Metodološki zvezki, Vol. 8, No. 1, 2011, 1-15

Clustering of Population Pyramids using Mallows' L^2 Distance

Katarina Košmelj[1] and Lynne Billard[2]

Abstract

In many real situations, data are collected/presented as histograms. Such examples are population pyramids, which present the age distribution of a population by gender for a particular country. The objective of this paper is to partition countries into homogenous groups according to the similarity of the shape of the population pyramids in each particular year and to observe the time-trend. We use a Mallows' L^2 distance for this purpose. A case study on East European countries in the period 1995-2015 is presented. The results reflect that the countries are becoming more and more similar and follow a pattern of aging populations. For the majority of countries, this process started long before 1990, for Kosovo, Albania and Macedonia it started after 1990.

1 Introduction

Often data come presented as histograms. Such examples are population pyramids which present the age distribution of a population by gender. Each age-variable is presented in the form of a histogram, the two gender histograms are plotted horizontally back-to-back, on the left for males and on the right for females. On the *y*-axis are age groups: the subintervals are usually five-year age groups; on the *x*-axis is the number of males/females or the corresponding proportion.

As an example, we shall consider data from the US Census Bureau for 14 Eastern European countries (EE) in the years 1995, 2000, 2005, 2010 and the "predicted data" for the year 2015 (http://www.census.gov/ipc/www/idb/informationGateway.php). The 14 EE countries are: Albania (AL), Bosnia and Herzegovina (BA), Bulgaria (BG), Czeck Republic (CZ), Croatia (HR), Hungary (HU), Kosovo (KO), Montenegro (ME), Macedonia (MK), Poland (PL), Romania (RO), Serbia (RS), Slovenia (SI), Slovakia (SK). These countries were chosen because very turbulent and dynamic changes were taking place in this region after the breakdown of the Eastern block in 1989 which changes were also revealed in their demographic status.

For illustration, population pyramids for Kosovo and Slovenia in 1995 are presented in Figure 1. For Kosovo a pyramid shape is apparent, indicating a high proportion of children and a low proportion of older people, thus high birth rate and high death rate and

[1] Biotechnical Faculty, University of Ljubljana, Slovenia; katarina.kosmelj@bf.uni-lj.si
[2] University of Georgia, Athens, USA; lynne@stat.uga.edu

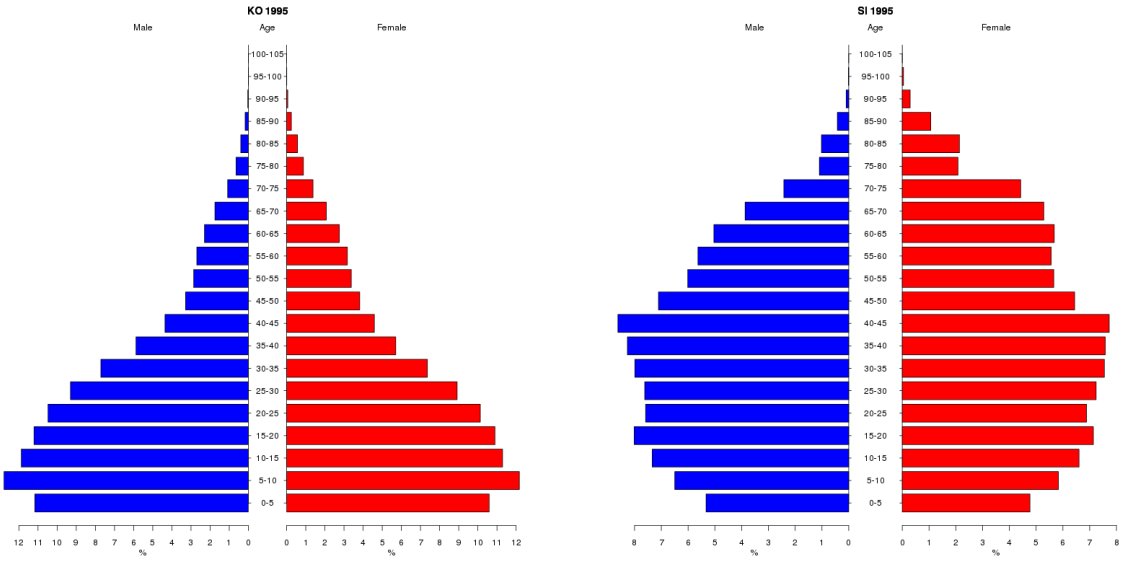

Figure 1: Population pyramid for Kosovo 1995 (left) and Slovenia 1995 (right).

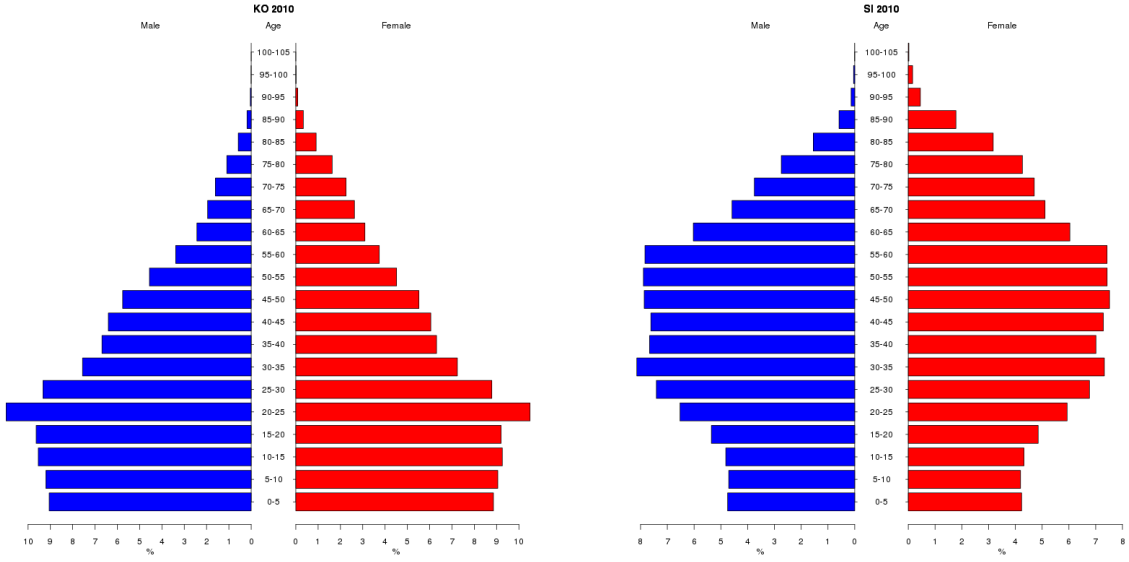

Figure 2: Population pyramid for Kosovo 2010 (left) and Slovenia 2010 (right).

