ACTAS DEL XVI CONGRESO
DR. ANTONIO A. R. MONTEIRO (2021)
2023, Páginas 149–159

METHODS OF SYMBOLIC DATA ANALYSIS TO OBTAIN INFORMATION FROM THE COVID-19 PANDEMIC

LEONEL GANGA AND ADRIANA MALLEA

ABSTRACT. Since the beginning of the Covid-19 pandemic, researchers have been interested in the different variations of the virus across various parts of the world. There exists a variety of studies using different mathematical models to attempt predictions in this pandemic. This paper is focused on the confirmed cases in countries of Central America published until March 8, 2021 on the website https://github.com/owid/covid-19-data. Techniques of symbolic data analysis (SDA) are applied to describe countries of South America with respect to the characteristics of the evolution of Covid-19. This allows us to visualize comparisons among them and, then, to make a supervised classification that reveals the positioning of each country against the pandemic as regards variables such as the values of the accumulated confirmed cases, the newly daily raising of confirmed cases and the relative number per million population. The main objective of the paper is to show the advantages of working with SDA that allows us to consider the inherent variability of the data on making a temporary grouping of the information since the beginning of the pandemic.

1. INTRODUCTION

Diday [8] introduces symbolic objects and presents a formalization that allows us to address knowledge richer than the common data and establishes a relationship with a classic model of data analysis. Symbolic objects represent an intension, a concept, and is defined, in general terms, as a conjunction of values or a set of values than can be analysed. It constitutes a description in intension of the class of individuals that forms the extension [6, 7].

Symbolic data analysis (SDA) is a generalization of the techniques of data analysis applied to matrices of symbolic data. The definitions and notations of symbolic variables and symbolic objects have been in constant evolution from the beginning [6, 7, 8, 9, 10, 3]. Data science, considered as a science itself, is, in general terms, the extraction of knowledge from data. Data mining is a powerful technology with a great potential to obtain such knowledge. However, from a statistics point of view, their tools have been developed to work with matrices of classic data, that is to say, where every unit is individual and the variables have a unique value for each individual. SDA offers a new way of thinking in data science when extending the standard input to a set of classes of individual entities. Therefore, the classes of a given population are considered higher level units. To take into account the variability between the members of each class, these are described by intervals, distributions, sets of categories or numbers that sometimes are weighted. New types of data are obtained, called 'symbolic' since they cannot be reduced to numbers without losing a lot of information. From a methodological point of view, SDA is a new paradigm that opens a vast domain of research and applications by providing complementary results to the classic methods applied to standard data [12].

2020 *Mathematics Subject Classification.* Primary 62H30.

2. METHODOLOGY

The paper consists of the analysis, using SDA techniques, of Covid-19 data published daily on the website https://github.com/owid/covid-19-data. Concretely, we have worked with the owid-covid-data database published on March 8, 2021. This database consists of the record of 59 variables related to Covid-19 in each country of the world, from December 31, 2020 to March 8, 2021. First, the data for Latin America have been filtered. Then, to build the table of symbolic data, a temporal grouping of the cases has been made, taking the country as a symbolic object and selecting only the variables (the source reference has been maintained):

(1) total_cases: Total of Covid-19 confirmed cases.
(2) new_cases: Covid-19 new confirmed cases.
(3) total_deaths: Total of deaths from Covid-19.
(4) new_deaths: New deaths from Covid-19.
(5) total_cases_per_million: Total of Covid-19 confirmed cases per 1,000,000 people.
(6) new_cases_per_million: New cases of Covid-19 per 1,000,000 people.
(7) total_deaths_per_million: Total of deaths from Covid-19 per 1,000,000 people.
(8) new_deaths_per_million: New deaths form Covid-19 per 1,000,000 people.
(9) total_tests: total of Covid-19 tests.
(10) total_tests_per_thousand: Total of Covid-19 tests per 1,000 people.
(11) Stringency_index: Stringency index on the Government's response towards the Covid-19 pandemic (rating scale from 0 to 100).
(12) population: The number of population in 2020.
(13) population_density: Population density, in the recent year.
(14) aged_65_older: Proportion of the population aged 65 or older, in the recent year.
(15) extreme_poverty: Proportion of the population who lives in extreme poverty, in recent years since 2010.
(16) diabetes_prevalence: Percentage of the diabetic population aged from 20 to 79, year 2017.
(17) female_smokers: Percentage of female smokers, in the most recent year available.
(18) male_smokers: Percentage of male smokers, in the most recent year available.

