INNOVATION, ENTREPRENEURSHIP AND MANAGEMENT SERIES

BIG DATA, ARTIFICIAL INTELLIGENCE AND DATA ANALYSIS SET

Volume 4

Advances in Data Science

Symbolic, Complex and Network Data

Edited by
Edwin Diday, Rong Guan
Gilbert Saporta and Huiwen Wang

iSTE

WILEY

Advances in Data Science

Big Data, Artificial Intelligence and Data Analysis Set

coordinated by
Jacques Janssen

Volume 4

Advances in Data Science

Symbolic, Complex and Network Data

Edited by

Edwin Diday
Rong Guan
Gilbert Saporta
Huiwen Wang

iSTE

WILEY

First published 2020 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2019951813

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78630-576-3

10

Clustering and Generalized ANOVA for Symbolic Data Constructed from Open Data

10.1. Introduction

Official statistics are very important sources of *open data* where National Statistical Offices play a vital role. More and more societies favor the idea of freely available data and, therefore, many governmental institutions have also established open data websites. At the international level, such sources of open data are, for example, the United Nations open data website [UN 17], The World Bank Open Data [WB 17], and The European Union Open Data Portal [EUR 17]. A commonly used technique to present their data in a transparent and compact way is *aggregation*. There are several important properties and advantages of data aggregation:

– it is usually the first step to make a large amount of data manageable;

– it extracts (first) information from *big data*;

– it protects the privacy of individuals (persons, companies etc.);

– it produces second-level units of data.

Aggregated data present original individual units at a higher level, which enables a different view of the data. *Symbolic Data Analysis (SDA)* provides tools for the analysis of such higher second-level units. *Second-level units* in SDA are called *concepts* or *classes* (Diday, inspired by Aristotle's collection of works on logic *The Organon* [ARI] in which he distinguishes between first-level objects called

Chapter written by Simona KORENJAK-ČERNE, Nataša KEJŽAR and Vladimir BATAGELJ.

Advances in Data Science: Symbolic, Complex and Network Data,
First Edition. Edited by Edwin Diday, Rong Guan, Gilbert Saporta and Huiwen Wang.
© ISTE Ltd 2020. Published by ISTE Ltd and John Wiley & Sons, Inc.

individuals and second-level *objects*). They represent a natural extension of aggregated descriptions of individuals.

SDA is an extension of the standard data analysis. Following the SDA approach, the aggregation process returns second-level units, *symbolic objects* (SOs), in which more information is usually preserved (e.g., a frequency distribution of individual values instead of just a mode). In order to find the answers to theoretical hypotheses, symbolic data tables with complex/structured data as table entries are the input to the SDA methods (several practical examples can be found in the SDA literature, for example, in [BIL 06], [NOI 11], [BRI 14] or [DID 16]).

In this chapter, we present a review of our contributions in one of such SDA topics, namely, clustering, adapted for symbolic data represention based on distributions of values. The adaptation of the classical methods was directly motivated by analyses of open data sets. It can be used with several dissimilarites [BAT 15a]. The usage is illustrated with applications on two different open data sets:

– TIMSS (Trends in International Mathematics and Science Study) by combining teachers' and students' data sets [KOR 11]; and

– countries' data descriptions based on their age–sex population distributions [KOR 15].

Furthermore, we present some basic ideas on how to generalize the well-known analysis of variance (ANOVA) for cases where no assumptions from classical ANOVA hold [BAT 15b]. The generalized method can be used on the described second-level units that we demonstrate on the example of population pyramids and HDI index.

10.2. Data description based on discrete (membership) distributions

With aggregation, a large set of (primary) units is partitioned into mutually disjoint sets/groups $P = \{P_j\}$. The representation of the group P_j is a second-level unit X_j. In this chapter, we discuss the case when an aggregated unit is represented by a distribution of values (with frequencies, relative frequencies or subtotals). We call this distribution a *discrete (membership) distribution*. Its categories are discrete values of primary units that have been aggregated.

More formally, to obtain such a representation, the domain of each variable $V_i (i = 1, \cdots, m)$ is partitioned into k_i subsets $\{\mathcal{V}_{ij},\ j = 1, \ldots k_i\}$. The *set of (second-level)* units $\mathbf{U}$ consists of *symbolic objects*. An SO X is described with a list of descriptions of variables $V_i, i = 1, \ldots, m$:

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m], \qquad [10.1]$$

where m denotes the number of variables, and $\mathbf{x}_i$ is a list of numerical values (usually, frequencies or subtotals over the corresponding groups)

$$\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ik_i}].$$

The same description (with the list of values for each symbolic variable) is used for a description of a cluster of symbolic objects C.

Such a representation of SO is based on *weighted modal* (or *histogram*, if $\mathcal{V}_{ij}, i = 1, \ldots m, j = 1, \ldots, k_i$, are intervals) type of symbolic variables. Let the sum of values of a variable V_i be denoted with n_{xi}

$$n_{xi} = \sum_{j=1}^{k_i} x_{ij}.$$

Then, the corresponding empirical probability distribution is

$$\mathbf{p}_{xi} = \frac{1}{n_{xi}} \mathbf{x}_i = [p_{xi1}, p_{xi2}, \ldots, p_{xik_i}].$$

A symbolic object X is in this way described with a list of couples

$$X = [(n_{x1}, \mathbf{p}_{x1}), (n_{x2}, \mathbf{p}_{x2}), \ldots, (n_{xm}, \mathbf{p}_{xm})]. \qquad [10.2]$$

The advantages of such a data description are:

– the description of each group has a fixed size;

– we can *deal with variables* that are *based on a different number of original (individual) units*;

– it *preserves more information* about the original first-level (primary) units and about groups than the usual one-value of an appropriate statistic – e.g., a mean value used in the classical approach;

– it produces *uniform descriptions* for all measurement types of variables;

– it is also *compatible* with the merging of disjoint clusters, i.e., knowing the descriptions of clusters C_1 and C_2, $C_1 \cap C_2 = \emptyset$, we can easily calculate the empirical probability distribution of their union as a weighted sum.

