Advances in Methodology and Statistics, 2022, *19*(1), 31–44. https://doi.org/10.51936/xhhj7682

Complementary results on sex-, age-, and cause-specific mortality in EU countries obtained by SYR symbolic data analysis software

Aleša Lotrič Dolinar[a,*], Filipe Afonso[b], Simona Korenjak-Černe[a], Edwin Diday[c]

[a]*University of Ljubljana, School of Economics and Business, Ljubljana, Slovenia*
[b]*Symbolic Data Lab, Roissy-Pôle, France*
[c]*Paris Dauphine University, Research Centre in Mathematics of Decision, Paris, France*

Abstract

Different mortality patterns across countries require different health and demographic policies. Positioning of the countries according to their characteristic mortality pattern can help allocate scarce resources appropriately. We use symbolic data analysis within SYR software to analyse 28 European Union countries' sex-, age-, and cause-specific mortality in 2015. There are two main advantages for using symbolic analysis: (i) it permits more transparent and informative data descriptions along with contextual relations, and (ii) advanced methods adapted for complex data representations can be employed to analyse such data, taking contextual relations into account. Clustering results based on symbolic data analysis show that groups of countries are strongly related to the geographical position of countries, with a clear east–west cut on the first-level partition and with an even more geographically consistent lower-level partition compared to the classical clustering result. Relations between the obtained clusters of countries and their external social and health indicators are well pronounced. We also identify the mortality rates as symbolic variables that discriminate the most between individual countries as well as between the resulting clusters. Knowledge of a country's mortality pattern and its position among comparable countries is valuable information for health and demographic policymakers and can be exploited to exchange good practices.

Keywords: cause of death, clustering, health policy, symbolic data, European Union

*Corresponding author
Email address: alesa.lotric.dolinar@ef.uni-lj.si (Aleša Lotrič Dolinar)
ORCID iD: 0000-0002-1574-5473 (Aleša Lotrič Dolinar)

1. Introduction

Countries with different sex- and age-specific mortality rate and structure suffer from different health problems and therefore face different health costs. It is important to know the country's position among other countries to be able to implement proper health policies and to distribute limited resources accordingly and potentially take advantage of good practices of the countries from a more favourably positioned group.

Because some illnesses or risk behaviours are more easily prevented or controlled, and consequential deaths therefore postponed (e.g., Crimmins et al., 2011; Heijink et al., 2013; Kerr et al., 2017; Lear et al., 2017; Stewart, 2012), it is essential to analyse what factors cause deaths in certain sex and age groups of the population. Analysis based on data supplemented by causes of death in our opinion produces more precise information about a country's health situation compared to the more common analysis of sex- and age-specific mortality.

We study mortality in European Union countries using new analytical tools for more complex data representations of sex-, age-, and cause-specific mortality rates. As causes of death are strongly related to sex and age, the input data are provided separately for each sex and age group. One of the aims of our study is to upgrade the results presented in Lotrič Dolinar et al. (2019), where only classical analyses were made. We argue that more relevant results can be achieved with more informative data descriptions, where each sex–age combination is described with two so-called symbolic variables: structure over causes of death, and mortality rate classified into four levels (low, mid-low, mid-high, and high). As such a representation requires appropriate analytical tools, we performed symbolic data analysis (e.g., Afonso, Diday, & Toque, 2018; Billard & Diday, 2006; Diday, 2016; Noirhomme-Fraiture & Brito, 2011), more precisely the clustering algorithms implemented in SYR software (Afonso, Haddad, et al., 2018).

Some other clustering analyses for mortality rates have been conducted previously. For example, Meslé and Vallin (2002) found that at the end of the 20th century the major aspect of European demography was the divergence between East and West, with the line of separation between the two geographical groups corresponding to the former Iron Curtain. Although they analysed life expectancy at birth, not taking the causes of death into account, they explained their findings by the inability of Eastern countries to follow Western countries in the so-called "cardiovascular revolution" (Vallin et al., 2002).

We, however, add the cause of death dimension and employ the symbolic data analysis approach in order to account for the variability of the data in a more appropriate way. We thus determine which groups of countries are formed and what the differences are between the groups. With this, we present our data as a symbolic object and identify groups of countries with similar characteristics with clustering methodology for symbolic data. Since clustering is a very important topic in symbolic data analysis, several methods have been developed (e.g., Billard & Diday, 2020; Brito & Dias, 2022). We conducted our analysis with program SYR that enables us to also identify the most discriminating mortality rates or sex, age, and cause combinations between individual countries as well as between resulting clusters. Using these results, we can provide policy makers with more detailed information for properly adjusting and implementing a country's health policy.

Our paper highlights two main points. First, we want to present the symbolic data analysis, which could be applied to the analysis of other inter-related multi-level data. And second, we believe that the contextual results for our specific example should be interesting and important for policy makers, which we elaborate in the concluding section.

The rest of the paper is organized as follows: in the second part we describe the original data and how they were transformed into the symbolic form, as well as the adapted methods

implemented in SYR software that we used for our study. In the third part we discuss our results for the mortality data of EU countries for 2015 in more detail and compare them with the results obtained with classical clustering methods. Finally, we conclude the paper with a discussion and conclusion.

2. Methodology

2.1. Methods

Intensity of mortality by certain cause of death heavily depends on sex and age. Therefore, our initial data consist of the number of deaths by sex, age, and cause of death (Eurostat, 2018a) and the corresponding size of the population (Eurostat, 2018c). In our study, we use 3-year-average standardized number of deaths by four causes of death within 36 sex–age groups (18 5-year age groups for each sex) for 28 EU countries in the year 2015. The analyzed causes are the three most important causes of death according to ICD-10 (World Health Organization, 2010), responsible for over 70 % of all deaths in the EU (Eurostat, 2018b): neoplasms, diseases of the circulatory system, and diseases of the respiratory system, plus the residual group "other causes". The same data were used in Lotrič Dolinar et al. (2019), where the study is based on the classical clustering methods. Since the data that relate to each sex–age combination represent structure of deaths over causes of death, we transform them into a representation that takes this relation into account and present them in the more informative symbolic data table. Instead of units with single observed values, symbolic data analysis deals with classes of individuals that are considered as higher-level units and thus constitute a new population of higher-level units with their own structure (Diday, 2016). To preserve their internal variability, intervals, histograms, probability distributions, bar charts, etc. are used for descriptions of such units. These types of data are called "symbolic" as they cannot be reduced to single numbers without a loss of much information (Diday, 2016).