The descriptions of each variable can be consulted on the website mentioned above. From the temporal aggregation of variables 2, 4, 6 and 8, the corresponding interval variables were obtained. The remaining continuous variables were added to this table. It should be noted that the corresponding value to March 8 (recent value) for the variables 1, 3, 5, 7, 9, 10 and 11 was indicated. The symbolic table was obtained with the DB2SO module, SODAS Software [11]. Then, with the VIEW and DSTAT modules of the same package, the symbolic description, Kiviat diagrams, frequency tables and a histogram of the symbolic variables were obtained.

3. DESCRIPTION AND SYMBOLIC VISUALIZATION

The visualization of an object is made by a graph called Zoom Star. This representation is based on the Kiviat diagrams, where each axis represents a variable. In the same graph, categorical variables, of intervals, with weight, taxonomy, etc. can be represented without overloading the graph. SODAS allows two types of representation: in 2D or 3D, which show different levels of detail. The representation in 2D allows a global impression of the object, meanwhile the representation in 3D gives more detailed information about it. In 2D the axes are joined by a line that connects the most frequent values of each variable. In case

of a tie of the most frequent values in many modalities, the line would join the two of them. When there exists a variable of interval, the line joins the minimum and maximum limits and the internal surface is coloured [2]. In this paper, the visualization in 2D of the SO of interest was chosen. The description and visualization of some countries are represented according to the analysed symbolic variables. So, in the description of the USA (Table 1), it can be observed that on May 8, 2021 there is a total of 29,038,600 confirmed cases of Covid-19, with a total amount of 525,752 deaths. The highest amount of daily cases in this period was 299,786 and of daily deaths was 4,465. The stringency index had a variation of 0 to 75.46 since the pandemic began in the country. The total of Covid-19 tests was 341,100,000, which represents a value of 1,030.51 tests for every 1,000 people (that is to say that there was more than one test per person). The percentage of 65-year-old people or older, of people living in extreme poverty, of diabetic people from 20 to 79 years old, and of women and men who smoke are, respectively, 15.41, 1.20, 10.79, 19.10 and 24.60. In an analogous way, the symbolic description of Chile, Argentina, Brazil and Uruguay is made. In order to make comparisons among the five countries, only the relative variables corresponding to each one are represented with Zoom Star graphs in Figures 1, 2, 3, 4 and 5. The visualization allows us to see that the shapes are very different. In this way, for instance, the USA presents a greater variation in the new cases per million inhabitants and a greater percentage of people of at least 65 years old. Figure 6 shows the superposition of the five graphs, where we can observe the difference among countries according to the considered variables, which leads us to see that although Argentina has a greater amount of total cases per million inhabitants than Chile, there does not exist much difference in the total per million inhabitants in both countries.

TABLE 1. Symbolic description of the countries: USA, Chile, Argentina.

USA	CHL	ARG
total_cases = 29038600.00	total_cases = 860533.00	total_cases = 2154690.00
new_cases = [0.00 : 299786.00]	new_cases = [0.00 : 13990.00]	new_cases = [0.00 : 18326.00]
total_deaths = 525752.00	total_deaths = 21163.00	total_deaths = 53121.00
new_deaths = [0.00 : 4465.00]	new_deaths = [0.00 : 1057.00]	new_deaths = [0.003351.00]
total_cases_per_mill = 87729.30	total_cases_per_mill = 45015.90	total_cases_per_mill = 47674.70
new_cases_per_mill = [0.00 : 905.69]	new_cases_per_mill = [0.00 : 731.84]	new_cases_per_mill = [0.00 : 405.48]
total_deaths_per_mill = 1588.36	total_deaths_per_mill = 1107.07	total_deaths_per_mill = 1175.35
new_deaths_per_mill = [0.00 : 13.49]	new_deaths_per_mill = [0.00 : 55.29]	new_deaths_per_mill = [0.00 : 74.14]
total_tests = 341100000.00	total_tests = 9721540.00	total_tests = 6465610.00
total_tests_per_thou = 1030.51	total_tests_per_thou = 508.55	total_tests_per_thou = 143.06
stringency_index = [0.00 : 75.46]	stringency_index = [0.00 : 87.50]	stringency_index = [11.11 : 100.00]
aged_65_older = 15.41	aged_65_older = 11.09	aged_65_older = 11.20
extreme_poverty = 1.20	extreme_poverty = 1.30	extreme_poverty = 0.6
diabetes_prevalence = 10.79	diabetes_prevalence = 8.46	diabetes_prevalence = 5.50
female_smokers = 19.10	female_smokers = 34.20	female_smokers = 16.20
male_smokers = 24.60	male_smokers = 41.50	male_smokers = 27.70