The second property is very useful when a data set is a combination of more than one initial data set, e.g., in the application on TIMSS data [KOR 11], or when we study demographic structures, e.g., age–sex structures [KOR 15] or causes of deaths by age and gender.

10.3. Clustering

Cluster analysis or *clustering* is the task of assigning a set of objects into groups called clusters so that the objects in the same cluster are *more similar* to each other than to those in other clusters. When we want to use clustering for solving the concrete research problem, the choice of a dissimilarity measure significantly affects the clustering result. It is, therefore, of crucial importance (1) how we choose a proper dissimilarity to reveal the structure that we are looking for, and (2) how we select a proper method to obtain optimal cluster representatives (that answers the initial research questions). We mainly focus on the latter issue in our adaptation of the clustering methods. We describe the issue of dissimilarity selection only in relation to the adaptation of methods (but for more, see [KEJ 11]).

We define a clustering problem as an optimization problem to find a partition $\mathbf{C}^*$ in a set of feasible *partitions* Φ for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C}\in\Phi} P(\mathbf{C}),$$

where $P(\mathbf{C})$ is a criterion function. $P(\mathbf{C})$ is based on the dissimilarities between units and/or cluster representatives.

For solving the clustering problem for SOs described with discrete distributions, we adapted the following classical clustering methods:

– the *leaders method* (a generalization of the k-means method [AND 73], [HAR 75] and dynamic clouds [DID 79]);

– the *agglomerative hierarchical clustering method* (for example, Ward's hierarchical clustering method [WAR 63]).

Besides a separate usage of each of them, one can combine both methods if they are based on the same criterion function (namely, use the same dissimilarity measure).The leaders method can be used with large data sets; however, the number of clusters has to be prespecified. An application of the compatible (based on the same criterion function) hierarchical method on the sample can be helpful to determine the number of expected clusters. A compatible hierarchical clustering method can also be used after the leaders method on its resulting clustering to uncover the structure of clustering and the number of clusters.

There are two basic choices in the leaders method:

– how we select a **representation** of units, clusters, and cluster representatives;

– which **dissimilarity measures** we use between units, clusters, and unit and cluster representative.

The main aim of our adapted methods is to obtain optimal clusters' representatives that resolve the following issues:

– **To consider demographic structure as SO**: *the optimal cluster representative* should be *meaningful (interpretable)*, namely, it should represent the demographic structure of the population of all units from the cluster. This was the motivation for the inclusion of weights into the representation of SOs and into the clustering criterion function.

– **Patents' citation data set**: the *error measure/dissimilarity* should consider all component values of a variable equally (for example, squared Euclidean distance favors the largest component value). This was the motivation for proposing alternative dissimilarities.

Our approach is based on the additive model. The criterion function $P(\mathbf{C})$ is the *sum of all cluster errors*. The error of a cluster $p(C)$ is the sum of dissimilarities of its units from the cluster's *optimal representative – leader* T_C.

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C) \qquad \text{where} \qquad p(C) = \sum_{X \in C} d(X, T_C).$$

The set of feasible partitions Φ is *a set of partitions* into k clusters of a finite set of units $\mathbf{U}$. We assume that a leader has the same structure of description as SOs (see [10.1]), i.e., it is represented with nonnegative vectors $\mathbf{t}_i$ of the size k_i for each variable V_i – its representation space is $\mathcal{T} = (\mathbb{R}_0^+)^{k_1} \times (\mathbb{R}_0^+)^{k_2} \times \cdots \times (\mathbb{R}_0^+)^{k_m}$.

For a given representative $T \in \mathcal{T}$ and a cluster C, we define the cluster error with respect to T:

$$p(C, T) = \sum_{X \in C} d(X, T),$$

where d is the selected dissimilarity measure. The best representative – leader T_C – is then the one that minimizes the sum of errors within the cluster

$$T_C = \arg\min_{T} p(C, T).$$

Then, we define

$$p(C) = p(C, T_C) = \min_{T} \sum_{X \in C} d(X, T).$$

A dissimilarity measure between SOs and T is defined as a weighted average (convex combination)

$$d(X, T) = \sum_{i=1}^{m} \alpha_i d_i(\mathbf{x}_i, \mathbf{t}_i), \quad \alpha_i \geq 0, \quad \sum_{i=1}^{m} \alpha_i = 1,$$

where α_i are weights for variables. They allow specifying the importance of the variables by the user. If not determined otherwise, they are all set to $\alpha_i = \frac{1}{m}$.

For each variable, we set

$$d_i(\mathbf{x}_i, \mathbf{t}_i) = \sum_{j=1}^{k_i} w_{xij}\delta(p_{xij}, t_{ij}), \quad w_{xij} \geq 0,$$

where w_{xij} are weights for each variable's component and δ is a basic dissimilarity.

The adapted clustering methods are implemented in the R package `clamix` [BAT 19] that supports the clustering of (very) large data sets of mixed (measured in different scales) units. Basic dissimilarities δ included in the R package `clamix` can be found by Batagelj *et al.* [BAT 15a]. For example, the selection of $\delta = (p_x - t)^2$ represents an extension of the squared Euclidean distance on SOs described with discrete distributions. Five other proposed basic dissimilarities for δ represent relative error measures proposed by Kejžar *et al.* [KEJ 11], extended on SOs.

New leader T_C of the cluster C is determined with

$$\begin{aligned} T_C &= \arg\min_T \sum_{X \in C} d(X, T) = \arg\min_T \sum_{X \in C} \sum_{i=1}^{m} \alpha_i d_i(X, T) = \\ &= \arg\min_T \sum_i \alpha_i \sum_{X \in C} d_i(\mathbf{x}_i, \mathbf{t}_i) = \left[\arg\min_{\mathbf{t}_i} \sum_{X \in C} d_i(\mathbf{x}_i, \mathbf{t}_i)\right]_{i=1}^{m}, \end{aligned}$$

where $\mathbf{t}_i = [t_{i1}, \ldots, t_{ik_i}]$. The solution $\mathbf{t}_i$ of the obtained optimization problem depends on the nature of the selected basic dissimilarity δ.

