

11th WORKSHOP ON **SYMBOLIC DATA ANALYSIS** **SDA 2026**

Programme and Book of Abstracts

22-24 October

Lisbon, Portugal



Maria Do Rosário Oliveira, Catarina P. Loureiro, Conceição Amado, Filipa Valdeira, Isabel M. Rodrigues, Lina Oliveira

(Editors)

**11th Workshop on
Symbolic Data Analysis
SDA 2026**

Programme and Book of Abstracts

22–24 October 2026

Instituto Superior Técnico
Lisbon, Portugal

Title:	<i>11th Workshop on Symbolic Data Analysis SDA 2026: Programme and Book of Abstracts</i>
Publisher:	Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
Editors:	Maria Do Rosário Oliveira, Instituto Superior Técnico Catarina P. Loureiro, Instituto Superior Técnico Conceição Amado, Instituto Superior Técnico Filipa Valdeira, Instituto Superior Técnico Isabel M. Rodrigues, Instituto Superior Técnico Lina Oliveira, Instituto Superior Técnico
Edition:	1st edition, Lisbon, October 2026
Format:	Electronic edition (PDF)
ISBN:	978-989-35910-3-1
Cover Design:	Catarina P. Loureiro

Photograph featured on the cover by Nuno Moreira.

Contents

Preface	4
Committees	5
Support and Sponsors	6
Program	7
 Abstracts	 11
I: Regression	12
Symbolic Interval-Valued Regression Model	13
Assessing the Impact of Riemannian Regression on Regression Methods for Symbolic Interval-Valued Data	15
Calibrated Bayesian composite likelihood approach to Irpino-Verde linear histogram re- gression	17
 II: Clustering	 18
Towards a Symbolic Data Extension of ClustOfVar: Clustering Ordinal Variables	19
Beyond Closed Interval Patterns in Symbolic Data Analysis	21
Dynamical Clustering Algorithm for Interval Data Based on Weighted Partitioning	24
Kernel-Based Fuzzy Co-Clustering for Interval-Valued Data	26
 III: Symbolic Profiles	 28
Assessing Adolescent Profile Coherence Using Classical and Distributed-Based Repre- sentations	29
From Publications to Symbolic Researcher Profiles: An SDA Approach to Exploring Interdisciplinarity	31
From Keywords to Themes: Distribution-Valued Symbolic Descriptions of Consensus Bibliometric Communities	33
 IV: Representations and Dimension Reduction	 36
Pairwise Logratio Functions: A Symbolic Data Analysis Perspective for Density-Valued Data	37
Masking Without Misleading: Learning Privacy-Preserving Representations against Re- construction Attacks	40
Biplots in Symbolic Data Analysis	42

V: Robust Statistics	44
The L1-Wasserstein Distance for Interval-Valued Data: Barycenter Computation and Robust Clustering	45
AIDA: An R Package for the Analysis of Interval DATA	46
Robust Interval Distributional regression model	48
Robust Principal Component Analysis of Symbolic Matrix-valued Data	50
VI: Networks	52
Monte Carlo Directed Interval Louvain (MC-DIL): Consensus Community Detection in Directed Networks with Interval-Valued Weights	53
Event Sequences and Networks: Part 1. Relational Events	55
Event Sequences and Networks: Part 2. Co-Presence Networks	57
VII: Time Series	59
Analysis of the Sea Level along the Portuguese Coast: A Symbolic Time Series Approach	60
Interval Time Series and Spatial Contiguity in the Analysis of Air-Pollution Patterns in Brazil	62
Clusterwise Autoregressive Modelling for Interval-Valued Time Series	64

Preface

The 11th Workshop on Symbolic Data Analysis (SDA 2026) is held in Lisbon, Portugal, from October 22nd to October 24th, 2026, organised by Instituto Superior Técnico, University of Lisbon. It continues a tradition of ten previous editions held across Europe and Asia, each bringing together the Symbolic Data Analysis community to share recent research, discuss ideas and explore new opportunities for collaboration.

Symbolic Data Analysis was born from the recognition that many phenomena are better described by intervals, histograms or distributions than by single values, preserving the variability within each unit. As data grow in volume and complexity, this perspective becomes ever more relevant. This edition reflects that vitality. Alongside established topics such as clustering and regression, the programme includes a session devoted to robust statistics, which addresses the challenge of outlying observations in symbolic data. It also features growing lines of research on networks and symbolic time series.

Over three days, the workshop features 23 oral presentations organised into seven thematic sessions. Contributions address methodological and theoretical advances, applications in areas as diverse as environmental sciences, bibliometrics, mobility and finance, and the development of new software. The chapters of this Book of Abstracts follow the sessions of the workshop: Regression, Clustering, Symbolic Profiles, Representations and Dimension Reduction, Robust Statistics, Networks, and Time Series.

We are sincerely grateful to all authors for sharing their research with the SDA community, and to the session chairs for guiding the discussions. We particularly thank the members of the Scientific Committee for their time and expertise, which ensured the scientific quality of this Book of Abstracts. We also gratefully acknowledge Instituto Superior Técnico and its Department of Mathematics for their support in organising SDA 2026, as well as the sponsors whose contributions made this event possible.

Beyond the scientific sessions, we hope participants will enjoy Lisbon, its light, its history and its cuisine, and the conversations that continue over dinner. These moments have always been part of the collaborative spirit of the SDA community. We warmly welcome all participants to Lisbon and look forward to three days of fruitful exchange.

Lisbon, October 2026

M. Rosário Oliveira
Catarina P. Loureiro
Conceição Amado
Filipa Valdeira
Isabel M. Rodrigues
Lina Oliveira

Committees

Scientific Committee

Boris Béranger, University of New South Wales, Australia

Lynne Billard, University of Georgia, USA

Paula Brito, University of Porto, Portugal

Sónia Manuela Mendes Dias, ESTG - Polytechnic University of Viana do Castelo, Portugal

Jasminka Dobša, University of Zagreb, Croatia

Liang-Ching Lin, National Cheng Kung University, Taiwan

Eufrásio de Andrade Lima Neto, Federal University of Paraíba, Brazil

M. Rosário Oliveira, Instituto Superior Técnico, Portugal

Rosanna Verde, University of Campania "Luigi Vanvitelli", Italy

Organizing Committee

Conceição Amado, Instituto Superior Técnico, Portugal

Catarina P. Loureiro, Instituto Superior Técnico, Portugal

Lina Oliveira, Instituto Superior Técnico, Portugal

M. Rosário Oliveira, Instituto Superior Técnico, Portugal

Isabel M. Rodrigues, Instituto Superior Técnico, Portugal

Filipa Valdeira, Instituto Superior Técnico, Portugal

Support and Sponsors

This event is partially funded by national funds through FCT- Fundação para a Ciência e a Tecnologia, I.P., within the framework of projects UID/04459/2025 (CAMGSD, IST-ID) and UID/04621/2025 (CEMAT, IST-ID).



Program

Program Overview

DAY 1		DAY 2		DAY 3	
<i>Thursday, 22 October</i>		<i>Friday, 23 October</i>		<i>Saturday, 24 October</i>	
11:00	Registration	08:30	Registration	09:00	Registration
12:00	Lunch	09:00	III - Symbolic Profiles	09:30	VII - Time Series
13:35	Opening Ceremony	10:30	Coffee Break	11:00	Discussion and Closing
14:00	I - Regression	11:00	IV - Representations and Dimension Reduction	11:30	Farewell and Lunch
15:30	Coffee Break	12:30	Group Photo		
16:00	II - Clustering	12:40	Lunch		
20:00	Dinner	14:00	V - Robust Statistics		
		16:00	Coffee Break		
		16:30	VI - Networks		
		20:00	Dinner		

Information for Participants

Participants can access Wi-Fi using their eduroam credentials or connect to the tecnico-guest network. Login credentials for tecnico-guest will be provided at the registration desk.

All scheduled presentations will be held in Room P3.10.