3.1. **Distribution of symbolic variables.** In Table 3 and Figure 7, we show the distribution of the variables total cases and total deaths per million inhabitants. Both of them show a distribution that presents a positive asymmetry. From the distribution of the first variable, we can see that approximately the first three classes gather 74,28% of the countries with a total of cases that varies from 8.89 to 26,324.88 and that has an average of 22,941.50 and

TABLE 2. Symbolic description of the countries: Brazil and Uruguay.

BRA	URY
total_cases = 11051700.00	total_cases = 64700.00
new_cases = [0.00 : 87843.00]	new_cases = [0.00 : 1514.00]
total_deaths = 266398.00	total_deaths = 658.00
new_deaths = [0.00 : 1986.00]	new_deaths = [0.00 : 17.00]
total_cases_per_mill = 51993.30	total_cases_per_mill = 18625.50
new_cases_per_mill = [0.00 : 413.26]	new_cases_per_mill = [0.00 : 435.84]
total_deaths_per_mill = [1253.29 : 1253.29]	total_deaths_per_mill = [189.42 : 189.42]
new_deaths_per_mill = [0.00 : 9.34]	new_deaths_per_mill = [0.00 : 4.89]
total_tests = 6421440.00	total_tests = 1065600.00
total_tests_per_thou = [0.30 : 30.21]	total_tests_per_thou = [306.76 : 306.76]
stringency_index = [5.56 : 81.02]	stringency_index = [23.15 : 72.22]
aged_65_older = 8.55	aged_65_older = 14.65
extreme_poverty = 3.40	extreme_poverty = 0.10
diabetes_prevalence = 8.11	diabetes_prevalence = 6.93
female_smokers = 10.10	female_smokers = 14.00
male_smokers = 17.90	male_smokers = 19.90

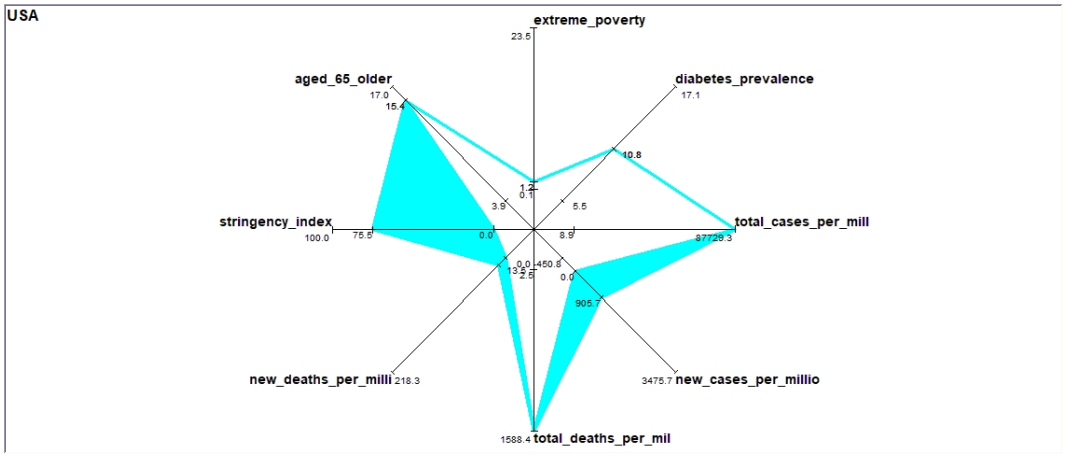

FIGURE 1. Symbolic visualization of the USA.

a spread of 20,305.64, not representatives according to the bias of distribution. The second one shows that almost 50% (focused on the first two classes) of the countries has a total of deaths per million that does not exceed 319.515.