a short life expectancy (i.e., "expansive" pyramid). The Slovenia pyramid reveals low birth rate, a low death rate and a long life expectancy with more females than males in the older age groups (i.e., "constrictive" pyramid).

In Figure 2, the population pyramids for the same two countries are presented for the year 2010. There are evident changes in their shapes. For Kosovo, each of the youngest age-cohort is smaller than the previous one, however the pyramid for Slovenia demonstrates a rapidly aging population with similar numbers of people in the youngest age-cohorts (Vertot, 2011).

The objective of this paper is to partition these 14 countries into homogenous groups according to the *similarity of the shape of the population pyramids* in each particular year and to observe the time-trend. Cluster analysis will be undertaken; therefore, an appropriate distance measure to meet this objective is needed. Countries described by population pyramids can be regarded as symbolic data objects (Billard and Diday, 2006) with two random variables, one presenting age for males and one for females. In the literature, several distances for histogram-type data can be found. Irpino and Verde (2006) proposed a new "Wasserstein based" distance (it is more correctly called Mallows' L^2 distance). Verde and Irpino (2007) analyzed different metrics in the dynamic clustering of histogram data. Korenjak-Černe et al. (2008) used Euclidean distance to cluster population pyramids. From the symbolic data setting, several distances for histograms are presented by Kim and Billard (2011). These distances are extended versions of Gowda-Diday, Ichino-Yaguchi, and De Carvalho distances for interval type data.

We have decided to apply the Mallows' L^2 distance for clustering of population pyramids for several reasons. As presented further on, this distance allows the constructions of a barycentric histogram which is an "optimal" cluster representative. It also allows to define a measure of total inertia which can be decomposed into the within and between inertia according to the Huygens theorem (we present the proof in the Appendix). Consequently, clustering results for different years can be compared and the time-trend assessed.

2 Methods

2.1 Wasserstein's distance and Mallows' distance

If F and G are the distribution functions of two random variables f and g, and F^{-1} and G^{-1} the corresponding quantile functions, the Wasserstein distance is defined as follows:

$$d_W(f,g) = \int_{-\infty}^{\infty} |F(t) - G(t)|dt = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt. \tag{2.1}$$

Mallows' L^2 distance (1972) is defined as follows:

$$d_M^2(f,g) = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt. \tag{2.2}$$

This distance can be considered as a natural extension of the Euclidean distance from point data to distribution data. More generally, Mallows' distance is an L^p distance defined for $p = [1, \infty]$, with $p = 2$ as in (2.2) being a special case. An interesting review of its properies is given in Levina and Bickel (2002). Historically, these distances were invented several times from different perspectives and can be found under different names (see review in Rüschendorf, 2001).

2.2 Mallows' L^2 distance for histograms

The histogram description $Y(u)$ of the object u is defined by H_u consecutive and non-overlapping intervals $I_{hu} = [\underline{y}_{hu}, \overline{y}_{hu})$, $h = 1, ..., H_u$, with the relative frequency π_{hu} as follows:

$$Y(u) = \{(I_{1u}, \pi_{1u}), ..., (I_{hu}, \pi_{hu}), ..., (I_{H_u u}, \pi_{H_u u})\}. \tag{2.3}$$

A cumulative relative frequency w_{hu} is associated with each interval: $w_{hu} = \sum_{j=1}^{h} \pi_{ju}$, $h = 1, ..., H_u$. Assuming a uniform density for each interval I_{hu}, we may describe the empirical distribution function and its inverse as piecewise linear functions:

$$\Psi_u(y) = w_{h-1,u} + \frac{w_{hu} - w_{h-1,u}}{\overline{y}_{hu} - \underline{y}_{hu}}(y - \underline{y}_{hu}), \quad \underline{y}_{hu} \leq y < \overline{y}_{hu}, \tag{2.4}$$

$$\Psi_u^{-1}(t) = \underline{y}_{hu} + \frac{\overline{y}_{hu} - \underline{y}_{hu}}{w_{hu} - w_{h-1,u}}(t - w_{h-1,u}), \quad w_{h-1,u} \leq t < w_{hu}. \tag{2.5}$$

In this context, the Mallows' L^2 distance (2.2) between the histograms of the objects u and v is written as follows:

$$d_M^2(Y(u), Y(v)) = \int_0^1 (\Psi_u^{-1}(t) - \Psi_v^{-1}(t))^2 dt. \tag{2.6}$$

In order to derive the Mallows' L^2 distance for the histogram setting, we follow the procedure proposed by Irpino and Verde (2006). The cumulative relative frequencies for $Y(u)$ and $Y(v)$ are set together : $\{w_{1u}, ..., w_{H_u u}, w_{1v}, ..., w_{H_v v}\}$. A zero value is added, then these values are sorted without repetitions and put in the vector $\mathbf{w}$ having $m + 1$ different components, $\mathbf{w} = [w_0 = 0, w_1, ..., w_l, ..., w_m = 1]$. Using subintervals $[w_{l-1}, w_l]$, $l = 1, ..., m$, we can decompose (2.6) into the following sum:

$$d_M^2(Y(u), Y(v)) = \sum_{l=1}^{m} \int_{w_{l-1}}^{w_l} (\Psi_u^{-1}(t) - \Psi_v^{-1}(t))^2 dt. \tag{2.7}$$