Day 1 - Thursday, 22 October 2026

Time	Programme
11:00–12:00	Registration
12:00–13:30	Lunch
13:35–14:00	Opening Ceremony
14:00–15:30	I - Regression Chair: Sónia Dias <i>Symbolic Interval-Valued Regression Model</i> , Liang-Ching Lin <i>Assessing the Impact of Riemannian Regression on Regression Methods for Symbolic Interval-Valued Data</i> , Oldemar Rodríguez Rojas <i>Calibrated Bayesian Composite Likelihood Approach to Irpino-Verde Linear Histogram Regression</i> , Andrej Srakar
15:30–16:00	Coffee Break
16:00–18:00	II - Clustering Chair: Jasminka Dobša <i>Towards a Symbolic Data Extension of ClustOfVar: Clustering Ordinal Variables</i> , Ana Felizardo Henriques <i>Beyond Closed Interval Patterns in Symbolic Data Analysis</i> , Mehdi Kaytoue <i>Dynamical Clustering Algorithm for Interval Data Based on Weighted Partitioning</i> , Renata M. C. R. de Souza <i>Kernel-Based Fuzzy Co-Clustering for Interval-Valued Data</i> , Francisco de A. T. de Carvalho
20:00	Dinner at Clube Militar Naval (Av. Defensores de Chaves 26, Lisboa)

Day 2 - Friday, 23 October 2026

Time	Programme
08:30–09:00	Registration
09:00–10:30	III - Symbolic Profiles Chair: Francisco de Carvalho <i>Assessing Adolescent Profile Coherence Using Classical and Distributed-Based Representations</i> , Jasminka Dobša <i>From Publications to Symbolic Researcher Profiles: An SDA Approach to Exploring Interdisciplinarity</i> , Maja Buhin Pandur <i>From Keywords to Themes: Distribution-Valued Symbolic Descriptions of Consensus Bibliometric Communities</i> , Carlo Drago
10:30–11:00	Coffee Break
11:00–12:30	IV - Representations and Dimension Reduction Chair: M. Rosário Oliveira <i>Pairwise Logratio Functions: A Symbolic Data Analysis Perspective for Density-Valued Data</i> , Rui Nunes <i>Masking Without Misleading: Learning Privacy-Preserving Representations against Reconstruction Attacks</i> , Telmo Silva Filho <i>Biplots in Symbolic Data Analysis</i> , Paula Brito
12:30–12:40	Group Photo
12:40–14:00	Lunch
14:00–16:00	V - Robust Statistics Chair: Renata Souza <i>The L1-Wasserstein Distance for Interval-Valued Data: Barycenter Computation and Robust Clustering</i> , Diogo Pinheiro <i>AIDA: An R Package for the Analysis of Interval DATA</i> , Catarina P. Loureiro <i>Robust Interval Distributional Regression Model</i> , Sónia Dias <i>Robust Principal Component Analysis of Symbolic Matrix-valued Data</i> , A. Pedro Duarte Silva
16:00–16:30	Coffee Break
16:30–18:00	VI - Networks Chair: Paula Brito <i>Monte Carlo Directed Interval Louvain (MC-DIL): Consensus Community Detection in Directed Networks with Interval-Valued Weights</i> , Hélder Alves <i>Event Sequences and Networks</i> , Vladimir Batagelj <i>Event Sequences and Networks, Part 2: Co-presence Networks</i> , Anuška Ferligoj
20:00	Dinner at Estórias na Casa da Comida (Travessa das Amoreiras 1, Lisboa)

Day 3 - Saturday, 24 October 2026

Time	Programme
09:00–09:30	Registration
09:30–11:00	VII - Time Series Chair: Liang-Ching Lin <i>Analysis of the Sea Level along the Portuguese Coast: A Symbolic Time Series Approach</i> , Maria Eduarda Borges <i>Interval Time Series and Spatial Contiguity in the Analysis of Air-Pollution Patterns in Brazil</i> , Luana Lúcia Alves de Azevêdo <i>Clusterwise Autoregressive Modelling for Interval-Valued Time Series</i> , Rosanna Verde
11:00–11:30	Discussion and Closing
11:30	Farewell and Lunch

Abstracts

Session I

Regression

Symbolic Interval-Valued Regression Model

Liang-Ching Lin^{1,*}

1. Department of Statistics and Institute of Data Science, National Cheng Kung University, Tainan, Taiwan.

*Contact author: lc.lin@ncku.edu.tw

Keywords: Asynchronous Observations, Order Statistics, Regression, Symbolic Data Analysis.

Symbolic data analysis can provide statistical inferences for macroscale data while preserving as much information as possible from microscale data. In this study, we focus on the symbolic interval-valued regression model. The microdata are reorganized into intervals by using the largest and smallest order statistics. In this study, we assume that the observed interval-valued datasets $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_m\}$ and $\mathbf{Y} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_m\}$ are the interval-valued observations, where $\mathbf{X}_i = [X_{u,i}, X_{l,i}]^T$ and $\mathbf{Y}_i = [Y_{u,i}, Y_{l,i}]^T$ with $X_{u,i} > X_{l,i}$ and $Y_{u,i} > Y_{l,i}$, $\forall i = 1, \dots, m$. A simple (univariable) interval-valued regression model can be constructed as follows:

$$\mathbf{Y}_i = \beta_0 + \beta_1 \mathbf{X}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, m, \quad (1)$$

where the residuals $\boldsymbol{\varepsilon}_i = (\varepsilon_{u,i}, \varepsilon_{l,i})^T$ are independent and identically distributed with $\varepsilon_{u,i} = \max\{\varepsilon_{1,i}, \dots, \varepsilon_{n,i}\}$ and $\varepsilon_{l,i} = \min\{\varepsilon_{1,i}, \dots, \varepsilon_{n,i}\}$, where $\varepsilon_{k,i}$ and $k = 1, \dots, n$, are independent and identically distributed normal random variables with a mean of zero and a variance of σ^2 for $n > 1$. A negative value of the coefficient (β_1) represents a negative correlation between the dependent and independent variables. Thus, the highest value of the dependent variable should be correlated with the lowest value of the independent variable. To align with intuition, we define

$$\beta_1 \mathbf{X}_i = \begin{cases} (\beta_1 X_{u,i}, \beta_1 X_{l,i})^T & \text{if } \beta_1 > 0, \\ (\beta_1 X_{l,i}, \beta_1 X_{u,i})^T & \text{if } \beta_1 < 0. \end{cases} \quad (2)$$

Here, if $\beta_1 > 0$,

$$Y_{u,i} = \beta_0 + \beta_1 X_{u,i} + \varepsilon_{u,i} > \beta_0 + \beta_1 X_{l,i} + \varepsilon_{l,i} = Y_{l,i}.$$

In contrast, if $\beta_1 < 0$,

$$Y_{u,i} = \beta_0 + |\beta_1|(-X_{l,i}) + \varepsilon_{u,i} > \beta_0 + |\beta_1|(-X_{u,i}) + \varepsilon_{l,i} = Y_{l,i}.$$

Therefore, model (1) mathematically maintains the natural order $Y_{u,i} > Y_{l,i} \forall i = 1, \dots, n$, regardless of the sign of β_1 .

Afterward, we derive the likelihood functions of aforementioned symbolic interval-valued regression model and derive the corresponding maximum likelihood estimators (MLEs). In addition, we describe the Fisher information matrix of the MLEs and show that the parameters demonstrate asymptotic normality.

Next, we extend the simple linear symbolic interval-valued regression model to a multiple linear symbolic interval-valued regression model. To simplify the representation, we denote $\beta_k \mathbf{X}_{k,i}^* = [\beta_k X_{u,k,i}^*, \beta_k X_{l,k,i}^*]$, where

$$\beta_k X_{u,k,i}^* = \begin{cases} \beta_k X_{u,k,i} & \text{if } \beta_k > 0 \\ \beta_k X_{l,k,i} & \text{if } \beta_k < 0 \end{cases} \quad \beta_k X_{l,k,i}^* = \begin{cases} \beta_k X_{l,k,i} & \text{if } \beta_k > 0 \\ \beta_k X_{u,k,i} & \text{if } \beta_k < 0 \end{cases} \quad (3)$$

Then, the multiple linear regression model for interval-valued variables $(\mathbf{Y}, \mathbf{X}_1, \dots, \mathbf{X}_p)$ is denoted as follows.

$$\mathbf{Y}_i = \beta_0 + \sum_{k=1}^p \beta_k \mathbf{X}_{k,i}^* + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, m. \quad (4)$$

The corresponding MLEs are again derived.