To make the adapted leaders method and the adapted agglomerative hierarchical clustering method compatible, the dissimilarity $D(C_u, C_v)$ in the agglomerative hierarchical clustering is determined by the following formula

$$D(C_u, C_v) = p(C_u \cup C_v) - p(C_u) - p(C_v).$$

The dissimilarity between the two clusters is the same as the cluster error of the merged cluster diminished by both the cluster errors.

For the dissimilarity $\delta = (p_x - t)^2$, we get the generalization of the Ward hierarchical method

$$D(C_u, C_v) = \sum_{i=1}^{m} \alpha_i \sum_{j=1}^{k_i} \frac{w_{uij} \cdot w_{vij}}{w_{uij} + w_{vij}} (u_{ij} - v_{ij})^2,$$

where u_{ij} and v_{ij} are the leader's components of the clusters C_u and C_v, respectively, i.e.,

$$u_{ij} = \frac{1}{w_{uij}} \sum_{X \in C_u} w_{xij} \cdot p_{xij}, \quad w_{uij} = \sum_{X \in C_u} w_{xij}, \quad \text{and}$$

$$v_{ij} = \frac{1}{w_{vij}} \sum_{X \in C_v} w_{xij} \cdot p_{xij}, \quad w_{vij} = \sum_{X \in C_v} w_{xij}.$$

A detailed definition of the methods' compatibility and the derivations for the leaders and for the dissimilarities $D(C_u, C_v)$ can be found by Batagelj *et al.* [BAT 15a].

10.3.1. *TIMSS – study of teaching approaches*

For studying teaching approaches, we used the TIMSS – Trends in International Mathematics and Science Study open data set [IEA 04], [TIM 04] for the years 1999 and 2003 (joint work with Barbara Japelj Pavešić, National Coordinator of the International Research of Trends in Knowledge of Mathematics and Science for Slovenia, The Educational Research Institute, Slovenia). The aim of the study was to find groups of teachers with similar teaching approaches where we combined the data set of teachers' answers with the data set of students' answers. The data set for the year 2003 includes a sample consisting of 6,552 teachers and 131,000 students, representing more than 10 million students of the 8th grade in 30 countries. All answers in the questionnaire were categorized (including age).

The data description used can be explained in a more general framework, i.e., as the so-called ego-centered or personal networks (see the basic scheme in Figure 10.1), that are rather common in social sciences. The ego-centered network consists of two related data sets: egos and alters. Each unit in the first data set, i.e., ego, can be related with different units from the second data set, i.e., alters.

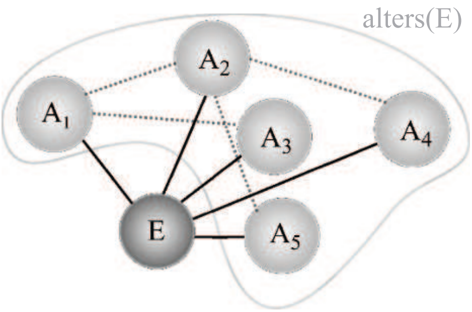


Figure 10.1. *Ego-centered network*

In our study, units of analysis were teachers, described by their variables: gender, age, education, their work in classes, pedagogical approaches used in the class, opinions about mathematics, classroom activities, the use of IT, and the issues on homework.

For each teacher, there was also a description for distributions of students' answers, describing students' attitudes toward mathematics, such as valuing math, enjoying learning, and self-confidence in mathematics, and activities in class such as students' use of IT, their participation in learning mathematics, and their strengths in mathematics. These values were collected with separate questionnaires for students.

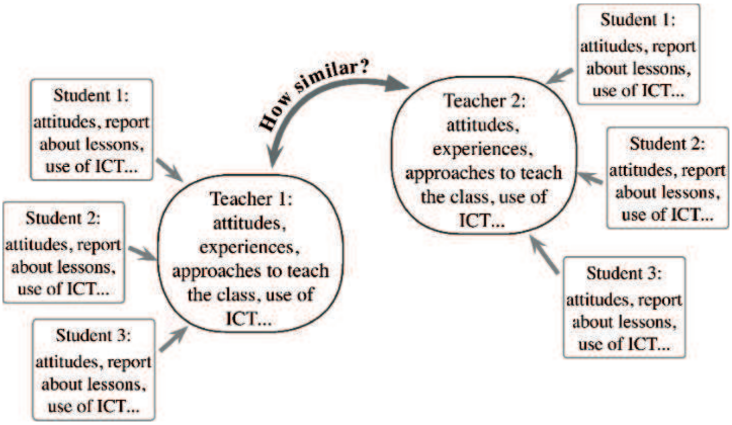

Student 1: attitudes, report about lessons, use of ICT...
Student 2: attitudes, report about lessons, use of ICT...
Student 3: attitudes, report about lessons, use of ICT...
Teacher 1: attitudes, experiences, approaches to teach the class, use of ICT...
How similar?
Teacher 2: attitudes, experiences, approaches to teach the class, use of ICT...
Student 1: attitudes, report about lessons, use of ICT...
Student 2: attitudes, report about lessons, use of ICT...
Student 3: attitudes, report about lessons, use of ICT...

Figure 10.2. *Teacher–students ego-centered network (Source: [KOR 11])*

We combined both data descriptions into one

$$SO(X) = [X, A(X)],$$

where the units were described with ego X and alters variables $A(X)$ as a symbolic object.