2.2. Data structure

Original data are presented with a classical data matrix that has 28 rows or units (representing 28 EU countries) and 144 columns or numerical variables (representing standardized number of deaths for both sexes, 18 age groups, and four categories for cause of death). Since the values for sex-, age-, and cause-specific mortality (i.e., mortality related to specific sex, specific age group, and specific cause of death, calculated for all possible combinations) are statistically and also contextually related (by each sex–age combination they represent structure of deaths over cause of death), we argue that such information cannot be clearly seen from the data representation in a classical data table form where each cell contains only a number or category. Therefore, we transform the classical data matrix into a symbolic data table where each cell contains a bar chart. This form enables the inclusion of contextual dependence through the so-called symbolic variables. The symbolic data still consist of 28 units or rows (representing EU countries), but the number of columns is reduced to 72 symbolic variables, where each of 36 sex–age combinations (18 age groups for each sex) is represented with:

1. a bar chart showing relative structure over causes of death, described by the relative frequency of the four cause-of-death categories, i.e., 1 = *Neoplasms*, 2 = *Circulatory*, 3 = *Respiratory*, and 4 = *Other*; and
2. a numerical value representing mortality rate (i.e., the number of deaths per certain number of people (100 000 in our case) per year) discretized into bar charts with four categories: low, mid-low, mid-high, and high level.

Mortality rate variable for each sex–age combination was separately discretized into these four categories (there are four categories in order to discretize in the same manner as for the four causes) using adapted Fisher algorithm (Diday et al., 2013) depending on the standardized number of deaths per 100 000 people; for every sex–age combination we got different boundaries between the four resulting categories in order to polarize the values as much as possible. To ensure cross-country comparability by controlling for different population structures by sex and age, and to compare our results with the classic results from Lotrič Dolinar et al. (2019), we used the standardized mortality rates, which we calculated by applying country-specific mortality rates to a standard population, using the combined actual population of all analysed countries as the standard population.

The transformation from classical to symbolic data is illustrated in Figure 1. The segment of the classical data table is presented in Table 1, while the produced symbolic data table for the symbolic variables of the type Sex[age-group]Cause is presented in Table 2.

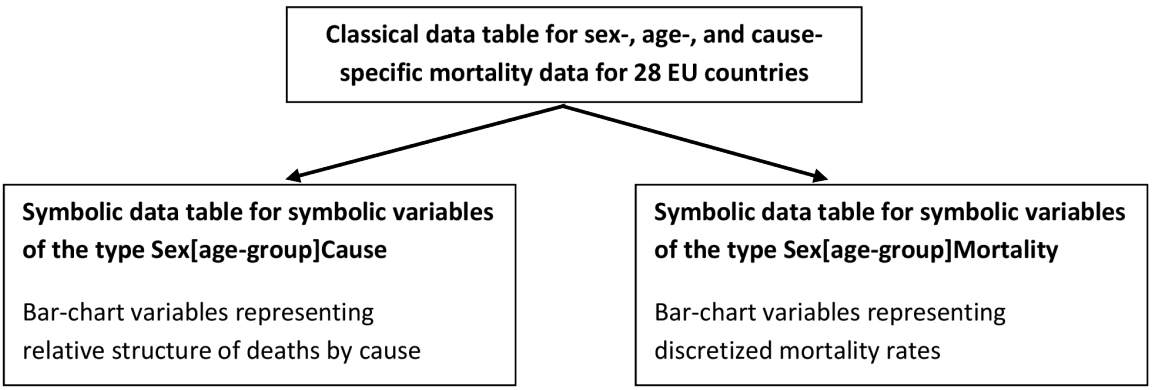


Figure 1: Transformation of the original sex-, age-, and cause-specific mortality rates data table to a symbolic data table

Table 1: A segment of the classical data table for sex-, age-, and cause-specific mortality data for 28 EU countries in 2015

Cause	Neo	Circ	Resp	Othr	Neo	Circ	Resp	Othr	Neo	⋯
Age	0–4	0–4	0–4	0–4	0–4	0–4	0–4	0–4	5–9	⋯
Sex	M	M	M	M	W	W	W	W	M	⋯
Country	M[0–4]				W[0–4]				⋯	
Austria (AT)	0.05	0.01	0.04	2.02	0.05	0.01	0.02	1.58	0.04	⋯
Belgium (BE)	0.06	0.04	0.08	2.13	0.06	0.04	0.04	1.62	0.07	⋯
Bulgaria (BG)	0.10	0.36	0.53	3.80	0.11	0.28	0.50	2.89	0.12	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Notes: Neo = *Neoplasms*, Circ = *Circulatory*, Resp = *Respiratory*, Othr = *Other*.

3. Application

3.1. Representation with symbolic data table

In this section we present some of the adapted methods of the SYR software that were used in our analysis.

One of the important advantages of symbolic data analysis (SDA) is the presentation of the input data and the resulting clusters. A symbolic data table includes contextual relations of the columns of the classical data matrix (i.e., for each sex–age combination we present structures of deaths by cause), while mortality rates are much easier to read if we use different colours for each of the four levels.