Monte Carlo simulations consider three scenarios for constructing interval-valued regression models:

1. We generate raw random samples for the independent variables and residuals and reorganize them into intervals. Then, the dependent variables are obtained from the sum of the intervals of the independent variables and residuals. This is designed to complement the proposed model in this paper.
2. We generate raw random samples for the independent variables and residuals and obtain the dependent variables by adding them. The dependent and independent variables are subsequently reorganized into intervals.
3. We generate raw random samples for the independent variables and residuals and obtain the dependent variables by adding them. The dependent variables are then uniformly sampled on each block. This is designed to complement the empirical analysis data.

In real-world applications, we use real cardiological interval and PM2.5 concentration data to demonstrate the adaptability of the proposed models. In the former case, the data are already depicted as intervals, and we attempt to fit the interval-valued regression models and assess performance with the model predictions. In the latter case, we want to assess the correlations between the PM2.5 concentration and the wind speed and temperature. However, the PM2.5 concentration was collected hourly, while speed and temperature were collected every 10 minutes. Owing to the asynchronous frequencies at which these three sets of data were collected, we reorganize the data to intervals, resembled the third scenario of simulation study. The fitted models and comparisons with the conventional method are presented to confirm the validity of the proposed method.

Assessing the Impact of Riemannian Regression on Regression Methods for Symbolic Interval-Valued Data

Oldemar Rodríguez Rojas^{1,*}

1. School of Mathematics and Research Center for Pure and Applied Mathematics (CIMPA), University of Costa Rica, San José, Costa Rica

*Contact author: oldemar.rodriguez@ucr.ac.cr

Keywords: Symbolic Data Analysis, Interval-Valued Data, Riemannian Regression, Shrinkage Regression, Nonlinear Regression.

Symbolic interval-valued data provide a natural representation for complex statistical units in which within-unit variability is intrinsically relevant. Several regression approaches have been developed for this type of data, including the Center Method and the Center and Range Method, together with constrained, shrinkage, robust, and nonlinear extensions (Lima Neto and de Carvalho, 2008, 2010; Rodríguez, 2018; Lima Neto and de Carvalho, 2018; Chacón and Rodríguez, 2021). These approaches, however, are generally formulated within a Euclidean framework and therefore do not explicitly exploit the local geometric structure of the observations.

The main objective of this work is to *assess and quantify the impact* of introducing Riemannian Regression into existing regression methods for symbolic interval-valued data. The proposed framework builds on a Riemannian-statistical construction that equips an arbitrary data table with local geometric information derived from neighborhood relationships and local distances, allowing classical statistical procedures to be reformulated in a geometry-aware setting (Rodríguez, 2024, 2025, 2026). In the regression setting, neighborhood information obtained through procedures such as UMAP, ISOMAP, or DBSCAN is transformed into data-dependent dissimilarities and incorporated through Riemannian centering and weighting around a Fréchet mean. In this way, existing interval regression procedures can be generalized so that their fitting mechanisms account for the local geometry of the observations.

Particular emphasis is placed on measuring how this Riemannian generalization changes predictive performance. First, the impact is evaluated for the classical Center Method and Center and Range Method (Lima Neto and de Carvalho, 2008). The analysis is then extended to shrinkage-based approaches, including Ridge Regression, LASSO, and Elastic Net, previously proposed to improve prediction, variable selection, and interpretability for symbolic interval-valued variables (Rodríguez, 2018). The same principle is applied to nonlinear approaches, including regression trees, random forests, boosting, K -nearest neighbors, support vector regression, and neural networks (Chacón and Rodríguez, 2021), as well as to nonlinear parametric regression (Lima Neto and de Carvalho, 2017). Robust interval regression based on exponential-type kernels (Lima Neto and de Carvalho, 2018) also provides a relevant benchmark for evaluating how geometry-aware weighting compares with robustness-oriented weighting.

The central contribution is therefore not only the construction of Riemannian versions of these methods, but a systematic empirical assessment of the *impact* produced by the Riemannian generalization. Classical and Riemannian versions will be compared on real and simulated interval-valued datasets using lower- and upper-bound root-mean-square errors and squared correlation measures. The study will quantify gains or losses attributable specifically to the incorporation of local Riemannian geometry and will identify which families of interval regression methods benefit

most from this additional information.

References

- Chacón, J. E. and O. Rodríguez (2021). Regression models for symbolic interval-valued variables. *Entropy* 23(4), 429.
- Lima Neto, E. d. A. and F. d. A. T. de Carvalho (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis* 52, 1500–1515.
- Lima Neto, E. d. A. and F. d. A. T. de Carvalho (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis* 54, 333–347.
- Lima Neto, E. d. A. and F. d. A. T. de Carvalho (2017). Nonlinear regression applied to interval-valued data. *Pattern Analysis and Applications* 20(3), 809–824.
- Lima Neto, E. d. A. and F. d. A. T. de Carvalho (2018). An exponential-type kernel robust regression model for interval-valued variables. *Information Sciences* 454–455, 419–442.
- Rodríguez, O. (2018). Shrinkage linear regression for symbolic interval-valued variables. *Revue MODULAD* 45, 19–38.
- Rodríguez, O. (2024). Riemannian statistics for any type of data. In A. Plaia, L. Egidi, and A. Abbruzzo (Eds.), *Proceedings of the Statistics and Data Science 2024 Conference: New Perspectives on Statistics and Data Science*, Palermo, Italy, pp. 121–126. Università degli Studi di Palermo.
- Rodríguez, O. (2025). Riemannian statistics for any type of data. In J. Trejos, T. Chadjipadelis, A. Grané, and M. Villalobos (Eds.), *Data Science, Classification, and Artificial Intelligence for Modeling Decision Making*, Studies in Classification, Data Analysis, and Knowledge Organization, Cham, pp. 151–160. Springer.
- Rodríguez, O. (2026). Riemannian principal component analysis. In P. Brito, F. Denti, F. Greselin, K. Jajuga, and M. Zenga (Eds.), *Navigating Complexity: Statistical Methods, Data Analysis, and Machine Learning for Actionable Insights*, Studies in Classification, Data Analysis, and Knowledge Organization, Cham, pp. 243–251. Springer.

Calibrated Bayesian composite likelihood approach to Irpino-Verde linear histogram regression

Andrej Srakar^{1,2,*}

1. Jozef Stefan Institute, UNESCO IRCAI

2. School of Economics and Business, University of Ljubljana

*Contact author: andrej.srakar@ijs.si

Keywords: Likelihood-based SDA, Irpino-Verde Histogram Regression, Calibrated Bayesian Composite Likelihood, Hybrid Gibbs Sampler, Posterior Contraction Rate.

Symbolic data analysis (SDA) aggregates large individual-level datasets into a small number of distributional summaries, such as intervals or histograms. Inference is carried out using these summaries in place of the original dataset, resulting in computational gains at the loss of some information. In likelihood-based SDA, the likelihood function is characterised by an integral with a large exponent and in some circumstances the likelihood function is known to produce biased parameter estimates. To remedy for this Yang and coauthors ([Yang et al., 2024](#)) have developed a Bayesian framework for SDA methods in these settings using pseudo-marginal Markov-chain Monte Carlo methods based on path sampling, block-Poisson estimator, and a much faster but approximate method based on Taylor expansion. In our article we extend their approach to the study of Bayesian likelihood approaches for Irpino-Verde two-component regression for histogram data. We propose a calibrated Bayesian composite likelihood approach which reduces the need for a fully specified likelihood function and allows for complex within-study correlations. To enable robust posterior inference we develop a hybrid Gibbs sampler algorithm with Open-Faced Sandwich post-sampling adjustment (see e.g. [Wang \(2025\)](#)). We derive a Bernstein-von Mises theorem and study contraction rates for the Bayesian estimator proposed. Through simulation studies we demonstrate that the approach yields unbiased estimates and maintains coverage probabilities close to the nominal levels. Our approach upgrades the previous approach of Yang and coauthors with allowing for a more complex and adequate SDA study design and shows a solid computational performance and improved asymptotic efficiency.