For example, the SO description of the teacher with id 4567 is

$$SO_{4567} = \Big[\underset{\overset{\uparrow}{T_1}}{(1,[0,0,0,1])},\ \underset{\overset{\uparrow}{T_2}}{(1,[0,0,0,0,1,0])},\ \underset{\dots}{\dots}\ \underset{\overset{\uparrow}{S_1}}{(100,[\,0.47,0.16,0.37,0,0])},\ \underset{\overset{\uparrow}{S_2}}{(100,[0,0,1,0])},\ \dots\Big]$$

where teacher variables $(T_1, T_2, \dots)$ have only a singular value, but the alters (students') variables $S_1, S_2, \dots$ contain distributions of students' answers. Most of them are distributed over the following four subsets: 1 = strongly agree, 2 = agree, 3 = disagree and 4 = strongly disagree, which express how much they agree/disagree with the statement that is considered as a student variable.

One hundred and one variables were included in the clustering process: 77 from teachers and 24 from the students' questionnaire. The adapted hierarchical method with squared Euclidean distance was used (without weights). We identified five main clusters. One of them contains units with mostly missing values. Teachers in other clusters differ in the usage of computers and calculators in their lectures, in assigning and monitoring homework and testing the knowledge of their students. We further observed also if there are links between the obtained clusters and other variables that were not included in the clustering process, like students' achievements and the teacher's country of origin. For example, in the TIMSS study, students are assigned to different benchmark levels of mathematical knowledge. The distribution of students reaching benchmarks for four clusters (cluster 2 with missing answers was omitted) is presented in Figure 10.3.

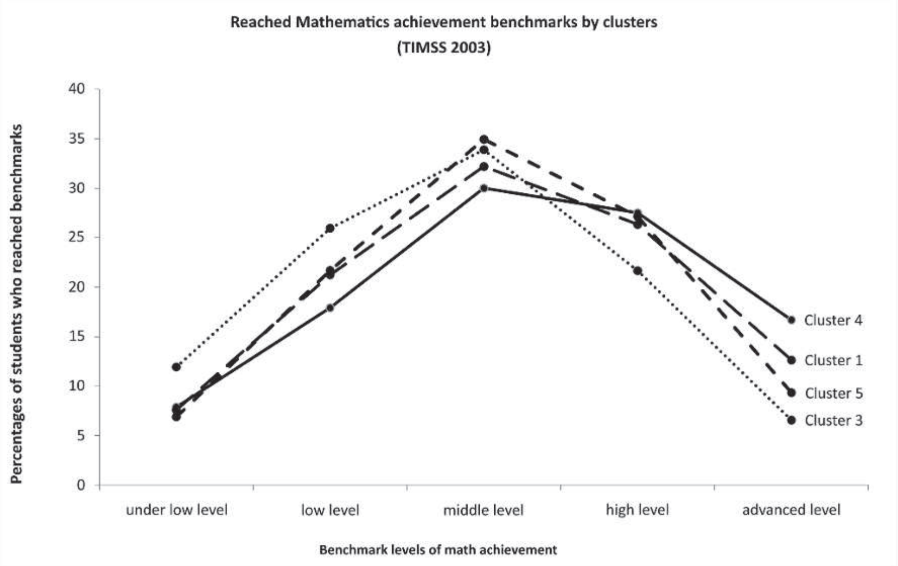


Figure 10.3. *Benchmark levels of mathematics achievement reached by students (source: [KOR 11])*

Additional details on the obtained results can be found in [KOR 11].

10.3.2. *Clustering countries based on age–sex distributions of their populations*

On the web, the data on age–sex distributions (population pyramids) for the countries and for many countries also for their administrative units are openly available. Although the population pyramid is simple and easy to understand, it well reflects characteristics of the observed time and region. It is mostly influenced by population processes (fertility, mortality and migration), and policies (social and political) can also have a strong influence on its shape, e.g., birth control in China, wars, and lifestyle. Because of this, population pyramids are often connected with the developing stage of the represented regions. The base for the graphical representation with the population pyramid is age–sex distribution of the population of the particular region in the particular time. We considered age–sex distribution as SO in the following way: each age–sex distribution (population pyramid) is described with two vectors of frequencies (one for each gender), representing the distributions of men/women by age.

A region X (world country, US county, municipality, and sub-national area) represented with population pyramids (age–sex distributions of the population) and cluster of regions C_u is described with two symbolic variables:

$$X = [(n_{xM}, \mathbf{p}_{xM}); (n_{xF}, \mathbf{p}_{xF})], \qquad C_u = [(n_{uM}, \mathbf{p}_{uM}); (n_{uF}, \mathbf{p}_{uF})]$$

where n_M is the number of men, $\mathbf{p}_M$ is the vector of relative frequencies of men over age groups, n_F is the number of women (female), and $\mathbf{p}_F$ is the vector of relative frequencies of women over age groups.

For example, the population of Ljubljana on July 1, 2011 was split into three economic age groups 0–19, 20–64 and 65+, where $n_{LjM} = 134{,}410$ men and $n_{LjF} = 145{,}488$ women, and the corresponding frequency distributions over the economic age groups are $[25{,}396, 90{,}466, 18{,}548]$ for men and $[24{,}204, 91{,}899, 29{,}385]$ for women.

The description of the corresponding SO (see expression [10.2]) is

$$X_{\mathrm{Lj}} = \Big[\underbrace{(\underset{n_{LjM}}{\underset{\uparrow}{134\,410}}, \underset{\mathbf{p}_{LjM}}{\underset{\uparrow}{[0.189, 0.673, 0.138]}})}_{\text{men}};\ \underbrace{(\underset{n_{LjF}}{\underset{\uparrow}{145\,488}}, \underset{\mathbf{p}_{LjF}}{\underset{\uparrow}{[0.166, 0.632, 0.202]}})}_{\text{women}}\Big]$$

For the cluster C_u, it holds

$$n_{ui} = \sum_{X \in C_u} n_{xi} \quad \text{and} \quad p_{ui} = \frac{1}{n_{ui}} \sum_{X \in C_u} n_{xi} \cdot p_{xi}, \quad i = M, F.$$