Similarly, Table 4 and Figure 8 show the positive asymmetric distribution of the variables that represent the new cases and new deaths per million people. Approximately 92% of the countries had, in the period considered, a maximum of 695 new daily cases of Covid-19 per

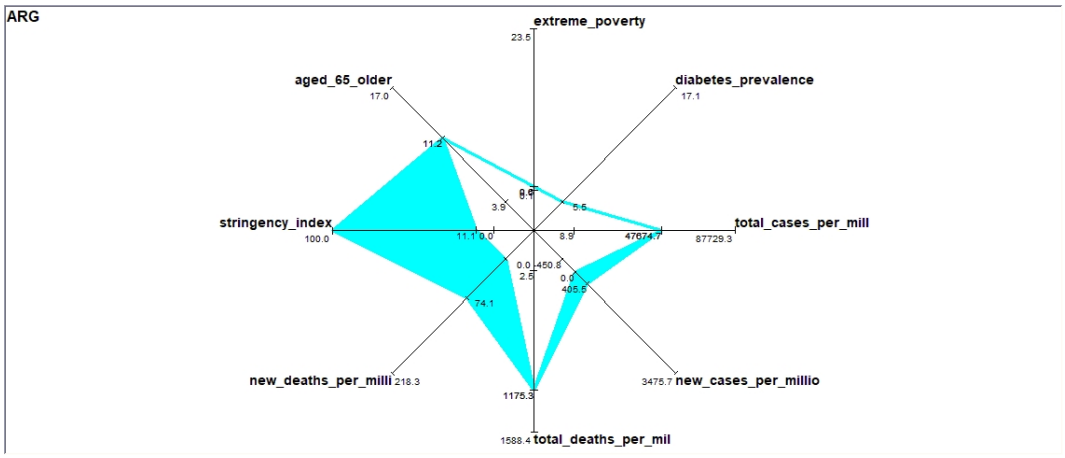

FIGURE 2. Symbolic visualization of Argentina.

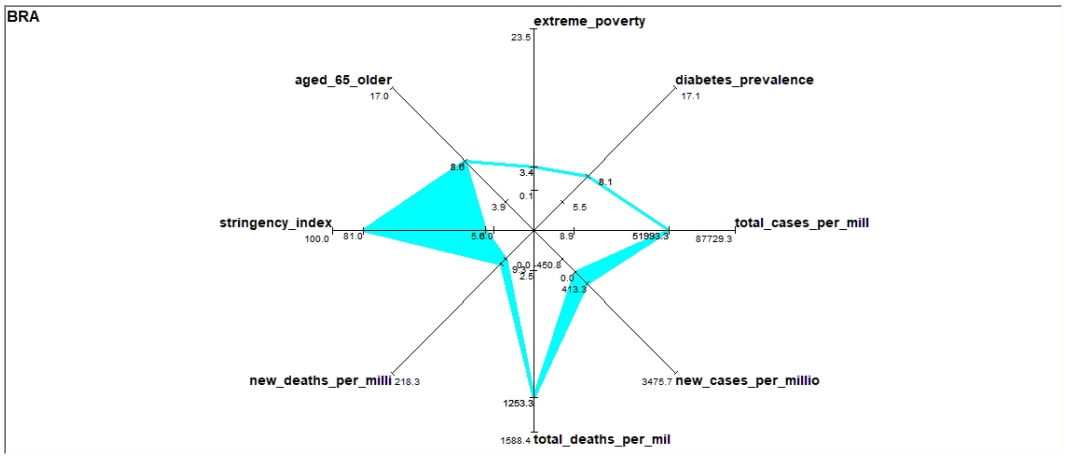

FIGURE 3. Symbolic visualization of Brazil.

million people; 83,58% of the countries present a maximum of 22 new deaths per million inhabitants.

4. CLASSIFICATION

A non-supervising classification is made using the dynamic clustering method [14, 5]. This method iteratively determines a series of partitions that improves on each step according to a grouping criterion. The algorithm is based on:

- prototypes to represent the clusters;
- proximity functions to assign the elements (symbolic objects) to the clusters in each stage.