In conclusion we discuss connections between Bayesian probability and statistics and symbolic data analysis. Both areas share some very interesting common features, in particular as related to Bayesian nonparametrics and concept of random measure. This promises highly interesting future pathways of study in SDA.

Our analysis is AI-assisted using ChatGPT 5.6 and Lean 4.

References

- Wang, Y. (2025). Advancing Statistical Methods for Multivariate and Network Meta-Analysis. Statistical Science Theses and Dissertations. 51.
- Yang, Y., M. Quiroz, B. Beranger, R. Kohn, and S. A. Sisson (2024). Analysing symbolic data by pseudo-marginal methods. [arXiv:2408.04419v2](#).

Session II

Clustering

Towards a Symbolic Data Extension of ClustOfVar: Clustering Ordinal Variables

Ana Felizardo Henriques^{1,*}, Adelaide Freitas^{1,2}, Fernando Sebastião³, João Marôco⁴

1. Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
2. Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
3. University of Leiria and Oeste, Leiria, Portugal; Laboratory of Separation and Reaction Engineering–Laboratory of Catalysis and Materials (LSRE-LCM), University of Leiria and Oeste, Leiria, Portugal; ALiCE – Associate Laboratory in Chemical Engineering, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
4. Intrepid Lab, ECEO, Lusófona University; CETRAD, University of Trás-os-Montes e Alto Douro

*Contact author: anac@ua.pt

Keywords: Ordinal Variables, Interval-Valued Data, RIDIT Transformation, Variable Clustering, Wasserstein Distance.

ClustOfVar, originally proposed by [Chavent et al. \(2012\)](#), is an R package for clustering a set of quantitative and qualitative variables. A hierarchical clustering algorithm has been implemented in the package. In what follows, we use the term ClustOfVar to also refer to the clustering method, rather than to the R package itself.

This work presents an initial step towards a symbolic data extension of ClustOfVar capable of handling quantitative, nominal, and ordinal variables while preserving the specific characteristics of each data type. As a first step towards this broader objective, we focus on the clustering of ordinal variables and propose an alternative representation based on symbolic data that preserves their intrinsic ordering without imposing unjustified numerical distances between categories.

As a variable clustering method, ClustOfVar aims to identify groups of variables with low within-cluster dissimilarity. The dissimilarity between two clusters C_1 and C_2 is defined as

$$d(C_1, C_2) = \mathcal{H}(C_1) + \mathcal{H}(C_2) - \mathcal{H}(C_1 \cup C_2),$$

where $\mathcal{H}(C)$ denotes the homogeneity of cluster C assessed through its association with a synthetic numerical variable representing the cluster. This synthetic variable is obtained as the first principal component resulting from Principal Component Analysis (PCA) for mixed data applied to the variables belonging to the cluster [[Chavent et al. \(2014\)](#); [Saracco and Chavent \(2016\)](#)]. Thus, $d(C_1, C_2)$ measures the loss of homogeneity resulting from merging the two clusters C_1 and C_2 .

Ordinal variables arise naturally in many applications, particularly in survey research where responses are collected in Likert-type scales. A relevant example is the Student Questionnaire – Grade 4 (SQ4), an international questionnaire administered within TIMSS to assess fourth-grade students’ attitudes towards mathematics, which has also been considered in our previous studies. Although ordinal variables are qualitative variables, their categories possess an intrinsic ordering that conveys information that should be preserved in the analysis. In the current implementation of ClustOfVar, ordinal variables can be treated as nominal variables. In this manner, the order of their attributes is not explicitly taken into account. An alternative approach is to assign consecutive numerical scores to the ordered categories and treat the resulting variable as quantitative. However, this approach implicitly assumes that the numerical distances between consecutive categories are meaningful and comparable, an assumption that is generally not justified for ordinal data.

We therefore explore an alternative symbolic data strategy for incorporating ordinal variables into ClustOfVar through a representation that preserves their ordinal nature while avoiding the assumption of equally spaced categories. These limitations motivate the development of a representation that incorporates the ordering information of ordinal variables without imposing arbitrary numerical distances between their categories.

We propose to represent ordinal variables as interval-valued symbolic data. Let X_i be an ordinal variable with ordered categories $j = 1, \dots, K_i$, let $p_j^{(i)} = P(X_i = j)$, and denote $F_0^{(i)} = 0$ and $F_j^{(i)} = \sum_{k=1}^j p_k^{(i)}$. We consider the RIDIT transformation, which assigns a numerical score to each category j according to its relative position in the empirical distribution of X_i , given by

$$\text{RIDIT}_j^{(i)} = F_{j-1}^{(i)} + \frac{p_j^{(i)}}{2}.$$

Rather than using the RIDIT values solely as point-valued numerical observations, we investigate their representation as interval-valued symbolic data. More specifically, we associate category j with the interval

$$I_j^{(i)} = [F_{j-1}^{(i)}, F_j^{(i)}]$$

and assume that its continuous representation is uniformly distributed over this interval. Hence, under the standard centre–radius representation of interval-valued symbolic data,

$$\text{center: } c_j^{(i)} = \text{RIDIT}_j^{(i)}, \quad \text{radius: } r_j^{(i)} = \frac{p_j^{(i)}}{2}.$$

The main methodological challenge is then to adapt the construction of the synthetic cluster variable and the corresponding homogeneity measure used in ClustOfVar to accommodate this interval-valued representation of ordinal variables. We use PCA for interval-valued data, retaining the first principal component to define the synthetic cluster variable, and consider an association measure based on the Wasserstein–Mallows metric under the uniform distributional assumption.

Preliminary results based on our SQ4 data are promising and suggest that the proposed strategy provides a meaningful way of incorporating the ordering information of ordinal variables into variable clustering. The results are discussed in terms of the resulting variable clusters and their interpretability, as well as the contribution of the ordinal variables to the corresponding synthetic cluster components. These results provide preliminary evidence that the proposed symbolic data representation can incorporate the ordering information of ordinal variables into ClustOfVar while avoiding the assumption of equally spaced categories.

Acknowledgments This work is supported by CIDMA (<https://ror.org/05pm2mw36>) under the Portuguese Foundation for Science and Technology (FCT, <https://ror.org/00snfq58>), Grants UID/04106/2025 (<https://doi.org/10.54499/UID/04106/2025>) and UID/PRR/04106/2025 (<https://doi.org/10.54499/UID/PRR/04106/2025>).

References

- Chavent, M., V. Kuentz-Simonet, A. Labenne, and J. Saracco (2014). PCAmixdata: Multivariate analysis of mixed data. *The R Journal* 6(1), 72–79.
- Chavent, M., V. Kuentz-Simonet, B. Liquet, and J. Saracco (2012). ClustOfVar: An R package for the clustering of variables. *Journal of Statistical Software* 50(13), 1–16.
- Saracco, J. and M. Chavent (2016). Clustering of variables for mixed data. In D. Fraix-Burnet and S. Girard (Eds.), *Statistics for Astrophysics: Clustering and Classification*, Volume 77 of *EAS Publications Series*, pp. 121–169. EDP Sciences.

Beyond Closed Interval Patterns in Symbolic Data Analysis

Mehdi Kaytoue^{1,*}, Egor Dudyrev², Sergei O. Kuznetsov³, David Loiret⁴, Miguel Couceiro⁵,
Amedeo Napoli²

1. Univ. Lyon, INSA Lyon, CNRS, LIRIS UMR 5205, F-69621, Lyon, France

2. Univ. Lorraine, CNRS, Loria, F-54000, Nancy, France

3. HSE University, 20 Myasnitskaya St, Moscow, 101000, Russian Federation

4. Kortexya, F-57007, Mey, France, www.kortexya.com

5. IST, Univ. Lisbon, INESC-ID, 1000-029 Lisboa, Portugal

*Contact author: mehdi.kaytoue@insa-lyon.fr

Keywords: Symbolic Data Analysis, Interval Domains, Formal Concept Analysis, Pattern Mining

Conceived by Edwin Diday, Symbolic Data Analysis (SDA) extends classical data analysis by representing objects through symbolic variables such as sets, intervals, or distributions. Billard and Diday (2003) show that this representation is particularly useful when observations describe groups rather than isolated individuals, since it preserves variability within each group. In her overview, Brito (2014) presents several methods in SDA, including clustering and Formal Concept Analysis (FCA). In its classical form, FCA structures the shared descriptions between data objects through Galois connections and concept lattices, as introduced by Ganter and Wille (1999).

