Robust multivariate analysis for mixed-type data: novel algorithm and its practical application in socio-economic research

Aurea Grané[a], Silvia Salini[b,*], Elena Verdolini[c]

[a] Universidad Carlos III de Madrid
[b] University of Milan
[c] University of Brescia and RFF-CMCC European Institute of Economics and Environment, Centro Euro-Mediterraneo sui Cambiamenti Climatici

Abstract

We propose a novel method and algorithm for the analysis and clustering of mixed-type data using a hierarchical approach based on Forward Search. In our procedure, the identification of groups is based on the identification of similar trajectories and then linked to very intuitive two-dimensional maps. The proposed algorithm can use different measures for the calculation of distance in the case of mixed-type data, such as Gower's metric and Related metric scaling. A key feature of our algorithm is its ability to discard redundant information from a given set of variables. The practical usefulness of the algorithm is illustrated through two applications of high relevance for empirical economic research. The first one focuses on comparing different indicators of environmental policy stringency in different countries. The second one applies our procedure to identify clusters of countries based on information regarding their institutional characteristics.

Keywords: Forward Search, Mixed Type Data, Outliers, Robustness, Redundant information, Clustering

*Corresponding author, Department of Economics, Management and Quantitative Methods, Via Conservatorio 7, 20122 Milano, Italy, Tel: +390250321538

Email addresses: agrane@est-econ.uc3m.es (Aurea Grané), silvia.salini@unimi.it (Silvia Salini), elena.verdolini@unibs.it (Elena Verdolini)

1. Introduction

Mixed-type data comprise both numeric and categorical features, and mixed datasets frequently occur in many domains, such as economics, health, finance, marketing, including data coming from socio-demographic surveys. Applied economists and social scientists are often faced with the necessity to deal with mixed-type data. For instance, mixed data indicators measuring a given economic or societal aspect often need to be compared to understand the extent to which they convey similar or different information, as in Galeotti et al. (2020). Furthermore, clustering often needs to be applied to mixed datasets to find structures and to group similar observations for further analysis, as in Nesta et al. (2019). These contributions highlight the challenges associated with the use of mixed-type data for socio-economic research. To begin with, one cannot rely on a simple distance measure, such as the Euclidean distance, because of the presence of categorical data. Moreover, in the statistical literature a few distance measures to deal with mixed data exist, such as Gower's similarity index (Ahmad and Khan, 2019) but they are plagued by important shortcomings, as highlighted in Grané and Romera (2018) and discussed more at length below. See also Foss et al. (2019) for clustering methods for mixed data and van de Velden et al. (2018) for distance based methods for mixed data.

A recent relevant methodological contribution in the context of mixed data is presented in Grané and Romera (2018), who construct robust profiles for mixed-type data using multidimensional scaling, which is one of the most extended methodologies to visualize the profile structure of mixed data. To this end, Grané and Romera (2018) compare different multidimensional scaling configurations (coming from different metrics) through a combination of sensitivity and robust analysis. They propose a robust joint metric combining different distance matrices, avoiding redundant information, via Related Metric Scaling (RelMS) as an alternative to classical Gower's metric.

The first (methodological) contribution of this paper to the literature on mixed data is the development of a novel robust algorithm for the explanatory data analysis of mixed datasets. This is achieved by combining the related metric scaling measure proposed by Grané and Romera (2018) with a Forward Search algorithm (Atkinson and

Riani (2004)). On the one hand, related metric scaling allows to overcome the main shortcomings of Gower's measure. On the other hand, Forward Search (FS) is a powerful general method which can be applied to many statistical models to make them robust. The FS algorithm was introduced by Atkinson and Riani (2000), (2004) in the context of robust regression models and has been extended to many other fields, such as financial models, cluster analysis, curve monitoring, robust inference, and such. In our context, Forward Search is useful because (a) it incorporates flexible data-driven trimming for the detection of outliers and unsuspected structure in the data and (b) it facilitates data visualization, in particular it allow us to visually represent how the procedure to calculate the related metric scaling joint metric unfolds rather than providing only a final picture of the outcome.

The second (practical) contribution of this paper is to demonstrate the usefulness of this novel algorithm for applied socio-economic analysis through two empirical applications of applied economic analysis. First, we show the usefulness of our approach for the comparison of mixed-data indicators of environmental policy which underline the analysis of Galeotti et al. (2020). We demonstrate the need of an alternative measure to Gower's metric in presence of mixed-type data by showing how RelMS can discard redundant information from different indicators. Furthermore, we use a stability analysis to show how the Multi Dimensional Scaling (MDS) configurations of RelMS are more stable than those using Gower.

Second, we apply our method to the widely known dataset described in La Porta et al. (1999) to show its usefulness in generating clusters for countries in terms of the key institutional dimensions. This procedure can be used to generate an index of similarity for potential use in applied research, such as the generation of instrumental variables similarly to what proposed in Nesta et al. (2019). The La Porta et al. (1999) database is a mixed dataset containing variables describing four key country-level institutional aspects: legal origin, political freedom, efficiency of institutions and interference with the private sector. These are important underlying characteristics of a country's institutional, legal and political framework.

The rest of the paper is organized as follows: Section 2 is devoted to describe the proposed algorithm. In Section 3 we present an alternative metric more robust than

Gower's. Section 4 presents the application to the two empirical applications related to environmental policy stringency and to institutional aspects: we describe the data, apply our algorithm, and comment the results and their usefulness for applied socio-economic research. Section 5 concludes, highlighting other potential application areas and discussing future research avenues.

2. Method

In this Section, we first describe the Forward Search Distance Based (FS-DB) algorithm. This novel approach combines the FS method with a distance-based tool, used in Grané and Romera (2018) to detect outliers in mixed-type datasets. While this distance-based tool can cope with any distance measure, the algorithm is initially described in terms of Gower's distance, since we are interested in mixed-type datasets. A more interesting alternative is given in Section 3, where the distance is tailored via RelMS.