The grouping criterion to be optimised is based on the addition of proximities between objects and the prototype of the clusters. The function of representation g associates with each partition $P = (C_1, \ldots, C_k)$ in k clusters its representation in k prototypes, that is to say, each class C_i is represented by the prototype $g(i)$. The function of assignment f assigns to the k-tuple of prototypes a k-tuple of classes or clusters of the partitions such that each class will be formed by the elements located at a minimum distance of a determined prototype

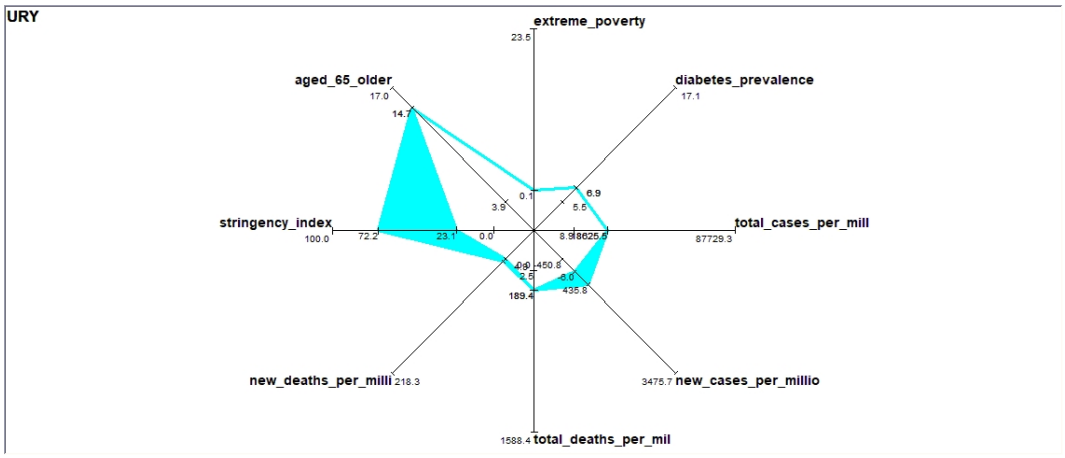

FIGURE 4. Symbolic visualization of Uruguay.

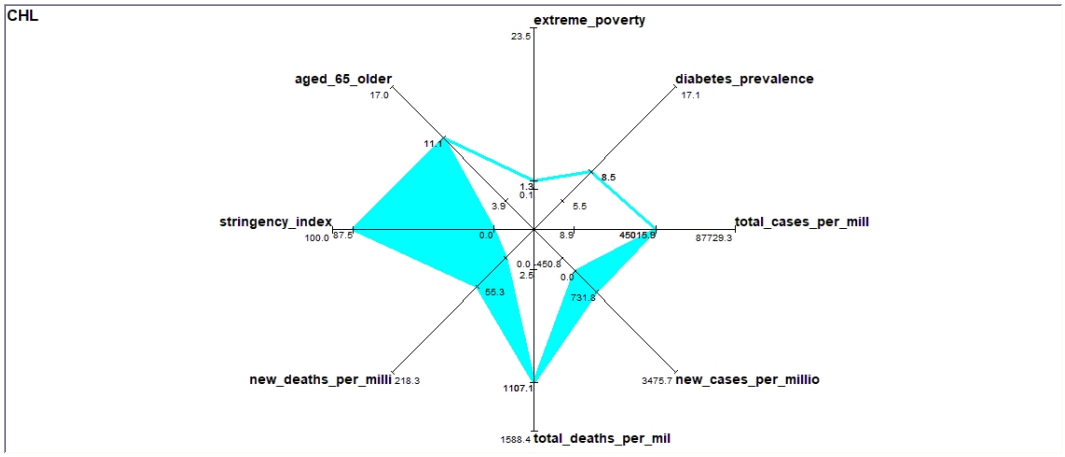

FIGURE 5. Symbolic visualization of Chile.

among all the considered prototypes in this stage:

$$C_l = \left\{ x \in E : d\left(x, g^{(l)}\right) \leq d\left(x, g^{(m)}\right),\ \forall m\ \in \{1, \ldots, k\} \right\}. \tag{1}$$

The optimization problem is to find a partition and its corresponding representation in prototypes that minimize the adequacy criterion W between the partition and its representation:

$$W(P,L) = \sum_{l=1}^{k} D(C_l, g^{(l)}) = \sum_{l=1}^{k} \sum_{x_i \in C_l} d^2(x_i, g^{(l)}), \tag{2}$$

where $g^{(l)}$ is a prototype of C^l; d being the distance chosen for the grouping. Then, the purpose of the method it to minimize the inertia within classes (W) and then to maximize the inertia between classes (B) of the partition. The dynamic clustering method has been programmed in the module SCLUST of the SODAS software. This module is applied to the symbolic table obtained, working with the Hausdorff distance. The output of the module offers the description of classes through its prototypes, the relation with the variables and the extension of the classes, that is to say, the forming countries.