Each interval $[w_{l-1}, w_l], l = 1, ..., m$, defines two uniformly dense *quantile-intervals* on the abscissa axis: IQ_{lu} and IQ_{lv}, for object u and v, respectively. They are obtained as follows:

$$IQ_{lk} = [\Psi_k^{-1}(w_{l-1}), \Psi_k^{-1}(w_l)], \quad k = u, v. \tag{2.8}$$

By considering each point in a quantile-interval as a function of the interval center c, $c = (a + b)/2$, and interval radius r, $r = (b - a)/2$: $IQ[a, b] \Leftrightarrow IQ(z) = c + r(2z - 1)$ for $0 \leq z \leq 1$, expression (2.7) can be simplified. Final integration gives the following formula:

$$d_M^2(Y(u), Y(v)) = \sum_{l=1}^{m} \pi_l^* [(c_{lu} - c_{lv})^2 + \frac{1}{3}(r_{lu} - r_{lv})^2] \tag{2.9}$$

where $\pi_l^* = w_l - w_{l-1}$. Given p histogram variables, expression (2.9) is generalized assuming variables are independent:

$$d_M^2(Y(u), Y(v)) = \sum_{k=1}^{p} \sum_{l=1}^{m} \pi_l^{*(k)} [(c_{lu}^{(k)} - c_{lv}^{(k)})^2 + \frac{1}{3}(r_{lu}^{(k)} - r_{lv}^{(k)})^2]. \tag{2.10}$$

When the variables are dependent, a Mahalanobis type of this distance should be used (Verde and Irpino, 2008).

This distance has several useful properties in the histogram setting: it is easily calculable, it works also for subintervals of different sizes. Moreover, it allows for identification of the *barycentric (centroid)* histogram. Let m_n define the number of distinct values in the vector **w** defined by the n histogram objects, as described above. The corresponding barycentric histogram $Y(b)$ can be expressed in terms of a vector of m_n pairs (c_{lb}, r_{lb}), $l = 1, ..., m_n$:

$$c_{lb} = n^{-1} \sum_{i=1}^{n} c_{li}, \qquad r_{lb} = n^{-1} \sum_{i=1}^{n} r_{li} \tag{2.11}$$

as follows:

$$Y(b) = \{([c_{1b} - r_{1b}; c_{1b} + r_{1b}], \pi_1); ...; ([c_{m_n b} - r_{m_n b}; c_{m_n b} + r_{m_n b}], \pi_{m_n})\}. \tag{2.12}$$

2.2.1 Illustration

Let us consider two histograms, $Y(A) = \{([0, 10); 0.6), [10, 20); 0.2), [20, 30); 0.2)\}$ and $Y(B) = \{([0, 10); 0.2), [10, 20); 0.6), [20, 30); 0.2)\}$, having common subintervals. There are five distinct values defining the vector $\mathbf{w}$: $\mathbf{w} = [0, 0.2, 0.6, 0.8, 1]$. The quantile intervals for histograms A and B obtained from the subintervals [0, 0.2], [0.2, 0.6], [0.6, 0.8] and [0.8, 1] are given in Table 1. The lower and upper bounds of the quantile intervals are calculated from Ψ^{-1} ; these two values define its center and radius. The last column presents the distance component of (2.9) which takes into account π^*. The distance between histograms is the sum of the components; $d_M^2 = 23.71$. Figure 3 illustrates the procedure graphically. The distribution function for the histogram A in given in blue and for the histogram B in red. Below the *x*-axis, the quantile intervals for A are given in blue and those for B in red.

Details for the calculation of the barycentric histogram are given in Table 2. The vector **w** defines four quantile intervals. For each barycenter quantile interval, the center is calculated as the average of the centers of the quantile intervals for A and B, and in the same way the radius is calculated; these two components determine the quantile interval

Table 1: Quantile intervals (lower bound, upper bound, center, radius) for histograms A and B defined by the vector **w** . The last column present the distance component which takes into account the weights π^*.The distance between histograms is $d_M^2 = 23.71$.

		Quantile intervals for A				Quantile intervals for B				Distance
w interval	π^*	lower	upper	center	radius	lower	upper	center	radius	component
[0,0.2]	0.2	0.00	3.33	1.67	1.67	0.00	10.00	5.00	5.00	2.96
[0.2,0.6]	0.4	3.33	10.00	6.67	3.33	10.00	16.67	13.33	3.33	17.78
[0.6,0.8]	0.2	10.00	20.00	15.00	5.00	16.67	20.00	18.33	1.67	2.96
[0.8,1.0]	0.2	20.00	30.00	25.00	5.00	20.00	30.00	25.00	5.00	0.00