2.1. The Foward Search philosophy of data analysis.

The FS is a data-driven strategy which is based on carefully chosen subsets of the data. The key difference with respect to other robust strategies for data analysis, is that the algorithm is not only based on one subsample, but on a sequence of subsets of the original data. It is an adaptive hard trimming method (Salini et al., 2016). In the words of their initial proponents "the FS is not a simple new algorithm but a new philosophy of looking at the data, which involves watching a film of the data rather than a snapshot". The crucial idea behind the FS approach is to monitor how a fitted model changes whenever a new statistical unit is added to the subset. The model of interest is initially fitted on a starting subset, whose units can change in each step of the algorithm. Thus, this approach helps to understand the effect that each unit (outlier or not, leverage point or not) exerts on the fitted model (see `http://rosa.unipr.it/FSDA/guide.html` for a more detailed description of the method).

2.2. The FS-DB algorithm

The idea behind our proposed approach is to help understand the structure of mixed-type datasets by identifying the subset of closest units (according to a user-selected distance measure) as well as those units that are the most distant from the set(s) of the data. Apart from the numerical outputs, there are two graphical outputs of our algorighm. First, the Forward-plot with the trajectories of the units which illustrate their performance along the steps of the algorithm. Second, the MDS-plot with the final MDS representations of the dataset. Graphical outputs are explained in Section 4. The algorithm implemented code has been submitted in order to be included in the next release of the common and flexible framework provided by the FSDA Toolbox of Matlab (Riani et al. (2012)[1])

The starting point of our procedure is a data matrix of mixed type of dimension $n \times p$. The steps we follows are:

1. Select a distance measure. In this first example we use Gower's similarity coefficient.Given two p-dimensional vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, Gower's similarity coefficient is defined as

$$s_{ij} = \frac{\sum_{h=1}^{p_1} \left(1 - |z_{ih} - z_{jh}|/R_h\right) + a + \alpha}{p_1 + (p_2 - d) + p_3}, \quad 0 \leq s_{ij} \leq 1, \tag{1}$$

   where $p = p_1 + p_2 + p_3$, p_1 is the number of continuous (or quantitative) variables, a and d are the number of positive and negative matches, respectively, for the p_2 binary variables, α is the number of matches for the p_3 multi-state categorical variables, and R_h is the range of the h-th continuous variable. Gower's distance is defined as $\delta^2(\mathbf{z}_i, \mathbf{z}_j) = 1 - s_{ij}$, which are the entries of the matrix of squared distances Δ.

2. Select a subset size ($m < n$). By default m is set as 10% of n.

3. Select the units inside the starting subset which have lowest distance measure.

4. Calculate the geometric variability of the subset $V_{\Delta(m)}$. Let $\{\mathbf{z}_i, 1 \leq i \leq m\}$ be m p-dimensional vectors containing the information of the m individuals in the

[1] The FSDA Toolbox of Matlab is freely available at `http://rosa.unipr.it/fsdadownload.html`.

subset and consider a matrix $\Delta_{(m)}$ of squared distances, with entries $\delta^2(\mathbf{z}_i, \mathbf{z}_j)$, for $1 \le i, j \le m$. The geometric variability of $\Delta_{(m)}$ is

$$V_{\Delta(m)} = \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \delta^2(\mathbf{z}_i, \mathbf{z}_j).$$

5. Calculate for each unit outside the subset the *distance-based proximity* function $\phi(i)$ to the subset. Given a new individual $\mathbf{z}_0 \in \mathbb{R}^p$, the distance-based proximity of $\mathbf{z}_0$ to the set $\{\mathbf{z}_i, 1 \le i \le m\}$ is

$$\phi(\mathbf{z}_0) = \frac{1}{m} \sum_{i=1}^{m} \delta^2(\mathbf{z}_0, \mathbf{z}_i) - V_{\Delta(m)}.$$

6. Include in the subset the unit with the minimum value of $\phi(i)$; set m equal to $m+1$.
7. Iterate the procedure from step 3 until all n units are included in the subset.
8. Monitoring $\phi(i)$ for each unit on the subset size.
9. Plot the trajectory in multidimensional scaling (MDS) maps and identify groups and outliers.

In this implementation we select the units inside the starting subset which lowest distance measure. However, note that step 3 allows the units to enter and exit the subset, since in each iteration the current subset is formed by those units with lowest distance measure. Another interesting approach, that we can explore for future development, is that detailed in Atkinson et al. (2006) where units in the initial subset are randomly chosen in order to check the stability to the starting point.

3. An alternative to Gower's metric: Related metric scaling

Gower's similarity coefficient is one of the most popular similarity measures and perhaps the easiest way to obtain a distance measure when working with mixed-type data. However, it presents two important drawbacks. The first one, pointed out long time ago by Gower (1992); Krzanowski (1994), is that, just like any distance function satisfying additivity with respect to variables, this coefficient ignores any association (correlation) between variables and, thus, is not able to discard any redundant information. The second drawback is the lack of robustness: this coefficient uses the standardized city block distance for quantitative variables (see equation (1)), which is not a robust measure. As a consequence, in the presence of outliers, the stability of the MDS configurations can be affected, as shown in Grané and Romera (2018). This second drawback may be solved by replacing standardized city block distance by, for instance, a robustified Mahalanobis distance. However, still the first drawback will remain in the new coefficient. Thus, our proposal is to overcome both shortcomings by obtaining a distance measure for mixed-type data via related metric scaling.

Related metric scaling (RelMS) is a multivariate technique that allows to obtain a unique representation of a set of observations from several distance matrices computed on the same set of observations. The method is based on the construction of a joint metric that satisfies several axioms related to the property of identifying and discarding redundant information (Cuadras (1998); Cuadras and Fortiana (1998)).

Given a set of $k \geq 2$ matrices of squared distances measured on the same group of n observations, $\{\Delta_\alpha\}_{\alpha=1,\ldots,k}$, the first requirement in the construction of the joint metric is that all matrices Δ_α have the same geometric variability. Note that the condition of equal geometric variability can always be assumed to hold, since multiplying a squared distances matrix by an appropriate constant amounts to a change of measurement unit.