Thus, the first tool that we used for this study is the data representation in the form of a symbolic data table. We present a segment of the symbolic data table in Table 2 where the ordering of the categories for causes of death is fixed and coloured as follows: *Neoplasms* (light gray), *Circulatory* (black), *Respiratory* (dark gray), *Other* (white). From the symbolic data table we can very easily notice, for example, that:

1. Bulgaria (BG) and Romania (RO) have rather similar bar charts for most of the (presented) sex–age combinations, especially in older age groups;
2. there is only one cause of death for Malta (MT) for 5–9-year-old boys;
3. the category *Neoplasms* (1) is more pronounced in the older age groups in Austria (AT), Belgium (BE), and Malta (MT) than in Bulgaria (BG) and Romania (RO), etc.

Clearly, such data representation is much more informative and intuitive than a classical data table containing just a number in each cell.

3.2. Data analysis with SYR software

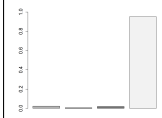
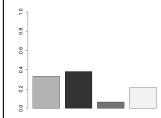
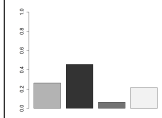
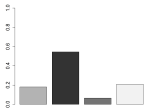
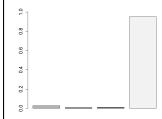
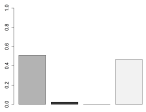
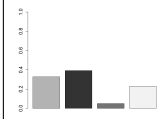
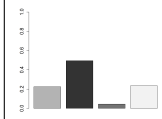
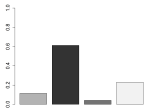
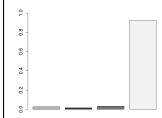
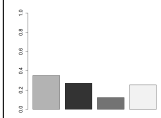
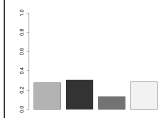
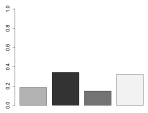
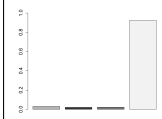
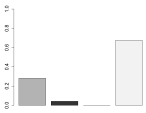
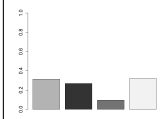
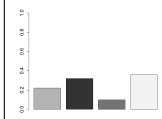
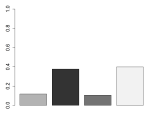
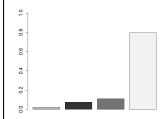
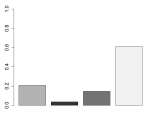
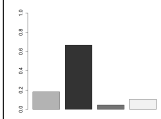
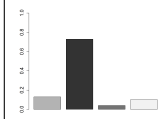
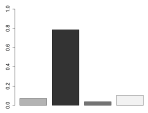
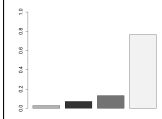
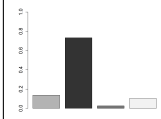
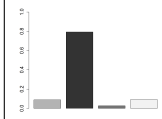
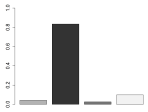
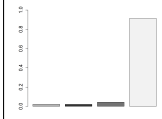
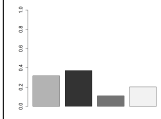
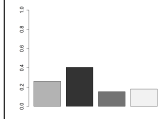
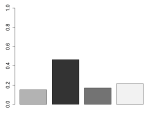
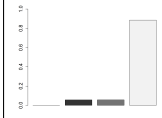
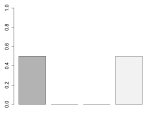
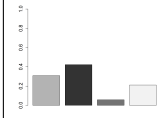
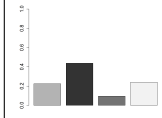
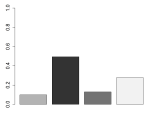
Since we have many countries and still too many symbolic variables to search for similarities and differences just by "manually" observing the complete symbolic data table, we therefore need additional tools to reduce the size of the data set into groups of similar countries. Because many variables are strongly correlated, we first reduce the number of variables with principal component analysis (PCA) and present the data on a factor plane. To obtain the explanatory power of the factor axes, we produce correlation circle and PCA values adapted for bar charts (Diday, 2013), see Figures 2 and 3. From here, the most discriminating input variable of the first and second principal component (PC) axis, PC1 and PC2, can also be detected as the one with the longest distance from the origin.

From such observations, it is also possible to make inferences about some correlations between input variables, indirectly via the principal components. For example, in the correlation circle for the input variables representing mortality rate (Figure 2) the sex–age combinations with high mortality expectedly appear on the opposite leg of the first axis rather than the sex–age combinations with low mortality. Observing categories of causes of death (Figure 3), we can see, for example, that the cause Circulatory appears most frequently on the left side of the factor plane of the first two axes, and the cause Neoplasms appears very frequently on the right side of the plane. The vertical axis (the upper part of the plane) is related to the cause Other. We also observe that along the first axis the longest distances from the origin appear for mortalities for older ages, while along the second axis the longest distances from the origin appear for mortalities for younger ages.

Symbolic cause and mortality variables can be presented with linear combinations of projections of relevant original variables on principal components with the angles of these projections representing weights (Diday, 2013). In such a way, we can detect correlations (indirectly through correlations with principal component axes) between our symbolic variables and also evaluate their importance in the sense of discriminating power between individual countries as well as between the resulting clusters.

Moreover, SYR software enables us to present each country on the factor plane, as well as to observe each symbolic variable in more detail.

Table 2: Several segments of the symbolic data table for sex-, age-, and cause-specific mortality data for 28 EU countries for symbolic variables of the type Sex[age-group]Cause

Country	Sex	Age group					
		0–4	5–9	⋯	75–79	80–84	85+
Austria (AT)	M			⋯			
	F			⋯			
Belgium (BE)	M			⋯			
	F			⋯			
Bulgaria (BG)	M			⋯			
	F			⋯			
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
Malta (MT)	M			⋯			
	F			⋯			
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
Romania (RO)	M			⋯			
	F			⋯			
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮

Notes: *Neoplasms* (light gray), *Circulatory* (black), *Respiratory* (dark gray), *Other* (white).