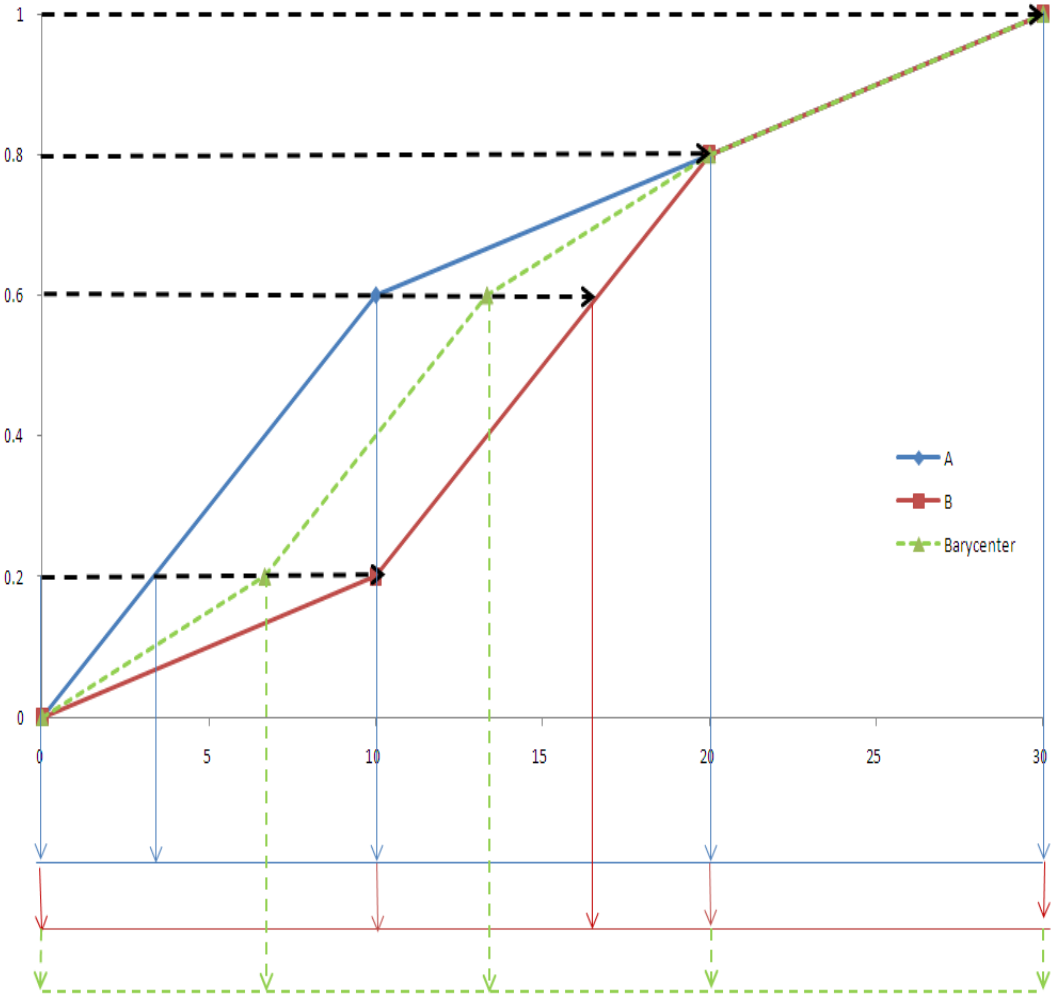

Figure 3: Distibution functions for A (in blue) and for B (in red) define the components of vector **w** on the y-axis. Quantile intervals are presented below the x-axis: for histogram A in blue and for histogram B in red. The dashed line presents the barycentric distribution function (in green). Barycentric quantile intervals are presented below the x-axis (in green).

Table 2: Quantile intervals for the barycentric histogram: center and radius are calculated from the center and radius of the quantile intervals for A and B, these two components determine the lower and the upper bounds.

w interval	**w**	π^*	center	radius	lower	upper
[0,0.2]	0.2	0.2	3.33	3.33	0.00	6.67
[0.2,0.6]	0.6	0.4	10.00	3.33	6.67	13.33
[0.6,0.8]	0.8	0.2	16.67	3.33	13.33	20.00
[0.8,1.0]	1.0	0.2	25.00	5.00	20.00	30.00

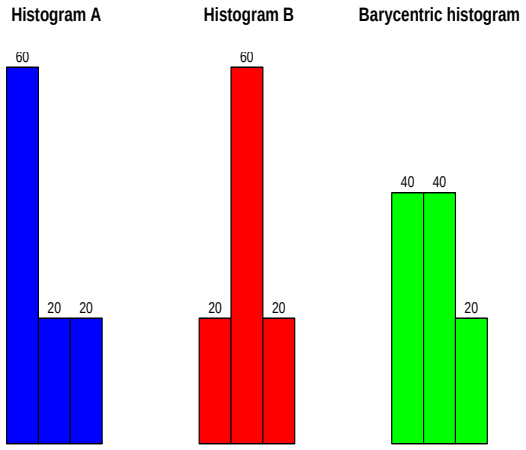

Figure 4: Histograms A and B and their barycentric histogram with the same subintervals as histograms A and B have.

lower and upper bounds. In Figure 3, the barycentric quantile intervals are presented in green below the *x*-axis, the barycentric distribution function is added as well (in green).

In Figure 4, histograms A (in red) and B (in blue) are shown. On the right, the barycentric histogram with the same subintervals as histograms A and B have is presented (in green); the subintervals' relative frequencies were obtained using linear interpolation (see green dotted line in Figure 3). In this case, the barycentric histogram is just the "average" of the original histograms A and B. A similar illustration is presented in Arroyo et al. (2011). The authors present the Mallows' and the Wasserstein's barycenter for two histograms. Notice the Mallows' barycentric histogram is an average of the position, range and shape of the corresponding histograms.

2.3 Inertia

Let us assume we have a partition of n histograms into K clusters, $P_K = \{C_1, C_2, ..., C_K\}$. In the cluster C_k, there are n_k histograms, $C_k = \{Y_k(1), ..., Y_k(u), ...Y_k(u_{n_k})\}$. The Mal-

lows' L^2 distance allows us to define a measure of *total inertia* TI using the notion of the global barycentric histogram $Y(b)$:

$$TI = \sum_{k=1}^{K} \sum_{u=1}^{n_k} d_M^2(Y_k(u), Y(b)). \tag{2.13}$$

The TI can be decomposed into *within inertia* WI and *between inertia* BI according to the *Huygens theorem*:

$$TI = WI + BI \tag{2.14}$$

where the *within inertia* WI and *between inertia* BI are defined as follows:

$$WI = \sum_{k=1}^{K} \sum_{u=1}^{n_k} d_M^2(Y_k(u), Y(b_k)) \tag{2.15}$$

$$BI = \sum_{k=1}^{K} n_k d_M^2(Y(b_k), Y(b)). \tag{2.16}$$

In (2.15) and (2.16), $Y(b_k)$ denotes the barycenter of the cluster C_k. Proof of (2.14) is given in the Appendix; it holds also for the multivariate case.