- First step: Standardize each matrix Δ_α with respect to its geometric variability $V_{\Delta_\alpha} = \frac{1}{2n^2}\sum_{i=1}^n \sum_{j=1}^n \delta^2(\mathbf{z}_i, \mathbf{z}_j)$, where $\delta^2(\mathbf{z}_i, \mathbf{z}_j)$ are the entries of matrix Δ, for $1 \leq i, j \leq n$. In an abuse of notation, we call Δ_α its standardized version.

- Second step: For each distance matrix Δ_α consider its doubly centered inner product matrix:
$$\mathbf{G}_\alpha = -\frac{1}{2}\mathbf{H}\Delta_\alpha\mathbf{H},$$
where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}'$, $\mathbf{I}_n$ is the identity matrix of order n and $\mathbf{1}$ is a $n \times 1$ vector of ones.

- Third step: Compute the inner product matrix of the joint metric as
$$\mathbf{G} = \sum_{\alpha=1}^{k} \mathbf{G}_\alpha - \frac{1}{k}\sum_{\alpha\neq\beta} \mathbf{G}_\alpha^{1/2}\mathbf{G}_\beta^{1/2}, \qquad (2)$$

where $\mathbf{G}_\alpha^{1/2}$ denotes the square root of $\mathbf{G}_\alpha$, which can be obtained through the singular value decomposition of $\mathbf{G}_\alpha$.

- Fourth step: The matrix of the joint metric can be computed as

$$\Delta = \mathbf{g}\mathbf{1}' + \mathbf{1}\mathbf{g}' - 2\mathbf{G}, \tag{3}$$

where $\mathbf{g} = diag(\mathbf{G})$ is a column vector containing the diagonal of matrix $\mathbf{G}$.

Note that, in order to obtain MDS configurations, it is enough to work with formula (2), although formula (3) is required for computing $\phi(i)$ in the Forward Search.

How do we interpret formula (2)? The first addend of this formula mimics Gower's similarity coefficient by adding the k metrics; the second one is responsible of discarding redundant information coming from different sources. This second addend provides more flexibility to the MDS configuration when working with mixed-type data. Thus, RelMS allows us to tailor a metric to reflect specific information of a mixed dataset. See Grané and Romera (2018) for the mathematical properties of the method. Another advantage of the method is that it does not require data preprocessing.

In our application, we construct the joint metric from $k = 3$ distance matrices, one for each variable type. In particular, Δ_1 contains the information related to quantitative variables and we use a robust version of Mahalanobis distance (Riani et al., 2009) to compute it. For multi-state categorical variables, we start by computing the similarity matrix $\mathbf{S}_2$ with Sokal-Michener's pairwise similarity coefficient (matching coefficient), and then we obtain $\Delta_2 = 2(\mathbf{1}\mathbf{1}' - \mathbf{S}_2)$; Finally, for binary variables, we compute the similarity matrix $\mathbf{S}_3$ with Jaccard's pairwise similarity coefficient and we get $\Delta_3 = 2(\mathbf{1}\mathbf{1}' - \mathbf{S}_3)$.

4. Application

The lack of a robust method for mixed-data is clearly an important limitation for applied research in environmental and sustainability issues, as argued above. In this section, we specifically illustrate this point through two examples. First, the lack of a robust approach to deal with mixed data is evident in the analysis of Galeotti et al.

(2020), who explore and compare different proxies of environmental policy stringency in a sample of 189 OECD countries over the years 1995-2009. Their analysis shows that different indexes of environmental policy stringency can give rise to significantly different country rankings (and, implicitly, country clustering) depending on whether they are based on data of environmental expenditures (i.e. inputs), emissions (i.e. outputs) or of the type of policy instrument implemented. The extent to which different indicators provide similar (and thus redundant information) is a matter of concern. Furthermore, the inability to identify outliers in a multivariate framework implies that any index or summary statistics (in this case, of environmental policy stringency) will not be robust, rather it will be influenced by such outliers and possibly mask their presence. This is due to the fact that composite indicators are necessarily data driven. Therefore, a major implication of the analysis presented in Galeotti et al. (2020) is the lack of a robust method to detect outliers and to analyze differences and similarities in a mixed data context.

Second, the recent contribution of Nesta et al. (2019) highlights the importance of being able to cluster countries based on several characteristics, some of which can be easily measured on continuous scales while others are necessarily summarized by categorical or binary proxies. Specifically, Nesta et al. (2019) use information on institutional policies to generate an Instrumental Variable in the context of their main research question, namely the impact of environmental policy on the direction of innovation. Specifically, Nesta et al. (2019) use information on the underlying legal and institutional characteristics of their sample of countries in order to cluster them. For each country, they then use the data on environmental policy stringency of all other observations in the cluster to generate the instrument used to address the endogeneity of their main variable of interest. A robust way to provide clear country clustering in this context would be really useful to address the issue of endogeneity in a more satisfactory way.

Using these two specific cases as illustrative, in the rest of this Section we show the main advantages of our novel algorithm, and its potential for improving the use of mixed-type data in socio-economic analysis.

4.1. Mixed data on environmental policy stringency: comparing different indicators

The aim of this subsection is to demonstrate both the need for and the usefulness of our approach for the comparison of mixed-data indicators of environmental policy stringency. To do so, we use the mixed data which underline the analysis of Galeotti et al. (2020), namely different indicators of environmental policy stringency for 18 OECD countries in the year 2009.[2] Descriptive statistics of the variables used are provided in Table 1, alongside the original data source. Figure 1 provides an overview of the values of the different indicators in our sample. As highlighted in Galeotti et al. (2020), the different indicators sometimes provide similar information regarding the level of environmental policy stringency in a given country (for instance, in the case of NOx Taxes and SOx Taxes for most of the countries), while in other cases the information provided is very different (for instance, this is the case with $FITs$ – Feed in Tariffs for solar and wind on the one hand and with CO_2 Tradable Certificates).