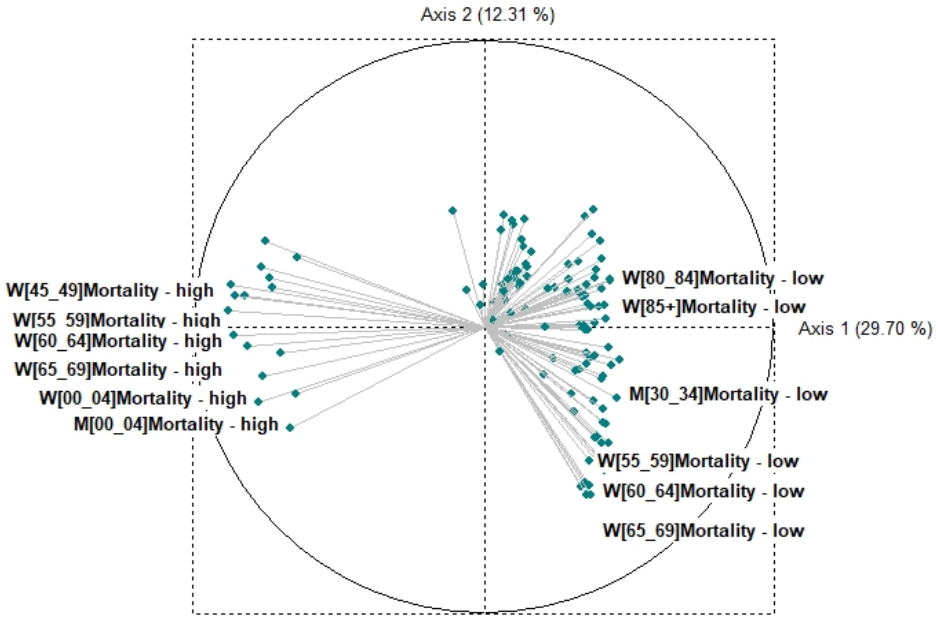


Figure 2: Correlation circle with some of the input variables representing mortality level

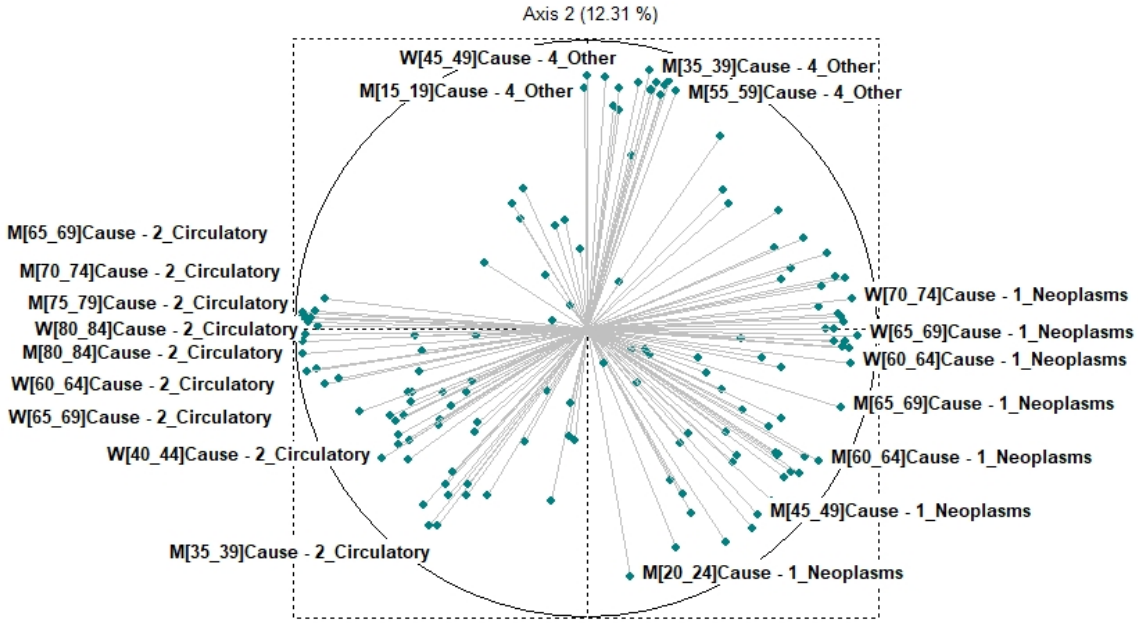


Figure 3: Correlation circle with some of the categories representing causes of death

From Figures 4 and 5 we can see that on the left side of the plane there are eastern EU countries and on the right side western EU countries. Roughly speaking, relating to the correlation circles we can say that in 2015 eastern EU countries had higher mortality rates and the most pronounced cause of death were circulatory diseases, while western EU countries had lower mortality rates, and the most pronounced cause of death were neoplasms. Because this claim is rather superficial, we were interested in observing the symbolic variables in more detail. We do this in two ways: by observing values for each symbolic variable separately, or by forming groups of similar countries and then finding characteristics of these groups to obtain common (symbolic) descriptions of the countries in each such group.

To demonstrate the first possibility, we present the more detailed results, for example, for women of age group 65–69 (this sex–age combination is chosen as one of the most discriminative symbolic variables of the "cause" type) in Figure 4. Here, we present structures over cause of death with pie charts, and we can notice that countries on the left side have the largest segments in light blue (*Circulatory*), but when we go from left to right the red segment (*Neoplasms*) becomes larger and larger.