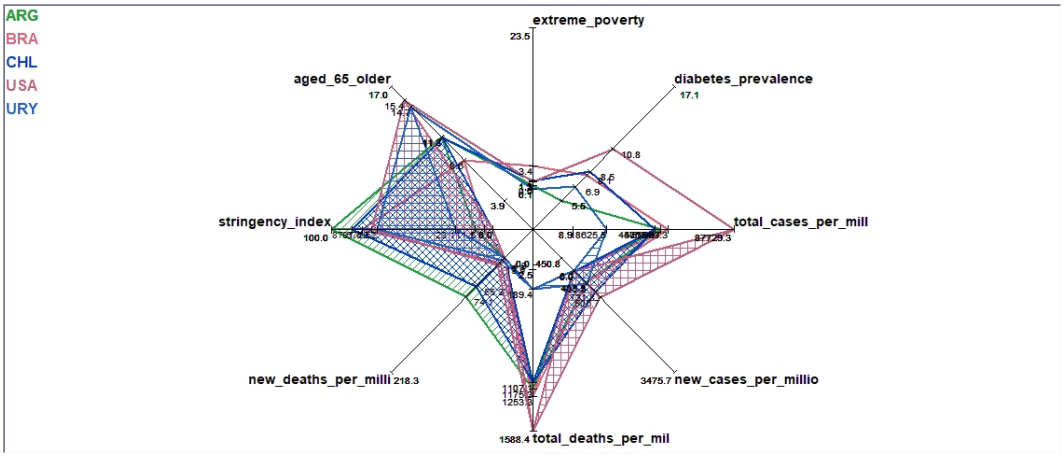

FIGURE 6. Superposition of symbolic objects Argentina, Brazil, Chile, USA, Uruguay.

TABLE 3. Distribution of the variables total cases and total deaths per million.

total_cases_per_million	total_deaths_per_million
limits: 8.887 - 87729.2 class width: 8772.0	limits: 2.515 - 1588.3 class width: 158.5
class 1 0.2571	class 1 0.2788
class 2 0.2571	class 2 0.2182
class 3 0.2286	class 3 0.1273
class 4 0.0286	class 4 0.0667
class 5 0.0571	class 5 0.0060
class 6 0.1143	class 6 0.0303
class 7 0.0000	class 7 0.0606
class 8 0.0000	class 8 0.0909
class 9 0.0000	class 9 0.0303
class 10 0.0571	class 10 0.0909
Central tendency: 22941.5084	Central tendency: 543.1047
Dispersion: 20305.6499	Dispersion: 492.8536

There are three indicators of the number of classes that produces the best grouping: the Calinski & Harabasz index [4], the C-index [13] and the Gamma index [1]. We show the best results for the grouping in four classes, according to Table 5.

The description of the variables is shown in Table 6. The quality is an indicator of the power of discrimination of the variable for the classification. As we can see in the table, the best rates of quality are the ones corresponding to the variables that represent the total of the cases and new cases of Covid-19, new deaths at a lower level and the corresponding relative per million. At the same time, these variables are the ones that contribute the most to the inertia, which is reinforced in the description of the groups. Group 1 is formed only by the USA, the second group is formed by Argentina, Bolivia, Brazil, Colombia, Mexico

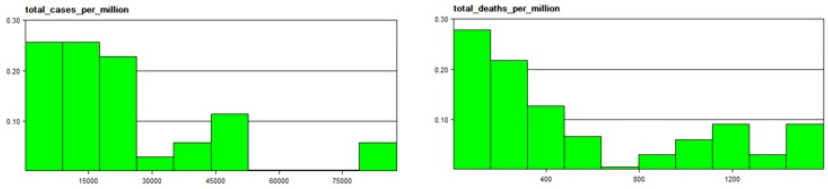

FIGURE 7. Histogram of the variables: Total of cases per million and Total of deaths per million.

