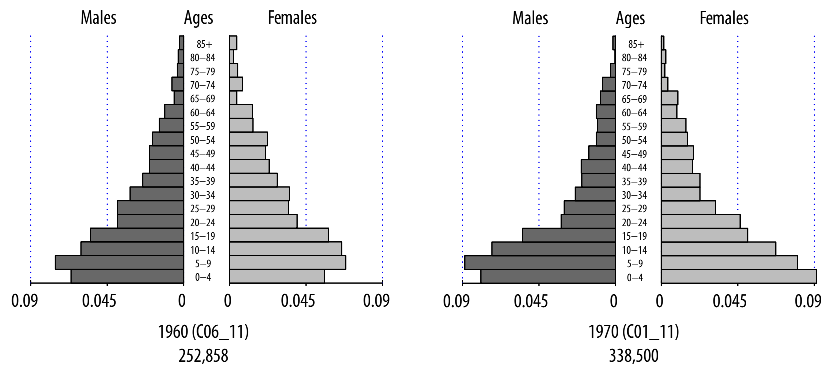
The large movement can be explained with the large time lap between data points (21 years) and the rapid reduction of fertility between the two censuses.

The largest movements to the left were detected in two states of Mexico between 1960 and 1970, Aguascalientes and Baja California Sur. Both moved for 5 clusters to the left from C06_11 into C01_11 in that single decade.

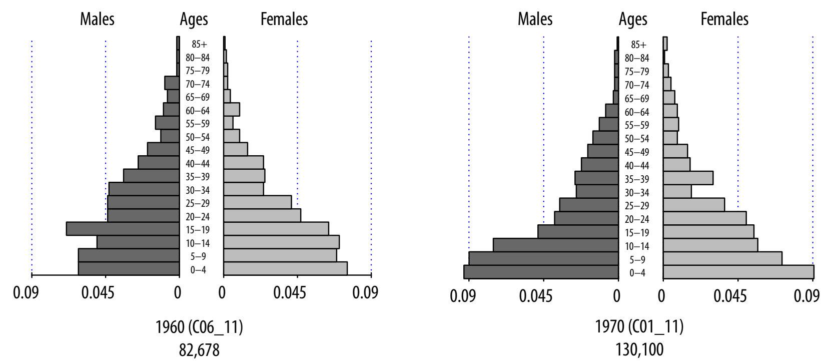


Figure 5
Aguascalientes and Baja California Sur [State: Mexico] in 1960 and 1970: Population pyramids for the sub-national areas with the biggest movements into the lower numbered clusters (-5) with the cluster to which they belong, and population size

Aguascalientes [State: Mexico]




Baja California Sur [State: Mexico]




Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.

From Figure 5 we see rather unusual shapes. We do not yet know the reason for these two changes, which may relate to specific migration or to errors in the census enumeration or processing.

In summary, with the clustering method we were able to detect the main stream of changes of age-sex structures. In most of the areas changes over time showed mainly improvements in the sense of a starting or deepening of the demographic transition of lower fertility and mortality. On the other hand, the clustering method also revealed some areas with unusual trajectories or pyramid shapes, which deserve to be studied separately with additional local knowledge.

## 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 censuses in two years in the 1990s and in the 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 in consistency with the clustering procedure and is also related with component variance. Component variance (calculated as an average of square distance of area percentage from country mean in this age-group) measures area deviation in each age-group. So the calculated average dissimilarity equals 2-times sum of variances in each component (each age-group). If it is divided by 2 and also by 36, it represents mean component variance. Its square root can be seen as mean standard deviation across the age groups.

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, 24.2 in 2000, and 18.5 in 2010. Since we don't have the same sub-national areas in each decade (see Figure 2) these values are not comparable. Therefore, we made comparisons for each pair of contiguous decades separately, for each pair using only the DAM that are recorded in both decades. We then calculated the average dissimilarity between the same sub-national areas in the first decade and then the second decade:

- between 1960 and 1970 (135 common sub-national areas) from 24.8 to 25.9,
- between 1970 and 1980 (181 common sub-national areas) from 29.8 to 26.4,
- between 1980 and 1990 (172 common sub-national areas) from 28.6 to 24.9,



- between 1990 and 2000 (267 common sub-national areas) from 24.3 to 21.1,
- between 2000 and 2010 (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 average dissimilarity 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 by 10,000 for comparison reasons) decreased from 27.8 in 1960 on 14.9 in 2010.

Similar comparisons were made for each country separately, and the results are given in Table 5. For some countries, census samples in decades before 1990 were incomplete because of missing data. For these entries in Table 5, the difference in average dissimilarity between the complete decade and the incomplete decade was computed from the areas that the two decades had in common, and the table completed using these differences between decades. For Latin America and the Caribbean, the census samples were different each decade; 2000 was taken as the 'anchor year' for the continental time series in the first line of Table 5 and the other years constructed from the differences between decades calculated from common areas as above.