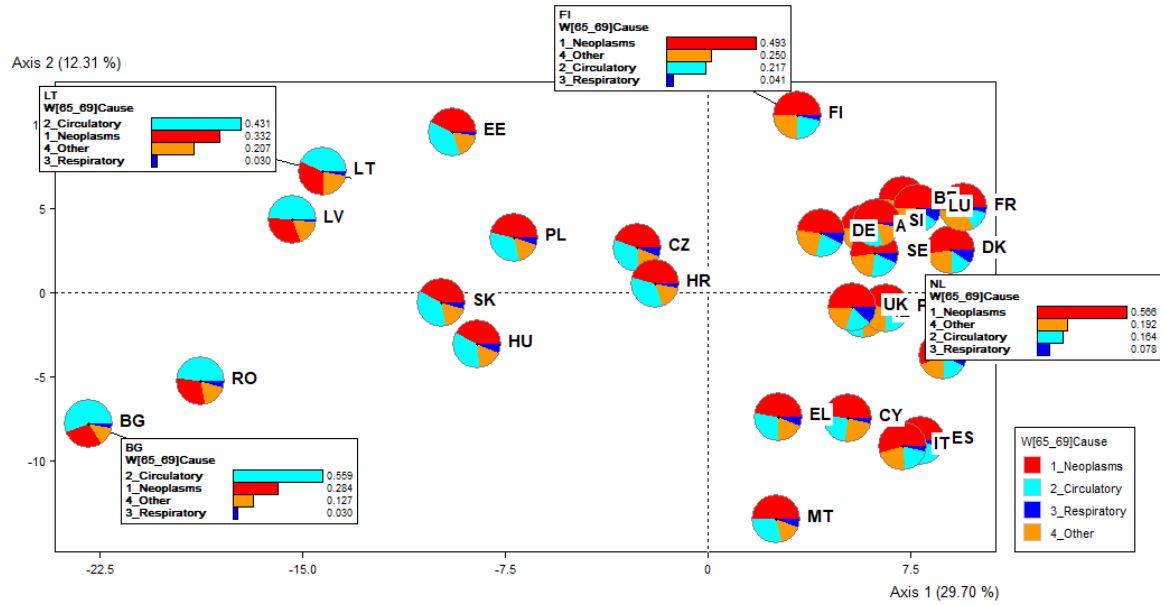


Figure 4: Positioning of individual countries according to the first two PC axes by structure over cause of death for symbolic variable W[65–69]Cause (structure of deaths by four cause categories for women aged 65–69 years)

To detect groups of similar EU countries, we performed an adapted *k*-means clustering method based on the first two factor axes. We identify four main groups. Obtained clusters are presented in Figure 5.

From Figure 5 it can be clearly seen that there are two groups of eastern EU countries on the left side, and two groups of western EU countries on the right. This implies that the groups of countries based on their mortality rates and main causes of death are very much related to the geographical position of the countries. To observe characteristics of the country clusters we can present them in a symbolic data table, in a similar way as it was done in Table 2 for individual countries. Figure 6 shows a segment of this presentation with bar-charts for the five most discriminating sex–age combinations across the countries in a certain cluster obtained with SYR program. In bar-charts, only the non-zero categories are

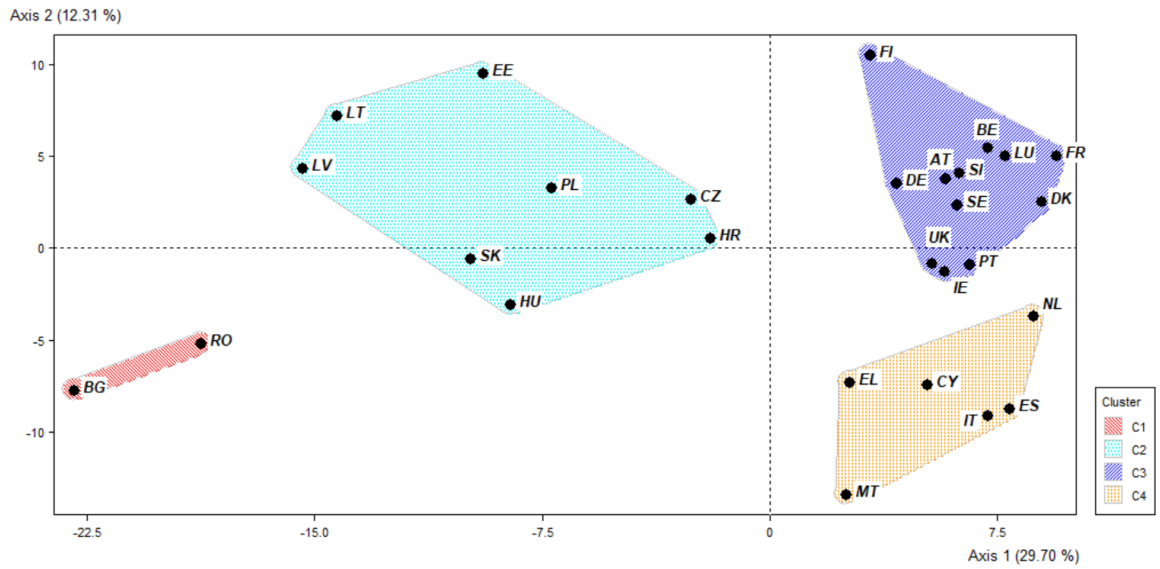


Figure 5: Four clusters of EU countries obtained with SYR clustering program based on the symbolic data description of sex-, age-, and cause-specific mortality in 2015

presented. The numbers below each column are related with the category value, and each category is of a different colour for easier observation.