Concept lattices support indeed the organization and interpretation of data, the discovery of shared properties, and the subsequent processing of represented knowledge, as emphasized by Ferré et al. (2020). In the SDA setting, Polaillon (1998) studied multi-valued and interval data using Galois lattices, while Brito and Polaillon (2005) defined Galois connections for distributional data using generalization by Maximum and Minimum. These operations provide the basis for union- or convexification-based generalization and their dual intersection operations. Ganter and Kuznetsov (2001) proposed pattern structures as an extension of FCA to complex descriptions beyond binary attributes. Following this principle, Kaytoue et al. (2011) formalized numerical pattern mining with interval pattern structures, while Agarwal et al. (2011) established its explicit connection with SDA for interval-valued data through symbolic Galois lattices.

Whether one starts from SDA or from FCA equipped with pattern structures, a natural representative is the closed interval pattern, which corresponds to the intent of a concept in the lattice. However, it is only one among potentially infinitely many patterns describing the same objects (the concept extent). The use of alternative representatives is not new: Li et al. (2006) argued, under a minimum-description-length criterion, that generators can be preferable to closed patterns when objects are described by itemsets while Kaytoue et al. (2011) formalized it for interval patterns.

As an illustration, we consider numerical data where each object g has an interval-valued description $\delta(g) = ([l_1, u_1], \dots, [l_d, u_d])$, with one interval for each data attribute. We consider here convexification-based generalization only. The partial order $\preceq$ is defined by componentwise inclusion: $P_1 \preceq P_2 \iff [l_1^{(1)}, u_1^{(1)}] \subseteq [l_1^{(2)}, u_1^{(2)}] \wedge \dots \wedge [l_d^{(1)}, u_d^{(1)}] \subseteq [l_d^{(2)}, u_d^{(2)}]$. The generalization operation is the componentwise convex hull $[l_1, u_1] \sqcap [l_2, u_2] = [\min(l_1, l_2), \max(u_1, u_2)]$. Let D be a finite dataset and let P be an interval pattern. Its extent is $\text{ext}_D(P) = \{g \in D \mid \delta(g) \preceq P\}$. Two patterns are equivalent on D when they have the same extent, and all equivalent patterns form an equivalence class: $[P]_D = \{Q \mid \text{ext}_D(Q) = \text{ext}_D(P)\}$. In continuous domains, this class may contain infinitely many patterns because boundaries can vary without covering an additional object. The closed pattern is the most specific representative (the smallest hyper-rectangle containing the

descriptions of the extent), whereas minimal generators are minimal descriptions whose closure is the closed pattern (the largest hyper-rectangles containing the descriptions of the extent).

Our position is that both closed patterns and generators are canonical for lattice construction, but should not automatically be treated as the best representations for pattern mining. In practice, a domain expert may prefer a pattern that is more robust to noise, more interpretable, or more useful for subsequent processing. In this talk, we properly define and then discuss this problem through existing approaches to interval pattern mining, including pattern-space projections and algorithms (Belfodil et al. (2018)) and extensions to more general convex descriptions (Belfodil et al. (2017)) among others. Addressing these questions would strengthen the link between SDA, FCA, and pattern mining while preserving the variability of symbolic representations.

Acknowledgements. French National Research Agency: SmartFCA project (ANR-21-CE23-0023). S.O. Kuznetsov is supported by the Basic Research Program of HSE University, Moscow.

References

- Agarwal, P., M. Kaytue, S. O. Kuznetsov, A. Napoli, and G. Polaillon (2011). Symbolic galois lattices with pattern structures. In *13th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC)*, Volume 6743 of *LNCS*, pp. 191–198. Springer.
- Belfodil, A., A. Belfodil, and M. Kaytue (2018). Anytime subgroup discovery in numerical domains with guarantees. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, Volume 11052 of *LNCS*, pp. 500–516. Springer. Best student paper award.
- Belfodil, A., S. O. Kuznetsov, C. Robardet, and M. Kaytue (2017). Mining convex polygon patterns with formal concept analysis. In *26th Int. Joint Conf. on Artificial Intelligence (IJCAI)*.
- Billard, L. and E. Diday (2003). From the statistics of data to the statistics of knowledge. *Journal of the American Statistical Association* 98(462), 470–487.
- Brito, P. (2014). Symbolic data analysis: another look at the interaction of data mining and statistics. *WIREs Data Mining Knowl. Discov.* 4(4), 281–295.
- Brito, P. and G. Polaillon (2005). Structuring probabilistic data by Galois lattices. *Mathématiques et sciences humaines* 169, 77–104.
- Ferré, S., M. Huchard, M. Kaytue, S. O. Kuznetsov, and A. Napoli (2020). Formal concept analysis: From knowledge discovery to knowledge processing. In *A Guided Tour of Artificial Intelligence Research: Volume II: AI Algorithms*, pp. 411–445. Springer.
- Ganter, B. and S. O. Kuznetsov (2001). Pattern structures and their projections. In *Conceptual Structures: Broadening the Base, 9th International Conference on Conceptual Structures (ICCS)*, Volume 2120 of *LNCS*, pp. 129–142. Springer.
- Ganter, B. and R. Wille (1999). *Formal Concept Analysis: Mathematical Foundations*. Springer.
- Kaytue, M., S. O. Kuznetsov, and A. Napoli (2011). Revisiting numerical pattern mining with formal concept analysis. In *Proc. of the 22nd Int. Joint Conf. on Artificial Intelligence*.
- Kaytue, M., S. O. Kuznetsov, A. Napoli, and S. Duplessis (2011). Mining gene expression data with pattern structures in formal concept analysis. *Inf. Sci.* 181(10), 1989–2001.

- Li, J., H. Li, L. Wong, J. Pei, and G. Dong (2006). Minimum Description Length Principle: Generators Are Preferable to Closed Patterns. In *21st Nat. Conf. on Artificial Intelligence (AAAI)*.
- Polailon, G. (1998). *Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme*. Ph. D. thesis, Université Paris IX-Dauphine.

Dynamical Clustering Algorithm for Interval Data Based on Weighted Partitioning

Eric C. Soares^{1,*}, Renata M. C. R. de Souza¹, Getúlio, J. A. do Amaral²

1. Centro de Informática, Universidade Federal de Pernambuco, Recife - PE, Brasil.

2. Departamento de Estatística, Universidade Federal de Pernambuco, Recife - PE, Brasil.

*Contact author: rmcrs@cin.ufpe.br

Keywords: Symbolic Data Analysis, Interval Data, Clustering, Weighted Partitioning

Symbolic Data Analysis (SDA) has developed a variety of methods for interval-valued data, including regression, classification, clustering, principal component analysis, and decision trees. In the clustering context, several partitioning methods have been proposed. The assignment of units to clusters depends on the adopted distance, and feature weights can be incorporated to identify clusters with different shapes and sizes [de Souza and de Carvalho \(2004\)](#); [de Carvalho and Lechevallier \(2009\)](#).

The choice of distance can affect the quality of the resulting partition. However, regardless of the adopted distance, partitioning algorithms may be sensitive to initialization and converge to poor local minima. A common strategy is to perform multiple random restarts to increase the probability of obtaining a good local minimum.

For classical data, [Tzortzis and Likas \(2014\)](#); [Liu et al. \(2019\)](#); [Golzari Oskouei et al. \(2021\)](#) proposed weighted partitioning methods to account for differences in within-cluster dispersion. In this work, we extend the cluster-weighting approach to interval-valued data using City-Block distances. The proposed method assigns a weight to each cluster according to its within-cluster dispersion and is also used as an initialization strategy for existing interval-valued clustering methods.