TABLE 4. Distribution of the variables: New cases per million and New deaths per million.

new_cases_per_million	new_deaths_per_million
limits: 0.0 - 3475.6 class width: 347.5	limits: 0.0 - 218.3 class width: 21.83
class 1 0.7651	class 1 0.8358
class 2 0.1489	class 2 0.0760
class 3 0.0401	class 3 0.0326
class 4 0.0183	class 4 0.0165
class 5 0.0082	class 5 0.0130
class 6 0.0075	class 6 0.0117
class 7 0.0033	class 7 0.0054
class 8 0.0029	class 8 0.0030
class 9 0.0029	class 9 0.0030
class 10 0.0029	class 10 0.0030
Central tendency: 327.7396	Central tendency: 19.7751
Dispersion: 391.8566	Dispersion: 26.4708

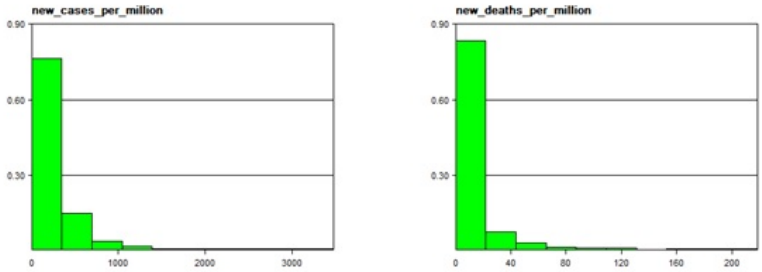

FIGURE 8. Histogram of the variables: New cases per million and New deaths per million.

and Peru. The fourth group is formed by Panama and the third group is formed by the rest of the American countries, mainly from Central America.

TABLE 5. Index to determine the best grouping.

ngps	C-H	C-Index	Gamma
4	103.37	0.01847	0.85145
3	6.60	0.14164	0.86644
2	13.59	0.14138	0.92312

TABLE 6. Quality measures and contribution of the variables.

Variable	Bj/Tj	Contribution	Quality
total_cases	70.02	16.51	1.65
new_cases	72.30	17.05	1.70
total_deaths	59.71	14.08	1.41
new_deaths	60.06	14.16	1.42
total_cases_per_million	46.48	10.96	1.10
new_cases_per_million	15.41	3.63	0.36
total_deaths_per_million	59.83	14.11	1.41
new_deaths_per_million	16.02	3.78	0.38
stringency_index	9.13	2.15	0.22
diabetes_prevalence	15.10	3.56	0.36

From the description and symbolic visualization of the prototypes (Table 7 and Figure 9), it is observed that the groups are formed according to the countries most affected by Covid-19 in terms of the variables that discriminate the most.

TABLE 7. Symbolic description of the prototypes corresponding to the partition.

Prototype_1 /4	Prototype_2 /4	Prototype_3 /4	Prototype_4 /4
total_cases = 29038600.00	total_cases = 2130480.00	total_cases = 12335.00	total_cases = 345236.00
new_cases = [0.00 : 299786.00]	new_cases = [0.00 : 21078.00]	new_cases =[0.00 : 480.00]	new_cases = [0.00 : 5186.00]
total_deaths = 525752.00	total_deaths = 53121.00	total_deaths = 251.00	total_deaths = 5934.00
new_deaths = [0.00 : 4465.00]	new_deaths = [0.00 : 1986.00]	new_deaths = [0.00 : 9.00]	new_deaths = [0.00 : 61.00]
total_cases_per = 87729.30	total_cases_per = 44786.40	total_cases_per = 11205.80	total_cases_per = 80012.60
new_cases_per_m =[0.00 : 905.69]	new_cases_per_m= [0.00 : 413.26]	new_cases_per_m =[0.00 : 236.27]	new_cases_per_m =[0.00 : 1201.92]
total_deaths_pe = 1588.36	total_deaths_pe = 1190.93	total_deaths_pe = 214.44	total_deaths_pe = 1375.28
new_deaths_per_ = [0.00 : 13.49]	new_deaths_per = [0.00 : 55.29]	new_deaths_per_ = [0.00 : 5.00]	new_deaths_per_ = [0.00 : 14.14]
stringency_inde = [0.00 : 75.46]	stringency_inde = 9.73	stringency_inde = [20.37 : 79.63]	stringency_inde = [8.33 : 93.52]
diabetes_preval = 10.79	diabetes_preval = 7.44	diabetes_preval = 10.71	diabetes_preval = 8.33

5. DEBATE

In this paper, we demonstrate that by employing temporal grouping, the data presented in a classical matrix can be transformed into symbolic data, thereby preserving their inherent variability. In particular, we work with the owid-covid-data published on March 8, 2021, from which only cases in American countries are filtered. In this way, we obtain symbolic description of objects, i.e. American countries, that shows the variations of the observed

FIGURE 9. Superposition of the prototypes of the classification.

characteristics form the beginning of the Covid-19 pandemic. We also focus on the possibility of exploring visualization by means of 2D graphs that, by superposition, allow us to make a quick comparison among countries in terms of the variables involved in the study.