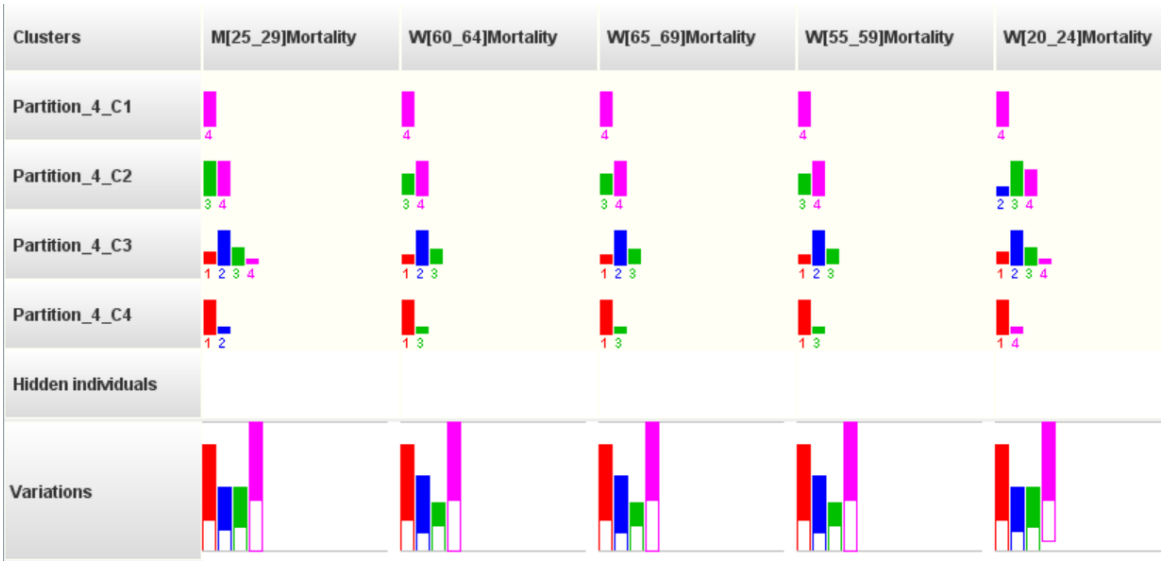


Figure 6: Symbolic data table for the four clusters of EU countries based on the sex-, age-, and cause-specific mortality in 2015, showing the first five most discriminating symbolic variables (columns), with the following mortality variable categories: 1 = *low*, 2 = *mid-low*, 3 = *mid-high*, and 4 = *high*

4. Results

Besides representing the input variables in a much more transparent manner, we focused also on the identification of groups of EU countries with similar mortality patterns, considering both dimensions: mortality rate and the mortality structure by main causes of death. At a glance, we can see clear division into eastern and western countries from a factor

plane in Figure 5. We compare our results with the classical analysis results (Lotrič Dolinar et al., 2019), where the authors applied classical Ward and k-means methods on the same original data and obtained these four clusters of countries:

1. West 1 (8 countries): Belgium (BE), Denmark (DK), France (FR), Luxembourg (LU), the Netherlands (NL), Portugal (PT), Spain (ES), and the United Kingdom (UK);
2. West 2 (10 countries): Austria (AT), Cyprus (CY), Finland (FI), Germany (DE), Greece (EL), Ireland (IE), Italy (IT), Malta (MT), Slovenia (SI), and Sweden (SE);
3. East 1 (6 countries): Croatia (HR), Czechia (CZ), Estonia (EE), Hungary (HU), Poland (PL), and Slovakia (SK); and
4. East 2 (4 countries): Bulgaria (BG), Latvia (LV), Lithuania (LT), and Romania (RO).

With the presented SDA method we obtain exactly the same division into eastern and western countries. However, the two groups within the eastern cluster and two within the western cluster are identified differently:

1. West 1 (12 countries): Austria (AT), Belgium (BE), Denmark (DK), Finland (FI), France (FR), Germany (DE), Ireland (IE), Luxembourg (LU), Portugal (PT), Slovenia (SI), Sweden (SE), and the United Kingdom (UK);
2. West 2 (6 countries): Cyprus (CY), Greece (EL), Italy (IT), Malta (MT), the Netherlands (NL), and Spain (ES);
3. East 1 (8 countries): Croatia (HR), Czechia (CZ), Estonia (EE), Hungary (HU), Latvia (LV), Lithuania (LT), Poland (PL), and Slovakia (SK); and
4. East 2 (2 countries): Bulgaria (BG) and Romania (RO).

More detailed inspection of the resulting clusters of EU countries regarding the causes of death shows that the cause *Circulatory* is much more pronounced in both eastern groups, while *Neoplasms* are more pronounced in both western groups, especially compared to the cluster consisting of Bulgaria and Romania, which have the fewest deaths from neoplasms in all ages. These two countries also have by far the largest number of deaths due to circulatory diseases in all ages. In these two countries we can also detect a much higher percentage of the *Respiratory* cause in the five youngest age groups (up to the age of 24), while this cause represents considerably more deaths in the three oldest age groups (above the age of 75) in the two western clusters. On the western side, there are more deaths due to neoplasms in the countries from cluster West 2 for men aged 55–69 and for women aged 30–59 compared to cluster West 1. Additionally, in cluster West 2 there are also more deaths from circulatory diseases for men up to age 69 and for women aged 30–39 compared to cluster West 1. The residual group *Other* causes dominates up to the age of 50 years for men and up to the age of 40 years for women in all four clusters and becomes very pronounced again after the age of 75 for both sexes of both western clusters, while in the Bulgaria-Romania cluster the share of deaths from other causes is considerably smaller compared to the other three clusters.

The most obvious difference in cluster characteristics when comparing the result obtained through the presented SYR method with the classical result is the number of deaths caused by respiratory diseases for children in the two eastern clusters. With only Bulgaria and Romania forming cluster East 2, its share of respiratory deaths for children is much higher compared to cluster East 1. Latvia and Lithuania, the two eastern countries that are placed differently than in the classical result, have overall mortality much closer to cluster East 1 obtained by SYR than to Bulgaria and Romania. Moreover, the *Neoplasms* pattern of Latvia and Lithuania for women of all ages and men aged 15–49 is closer to cluster East 1, and that also holds for deaths due to respiratory diseases and the cause *Other* for both sexes and almost all ages, while deaths from circulatory diseases are somewhere in between the

two eastern clusters.