The first step to demonstrate the need for and usefulness of our approach as alternative to Gower's for mixed-type data is to compare the MDS configurations computed from Gower's and RelMS metrics in this context. As Figure 2 shows, the percentage of explained variability is greater when using RelMS metric. In both configurations there are four countries quite far from the others (Canada, USA, Austria and Turkey). However, the relative positions of the other countries are different in the two configurations. For example, with Gower's metric Japan looks close to Denmark, whereas when using RelMS Japan is more isolated.

4.1.1. Percentage of redundant information

One of the attractive properties of RelMS is the ability to discard redundant information. For this reason, we calculate the percentage of redundant information in the

[2]Note that this application is done on a sub-set of the original Galeotti et al. (2020) data, namely a cross-section of the original database which contained data for the 19 OECD countries over the years 1995–2009. The countries included in this analysis are Austria, Australia, Canada, Germany, Denmark, Spain, Finland, France, the United Kingdom, Greece, Hungary, Italy, Japan, Netherlands, Portugal, Sweden, Turkey, and the United States

Variable	Name	Mean	Std. Dev.	Min.	Max.	Type	Data Source
CO2 Tax	$CO2-Tax$	0.333	1.414	0	6	categorical, ordered	Botta and Kozluz (2014)
NOx Tax	$NOx-Tax$	1.667	2.169	0	6	categorical, ordered	Botta and Kozluz (2014)
SOx Tax Indicator	$SOx-Tax$	1.444	1.977	0	6	categorical, ordered	Botta and Kozluz (2014)
CO2 Certificates	$CO2-TraS$	3.222	2.157	0	6	categorical, ordered	Botta and Kozluz (2014)
Green Certificates	$Green-TraS$	1.167	1.855	0	6	categorical, ordered	Botta and Kozluz (2014)
White Certificates	$White-TraS$	0.611	1.29	0	5	categorical, ordered	Botta and Kozluz (2014)
FIT Wind Indicator	$Wind-FIT$	1.833	1.581	0	5	categorical, ordered	Botta and Kozluz (2014)
FIT Solar Indicator	$Solar-FIT$	2.944	2.235	0	6	categorical, ordered	Botta and Kozluz (2014)
Sulphur Content Indicator	$Sulph-cont$	5.833	0.383	5	6	categorical, ordered	Botta and Kozluz (2014)
R&D Indicator	$RD-indicator$	2.722	1.904	0	6	categorical, ordered	Botta and Kozluz (2014)
Diesel Tax	$Diesel-tax$	4.111	1.132	2	6	categorical, ordered	Botta and Kozluz (2014)
DRS Indicator	$DRS-indicator$	0.556	0.511	0	1	binary	Botta and Kozluz (2014)
NOx Limits	$NOx-Limits$	4.333	1.749	1	6	categorical, ordered	Botta and Kozluz (2014)
SOx Limits	$SOx-Limits$	4.389	1.614	0	6	categorical, ordered	Botta and Kozluz (2014)
PM Limits	$PM-Limits$	2.222	1.263	1	6	categorical, ordered	Botta and Kozluz (2014)
Levinson Indicator EM	$BL-EM$	1.151	0.306	0.685	1.719	continuous	Galeotti et al. (2019)
Levinson Indicator CO2	$BL-CO2$	1.23	0.454	0.721	2.245	continuous	Galeotti et al. (2019)
Levinson Indicator SOX	$BL-SOX$	2.492	2.745	0.211	11.958	continuous	Galeotti et al. (2019)
Levinson Indicator NOX	$BL-NOX$	1.43	1.318	0.197	6.263	continuous	Galeotti et al. (2019)
Levinson Indicator NMVOC	$BL-NMVOC$	1.748	1.172	0.369	4.466	continuous	Galeotti et al. (2019)
Levinson Indicator NH3	$BL-NH3$	1.288	0.875	0.574	4.302	continuous	Galeotti et al. (2019)
Levinson Indicator N20	$BL-N2O$	1.43	1.098	0.482	4.25	continuous	Galeotti et al. (2019)
Levinson Indicator CO	$BL-CO$	2.635	1.541	0.118	5.885	continuous	Galeotti et al. (2019)
Levinson Indicator CH4	$BL-CH4$	1.699	1.352	0.581	6.761	continuous	Galeotti et al. (2019)
Energy R&D Intensity	$RDD-GDP$	0.519	0.48	0.029	1.898	continuous	IEA (2015)

Table 1: Variable description and descriptive statistics for Environmental Stringency data. *Categorical and binary variables come from Botta and Kozluz (2014). Continuous variables come from Galeotti et al. (2019) and are computed following the approach proposed in Brunel and Levinson (2013) using data on emissions and value added from WIOD 2013 (Timmer 2015). The last variable is computed using data on Energy R&D Investments from the IEA (2015) and data on GDP from the National Accounts of OECD Countries (both in constant 2013 PPP US dollars)*