The power of SDA to obtain homogeneous classes of countries is shown to apply methods of symbolic clusterig. We make a non-supervising classification of symbolic objects using the dynamic clustering method, showing that the countries are grouped according to the total number of cases and new cases of Covid-19 and, to a lesser degree, by new deaths and new deaths per million inhabitants.

It is worth mentioning that the results obtained may not be relevant in the future due to the dynamic evolution of the pandemic in most countries. However, we intend to show the advantages of these methods, which allow the intrinsic variability of the data to be taken into account.

REFERENCES

[1] F. B. Baker and L. J. Hubert, Measuring the power of hierarchical cluster analysis, *J. Amer. Statist. Assoc.* **70** (1975), no. 349, 31–38. https://doi.org/10.2307/2285371.

[2] H. Bock and E. Diday (editors), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer, Berlin, 2000.

[3] P. Brito, Symbolic data analysis: another look at the interaction of data mining and statistics, *WIREs Data Mining Knowl. Discov.* **4** (2014), no. 4, 281–295. https://doi.org/10.1002/widm.1133.

[4] T. Caliński and J. Harabasz, A dendrite method for cluster analysis, *Comm. Statist.* **3** (1974), 1–27. MR 0375641.

[5] M. Chavent, F. de A. T. de Carvalho, Y. Lechevallier, and R. Verde, Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle, *Rev. Statist. Appl.* **51** (2003), no. 4, 5–29. http://www.numdam.org/item/?id=RSA_2003__51_4_5_0.

[6] E. Diday, Introduction à l'approche symbolique en analyse des données, in: *Journées "Symbolique-Numérique" pour l'apprentissage des connaissances à partir de données*, 21–56, CEREMADE, Université Paris IX - Dauphine, 1987.

[7] E. Diday, Introduction à l'analyse des donnéss symboliques: objets symboliques modaux et implicites, in: *Deuxièmes journées "Symbolique-Numérique" pour l'apprentissage de connaissances à partir de données*, 127–139, LRI, Université d'Orsay, 1988.

[8] E. Diday, Des objets de l'analyse de données à ceux de l'analyse des connaissances, in: Y. Kodratoff and E. Diday (editors), *Induction symbolique et numérique à partir de données*, 7–95, Cépaduès, Toulouse, 1991.

[9] E. Diday, *An Introduction to Symbolic Data Analysis*, Tutorial at IFCS '93, Rapport de recherche no. 1936, INRIA, Roquencourt, 1993. https://inria.hal.science/inria-00074738.
[10] E. Diday, *Quelques Aspects de l'Analyse des Données Symboliques*, Rapport de recherche no. 1937, INRIA, Rocquencourt, 1993. https://inria.hal.science/inria-00074737.
[11] E. Diday and M. Noirhomme-Fraiture (editors), *Symbolic Data Analysis and the SODAS Software*, Wiley, 2008.
[12] E. Diday, Thinking by classes in data science: the symbolic data analysis paradigm, *Wiley Interdiscip. Rev. Comput. Stat.* **8** (2016), no. 5, 172–205. MR 3544255.
[13] L. J. Hubert and J. R. Levin, A general statistical framework for assessing categorical clustering in free recall, *Psychol. Bull.* **83** (1976), no. 6, 1072–1080. https://doi.org/10.1037/0033-2909.83.6.1072.
[14] R. Verde, F. A. T. de Carvalho, and Y. Lechevallier, A dynamical clustering algorithm for symbolic data, in *Tutorial on Symbolic Data Analysis*, held during the 25th Annual Conference of the Gesellschaft für Klassifikation (GfKl), University of Munich, 2001.

(Leonel Ganga) FACULTAD DE CIENCIAS EXACTAS, FÍSICAS Y NATURALES, UNIVERSIDAD NACIONAL DE SAN JUAN, SAN JUAN, ARGENTINA
Email address: leonel.ganga@unsj-cuim.edu.ar

(Adriana Mallea) FACULTAD DE FILOSOFÍA, HUMANIDADES Y ARTES, UNIVERSIDAD NACIONAL DE SAN JUAN, SAN JUAN, ARGENTINA
Email address: lamallea@ffha.unsj.edu.ar