In Argentina, Costa Rica, and Uruguay the average dissimilarity among sub-national areas decreased in each decade. The same can be noticed for Paraguay and Venezuela from 1980 on, and for Brazil and Mexico from 1990 on. On the other hand, we can see increasing average dissimilarities among sub-national areas in each decade for Bolivia and Nicaragua, and also for Panama from 1970 on. In some countries, the average dissimilarity varies a lot between years (e.g. Haiti has especially high differences in 1980, while Chile saw the average dissimilarity between its regions drop in 1990 to one third of its value in previous decades.). The explanation for changes within each country will require more detailed consideration of the data with additional local knowledge.

Table 5
Latin America and the Caribbean from 1960 to 2011: The average dissimilarity between age-sex structures of sub-national areas within each country in each decade

| | No. of areas in 2000 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010 |
|---|---|---|---|---|---|---|---|
| Latin America and Caribbean | | 33.4 | 34.5 | 31.1 | 27.4 | 24.2 | 19.8 |
| Argentina | 24 | | 29.88 | 21.86 | 17.24 | 13.29 | 9.31 |
| Bolivia | 8 | | 5.50 | | 5.84 | 6.31 | |
| Brazil | 25 | 25.37 | 8.16 | 13.51 | 13.65 | 12.21 | 10.75 |
| Chile | 8 | 12.02 | 8.97 | 9.51 | 3.04 | 2.47 | |
| Colombia | 24 | 8.78 | 9.15 | 9.72 | 11.06 | 8.21 | |
| Costa Rica | 7 | 6.63 | 5.82 | 5.05 | | 4.21 | 3.85 |
| Cuba | | | | | | 3.41 | |
| Dominican Republic | 25 | 8.87 | 7.84 | 9.03 | | 9.67 | 9.13 |
| Ecuador | 14 | 12.82 | 10.21 | 10.75 | 9.86 | 10.51 | 10.78 |
| El Salvador | 14 | | | | 6.09 | 6.06 | |
| Haiti | 4 | | 2.56 | 12.53 | | 5.43 | |
| Jamaica | 14 | | | 7.37 | 6.23 | 6.40 | |
| Mexico | 32 | 8.34 | 5.39 | | 6.30 | 5.86 | 4.92 |
| Nicaragua | 15 | | 4.78 | | 6.85 | 10.63 | |
| Panama | 7 | 14.10 | 11.76 | 15.86 | 16.35 | 17.62 | 19.08 |
| Paraguay | 10 | | 15.75 | 20.65 | 15.95 | 11.88 | |
| Peru | 25 | | | | 15.45 | 11.81 | |
| Puerto Rico | 6 | | | 12.16 | 7.76 | 5.70 | 6.70 |
| Uruguay | 19 | 12.76 | 9.85 | 7.41 | 6.86 | 6.51 | 6.18 |
| Venezuela | 22 | | 8.10 | 8.45 | 7.94 | 6.74 | |

Source: Analysis of age-sex structures of DAM, from national censuses held in IPUMS.
Note: All countries include a census sample in the 2000 decade. The number of DAM in that decade is shown. The extra censuses in 1995 (Mexico) and 2005 (Mexico and Puerto Rico) are not included. Where the areas in the dataset vary from decade to decade, a time series has been constructed (see text). The constructed values are given in the table in italics. Blank cells indicate no census sample in that decade.

Differences in average dissimilarity *between* countries are also of interest from Table 5 but should be treated with caution. They reflect not only demographic inequality, but administrative boundaries. As an illustration, the administrative merging of two areas with very different age-sex structures would produce a single area with intermediate age-sex structure, and



reduce the average dissimilarity in the country without changing any of its demographic circumstances.

Nonetheless, the average dissimilarity between DAM age-sex structure in Argentina, Brazil, Panama, Paraguay and Uruguay in 2000, were each well over twice as high as the dissimilarities between the DAM of Costa Rica, Chile or Cuba. Again, further analysis would be needed to understand the extent to which these country differences could be due to the impact of administrative boundaries in each country, to socio-economic inequality since we have shown a close relationship between socio-economic indicators and age-sex structure, and to different trajectories of the demographic transition within each country.

## Discussion

This analysis has taken a novel approach to understanding variation and demographic trends by grouping area age-sex pyramids using cluster analysis. The focus of our observations was diversity of the age-sex structures of sub-national areas (relative to the whole population in the observed area). We chose Ward's agglomertive hierarchical method based on squared Euclidean distance since this clustering method has the analysis of variance perspective. Four main clusters are clearly detected. The four clusters are clearly ordered and distinguished by a reducing percentage of young, an increasing proportionof elderly, a reduction of the percentage of workers in agriculture, and increases in 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. We extended observation to eleven more detailed sub-clusters that also shows the same pattern of changing age-sex structures.