It can be shown that the inertia of the union of two disjoint clusters C_s and C_t is computed as follows:

$$TI(C_s \cup C_t) = TI(C_s) + TI(C_t) + \frac{n_s \cdot n_t}{n_s + n_t} d_M^2(Y(b_s), Y(b_t)). \tag{2.17}$$

The last term of (2.17) is recognized as Ward's distance between the two clusters C_s and C_t. Since d_M^2 is a form of Euclidean distance, it can be used with Ward hierarchical clustering method (Batagelj, 1988); dynamic clustering procedure can be used as well (Irpino et al., 2006; Verde and Irpino, 2007).

3 Results

Countries described by population pyramids can be regarded as symbolic data objects (Billard and Diday, 2006) with two random variables, one presenting age for males and one for females; for simplicity, we shall assume these two variables are independent. The objective of our study is to partition the 14 EE countries into homogenous groups according to the similarity of the shape of their population pyramids in each particular year. Ward's agglomerative clustering method based on the Mallows' distance presented in (2.10) was used. Summary results are presented in Table 3.

The dendrogram for 1995 is in Figure 5a). It shows that the 14 countries are clustered into two clusters on the first level: Cluster 1 contains Albania (AL) and Kosovo (KO), while the remaining 12 countries are in the second cluster. The second cluster splits into two clusters; Cluster 2 has 5 countries: Bosnia and Herzegovina (BA), Macedonia (MK),

Table 3: Total inertia TI, Between inertia BI and Within inertia WI for the obtained partitions in 1995, 2000, 2005, 2010 and 2015 (in bold). For each cluster, the following information is given: identification, members, size, cluster Between inertia and Within inertia contribution.

Year	Cluster id	Cluster members	Size	TI	BI	WI
1995				**448.78**	**411.15**	**37.63**
	1	AL KO	2		283.00	9.52
	2	BG CZ HR HU RO RS SI	7		121.89	18.66
	3	BA ME MK PL SK	5		6.29	9.45
2000				**422.29**	**387.82**	**34.48**
	1	AL KO	2		266.53	8.95
	2	BG CZ HR HU RS SI	6		118.82	7.45
	3	BA ME MK PL RO SK	6		2.46	8.08
2005				**394.63**	**356.34**	**38.29**
	1	AL KO	2		256.06	14.03
	2	BG CZ HR HU RS SI	6		99.32	5.68
	3	BA ME MK PL RO SK	6		0.96	18.58
2010				**362.20**	**337.64**	**24.56**
	1	KO	1		190.61	0.00
	2	AL MK	2		62.70	11.91
	3	BG CZ HR HU RS SI	6		81.34	5.26
	4	BA ME PL RO SK	5		2.98	7.39
2015				**342.22**	**321.19**	**21.02**
	1	KO	1		190.55	0.00
	2	AL MK	2		55.17	8.73
	3	BA BG CZ HR HU RS SI	7		73.65	8.29
	4	ME PL RO SK	4		1.83	4.01

Montenegro (ME), Poland (PL) and Slovakia (SK). In Cluster 3, there are the remaining 7 countries: Bolgaria (BG), Czech Republic (CZ), Croatia (HR), Hungary (HU), Romania (RO), Serbia (RS) and Slovenia (SI). Figure 5b), c) and d) present the barycentric histograms for Cluster 1, 2 and 3, respectively. The shape of the barycentric pyramid for Cluster 1 is expansive; for the other two clusters, it is constrictive. Table 3 shows that the total inertia equals 448.78, the between inertia 411.15 and the within inertia 37.63. Cluster 1 and Cluster 3 are similarly homogenous: their within inertia contribution is 9.52 and 9.45, respectively. However, the barycenter of Cluster 3 is the nearest to the global barycenter, since its between inertia contribution is the lowest (6.26).

Results for the years 2000 and 2005 are the same as the results for 1995 except that RO is moved from Cluster 2 to Cluster 3. However, total inertia decreased from 448.78 in 1995 to 422.29 in 2000 and to 394.63 in 2005; similarly, the between inertia decreased from 411.15 to 387.82 to 356.34.

The dendrogram for 2010 (see Figure 6a)) reflects a substantial change: MK is moved