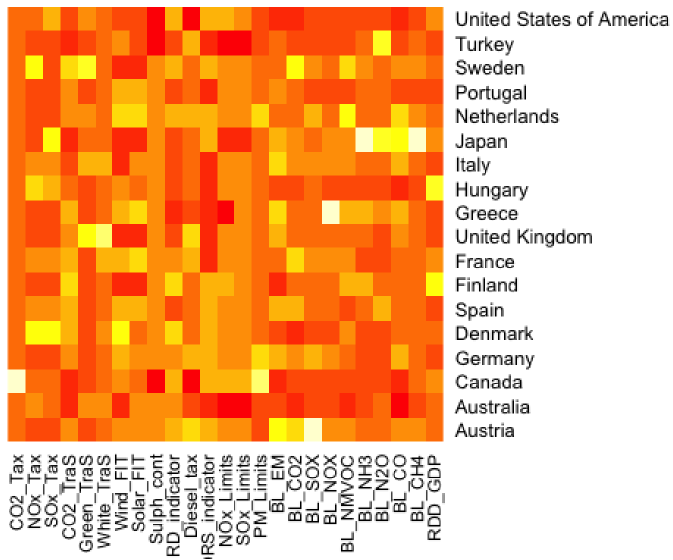


Figure 1: Heatmap of Environmental Stringency data

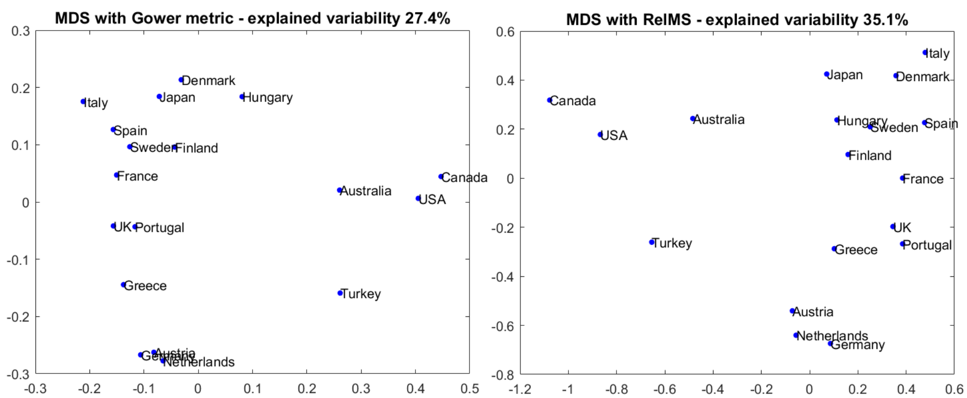
indicators for the countries in our sample. This will also help to explain why the MDS configurations obtained with Gower's metric or RelMS are rather different.

We first of all need to understand whether matrices Δ_1, Δ_2, and Δ_3 share a high percentage of information. To answer this question we compute the following measure of agreement between two matrices:

$$\rho(\alpha,\beta) = \frac{V(\Delta_\alpha,\Delta_\beta)}{\|\Delta_\alpha\|\,\|\Delta_\beta\|},$$

where $V^2(\Delta_\alpha,\Delta_\beta) = \frac{1}{n}|\mathrm{tr}(\mathbf{G}_\alpha) - \mathrm{tr}(\mathbf{G}_\beta)|$, $\|\Delta_\alpha\|^2 = \mathrm{tr}(\mathbf{G}_\alpha)$. Coefficient ρ takes values in $[0,1]$, being equal to one in case of orthogonality (that is, the Euclidean configurations associated to Δ_α and Δ_β generate orthogonal subspaces on $\mathbb{R}^n$) and equal to zero in case of equality (that is, in case $\Delta_\alpha = \Delta_\beta$). Then, the percentage of information shared by two distance matrices is obtained as $(1-\rho)100$ percent.

In our case, the percentages of shared information by matrices Δ_1, Δ_2, and Δ_3 are shown in Table 2, where we can see that these matrices contain redundant information (indeed they share more than 85% of the information). This is one of the reasons why the MDS configurations look different when using RelMS metric or Gower's. A second



(a) (b)

Figure 2: MDS configurations obtained from (a) Gower's metric and (b) RelMS metric

reason explaining why the configurations look different is related with the stability of those configurations, as we discuss next.

Table 2: Percentages of information shared by matrices Δ_1, Δ_2, and Δ_3

	Δ_1	Δ_2	Δ_3
Δ_1	100%	90.0%	90.4%
Δ_2		100%	86.2%
Δ_3			100%

4.1.2. Stability analysis

Here we are interested in evaluating the influence of the ith observation on the other $n-1$ observations, in the sense that how the exclusion of the ith observation from the original dataset can affect the MDS configuration of the $n-1$ remaining points. To solve this question we apply the leave-one-out cross-validation procedure proposed by Krzanowski (2006).

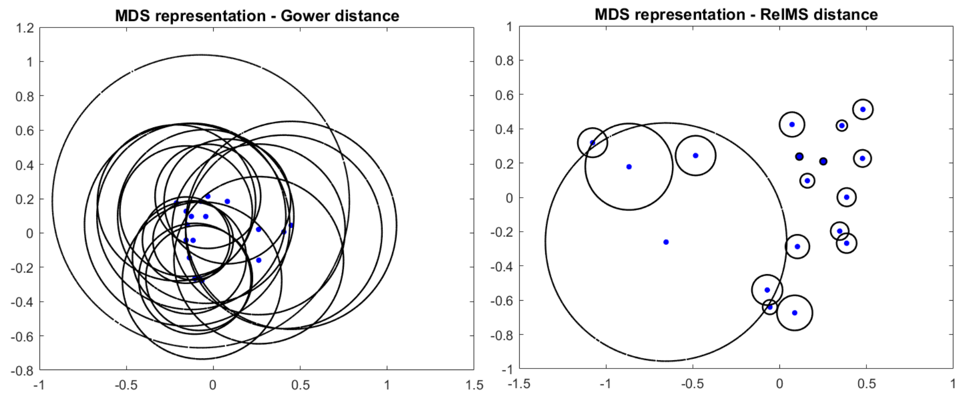
The idea of this method is to leave out each observation (that is, each row of the dataset) in turn and to compute the MDS configuration with the remaining $n-1$ observations. Then, the excluded observation is projected onto the MDS configuration using Gower's interpolation formula (Gower, 1968) leading to an "augmented" configuration. Finally, the n "augmented" configurations are compared with the original one (that is, that obtained from the whole dataset) by superimposing them. More specifically, they are just put on top each other (correctly aligned), as described in Krzanowski (2006), with no Procrustean rotation. Since sometimes the $n(n+1)$ points may overload the diagram, it is recommended to surround each point with the smallest hypersphere that contains a given percentage (e.g., 95 percent) of the cross-validatory replicate points. Hence, small hyperspheres indicate a very stable point, whereas large ones a very unstable one.