Let $U = \{1, \dots, n\}$ be a set of n units described by p symbolic interval-valued variables. Each unit i is represented by a vector of interval-valued variables $\mathbf{x}_i = (x_i^1, \dots, x_i^p)$, where $x_i^j = [a_i^j, b_i^j] \in I$, with $I = \{[a, b] : a, b \in \mathbb{R}, a \leq b\}$. Following the dynamic clustering algorithm proposed, the method searches for a partition $P = (C_1, \dots, C_K)$ of U and a set of prototypes $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_K\}$ that locally minimize an adequacy criterion J . The prototype of cluster C_k is represented by $\mathbf{y}_k = (y_k^1, \dots, y_k^p)$, where $y_k^j = [\alpha_k^j, \beta_k^j] \in I$. Let $W = \{\omega_1, \dots, \omega_K\}$ denote the cluster weights, subject to $\omega_k \geq 0$ and $\sum_{k=1}^K \omega_k = 1$, and let $t \in [0, 1)$ control the influence of the cluster weights on the adequacy criterion. The cluster-weighted criterion based on the City-Block distance for interval-valued data is defined as

$$J_W = \sum_{k=1}^K \omega_k^t \sum_{i \in C_k} d(\mathbf{x}_i, \mathbf{y}_k).$$

where $d(\mathbf{x}_i, \mathbf{y}_k)$ denotes the City-Block distance between the interval-valued unit $\mathbf{x}_i$ and the prototype $\mathbf{y}_k$. The cluster-weighting strategy is also incorporated into criteria with global and cluster-specific feature relevance weights, denoted by J_{WV} and J_{WVC} , respectively. For a fixed partition and fixed prototypes, the cluster weight is given by

$$\omega_k = \left[\sum_{h=1}^K \left(\frac{\sum_{i \in C_k} d(\mathbf{x}_i, \mathbf{y}_k)}{\sum_{i \in C_h} d(\mathbf{x}_i, \mathbf{y}_h)} \right)^{\frac{1}{t-1}} \right]^{-1}. \quad (5)$$

Following Celebi et al. (2013), the exponent $t \in [0, 1)$ controls the influence of the cluster weights on the adequacy criterion. According to Eq. (5), the cluster weights are determined by the within-cluster dispersions. Clusters with larger within-cluster dispersion receive larger weights and consequently contribute more strongly to the adequacy criterion, thereby penalizing highly dispersed clusters.

Eight synthetic interval-valued datasets in $\mathbb{R}^2$ with three clusters were considered, regarding balanced and unbalanced scenarios with equal or different variances for variables and clusters are considered. Clustering methods based on City-Block distance were evaluated using 10 Monte Carlo replications with 10 initializations each. We use several criteria: without feature weighting (J), with global feature weighting (J_V), and with cluster-specific feature weighting (J_{VC}) with their cluster-weighted versions J_W , J_{WV} , and J_{WVC} , respectively. In addition, hybrid strategies $J_W + J$, $J_{WV} + J_V$, and $J_{WVC} + J_{VC}$ were also evaluated, in which the partition obtained by the corresponding cluster-weighting method is used to initialize a respective literature method J , J_V or J_{VC} .

For each dataset, the algorithms were ranked according to their Adjusted Rand Index (ARI) values. The final performance was assessed by the average of the ranks regarding eight datasets. The lower ranks indicate better performances. The hybrid approaches achieved the lowest (best) ranks within their respective groups: 14.5 for criteria $J_W + J$, 11.5 for criteria $J_{WV} + J_V$ and 12.0 for criteria $J_{WVC} + J_{VC}$.

Eight real interval-valued datasets of SDA literature are also used in the experimental evaluation. The average performance ranking obtained from 200 initializations is calculated. This ranking provides a global assessment of the algorithms and the results show that, in general, the hybrid methods achieved the best results for ARI index. Thus, we can conclude the effectiveness of the proposed cluster-weighting approach as an initialization strategy for interval-valued clustering methods.

References

- Celebi, M. E., H. A. Kingravi, and P. A. Vela (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications* 40, 200–210.
- de Carvalho, F. A. T. and Y. Lechevallier (2009). Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition* 42, 1223–1236.
- de Souza, R. M. C. R. and F. A. T. de Carvalho (2004). Clustering of interval data based on city–block distances. *Pattern Recognition Letters* 25, 353–365.
- Golzari Oskouei, A., M. A. Balafar, and C. Motamed (2021). Fkmawcw: Categorical fuzzy k-modes clustering with automated attribute-weight and cluster-weight learning. *Chaos, Solitons & Fractals* 153, 111494.
- Liu, J., Y. Guo, D. Li, Z. Wang, and Y. Xu (2019). Kernel-based minmax clustering methods with kernelization of the metric and auto-tuning hyper-parameters. *Neurocomputing* 359, 173–184.
- Tzortzis, G. and A. Likas (2014). The minmax k-means clustering algorithm. *Pattern Recognition* 47, 2505–2516.

Kernel-Based Fuzzy Co-Clustering for Interval-Valued Data

José Nataniel A. de Sá¹, Marcelo R.P. Ferreira², Francisco de A. T. de Carvalho^{1,*}

1. Universidade Federal de Pernambuco, Centro de Informática - CIn/UFPE, Av. Jornalista Aníbal Fernandes, s/n, CEP 50.740-560 - Recife, PE, Brazil

2. Universidade Federal da Paraíba, Departamento de Estatística - DE/UFPB, Conj. Pres. Castelo Branco III, CEP 58.050-585 - João Pessoa, PB, Brazil

*Contact author: fatc@cin.ufpe.br

Keywords: Interval-Valued Data, Co-Clustering, Gaussian Kernel.

Symbolic Data Analysis (SDA) is a field of statistics and artificial intelligence that focuses on the analysis of multi-valued data, such as intervals, histograms, and categories [Bock and Diday \(2012\)](#). This type of data arises when the objective is to analyze groups of individuals rather than single observations. For instance, while data may be collected at the level of students, the analytical goal may be to study universities. Traditional aggregation methods, such as using the mean or median for numerical features and the mode for categorical features, often lead to a loss of information about data variability. Symbolic data emerges as an alternative to these conventional strategies for better preserving variability, enabling a more informative representation of aggregated data. The focus of this paper is to study interval-valued data.

Clustering is a well-known area of unsupervised learning whose goal is to identify clusters of similar samples. Among clustering algorithms, kernel-based methods are particularly suited to handling nonlinear data by transforming the original data space into a higher-dimensional kernel space. This transformation mitigates a major limitation of two of the most widely used clustering algorithms, K-means and Fuzzy C-Means, which often fail to correctly identify non-spherical and nonlinearly separable clusters.

Traditional clustering methods group only the samples, assume no cluster structure in the features, and implicitly assume that all clusters are defined across all features. In contrast, co-clustering algorithms simultaneously group samples and features based on the associations between them, forming data blocks in which samples and features are correlated. By grouping data in two dimensions, these methods provide a partition of the samples, as in traditional clustering, while also yielding a partition of the features, and have demonstrated competitive performance compared to classical clustering techniques. Co-clustering methods have been successfully applied across several domains, including recommender systems [Iftikhar et al. \(2024\)](#), document clustering [Fettal et al. \(2023\)](#), and gene expression analysis [Husseini et al. \(2024\)](#).

Kernel-based co-clustering algorithms have also been proposed in the literature. Ref. [de Sá et al. \(2022\)](#) introduced the first approach combining co-clustering with kernel functions, employing a Gaussian kernel and proposing an algorithm that computes feature weights for sample clusters. Subsequently, Ref. [de Sá et al. \(2025\)](#) presented hard and fuzzy Gaussian kernel co-clustering methods with automatic computation of the kernel width parameter. Finally, Ref. [de Sá et al. \(2025\)](#) proposed a method that computes weights for both samples and features, demonstrating strong performance in scenarios with noise in both dimensions.

Despite the broad applicability of co-clustering algorithms, to the best of our knowledge, no existing work has addressed co-clustering methods for interval-valued data. This presentation is concerned with novel co-clustering models for interval-valued data. More precisely, We introduce

the *Interval Fuzzy Double K-means (IFDK)* algorithm, as well as its kernelized extensions, namely the *Interval Kernel Fuzzy Double K-means in the Original Space (IKFDK-O)* and the *Interval Kernel Fuzzy Double K-means in the Kernel Space (IKFDK-K)*, which employ the Gaussian kernel to identify nonlinear cluster structures. Experiments conducted on synthetic and benchmark interval-valued datasets demonstrate the effectiveness of the proposed methods compared to traditional interval fuzzy clustering approaches, particularly in scenarios involving noisy data. On the benchmark datasets, the proposed kernel-based algorithms achieved superior or competitive results, whereas the IFDK method was outperformed by the other methods.