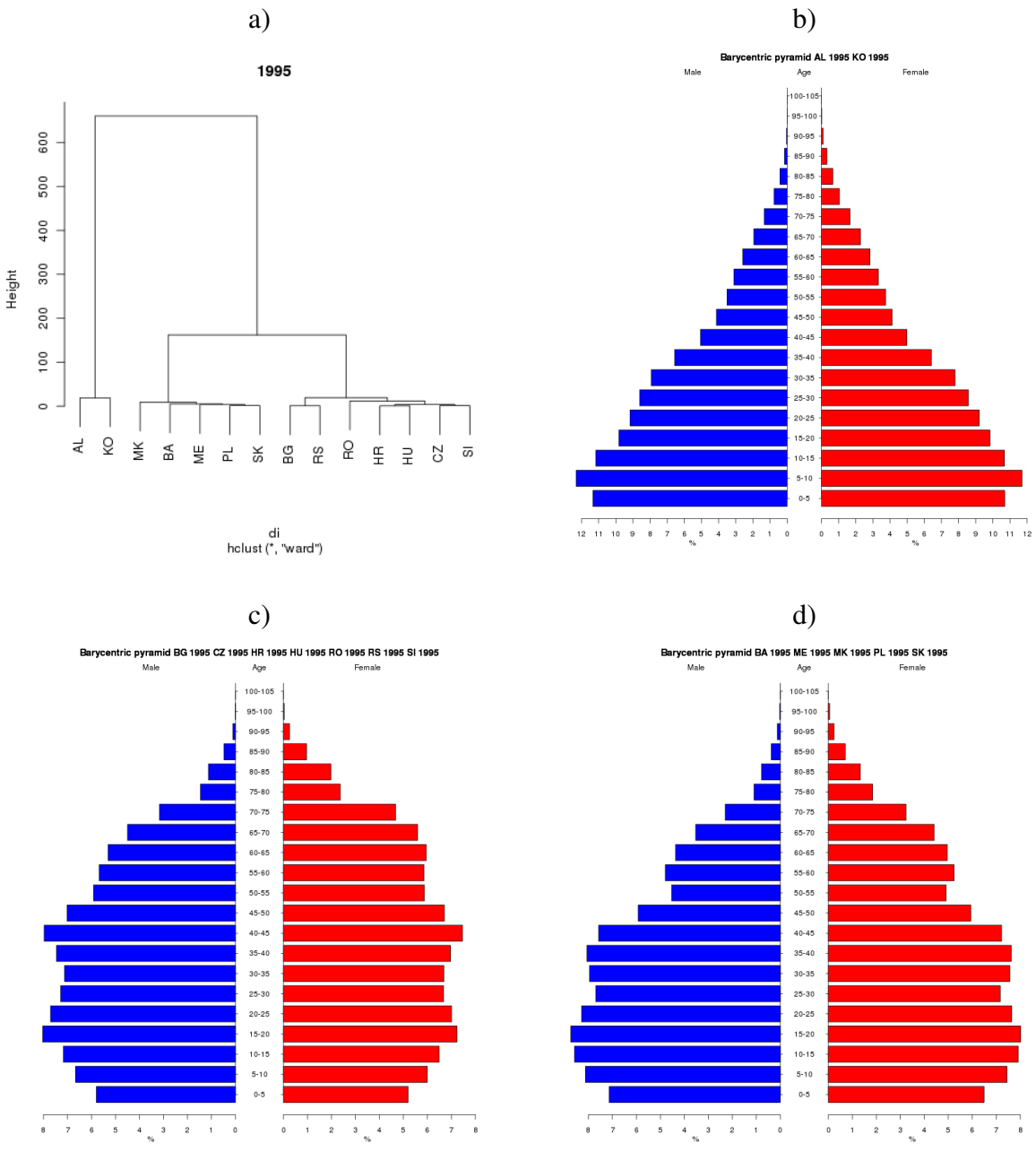

Figure 5: a) Dendrogram for the 14 East European countries for 1995 obtained by Ward's method; b) barycentric histograms for Cluster 1 (AL, KO); c) barycentric histograms for Cluster 2 (BG, CZ, HR, HU, RO, RS and SI; and d) barycentric histogram for Cluster 3 (BA, ME, MK, PL, SK).

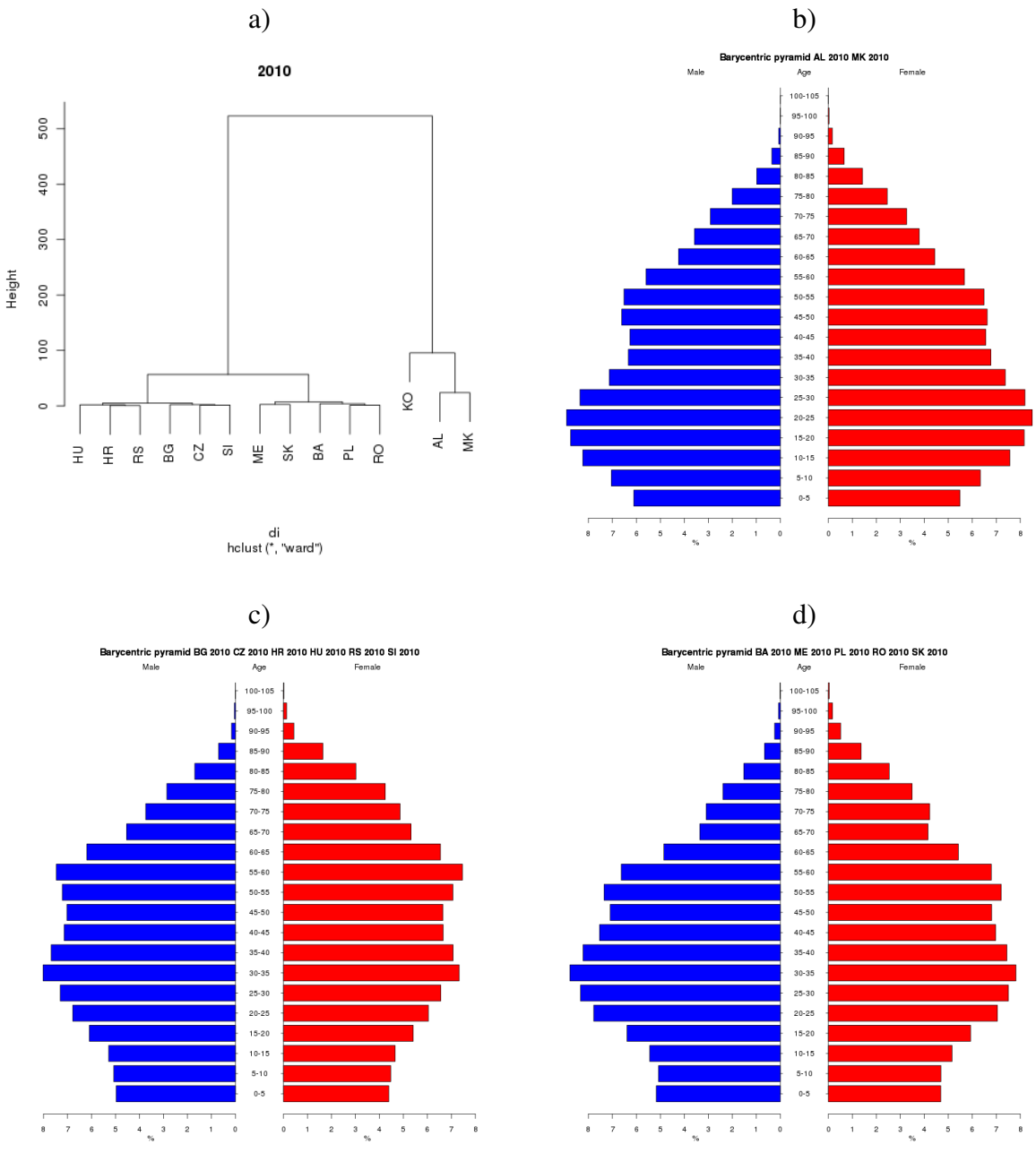

Figure 6: a) Dendrogram for the 14 East European countries for 2010 obtained by Ward's method; b) barycentric histograms for Cluster 2 (AL, MK); c) barycentric histograms for Cluster 3 (BG, CZ, HR, HU, RS and SI; and d) barycentric histogram for Cluster 4 (BA, ME, PL, RO, SK).