In Figure 3 we depict the 95 percent-stability regions for the MDS configurations using Gower's metric (a-panel) and RelMS metric (b-panel). The radius of each hypersphere is given by the squared root of the 95th quantile of the l^2 distances between the original coordinates of the point and the coordinates of its replicates. We can see that in panel (b) there is only one point very unstable (influential), whereas in panel (a) most of the points are unstable (and thus, influential). Additionally, we can compute some descriptive statistics for the radii. For example, the mean (and SD) are 0.1617 (0.1683) for Gower's metric and 0.0096 (0.0684) for RelMS metric. Hence, our conclusion is that MDS configuration computed from RelMS distance is more stable than Gower's.

4.1.3. Forward Search trajectory of the two metrics

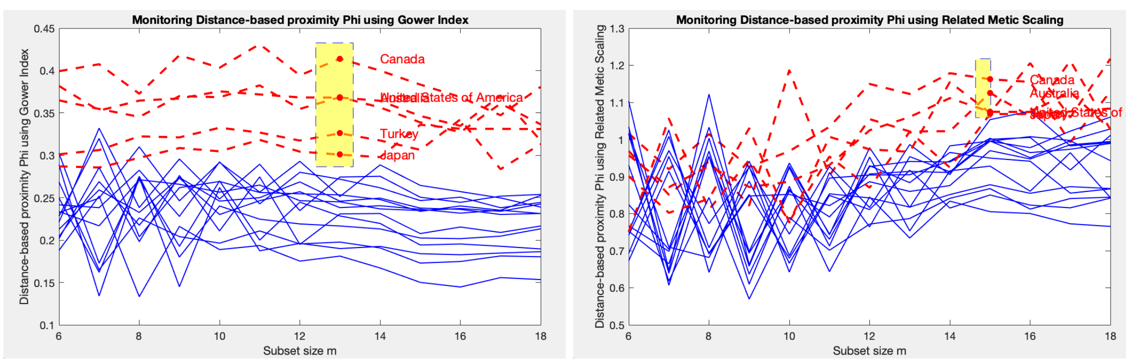
Figure 4 below shows the Forward plot using Gower's index (a-panel) and RelMS (b-panel). In this figure, trajectories which end close to one another represent countries which are similar among themselves, but different from others.

As expected the two measures produce different trajectories: we know from the previous analysis that Gower's metric does not discard redundant information. The brushing units, highlighted in red, are the units that enter at the end of the search, so the most distant from the bulk of the data. In this case, it is apparent that relying on Gower's metric suggests that Australia, Canada, Japan, Turkey and United States of America may cluster together, and separately from other countries (a panel). When accounting for redundant information, on the other hand, these countries do not appear



(a) (b)

Figure 3: Sensitivity analysis of MDS configurations. In (a) sensitivity analysis of the MDS configuration from Gower's metric; in (b) sensitivity analysis of the MDS configuration from RelMS metric



(a) (b)

Figure 4: Forward plots: monitoring the ϕ distance based proximity measure. In (a) you can see the trajectories using Gower's measure, in (b) you can see the trajectories using Related Metric Scaling

as different from the others (b panel). We further discuss the usefulness of our approach for clustering purposed in the next subsection, which contains our second application.

4.2. Mixed data on countries' institutional structure: clustering

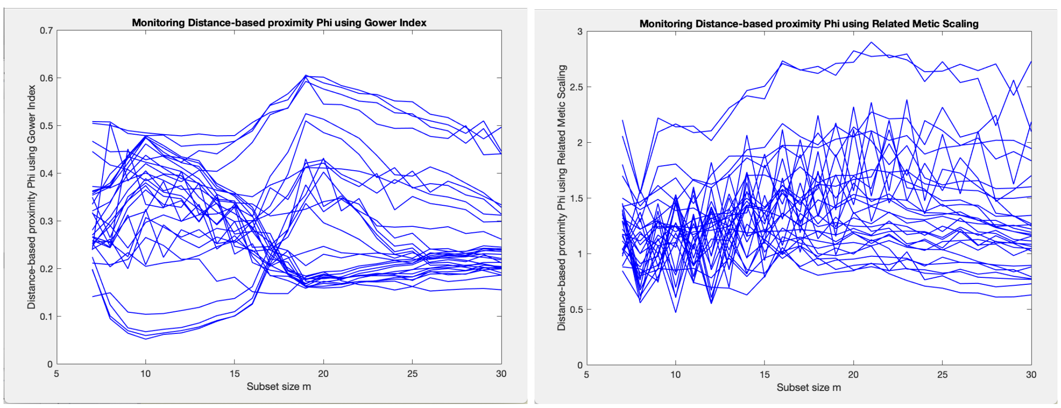
In order to demonstrate the usefulness of our novel algorithm for clustering observations, we rely on data on institutional aspects which we source our data from the widely known dataset described in La Porta et al. (1999). This mixed dataset contains several variables describing four key country-level institutional aspects: legal origin, political freedom, efficiency of institutions and interference with the private sector. The "Legal origin" variables identify the legal origin of the Company Law or Commercial law of a given country. They are a set of binary indicators identifying if a country is of either British, French, Socialist, German and Scandinavian legal origin. "Political Freemdom" is measured with two proxies: a democracy index and a political freedom index, both ranging from 0 to 10. Lower values indicate lower levels of political freedom. "Efficiency of Institutions" is measured through three variables: corruption, bureaucratic delays and tax compliance. Corruption and bureaucratic delays range from 0 to 10, while tax compliance is measured on the scale from 0 to 6. In all three cases, the index increase when efficiency increases. Lastly, interference with the private sector is measured with an index of property rights and a business regulation index, both raging from 1 to 5. For a more detailed description of the database, please refer to La Porta et al. (1999). For tractability, we limit our dataset to a sample of 35 countries.[3]

All these variables selected for our analysis were chosen because they can help to identify economies with similar underlying institutional structures. Identifying clusters of countries is indeed potentially relevant for the study of economic outcomes, on the one hand, and for use to generate instrumental variables in economic analyses. Indeed, each of these variables separately has been proposed as, or used in the computation of, instrumental variable in previous literature (as, for instance, in Nesta et al. (2019)). Yet, the lack of an aggregation technique appropriate for mixed data meant that each variable could only be used separately, and that the usefulness of focusing on different institutional aspects has remained largely unexplored.