On the western side, the grouping is even more different compared to the classical result. Cluster West 2 seems predominantly Mediterranean, apart from the Netherlands. The neoplasms mortality rates in the Netherlands, however, resemble much more the Mediterranean cluster than the rest of the western countries, while for circulatory diseases the opposite is true.

In short, the division between eastern and western countries is the same as with the classical approach, but the division into two lower-level groups for the western countries is quite different. Although the new partition of the 28 EU countries into four clusters is not directly related to some known grouping of these countries, it however still represents geographically consistent clusters, even more so than with the classical result. Contrary to the classical result where the same two larger clusters (eastern and western) were each further divided in an east–west sense on a lower level, the partition within the same two larger clusters is now basically in the north–south direction. Moreover, we also performed some matching between the resulting clusters and clusters of countries based on different social and health indicators. We separately grouped individual countries for several social indicators: social system (Sapir, 2006), health expenditure, alcohol consumption, tobacco smoking (Eurostat, 2016; World Health Organization, 2015a, 2015b), and the EuroHealth Consumer Index (Björnberg, 2016). Based on these results, we evaluated the correspondence between each social indicator clustering result and our mortality-rate–cause-of-death clustering result using the adjusted Rand index (Hubert & Arabie, 1985). Comparing the findings of the same procedure also for the classical clustering result, the correspondence between social indicator clusters and the symbolic data clustering result is better than with the classical result.

Besides forming and comparing clusters of countries, SYR program enables us to identify those symbolic variables that discriminate the most between the four resulting clusters, as well as between the individual countries. Based on the cluster representation (Figure 6), it turns out that the five most discriminating symbolic variables are all related to the mortality rate, not to a certain cause of death. The largest differences among the four clusters are in the mortality rate for men in the 25–29-year age group, the highest one being in the first cluster (Bulgaria and Romania). The same is true also for the mortality rate for women in the five-year age groups from 55 to 69, and for young women in the 20–24-year age group.

5. Discussion

The main aim of the paper is to show the usefulness of a new tool for dealing with complex data using symbolic analysis. This approach is presented on a case study of sex-, age-, and cause-specific mortality data for EU countries in 2015. The main advantages of the new methods are these:

1. more intuitive and informative presentation of the countries based on the input data with the symbolic data table that enables us to also present contextual relations;
2. use of adapted classical statistical and machine learning methods (e.g., PCA, clustering) for this type of data representation (Diday, 2020);
3. more informative presentation of the resulting clusters of countries in a symbolic data table for clusters; and
4. column ordering (in our case mortality rates and structures of deaths by cause at different sex–age combinations) from the most to the least discriminating.

Since the values for sex-, age-, and cause-specific mortality are statistically and also contextually related (for each sex–age combination they represent structure of deaths by

cause of death), we argue that such information cannot be clearly observed from data representation in a classical data matrix form. Therefore, we suggest displaying the data in a symbolic table, as this enables the presentation of contextual dependence through symbolic variables. Such a presentation offers a much more intuitive and therefore relevant view of the data. To be able to also include these relations in a further analysis of the data, we used adapted tools implemented in the SYR software. We argue that based on the more informative data description, the results obtained using a symbolic data analysis approach are also more relevant compared to the classical result.

We observe countries in the plane of the first two obtained principal component axes where the contrast between mortality due to circulatory diseases and mortality due to neoplasms is particularly pronounced. This division is clearly in line with the major east–west partition of the analyzed countries, with higher mortality from the diseases of the circulatory system in the eastern countries and higher mortality from neoplasms in the western countries. We identify the variables that best discriminate between the single countries, as well as between the clusters of countries; these variables are the mortality rates (not structures by cause) in both cases. The clustering result shows that the two-cluster partition resulting from the presented approach is exactly the same as with classical clustering, but the lower-level partitions differ substantially. The new four resulting groups are even more geographically consistent, with a more north–south division within each of the two larger first-level clusters (East, West). Additionally, the new partition also corresponds better to different social and health indicators of the analysed countries and is still in line with findings about the East lagging in the "cardiovascular revolution" compared to the West (Meslé & Vallin, 2006); specifically, we found that lower mortality (from cardiovascular diseases) is related to a lower population share of individuals who smoke, lower alcohol consumption, and higher health expenditure per capita (Eurostat, 2016; World Health Organization, 2015a, 2015b).

However, because the symbolic representation of the data is complex, it is very difficult to observe correlations between the symbolic variables, and to the best of our knowledge there is not yet a definite method of adapting this concept to a symbolic approach. Therefore, we were only able to overcome this limitation by making inferences about the correlations indirectly via the principal components.

The symbolic data procedure used in the presented analysis could be also applied to analysis of any other inter-related multi-level data, especially if also using clustering methods. Concerning contextual findings for our specific example, we believe that they should be interesting and important for policy makers, which we elaborate in the following concluding section.

6. Conclusion

In light of ever-aging populations and related increasing health costs, it is crucial to distribute limited health resources as optimally as possible. Looking toward better-performing countries can help health and demographic decision makers, especially in the area of preventable deaths, where policymakers can take advantage of the information we provide based on the presented analysis. When analysing data in a more complex sense using symbolic analysis, we still detect a clear East–West division along the former Iron Curtain as with the classical result, with the West performing better in the area of circulatory diseases and overall mortality rate, and with the East performing better in the area of neoplasms. Further partition, however, is substantially different from the classical result.

Within each of the two broad groups of countries, eastern and western, we now observe a more geographically consistent division in the north–south direction. Symbolic data analysis methods also reveal best discriminating sex–age combinations among countries and also among the four obtained clusters of countries for mortality rates and for distributions of deaths over cause of death. In addition, matching between the resulting clusters and clusters of countries based on different social and health indicators confirmed that the SYR clusters are matched better than the classical clusters from the aspects of social system, health expenditure, alcohol consumption, tobacco smoking, and the EuroHealth Consumer Index. As these are all factors that can be directly influenced by appropriate policies, this finding can represent a strong incentive for the countries to look toward, and thus aspire to replicating the health indicators of, the countries from the better-performing clusters.