References

- Bock, H.-H. and E. Diday (2012). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Heidelberg: Springer Science & Business Media.
- de Sá, J. N. A., M. R. Ferreira, and F. d. A. de Carvalho (2022). Kernel-based fuzzy co-clustering in feature space with automated variable weighting. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Padua, Italy, pp. 1–8. IEEE.
- de Sá, J. N. A., M. R. P. Ferreira, and F. A. T. de Carvalho (2025). Fuzzy and crisp gaussian kernel-based co-clustering with automatic width computation. *IEEE Transactions on Fuzzy Systems* 33(6), 1977–1991.
- de Sá, J. N. A., M. R. P. Ferreira, and F. A. T. de Carvalho (2025). A dual-weighted gaussian kernel-based method for fuzzy co-clustering. In *2025 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Reims, France, pp. 1–6. IEEE.
- Fettal, C., L. Labiod, and M. Nadif (2023). Boosting subspace co-clustering via bilateral graph convolution. *IEEE Transactions on Knowledge and Data Engineering* 36, 960–971.
- Husseini, Z. M., M. H. F. Zarandi, and A. Ahmadi (2024). Type2 soft biclustering framework for alzheimer microarray. *Applied Soft Computing* 152, 111227.
- Ifthikhar, A., M. A. Ghazanfar, M. Ayub, S. A. Alahmari, N. Qazi, and J. Wall (2024). A reinforcement learning recommender system using bi-clustering and markov decision process. *Expert Systems with Applications* 237, 121541.

Session III

Symbolic Profiles

Assessing Adolescent Profile Coherence Using Classical and Distributed-Based Representations

Jasminka Dobša^{1*}, Simona Korenjak-Černe^{2,3}, Miranda Novak⁴, Maja Buhin Pandur¹,
Lucija Šutić Barbarić⁴

1. University of Zagreb Faculty of Organization and Informatics, Varaždin, Croatia

2. University of Ljubljana, School of Economics and Business

3. Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia

4. University of Zagreb Faculty of Education and Rehabilitation Sciences, Zagreb, Croatia

*Contact author: jasminka.dobsa@foi.hr

Keywords: Adolescent Profiles, Clustering, Distribution-Based Data Representation, Classification, Wasserstein Distance.

This study investigates the coherence of adolescent profiles identified from ecological momentary assessment (EMA) data and compares classical and distribution-based data representations. The profiles are further characterized using passive mobile sensing data on application usage and survey data including demographic characteristics and symptoms of depression, anxiety, and stress.

The dataset comprises 131 Croatian high school students who evaluated the quality of their close friendships and their affect seven times a day for one week (up to 49 assessments per participant). Data were collected using the Effortless Assessment Research System (EARS) from Ksana Health, University of Oregon ([Lind et al. \(2018a,b\)](#)). Based on factor analysis, the 25 EMA variables were grouped into six constructs reflecting positive and negative affect, positive and negative emotions towards friends, confidence, and attentiveness. Passive mobile sensing data included the use of 927 applications, which were semi-automatically classified into 16 application categories using generative artificial intelligence.

Adolescent profiles were identified using conventional Ward's hierarchical clustering and k-means clustering, as well as their adaptations for distribution-valued data within the framework of Symbolic Data Analysis (SDA) ([Billard and Diday \(2006\)](#); [Diday \(2016\)](#)). In the classical approach, each construct is represented by its mean value across repeated EMA measurements. In the SDA approach, each construct is represented by a distribution, thereby preserving within-individual variability in responses over the observation period. Distribution-valued data were clustered using methods based on the Wasserstein distance ([Irpino and Verde \(2015\)](#); [Verde and Irpino \(2022\)](#)). The two approaches produced similar cluster structures for smaller numbers of clusters ($k = 2, 3$), whereas more substantial differences emerged for $k = 4$ and $k = 5$. A four-cluster solution was considered the most appropriate based on the dissimilarity matrix, the dendrogram, and expert judgement. External validation using survey data and passive mobile sensing variables indicated better separation of profiles obtained from distribution-valued data, particularly for application usage variables.

As an additional assessment of profile coherence, we examine how accurately cluster membership can be predicted using supervised machine learning. Two classification settings corresponding to the two data representations are compared: classification based on conventionally aggregated data and classification based on distribution-valued data. For the latter, a k -nearest neighbours classifier adapted to distribution-valued data using the Wasserstein distance is applied. Classification performance provides an additional criterion for evaluating the separability and reproducibility of

adolescent profiles obtained using classical and distribution-based representations.

References

- Billard, L. and E. Diday (2006). *Symbolic Data Analysis. Conceptual Statistics and Data Mining. Wiley Series in Computational Statistics*. John Wiley & Sons, Inc.
- Diday, E. (2016). Thinking by classes in data science: the symbolic data analysis paradigm. *WIREs Comput Stat* 8, 172–205.
- Irpino, A. and R. Verde (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. *Adv Data Anal Classif* 9, 143–175.
- Lind, M., M. Byrne, G. Wicks, A. Smidt, and N. Allen (2018a). The effortless assessment of risk states (ears) tool: An interpersonal approach to mobile sensing. *JMIR Ment. Health* 5.
- Lind, M., M. Byrne, G. Wicks, A. Smidt, and N. Allen (2018b). The effortless assessment of risk states (ears) tool: An interpersonal approach to mobile sensing. *JMIR Ment. Health* 5.
- Verde, R. and A. Irpino (2022). Partitive and hierarchical clustering of distributional data using the wasserstein distance. chapter 5. In *Analysis of Distributional Data*, pp. 105–126. CRC Press.

From Publications to Symbolic Researcher Profiles: An SDA Approach to Exploring Interdisciplinarity

Maja Buhin Pandur^{1,*}, Luka Kronegger², Jasminka Dobša¹

1. University of Zagreb Faculty of Organization and Informatics, Varaždin, Croatia

2. University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia

*Contact author: mbuhin@foi.unizg.hr

Keywords: Symbolic Data Analysis, Symbolic Clustering, Interdisciplinarity, Disciplinary Diversity, Symbolic Researcher Profiles

Bibliometric studies often examine interdisciplinarity through disciplinary diversity and related indicators (Porter and Rafols, 2009). Diversity is, however, a multidimensional concept that includes variety, balance, and disparity (Wang et al., 2015; Leydesdorff, 2018). Bibliometric indicators must also be examined with caution, because they may not exactly present researchers' views on interdisciplinarity (Aksnes et al., 2026). Although interdisciplinarity has been extensively studied using traditional bibliometric indicators, there is still a lack of research on the symbolic representation of researchers' publication sets, which also contributes to between- and within-researcher diversity.

At the same time, Symbolic Data Analysis (SDA) provides a method for analyzing complex and aggregated units while conserving information about their internal variability. Instead of representing an aggregated unit by a single value, symbolic variables may be intervals, groups of categories, or distributions (Billard and Diday, 2003; Brito, 2014; Diday, 2016). This makes SDA particularly suitable for bibliometric portfolios, in the case of which a given researcher's set of publications may vary in terms of publication year, citation features, affiliation, and disciplinary fields.

This study proposes an SDA-based framework in which individual researchers are represented as symbolic researcher profiles based on their publication-level bibliographic information. The initial population consists of 6,289 unique researchers identified in the Slovenian Current Research Information System (SICRIS), of which 5,954 are in Social Sciences and 335 in Interdisciplinary Research. SICRIS defines the researcher population and also serves as an external reference point instead of being included in the clustering methodology. Publication data are retrieved from the Web of Science (WoS) Core Collection for a 30-year observation period (1996 - 2025). The final analysis will only include researchers associated with WoS records whose portfolios satisfy the final inclusion criteria, using all available combinations of author names, affiliations, and persistent identifiers. Publications by matched researchers will include all eligible publications in the WoS Core Collection, not only those belonging to Social Sciences, thus preserving interdisciplinarity without mixing up the population definition and the disciplinary measurement.