Figure 5 below shows the Forward plot using Gower index (a-panel) and using

[3]The countries included in our sample are: Australia, Austria, Belgium, Brazil, Canada, Switzerland, China, Czech Republic, Germany, Denmark, Spain, Finland, France, United Kingdom, Greece, Hungary, Indonesia, India, Ireland, Iceland, Italy, Japan, Republic of Korea, Mexico, Netherlands, Norway, Poland, Portugal, Russian Federation, Slovak Republic, Slovenia, Sweden, Turkey, United States, South Africa.

related metric scaling (b-panel). In this figure, trajectories which end close to one another represent countries which are similar among themselves, but different from others. The two measures produce different trajectories. A main reason for this is the fact that Gower's metric does not discard redundant information and it is a less robust measure with respect to related metric scaling. That is, individuals that look close with Gower's metric, may not look so close when using a robust metric that can discard redundant information.

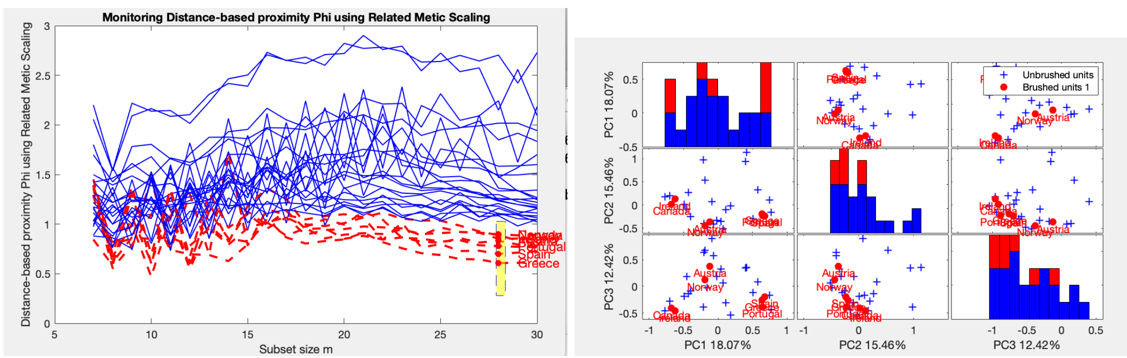


(a) (b)

Figure 5: Forward plots: monitoring the ϕ distance based proximity measure. In (a) you can see the trajectories using Gower's measure, in (b) you can see the trajectories using Related Metric Scaling

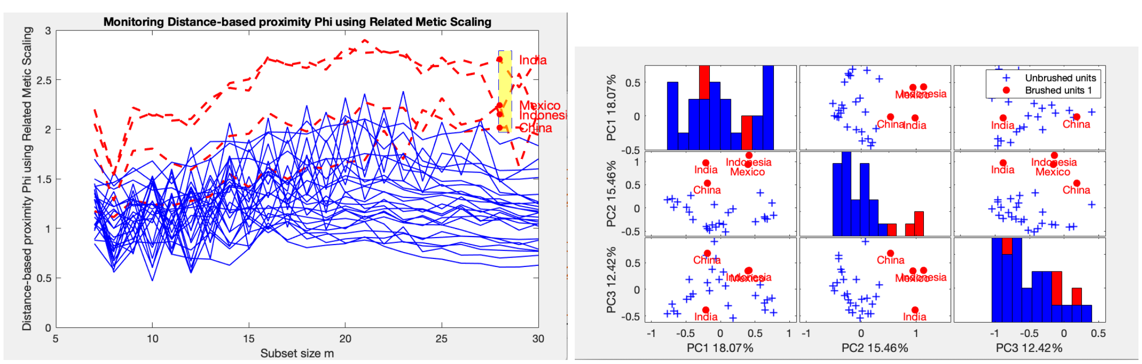
In Figure 6 and Figure 7 we show some groups of countries with similar trajectories in the related metric scaling configuration, the more robust one. Using the implemented algorithms it is possible to brush the trajectory and to show the selected countries in the multidimensional scaling maps. In this example we consider three coordinates but it is possible to plot the units in a different space. For example in the scatter-plot matrix of the original quantitative variable, coloring of splitting for level of categorical or binary variables.

Figure 6 highlights the units that enter at the begin of the search, so the units which are nearest to one another in terms of the different indicators of environmental policy stringency. These countries are Austria, Canada, Spain, Greece, Ireland, Norway, Portugal. Figure 7 highlights the units that enter at the end of the search, so the most distant from the bulk of the data, China, Indonesia, India and Mexico. This evidence confirm that these fast developing countries are very close when it comes to their environmental policy stringency. Importantly, note that the order in which observations enter the search are not determined by the level of the mixed data considered, rather by the distance of difference observations from one another.



(a) (b)

Figure 6: Output of the algorithm using the joint metric obtained via related metric scaling. In (a) you can see Forward plot, in (b) you can see the MDS Plot



(a) (b)

Figure 7: Output of the algorithm using the Related Metric Scaling. In (a) you can see Forward plot with highlighted some units that enter at the end of the search using the brushing option, in (b) you can see the corresponding MDS Plot

This algorithm and the procedure to handle the clustering of various observations over step in the forward search clearly offer the practical advantages of providing a clustering when in presence of mixed data. It also speaks to the potential of using an underlying score of similarity against these broad-based aspects of institutional design and quality. This could potentially be of great use for applied socio-economic researchers to generate exogenous instrumental variables using information about other countries in the cluster to instrument for one's own variable of interest, following the procedure of Nesta et al. (2019)

5. Conclusions

This paper develops a novel robust algorithm for the explanatory data analysis of mixed datasets. We code this approach in the common and flexible computational framework provided by the FSDA Toolbox of Matlab by combining the RelMS joint metric proposed by Grané and Romera (2018) with a Forward Search algorithm.[4]. From the methodological point of view, this is a significant improvement for two reasons. On the one hand, the related metric scaling allows to overcome the main shortcomings of Gower's measure, which up to recently has been the most common approach to the analysis of mixed datasets. On the other hand, applying the Forward Search method we incorporate in our procedure flexible data-driven trimming for the detection of outliers and unsuspected structure in the data and we facilitate data visualization. Another advantage is that the method does not requires data preprocessing.