The dissimilarity between clusters C_u and C_v is in this case rewritten as

$$D(C_u, C_v) = \frac{1}{2}\Big(\frac{n_{uM} \cdot n_{vM}}{n_{uM} + n_{vM}} ||\mathbf{p}_{uM} - \mathbf{p}_{vM}||^2 + \frac{n_{uF} \cdot n_{vF}}{n_{uF} + n_{vF}} ||\mathbf{p}_{uF} - \mathbf{p}_{vF}||^2\Big).$$

We used the adapted hierarchical clustering method with weighted squared Euclidean distance as dissimilarity for the applications on several open data sets:

– *Slovenian municipalities* on July 1, 2011, where data were obtained from the National Statistical Office of the Republic of Slovenia. Analyses were made with the original data with 21 five-year groups (0–4 years, 5–9 years, 10–14 years, ..., 95–99 years, 100+) and also with data aggregated into three economic age groups (0–14 years, 15–64 years, and 65+ years) (joint work with Jože Sambt, University of Ljubljana, Faculty of Economics [KOR 12]);

– *population pyramids of the world countries* obtained from the International Data Base (IDB) [US 08]. The data are divided into 17 five-year groups (0–4 years, 5–9 years, 10–14 years, ..., 75–79 years, 80+), where we also observed time changes with the 5-year time-lag (for years 1996, 2001 and 2006);

– *US counties* from US Census 2000 Summary File 1, prepared by U.S. Census Bureau [US 11]. Data for the year 2000 include $3,219$ US counties. The data for the year 2010 include $3,221$ US counties with the additional variable ethnicity that was included in our analysis;

– *Brazilian municipalities* with IBGE – Brazilian Institute of Geography and Statistics data [BRA 14], where we analyzed $5,570$ municipalities for 2010. We used data descriptions based on the age–sex structures and also on age–area (urban/rural) structures;

– *sub-national areas in Latin America and the Caribbean* with IPUMS dataset of census microdata from 1960 to 2011 (joint work with Ludi Simpson, University of Manchester, UK).

The main characteristic of the adapted clustering method based on squared Euclidean distance is that with the inclusion of sizes as weights for each variable (the number of men/women) into the clustering process, the obtained optimal cluster representative is again age–sex distribution of the region determined by the corresponding cluster (thus, we get meaningful cluster representative). Note, however, that age groups are considered as categories (not intervals and without ordering).

The main aim of the analysis of the countries based on their age–sex distributions, obtained with the adapted clustering methods, was to identify groups of countries with similar age–sex structures and to identify groups of countries with similar structural changes over time [KOR 15]. In order to achieve a relevant comparison, 215 of the countries for which data were available at all three time points (1996, 2001, and 2006) were included in the analysis. With the symbolic data descriptions of the population age–sex distributions, we save complete information about distributions, and with the inclusion of the sizes as weights into the clustering process, we obtained meaningful optimal cluster representatives, i.e., age–sex distributions of the population included in the countries inside the clusters. We identified four main clusters for each of the observed years. Their shapes rather well reflect basic demographical developing stages. Clusters are studied in detail also for partitions at lower levels.

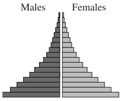
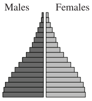
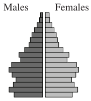
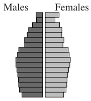
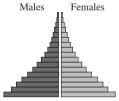
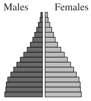
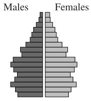
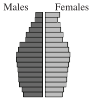
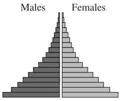
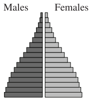
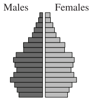
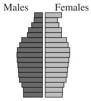
Year	Description	A	B	C	D
1996	age-sex dist.				
	# countries	69	60	30	56
	population size	829,142,426	2,244,626,554	1,497,140,509	1,183,906,689
	% all population	**14.4**	**39.0**	**26.0**	**20.6**
2001	age-sex dist.				
	# countries	77	49	40	49
	population size	991,132,385	2,381,286,016	1,578,815,287	1,188,761,765
	% all population	**16.1**	**38.8**	**25.7**	**19.4**
2006	age-sex dist.				
	# countries	50	59	40	66
	population size	776,127,063	2,356,718,043	2,080,538,430	1,303,900,362
	% all population	**11.9**	**36.2**	**31.9**	**20.0**

Figure 10.4. *Four main clusters obtained from 215 of the countries for the years 1996, 2001 and 2006 with their population pyramids, number of countries in each cluster, and overall population (source: [KOR 15])*

Observation over time showed that the shapes of the sex–age structures of the countries mostly changed from the more expansive shape (with a large number of people in young ages and fast decline for older people) to stationary or even constrictive shape that usually express a more developed stage (with lower fertility and mortality rate and longer life expectancy). Further observation of the time changes revealed five main clusters of similar time changes of population age–sex distributions over the observed time. More details about the results can be found in [KOR 15].

The application to sub-national age–sex distributions from Latin America and the Caribbean [KOR 17] was a part of a wider project compared to sub-national demographic development in Latin America and the Caribbean [SIM 16]. The main focus of the study was to examine sub-national time series of age–sex structures for many countries in Latin America and the Caribbean, to summarize the diversity and the socio-demographic associates of changing age–sex structures, and to identify and characterize the development of those age–sex structures over time, useful to the practice of demographic projections. As a clustering result, we identified four main shapes for the population pyramids that are strongly related to the additional socio-demographic indicators for clusters' descriptions. Most of the time movements of the observed regions were from clusters with indicators expressing less developed stages to more developed ones. Observation of population pyramids over time revealed that the shape of the age–sex structure of some of the areas significantly changed over the observed time period from 1960 to 2010 [e.g., Federal District (Brazil)]. The changes can be explained with the additional knowledge of special circumstances in this area. The clustering method also revealed some areas with rather unusual shapes that require a more detailed study of the data and of social and political situations in the observed area and time. Dissimilarities among structures in different decades indicate that age–sex structures of the observed areas become more similar over time.