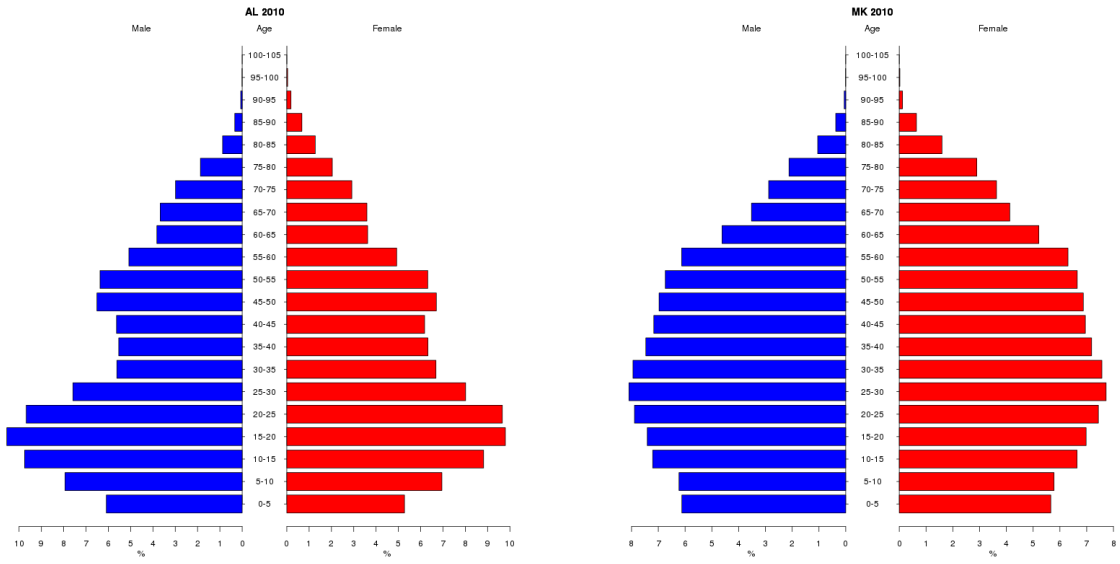

Figure 7: Population pyramid for Albania 2010 (left) and Macedonia 2010 (right).

to the cluster with AL and KO, suggesting two subclusters: one consisting of KO and the other of AL and MK. How many clusters are to be adopted? Total inertia is 362.20; in the case of 3 clusters, it splits into Between inertia and Within inertia as follows: 362.20 = 309.37 + 52.82; in the case of 4 clusters, 362.20 =337.64 + 24.56. In order to obtain the optimal number of clusters, the Calinski-Harabasz pseudo F-statistic for the number of clusters k (Calinski and Harabasz, 1974) can be used:

$$CH(k) = \frac{BI(k)/(k-1)}{WI(k)/(n-k)} \tag{3.1}$$

with higher values of $CH(k)$ preferred. In our case, $CH(3) = 32.21$, $CH(4) = 45.282$. Therefore, the partition into 4 clusters was adopted. Cluster 1 consists of KO only, Cluster 2 of AL and MK, Cluster 3 of 6 countries: BG, CZ, HU, HR, RS and SI, and Cluster 4 of 5 countries: BA, ME, PL, RO and SK. The barycentric histograms for Cluster 2, 3 and 4 are presented in Figure 6 b), c) and d), respectively. The KO pyramid is changing from an expansive shape (Figure 1 left) to a constrictive shape (see Figure 2 left).The shape of the barycentric pyramid for AL and MK can be better understood as the average of the population pyramids of Figure 7: for the younger age cohorts (age 0-20) each age group consists of a smaller proportion than for the previous group; but the AL pyramid has a part of its population in the age groups 20-40 years which is missing (presumably because of migrations). Cluster 3 and Cluster 4 have similar constrictive shapes. They are similarly homogenous (their within inertia contributions are 5.26 and 7.39, respectively).

The results for 2015 are based on the predicted data and are given in Table 3. They are similar to the results for 2010, except that BA is moved from Cluster 4 to Cluster 3. The tendency of decreasing inertia remains.

To sum up: the results reflect demographic changes in this short time-interval. In general, we observe a pattern of aging populations: a decline in the number of births and an

increase in the number of elderly persons. For the majority of countries considered in the dataset, this has been going on long before 1990 and their pyramids reflect a constrictive shape within the observed period. For KO, AL and MK, this process started after 1990. The results reflect that the countries are becoming more and more similar and follow a pattern of aging populations.

4 Conclusion

We chose to use the Mallows' L^2 distance to cluster the population pyramids according to the similarity of their shapes. Its calculation is simple, even when the number and length of histograms' subintervals may differ. However, two assumptions are taken into account in its derivation for the histogram setting: the distribution within each histogram-subinterval is uniform, the variables presenting age for males and females are independent. The first assumption is a standard one when histograms are under consideration. On the other hand, age for males and females is dependent and ignoring this dependency presents a simplification. The results we obtained are satisfactory - countries with similar shapes of the pyramids were successfully detected for each year under the analysis. The global and the local barycentric histograms offer a deeper understanding of the yearly results. Additional insight into the results can be obtained from the total inertia and its decomposition into the within and between inertia; in such a way the time-trend can be assessed as well.

References

[1] Arroyo, J., González-Rivera, G., Maté, C., and Muñoz San Roque, A. (2011): Smoothing methods for histogram-valued time series: an application to value-at-risk. *Statistical Analysis and Data Mining*, **4**(2), 216-228.