Our analysis also points to fruitful avenues of future methodological research. These include, as just mentioned, the possibility to select the units inside the starting subset in the Forward Search randomly in order to check the stability to the starting point, as suggester in Atkinson et al. (2006). Moreover new interactive options for data visualization to improve the brushing of the units, for instance to produce the scatter plot matrix of the original quantitative variables or to color the dots differently based on nature of the variables selected; the implementation of other robust distances, such

[4]The code is available under request to the authors and is in the process of optimization and checking, with the aim of adding it to the next FSDA release

as a robust Gower measure (Mahalanobis instead of Manhattan) in the second step of the algorithm; the optimization of the proposed methods for a larger datasets, including tests for anomalous data and contaminations. We are also exploring the possibility to develop the classical hierarchical cluster using Releted Metric Scaling as a distance measure inside the common and flexible computational framework provided by the FSDA Toolbox of Matlab, including the generation of a dendrogram. Finally, in section 4.1.2 we use the leave-one-out cross-validation procedure proposed by Krzanowski (2006) to check the stability of MDS. An interesting future development, also to be implemented in the FSDA toolbox, could be to apply the FS for the same purpose, with the aim to avoid the typical masking effect of the outliers.

The usefulness of this new method is illustrated through two applications relevant for applied socio-economic analysis. First, we build on Galeotti et al. (2020) and use our method to compare different indicators of environmental policy stringency. Second, we apply our novel approach to the widely known dataset of La Porta et al. (1999): we use data on institutional characteristics of a given country to generate country clusters which account for several complementary aspects, namely legal origin, political freedom, efficiency of institutions and interference with the private sector. These examples confirm the high potential applicability of our novel approach beyond current applications for the clustering of observations and the generation of similarity indexes, including the generation of instrumental variables.

6. Acknowledgments

Elena Verdolini gratefully acknowledges funding from the Horizon 2020 Research and Innovation Programme under grant agreement No 730403 (INNOPATHS).

References

Ahmad A, Khan S (2019) Survey of State-of-the-Art Mixed Data Clustering Algorithms. IEEE Access 7: 31883-31902

Atkinson AC and Riani M (2000) Robust Diagnosis Regression Analysis. New York: Springer.

Atkinson A and Riani M (2004) The forward search and data visualisation, Computational Statistics 19: 29–54

Atkinson A, Riani M, Cerioli A (2006) Random start forward searches with envelopes for detecting clusters in multivariate data. In Data analysis, classification and the forward search. Springer, Berlin, Heidelberg. 163–171)

Atkinson AC, Riani M, Cerioli A (2010) The forward search: Theory and data analysis. Journal of the korean statistical society 39(2):117–134

Botta, E, Koźluk, T (2014) Measuring environmental policy stringency in OECD countries: a composite index approach. OECD Economics Department Working Paper N. 1177, http://dx.doi.org/10.1787/5jxrjnc45gvg-en

Brunel C, Levinson A (2013) Measuring environmental regulatory stringency. OECD Working Paper N. 2013/05. http://dx.doi:10.1787/18166881.

Cuadras CM (1998) Multidimensional dependencies in classification and ordination. Analyses Multidimensionelles des Données pp 15–25

Cuadras CM, Fortiana J (1998) Visualizing categorical data with related metric scaling. In: Visualization of Categorical Data, Elsevier, pp 365–376

Cuadras CM, Fortiana J, Oliva F (1997) The proximity of an individual to a population with applications in discriminant analysis. Journal of Classification 14(1):117–136

Foss HA, Markatou M, Bonnie, R (2019) Distance Metrics and Clustering Methods for Mixed-type Data. International Statistical Review 87(1): 80–109.

Galeotti M, Salini S, Verdolini E (2020) Measuring Environmental Policy Stringency: Approaches, Validity, and Impact on Energy Efficiency. Energy Policy 136:111052.

Gower JC (1992) Generalized Biplots. Biometrika 79(4)75–93

Gower JC (1968) Adding a Point to Vector Diagrams in Multivariate Analysis. Biometrika 55:582–85.

Grané A, Romera R (2018) On visualizing mixed-type data: A joint metric approach to profile construction and outlier detection. Sociological Methods & Research 47(2):207–239

IEA - International Energy Agency (2015). Energy Technology RD&D Statistics Database, available at www.oecd.org.

Krzanowski WJ (1994) Ordination in the Presence of Group Structure for General Multivariate Data. Journal of Classification 11:195–297

Krzanowski WJ (2006) Sensitivity in Metric Scaling and Analysis of Distance. Biometrics 62:239–44.

La Porta R, Lopez-de Silanes F, Shleifer A, Vishny R (1999) The quality of government. The Journal of Law, Economics, and Organization 15(1):222–279

Nesta, L, Verdolini E, Vona F (2019) Threshold policy effects and directed technical change in Energy Innovation. *Updated version of Nesta et al. 2018, Sciences Po publications No 2018–05*.

Riani M, Perrotta D, Torti F (2012) FSDA: a matlab toolbox for robust analysis and interactive data exploration. Chemometrics and Intelligent Laboratory Systems 116:17–32

Riani M, Atkinson A, Cerioli A (2009) Finding an unknown number of multivariate outliers. Journal of the Royal Statistical Society: series B (statistical methodology) 71.2:447–466

Salini S, Cerioli A, Laurini F, Riani M (2016) Reliable robust regression diagnostics. International Statistical Review 84(1):99–127

Timmer MP, Dietzenbacher E, Los B, Stehrer R, de Vries G.J. (2015) An Illustrated User Guide to the World Input–Output Database: the Case of Global Automotive Production. Review of International Economics 23:575–605

van de Velden M, Iodice D'Enza A and Markos A (2018) Distance-based clustering of mixed data. Wires Computational Statistics, 11(3).