10.4. Generalized ANOVA

ANalysis Of VAriance (ANOVA) is one of the most common statistical approaches for detecting "differences" among groups/clusters. It is based on

– squared Euclidean distance;

– sum-of-squares decomposition equality: $SS_T = SS_B + SS_W$ where SS_T stands for *total sum of squares* (deviations of the values around the total mean), SS_B for *between-group sum of squares* (deviations of group means around total mean), and SS_W for *within-group sum of squares* (sum of the deviations of values around group mean - sum of the group errors);

– assumptions about distributions — normal distributions with equal variances.

We are interested in an extension/adaptation of the ANOVA method with a general measure of spread. We present here our basic ideas about a possible generalization of the standard approach that enables a more general usage. We propose to combine some available theoretical results for each of the following three main steps: 1) selection of an appropriate measure of spread; 2) construction of a test statistic for non-parametric multivariate analysis; and 3) calculation of a P-value.

The sum of squares of the group can also be viewed as the *error of the group*, denoted by $p(C)$ (see the previous section), or the *inertia*, sometimes denoted by $I(C)$. From mechanics, we know the Huygens theorem

$$I_T = I_B + I_W,$$

where I_T stands for total inertia, I_B for between-group inertia, and I_W for within-group inertia.

For the basic dissimilarity $\delta = (p_x - t)^2$, the total inertia is

$$I_T = \sum_{X \in \mathbf{U}} d(X, T_U), \text{ where}$$

$d(X, T_U) = \sum_i \alpha_i \; d_i(\mathbf{x}_i, \mathbf{t}_{Ui}) = \sum_i \alpha_i \; w_{xi} ||\mathbf{p}_{xi} - \mathbf{t}_{Ui}||^2$, and $\mathbf{t}_{Ui} = \dfrac{1}{\sum_{X \in \mathbf{U}} w_{xi}} \sum_{X \in \mathbf{U}} w_{xi} \cdot \mathbf{p}_{xi}$, where $\mathbf{U}$ denotes the whole set of units.

The between inertia is

$$I_B = \sum_{C \in \mathbf{C}} d(T_C, T_U),$$

with $d(T_C, T_U) = \sum_i \alpha_i \; w_{Ci} ||\mathbf{t}_{Ci} - \mathbf{t}_{Ui}||^2$.

And the within inertia is

$$I_W = P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C) = \sum_{C \in \mathbf{C}} \sum_{X \in C} d(X, T_C),$$

where $d(X, T_C) = \sum_i \alpha_i \; d_i(\mathbf{x}_i, \mathbf{t}_{Ci}) = \sum_i \alpha_i \; w_{xi} ||\mathbf{p}_{xi} - \mathbf{t}_{Ci}||^2$, and the representative of variable i: $\mathbf{t}_{Ci} = \frac{1}{w_{Ci}} \sum_{X \in C} w_{xi} \cdot \mathbf{p}_{xi}$, where $w_{Ci} = \sum_{X \in C} w_{xi}$.

With a general dissimilarity, using the ideas from [BAT 88]:

1. Define the *cluster error* $p(C)$

$$p(C) = \frac{1}{2 \cdot w(C)} \sum_{X \in C} \sum_{Y \in C} w(X) \cdot w(Y) \cdot d(X, Y)$$

which is a generalization of the classical formula for the squared Euclidean distance

$$p(C) = \frac{1}{2 \cdot n_C} \sum_{X \in C} \sum_{Y \in C} ||X - Y||^2.$$

2. Introduce a generalized (possible imaginary) center $\tilde{C}$ of a cluster C defined with the extension of a dissimilarity to units and cluster centers

$$d(Y, \tilde{C}) = d(\tilde{C}, Y) = \frac{1}{w(C)} \left(\sum_{X \in C} w(X) \cdot d(X, Y) - p(C) \right),$$

where Y is a unit or a cluster center. Definition of the generalized center is based on the classical formula for the center $\bar{C} = \arg\min_Y \sum_{X \in C} ||X - Y||^2$ with the squared Euclidean distance and equality

$$||Y - \bar{C}||^2 = \frac{1}{n_C} \sum_{X \in C} \left(||X - Y||^2 - ||X - \bar{C}||^2 \right).$$

3. The generalized Huygens theorem holds:

$$I_T = I_B + I_W,$$

where

$$I_T = p(\mathbf{U}) = \frac{1}{2 \cdot w(\mathbf{U})} \sum_{X,Y \in \mathbf{U}} w(X) \cdot w(Y) \cdot d(X,Y),$$

$$I_W = \sum_{C \in \mathbf{C}} p(C),$$

$$I_B = \sum_{C \in \mathbf{C}} w(C) \cdot d(\tilde{C}, \tilde{\mathbf{U}}) = I_T - I_W.$$

The problem might occur since the extended "dissimilarity" between (imaginary) center and each unit $d(Y, \tilde{C})$ is not necessary nonnegative for *every dissimilarity*. In [BAT 88], it is shown that the triangle inequality is a sufficient condition for the extended dissimilarity $d(Y, \tilde{C})$ to be nonnegative. Therefore, in the next step, we show how it is possible to produce a dissimilarity from a general one that can be used in the generalized ANOVA process.

In [JOL 86], for general dissimilarity measure d, there exists a unique nonnegative real number p, called metric index, such that d^α is a metric[1] for all $\alpha \leq p$, and d^α is not a metric for all $\alpha > p$. If a dissimilarity d is not a metric, it can be transformed into it using the power transformation. Therefore, we can first find metric index p of arbitrarily chosen dissimilarity d and in the generalized Huygens theorem

– use d if $p \geq 1$;

– otherwise (if $p < 1$) use d^p.

The test statistic for the generalized ANOVA that we used is in line with the approach of Anderson and McArdle [AND 01], [MCA 01], which was applied to ecology data, and the approach of Studer *et al.* [STU 11], applied to the life trajectory analysis. The construction of their test statistic is based on the ratio of sums of squares as in the classical ANOVA.

$$F = \frac{I_B/(m-1)}{I_W/(n-m)},$$

where m is the number of clusters and n is the number of units. The sums of squares are substituted by the generalized inertias I_B and I_W, respectively. Since the distribution of F is in the case of different dissimilarities not necessarily the F-distribution, P-values are calculated by a nonparametric (permutation) method.