While most areas have changed their age structure over time in the period between 1960 and 2011 in line with starting or deepening the demographic transition of lower fertility and mortality, the clustering method also reveals some areas with unusual trajectories or pyramid shapes. These require local knowledge for their explanation (for example, the Brazilian sub-national area Distrito Federal developed as the Brazilian capital). It is tempting to suggest that age structure might be predicted on the basis of the clear time-pattern shown in this study. In general diagnostic terms this is certainly a helpful observation. Planning can expect a continuation of ageing in subnational areas. However, the analysis is in the nature of a powerful demonstration of demographic trends, and has not been developed into a prognostic tool. The priority for projections of age structure should remain the improvement of vital statistics and censuses, and the development of sub-national cohort component demographic projections. The unusual cases of discontinuous development of age structures are a warning however, firstly that local knowledge can improve demographic projections if shocks of migration can be foreseen, but also that catastrophes or other shocks can often not be foreaseen but significantly affect the localage composition.

Some countries are very homogenous (Cuba, Chile, Costa Rica), and others are very diverse in their sub-national age-sex structure (Argentina, Brazil, Ecuador, Panama).

Our study has used data available for twenty countries, with a different selection of countries in each decade dependent on those national censuses which have been deposited with the IPUMS. In order to compare across decades the study has computed the average dissimilarity of age-sex structures among sub-national areas that were common in the study for adjacent decades. This indicates that age-sex structures have become steadily more similar since the 1970s. According to the evidence of these data, one can confirm a convergence in sub-national age-sex structures during the last 4 decades.The close association of other variables with age-sex structures suggests that variation in age composition fairly measures social inequalities in a country and over time.

We believe that obtained statistical results help as to identify some relations that in combination with additional local expert knowledge provide more in depth understanding of time changes of these areas.



## References:


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

Batagelj, V. and Kejžar, N. (2010), *clamix - Clustering symbolic objects R package*, in <https://r-forge.r-project.org/projects/clamix/>, accessed 31st May 2017.

González, L. y Torres, E. (2012), "Estimaciones de Población en Áreas Menores en América Latina: revisión de métodos utilizados", en 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.

IPUMS (2015), *Integrated Public Use Microdata Series, International: Version 6.4 [Machine-readable database]*. Minneapolis: Minnesota Population Center, University of Minnesota, in <https://international.ipums.org/international/>, accessed 31st May 2017.

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

Korenjak-Černe S., Kejžar, N. and Batagelj, V. (2015), "A weighted clustering of population pyramids for the world's countries, 1996, 2001, 2006", in *Population Studies: A Journal of Demography* Vol. 69, No. 1, p. 105-120.

ALAP (2017), *Análisis comparativo de las tendencias demográficas subnacionales en América Latina y el Caribe (s-ALyC)*, Asociación Latinoamericana de Población (website), <http://www.alapop.org/alap/index.php?option=com_content&view=article&id=1427&Itemid=744>, accessed 31st May 2017.

# Appendix A

Table 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”

| | 1960-1962 | 1963-1964 | 1970-1972 | 1973-1976 | 1980-1982 | 1984-1985 | 1990-1992 | 1993-1996 | Total |
|---|---|---|---|---|---|---|---|---|---|
| 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 (DAM).



Table A2:
Number of sub-national areas – DAM by groups of years for 1396 areas

| | 1960-1962 | 1963-1964 | 1970-1972 | 1973-1976 | 1980-1982 | 1984-1985 | 1990-1992 | 1993-1996 | 2000-2002 | 2003-2007 | 2010-2011 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

Table A3:
Number of sub-national areas – DAM in 11 clusters of 1396 areas

Detected with Ward agglomerative hierarchical clustering (relıvar)

| | C01 | C02 | C03 | C04 | C05 | C06 | C07 | C08 | C09 | C10 | C11 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 representing age-sex structure with one vector

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:

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

*Dissimilarity between sub-national areas X and Y measured with*

- squared Euclidean distance[4] $d(X,Y) = \|\vec{p}_X - \vec{p}_Y\|^2 = \sum_{j=1}^{36} (p_{Xj} - p_{Yj})^2$

- Euclidean distance $d(X,Y) = \|p_X - p_Y\| = \sqrt{\sum_{j=1}^{36} (p_{Xj} - p_{Yj})^2}$

(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 $d(X,Cl) = \|\vec{p}_X - \vec{p}_{Cl}\|^2 = \sum_{j=1}^{36} (p_{Xj} - p_{Clj})^2$

- Euclidean distance $d(X,Cl) = \|p_X - p_{Cl}\| = \sqrt{\sum_{j=1}^{36} (p_{Xj} - p_{Clj})^2}$

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

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

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

- Euclidean distance $d(X, Co) = \left\| p_X - p_{Co} \right\| = \sqrt{\sum_{j=1}^{36} (p_{Xj} - p_{Coj})^2}$

*(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) = $\frac{1}{nCo} sum\_d(X, 11Cl)$

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