References

Afonso, F., Diday, E., & Toque, C. (2018). *Data science par analyse des donées symboliques*. Editions Technip.

Afonso, F., Haddad, R., Toque, C., Eliezer, E.-S., & Diday, E. (2018). *User manual of the SYR software*. https://www.symbad.co/le-logiciel-syr/

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

Billard, L., & Diday, E. (2020). *Clustering methodology for symbolic data*. Wiley. https://doi.org/10.1002/9781119010401

Björnberg, A. (2016). *EuroHealth Consumer Index 2015: Report*. Health Consumer Powerhouse. https://healthpowerhouse.com/media/EHCI-2015/EHCI-2015-report.pdf

Brito, P., & Dias, S. (Eds.). (2022). *Analysis of distributional data*. Chapman & Hall. https://doi.org/10.1201/9781315370545

Crimmins, E. M., Preston, S. H., & Cohen, B. (Eds.). (2011). *Explaining divergent levels of longevity in high-income countries*. National Academies Press. https://doi.org/10.17226/13089

Diday, E. (2013). Principal component analysis for bar charts and metabins tables. *Statistical Analysis and Data Mining*, *6*(5), 403–430. https://doi.org/10.1002/sam.11188

Diday, E. (2016). Thinking by classes in data science: The symbolic data analysis paradigm. *Wiley Interdisciplinary Reviews: Computational Statistics*, *8*(5), 172–205. https://doi.org/10.1002/wics.1384

Diday, E. (2020). Explanatory tools for machine learning in the symbolic data analysis framework. In E. Diday, R. Guan, G. Saporta, & H. Wang (Eds.), *Advances in data science: Symbolic, complex and network data* (pp. 1–30). Wiley. https://doi.org/10.1002/9781119695110.ch1

Diday, E., Afonso, F., & Haddad, R. (2013). The symbolic data analysis paradigm, discriminant discretization and financial application. In R. Guan, Y. Lechevallier, G. Saporta, & H. Wang (Eds.), *Revue des nouvelles technologies de l'information: Vol. RNTI-E-25. Advances in theory and applications of high dimensional and symbolic data analysis* (pp. 1–14). Editions RNTI.

Eurostat. (2016). *Smoking of tobacco products by sex, age and educational attainment level* [Data Set]. European Commission. http://data.europa.eu/88u/dataset/varzwbq6vy3cfkw2ubitza

Eurostat. (2018a). *Causes of death: Deaths by NUTS 2 region of residence and occurrence, 3 year average* [Data Set]. European Commission. http://data.europa.eu/88u/dataset/m0ppsecjhfgxfmvng9gdg

Eurostat. (2018b). *Causes of death: Deaths by country of residence and occurrence* [Data Set]. European Commission. http://data.europa.eu/88u/dataset/uiak4pd0lanocottq4ebq

Eurostat. (2018c). *Population on 1 January by age and sex* [Data Set]. European Commission. http://data.europa.eu/88u/dataset/wjwcoscim2vainua6qufq

Heijink, R., Koolman, X., & Westert, G. P. (2013). Spending more money, saving more lives? The relationship between avoidable mortality and healthcare spending in 14 countries. *The European Journal of Health Economics*, *14*, 527–538. https://doi.org/10.1007/s10198-012-0398-3

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218. https://doi.org/10.1007/BF01908075

Kerr, J., Anderson, C., & Lippman, S. M. (2017). Physical activity, sedentary behaviour, diet, and cancer: An update and emerging new evidence. *The Lancet Oncology*, *18*(8), E457–E471. https://doi.org/10.1016/S1470-2045(17)30411-4

Lear, S. A., Hu, W., Rangarajan, S., Gasevic, D., Leong, D., Iqbal, R., Casanova, A., Swaminathan, S., & Yusuf, S. (2017). The effect of physical activity on mortality and cardiovascular disease in 130 000 people from 17 high-income, middle-income, and low-income countries: The pure study. *Lancet*, *390*(10113), 2643–2654. https://doi.org/10.1016/S0140-6736(17)31634-3

Lotrič Dolinar, A., Sambt, J., & Korenjak-Černe, S. (2019). Clustering EU countries by causes of death. *38*, 157–172. https://doi.org/10.1007/s11113-019-09518-1

Meslé, F., & Vallin, J. (2002). Mortality in Europe: The divergence between East and West. *Population*, *57*(1), 157–197. https://doi.org/10.3917/popu.201.0171

Meslé, F., & Vallin, J. (2006). The health transition: Trends and prospects. In G. Caselli, J. Vallin, & G. Wunsch (Eds.), *Demography: Analysis and synthesis* (pp. 247–602). Academic Press.

Noirhomme-Fraiture, M., & Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, *4*(2), 157–170. https://doi.org/10.1002/sam.10112

Sapir, A. (2006). Globalization and the reform of European social models. *Journal of Common Market Studies*, *44*(2), 369–390. https://doi.org/10.1111/j.1468-5965.2006.00627.x

Stewart, B. W. (2012). Priorities for cancer prevention: Lifestyle choices versus unavoidable exposures. *The Lancet Oncology*, *13*(3), e126–e133. https://doi.org/10.1016/S1470-2045(11)70221-2

Vallin, J., Meslé, F., & Valkonen, T. (2002). *Trends in mortality and differential mortality*. Council of Europe.

World Health Organization. (2010). *International statistical classification of diseases and related health problems* (10th ed.). https://icd.who.int/

World Health Organization. (2015a). *WHO global report on trends in prevalence of tobacco smoking 2015*. https://apps.who.int/iris/handle/10665/156262

World Health Organization. (2015b). *World health statistics 2015*. https://apps.who.int/iris/handle/10665/170250