[1] Dissimilarity d is metric if besides non-negativity, identity and symmetry, also triangle inequality holds, i.e., for each triple of units X, Y and Z, it holds $d(X,Z) \leq d(X,Y) + d(Y,Z)$.

McArdle and Anderson [MCR 01] showed that their method can be used with an arbitrary semimetric measure. Here, we add that the method can be used with a general dissimilarity measure as long as the dissimilarity between (imaginary) center $\tilde{C}$ and each unit is nonnegative. The nonnegativity can be achieved by the application of the metric index on the dissimilarity matrix just before using the nonparametric method.

The computations for the generalized ANOVA from dissimilarity measures were made by using the procedure `dissassoc` from the *R* package `TraMineR` by Studer *et al.* [STU 11]. It computes and tests the rate of discrepancy (defined from a dissimilarity matrix) explained by categorical variable(s).

To demonstrate the proposed approach, we performed these steps on the data of the countries described with the age–sex structures of their population for the year 2005. The data were obtained from the International Data Base (IDB) for the year 2005 [US 08]. Populations are divided into 17 five-year groups (0–4 years, 5–9 years, 10–14 years, ..., 75–79 years, 80+) for each gender. Unit representation is based on the same symbolic data analysis approach as in the application of population pyramids: data representation with two vectors, i.e., distributions of men/women over age-groups.

Groups were determined by the Human Development Index (HDI) found in The United Nations Development Program (UNDP) [UN 15b], which was developed by Pakistani economist Mahbub ul Haq to emphasize the importance of people, not only economy, for human development. It is a summary measure of the average achievement in key dimensions of human development: a long and healthy life, indicated by life expectancy at birth, being knowledgeable, considering mean years of schooling and expected years of schooling, and having a decent standard of living, where measurement is based on GNI (Gross National Income) per capita.

We calculated the dissimilarity between countries X and Y with the formula

$$d(X,Y) = \frac{1}{2}\Big(\frac{n_{xM} \cdot n_{yM}}{n_{xM} + n_{yM}}||\mathbf{p}_{xM} - \mathbf{p}_{yM}||^2 + \frac{n_{xF} \cdot n_{yF}}{n_{xF} + n_{yF}}||\mathbf{p}_{xF} - \mathbf{p}_{yF}||^2\Big).$$

Since we used squared Euclidean distance, the extended dissimilarity $d(U, \tilde{C})$ is nonnegative and we would not need to calculate the metric index. We do this here for demonstration purposes. The metric index for the obtained dissimilarity matrix is $p = 0.06438$. We used in the process d^p instead of d.

The HDI is used to rank countries by human development in the annual Human Development Reports prepared and published by The United Nations Development Program (UNDP). Data for the year 2005 from Table 2: Human Development

Index trends, 1980–2013, include 173 world's countries [UN 15a]. Three classes for the year 2005 (The Human Development Report 2007/2008) were determined as:

– high (HDI 0.800 or more);

– medium (HDI from 0.500 to 0.799);

– low (HDI below 0.500).

The results of the generalized ANOVA for 173 countries for the year 2005, based on d^p and three HDI classes (obtained with the usage of the `clamix` program for the calculation of the dissimilarity between countries and using `TraMineRs` procedure `dissassoc` to compute inertias and F-value), are

$$I_T = 155.05, I_W = 144.48, I_B = I_T - I_W = 10.57$$

$$F = \frac{I_B/(m-1)}{I_W/(n-m)} = \frac{10.57/(3-1)}{144.48/(173-3)} = 6.22$$

The results obtained show a larger discrepancy between groups than within them (P-value = 0.0002, P-value used with or without the metric index for the world country examples is the same). This indicates that there are noticeable differences between groups of countries determined with their HDI index according to the age–sex structure of the population.

10.5. Conclusion

Open data are very often available in aggregated form and can be considered as so-called second-level units. To preserve internal variation of the original (primary) units, these second-level units need to be represented with a more complex representation of aggregated values than the usual single mean value. SOs provide such a description, and SDA methods can be used to analyze them. The institutions that offer open data are, therefore, invited to produce/release the aggregated data in the form of SOs.

In this chapter, we presented adapted clustering methods for second-level units that were motivated by analysis of some open data sets. The main aim of the presented methods is to produce meaningful (informative) optimal cluster representatives. In order to obtain the desired properties of optimal cluster representatives, we have proposed some alternative dissimilarity measures between second-level units represented with empirical discrete (membership) distributions and the inclusion of weights. We demonstrated their usage with applications on TIMSS open data base and demographic age–sex structures on different sets of teritorial units.

In order to study differences among pre-specified groups of units, we presented an approach to generalize ANOVA with the following two main advantages: (1) it can be used with any dissimilarity measure and (2) it is nonparametric – it has no *a priori* assumptions about variable distributions.

NOTE.– This work was partially supported by the Slovenian Research Agency, Programmes P1-0294 and P3-0154 and by Russian Academic Excellence Project ‘5-100’.

10.6. References

[AND 73] ANDERBERG M., *Cluster Analysis for Applications*, Academic Press, New York, 1973.

[AND 01] ANDERSON M.J., “A new method for non-parametric multivariate analysis of variance”, *Austral Ecology*, vol. 26, no. 1, pp. 32–46, 2001.

[ARI] ARISTOTLE, “The Organon”. Available at: https://archive.org/details/AristotleOrganon.

[BAT 88] BATAGELJ V., “Generalized ward and related clustering problems”, in BOCK H.H. (ed.), *Classification and Related Methods of Data Analysis*, North-Holland, Amsterdam, pp. 67–74, 1988.

[BAT 19] BATAGELJ V., KEJŽAR N., “Clamix—clustering symbolic objects R package”. Available at: https://r-forge.r-project.org/projects/clamix/, 2019.