[2] Batagelj, V. (1988): Generalized Ward and related clustering problems. In H.H. Bock (Ed): *Classification and Related Methods of Data Analysis*, 67-74. Amsterdam: North-Holland.

[3] Billard, L. and Diday, E. (2006): *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley Series in Computational Statistics.

[4] Calinski, R.B. and Harabasz, J. (1974): A dendrite method for cluster analysis, *Communications in Statistics*, **3**, 1-27.

[5] Irpino, A., Lechevallier, Y., and Verde, R. (2006): Dynamic clustering of histograms using Wasserstein metric. In A. Rizzi, and M. Vichi (Ed): *COMPSTAT 2006*, 869-876. Berlin: Physica-Verlag.

[6] Irpino, A. and Verde, R. (2006): A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In V. Batagelj, H.H. Bock, A. Ferligoj, and A. Žiberna (Ed): *Data Science and Classification*, 185-192. Berlin: Springer.

[7] Kim, J. and Billard, L. (2011): Dissimilarity measures for histogram-valued observations. *Communications in Statistics*: Theory and Methods, in press.

[8] Korenjak-Černe, S., Kejžar, N., and Batagelj, V. (2008): Clustering of population pyramids. *Informatica*, **32**(2), 157-167.

[9] Levina, E. and Bickel, P. (2002): The EarthMover's Distance is the Mallows Distance: Some Insights from Statistics (http://www.stat.lsa.umich.edu/˜elevina/EMD.pdf).

[10] Mallows, C.L. (1972): A note on asymptotic joint normality. *Annals of Mathematical Statistics*, **43**(2), 508-515.

[11] Rüschendorf, L. (2001): Wasserstein metric. In M. Hazewinkel (Ed.): *Encyclopaedia of Mathematics*. Berlin: Springer.

[12] Verde, R. and Irpino, A. (2007): Dynamic clustering of histogram data: using the right metric. In P. Brito, P. Bertrand, G. Cucumel, and F. de Carvalho (Ed.): *Selected Contributions in Data Analysis and Classification*, 123-134, Berlin: Springer.

[13] Verde, R. and Irpino, A. (2008): Comparing histogram data using a Mahalanobis-Wasserstein distance. In: P. Brito (Ed.): *COMPSTAT 2008*, 77-89. Berlin: Physica-Verlag.

[14] Vertot, N. (2011): The elderly in Slovenia. Ljubljana: Statistical Office of the Republic of Slovenia, 2011 (http://www.stat.si/doc/pub/Stari2010-ANG.pdf).

A Appendix

We can represent the histogram $Y_k(u)$ in the cluster C_k, $u = 1, ..., n_k$, $k = 1, ..., K$, by the quantile-interval centers and radii (see Section 2.1) as

$$Y_k(u) = \{([c_{l,ku} - r_{l,ku}; c_{l,ku} + r_{l,ku}], \pi_{l,ku}), l = 1, ..., m_k\}. \tag{A.1}$$

The center and the radius for the l-th quantile-interval of the barycentric histogram for the cluster C_k is calculated as follows:

$$c_{l,b_k} = n_k^{-1} \sum_{u=1}^{n_k} c_{l,ku}, \quad r_{l,b_k} = n_k^{-1} \sum_{u=1}^{n_k} r_{l,ku}; \tag{A.2}$$

similarly, for the global barycentric histogram:

$$c_{lb} = n^{-1} \sum_{k=1}^{K} \sum_{u=1}^{n_k} c_{l,ku}, \quad r_{lb} = n^{-1} \sum_{k=1}^{K} \sum_{u=1}^{n_k} r_{l,ku} \tag{A.3}$$

where $n = \sum_{k=1}^{K} n_k$. Then from (2.13) and (2.9), TI is expressed as follows:

$$TI = \sum_{k=1}^{K} \sum_{u=1}^{n_k} d_M^2(Y_k(u), Y(b)) = \sum_{k=1}^{K} \sum_{u=1}^{n_k} \sum_{l=1}^{m_n} \pi_l^* [(c_{l,ku} - c_{lb})^2 + \frac{1}{3}(r_{l,ku} - r_{lb})^2].$$

Hence, we can write

$$\begin{aligned} TI &= \sum_{k=1}^{K}\sum_{u=1}^{n_k}\sum_{l=1}^{m_n}\pi_l^*[(c_{l,ku}-c_{l,b_k}+c_{l,b_k}-c_{lb})^2+\frac{1}{3}(r_{l,ku}-r_{l,b_k}+r_{l,b_k}-r_{lb})^2] \\ &= \sum_{k=1}^{K}\sum_{u=1}^{n_k}\sum_{l=1}^{m_n}\pi_l^*[((c_{l,ku}-c_{l,b_k})^2+(c_{l,b_k}-c_{lb})^2+\frac{1}{3}(r_{l,ku}-r_{l,b_k})^2+(r_{l,b_k}-r_{l,b})^2] \end{aligned} \tag{A.4}$$

since the two cross-product terms become zero: $\sum_{u=1}^{n_k}(c_{l,ku}-c_{l,b_k})=n_kc_{l,b_k}-n_kc_{l,b_k}=0$; and similarly for the radii. Hence, the expression for TI can be rewritten as follows:

$$TI = WI + BI \tag{A.5}$$

where WI and BI are as follows:

$$WI=\sum_{k=1}^{K}\sum_{u=1}^{n_k}\sum_{l=1}^{m_n}\pi_l^*[(c_{l,ku}-c_{l,b_k})^2+\frac{1}{3}(r_{l,ku}-r_{l,b_k})^2]=\sum_{k=1}^{K}\sum_{u=1}^{n_k}d_M^2(Y_k(u),Y(b_k)), \tag{A.6}$$

$$BI=\sum_{k=1}^{K}\sum_{l=1}^{m_n}\pi_l^*n_k[(c_{l,b_k}-c_{lb})^2+\frac{1}{3}(r_{l,b_k}-r_{lb})^2]=\sum_{k=1}^{K}n_kd_M^2(Y(b_k),Y(b)). \tag{A.7}$$

Hence, the result (2.14) follows.