[BAT 15a] BATAGELJ V., KEJŽAR N., KORENJAK-ČERNE S., “Clustering of modal valued symbolic data”, *ArXiv e-prints 1507.06683*, July 2015.

[BAT 15b] BATAGELJ V., KEJŽAR N., KORENJAK-ČERNE S., “Generalized ANOVA for SDA”, *SDA Workshop 2015*, University of Orléans, November 17–19, 2015. Available at: http://www.univ-orleans.fr/mapmo/colloques/sda2015/, pp. 45–46, 2015.

[BIL 06] BILLARD L., DIDAY E., *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley, Chichester, 2006.

[BRA 14] BRAZILIAN INSTITUTE OF GEOGRAPHY AND STATISTICS (IBGE), “Census 2000 Summary File 1 [Data file]”. Available at: http://www.cidades.ibge.gov.br/xtras/home.php, accessed 2014.

[BRI 14] BRITO P., “Symbolic data analysis: another look at the interaction of data mining and statistics”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 4, pp. 281–295, 2014.

[DID 79] DIDAY E., Optimisation en classification automatique, Institut national de recherche en informatique et en automatique, Rocquencourt, 1979.

[DID 16] DIDAY E., “Thinking by classes in data science: the symbolic data analysis paradigm”, *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 8, no. 5, pp. 172–205, 2016.

[EUR 17] EUROPEAN UNION, “The European Union open data portal”. Available at: https://data.europa.eu/euodp/en/home, 2017.

[HAR 75] HARTIGAN J., *Clustering Algorithms*, Wiley-Interscience, New York, 1975.

[IEA 04] IEA – INTERNATIONAL ASSOCIATION FOR EVALUATION OF EDUCATIONAL ACHIEVEMENT, “IEA website, data repository for TIMMS data”. Available at: http://www.iea.nl/data.html, accessed 2004.

[JOL 86] JOLY S., LE CALVÉ G., “Etude des puissances d’une distance”, *Statistique et Analyse de Données*, pp. 30–50, North-Holland, Amsterdam, 1986.

[KEJ 11] KEJŽAR N., KORENJAK-ČERNE S., BATAGELJ V., “Clustering of distributions: a case of patent citations”, *Journal of Classification*, vol. 28, no. 2, pp. 156–183, 2011.

[KOR 11] KORENJAK-ČERNE S., BATAGELJ V., JAPELJ PAVEŠIĆ B., “Clustering large data sets described with discrete distributions and its application on TIMSS data set”, *Statistical Analysis and Data Mining*, vol. 4, no. 2, pp. 199–215, 2011.

[KOR 12] KORENJAK-ČERNE S., BATAGELJ V., SAMBT J. *et al.*, “Hierarchical clustering method for discrete distributions with the case of clustering population pyramids of Slovenian municipalities”, *Facing Demographic Challenges : Proceedings of the 15th International Multiconference Information Society IS 2012*, October 8–9, 2012, Ljubljana, Slovenia, volume B, (Informacijska družba, ISSN 1581-9973), Institut Jožef Stefan, Ljubljana, pp. 31–35, 2012 (in Slovenian).

[KOR 15] KORENJAK-ČERNE S., KEJŽAR N., BATAGELJ V., “A weighted clustering of population pyramids for the world’s countries, 1996, 2001, 2006”, *Population Studies*, vol. 69, no. 1, pp. 105–120. Available at: http://www.tandfonline.com/doi/full/10.1080/00324728.2014.954597, 2015.

[KOR 17] KORENJAK-ČERNE S., SIMPSON L., “Clustering age-sex structures to monitor their development over time: Latin America and the Caribbean sub-national areas 1960-2011. In ALAP (La Asociación Latinoamericana de Población) Project report”. Available at: http://www.cmist.manchester.ac.uk/research/projects/s-alyc/, 2017.

[MCA 01] MCARDLE B.H., ANDERSON M.J., “Fitting multivariate models to community data: a comment on distance-based redundany analysis”, *Ecology*, vol. 82, no. 1, pp. 290–297, 2001.

[NOI 11] NOIRHOMME-FRAITURE M., BRITO P., “Far beyond the classical data models: symbolic data analysis”, *Statistical Analysis and Data Mining*, vol. 4, no. 2, pp. 157–170, 2011.

[SIM 16] SIMPSON L., GONZALES L., “Comparative subnational demographic development in Latin America and the Caribbean (s-ALyC)”. Available at: http://www.cmist.manchester.ac.uk/research/projects/s-alyc/, 2016.

[STU 11] STUDER M., RITSCHARD G., GABADINHO A. *et al.*, "Discrepancy analysis of state sequences", *Sociological Methods & Research*, vol. 40, no. 3, pp. 471–510, 2011.

[TIM 04] TIMSS & PIRLS. INTERNATIONAL STUDY CENTER. BOSTON COLLEGE, LYNCH SCHOOL OF EDUCATION. USA, "TIMSS – Trends in International Mathematics and Science Study open data set. TIMSS 1999 and TIMSS 2003 [Data files]". Available at: http://timss.bc.edu, accessed 2004.

[UN 15a] UNITED NATIONS, "Human Development data, The United Nations Development Program (UNDP). Available at: http://hdr.undp.org/en/data, accessed 2015.

[UN 15b] UNITED NATIONS, "Human development reports, The United Nations Development Program (UNDP)". Available at: http://hdr.undp.org/en/content/human-development-index-hdi, accessed 2015.

[UN 17] UNITED NATIONS, "The United Nations open data website". Available at: http://data.un.org/, 2017.

[US 08] U.S. CENSUS BUREAU, "IDB: International Data Base". Available at: http://www.census.gov/ipc/www/idbnew.html, accessed 2008.

[US 11] U.S. CENSUS BUREAU, "Census 2000 Summary File 1 [Data file]". Available at: http://factfinder.census.gov/, accessed 2011.

[WAR 63] WARD J., "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[WB 17] THE WORLD BANK, "The World Bank Open Data". Available at: http://data.worldbank.org, 2017.