Information at the publication level is aggregated at the researcher level using symbolic variable types that are appropriate for each characteristic. Publication years, citation counts, and references characteristics are represented as interval-valued variables, affiliations as a multi-valued variables, and disciplinary composition as multi-valued modal variable describing the relative distribution of WoS categories according to a predefined counting rule for publications in more than one category. Symbolic dissimilarities between these profiles will provide as the basis for symbolic clustering, whose clusters will be interpreted mainly in terms of disciplinary diversity. The diversity of disciplines will be defined by the number of different disciplinary categories used in a researcher portfolio, while the balance will be assessed using normalized Shannon entropy taking into account the

relative distribution across those categories (Yegros-Yegros et al., 2015). Measures accounting for disparities between disciplines, such as Rao-Stirling diversity, may be considered in subsequent extensions of the analysis (Leydesdorff, 2018)

Two main research questions are considered. First, can the researchers' publication portfolios be represented as symbolic objects that preserve the heterogeneity within a researcher and show the disciplinary diversity of the publication portfolio through symbolic clustering? Second, among researchers that have efficient links to publication records on WoS, are the clusters that are more disciplinarily diverse populated with more researchers belonging to the category of Interdisciplinary Research on SICRIS? The SICRIS classification provides as the external benchmark for evaluating whether clusters with more disciplinary diversity include more researchers classified under Interdisciplinary Research. As an institutional classification independent of WoS-based symbolic profiles, SICRIS enables descriptive comparison without being treated as a ground-truth measure of interdisciplinarity.

At submission, the SICRIS-based population has been established, while the process of gathering the publication records from WoS, matching these records, and creating the final sample to analyze is still in progress. Thus, the number of units in the sample and the outcome of empirical clustering are yet to be known at this stage. The proposed methodology does not regard disciplinary diversity as interdisciplinarity, but seeks to test whether the SDA approach would enable researchers to grasp this aspect of research better due to the structural differentiation of the publication portfolio rather than its scalar aggregation. This can be a pilot project that could be later supplemented with other methodologies like social network analysis or topic modeling.

References

- Aksnes, D. W., H. Karlstrøm, and F. N. Piro (2026). Self-reported and bibliometric interdisciplinarity measures rarely correspond: A survey-based comparative analysis of indicators and researcher perceptions. *Scientometrics* 131, 189–208.
- Billard, L. and E. Diday (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association* 98(462), 470–487.
- Brito, P. (2014). Symbolic data analysis: Another look at the interaction of data mining and statistics. *WIREs Data Mining and Knowledge Discovery* 4(4), 281–295.
- Diday, E. (2016). Thinking by classes in data science: The symbolic data analysis paradigm. *WIREs Computational Statistics* 8(5), 172–205.
- Leydesdorff, L. (2018). Diversity and interdisciplinarity: How can one distinguish and recombine disparity, variety, and balance? *Scientometrics* 116, 2113–2121.
- Porter, A. L. and I. Rafols (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* 81(3), 719–745.
- Wang, J., B. Thijs, and W. Glänzel (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLOS ONE* 10(5), e0127298.
- Yegros-Yegros, A., I. Rafols, and P. D'Este (2015). Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. *PLOS ONE* 10(8), e0135095.

From Keywords to Themes: Distribution-Valued Symbolic Descriptions of Consensus Bibliometric Communities

Carlo Drago^{1,*}, Angelo Leogrande², Gentian Hoxhalli³

1. Department of Economics, Psychology, Communication, Education, and Motor Sciences, University Niccolo Cusano, Rome, Italy

2. LUM University, Giuseppe Degennaro, Casamassima, Bari, Italy

3. Luarasi University, Tirana, Albania & Academy of Armed Forces, Tirana, Albania

*Contact author: carlo.drago@unicusano.it

Keywords: Symbolic Data Analysis, Distribution-Valued Data, Wasserstein Distance, Bibliometrics, Migration Studies.

Bibliometric science mapping transforms documents into networks of co-occurring terms and summarizes them into thematic clusters. This is useful for scientific literatures, but the process often ends once clusters have been labeled, treating each theme as an output based on single points. We propose a different symbolic protocol in which consensus bibliometric communities become the statistical units of analysis and are described by distribution-valued variables. The contribution relies on the fact that it is possible to change the units from individual keywords to meso-level thematic classes. Each class is then represented by the empirical distribution of node centralities within the community.

This follows the SDA principle that classes may be analyzed through symbolic descriptions that preserve internal variability (Billard and Diday, 2003; Diday, 2016; Brito, 2014).

The workflow has three steps. First, author keywords are normalized and transformed into a weighted co-occurrence network.

Second, stochastic community detection is stabilized by an algorithm-balanced consensus procedure: Leiden, Louvain, Walktrap, and Infomap are each run 20 times, their partitions are averaged within algorithm and then across algorithms into a co-assignment matrix, a strict majority rule (co-assignment above 0.50) defines a weighted consensus graph, and a final Leiden partition on that graph defines the thematic classes, following standard consensus practice (Lancichinetti and Fortunato, 2012). This retains the methodological core of co-word analysis while reducing dependence on a single algorithmic partition (Callon et al., 1983; Cobo et al., 2011; Aria and Cuccurullo, 2017). Third, each consensus community is re-described as a distribution-valued symbolic unit.

Let C_i be a community and, for $v \in \{d, b\}$, let $\mu_{iv} = n_i^{-1} \sum_{u \in C_i} \delta_{x_{uv}}$ be the empirical distribution of Freeman degree (d) or betweenness (b). The univariate dissimilarity is

$$W_{2,v}^2(i, j) = \int_0^1 \{Q_{iv}(p) - Q_{jv}(p)\}^2 dp,$$

where Q_{iv} is the quantile function. Variables are standardized by their between-community Wasserstein variances, $s_v^2 = m^{-1} \sum_i W_{2,v}^2(\mu_{iv}, \bar{\mu}_v)$, where $\bar{\mu}_v$ is the Wasserstein barycenter, that is, the average quantile function (Irpino and Verde, 2015). The combined distance is $D^2(i, j) = \sum_v \omega_v W_{2,v}^2(i, j) / s_v^2$, with $\omega_d = \omega_b = 0.5$. Communities are clustered by Ward.D2 on D . The weights are fixed a priori rather than estimated; this makes the present specification a non-adaptive baseline in the sense of Irpino et al. (2014).

This construction addresses a limitation of interval-valued summaries. Min–max intervals are transparent but may be dominated by hub keywords; histogram-valued or distribution-valued descriptions retain the internal shape of the community, including the mass of peripheral terms and the upper tail of bridging terms. It also connects bibliometric science mapping with existing SDA work on networks as symbolic data and with Wasserstein approaches for distributional data (Giordano and Brito, 2014; Dias and Brito, 2015). A histogram approximation computed with HistDAWass (Irpino and Verde, 2015; Irpino, 2024) is used as a sensitivity check on the discretization of these distributions, reproducing the reported partition (adjusted Rand index 1.00).

The illustration concerns migration research indexed in Scopus. From 9,543 deduplicated records, 8,640 contain usable author keywords; the final network contains 267 keyword nodes and 888 edges. The consensus procedure identifies 40 communities, with modularity 0.662 on the original graph. All communities, including small and singleton communities, are retained as empirical distributions. The distributional clustering selects two clusters by maximum Ward.D2 silhouette (average silhouette 0.541). The first contains five field-structuring communities centered on migration, refugees, mental health/war, immigration, and refugee/youth. The second contains 35 more specialized or compact themes, including forced displacement, diaspora, education and inclusion, migrant health, climate migration, intersectionality, migrant workers, Rohingya displacement, deportation, remittances, and xenophobia.

The Wasserstein representation changes the interpretation of the result. Field-structuring themes are characterized by quantile profiles with a long upper tail in degree and betweenness, while most of their keywords remain peripheral. Conversely, specialized communities show lower and more compact centrality distributions.

References

- Aria, M. and C. Cuccurullo (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11(4), 959–975.
- Billard, L. and E. Diday (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association* 98(462), 470–487.
- Brito, P. (2014). Symbolic data analysis: Another look at the interaction of data mining and statistics. *WIREs Data Mining and Knowledge Discovery* 4(4), 281–295.
- Callon, M., J.-P. Courtial, W. A. Turner, and S. Bauin (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information* 22(2), 191–235.
- Cobo, M. J., A. G. Lopez-Herrera, E. Herrera-Viedma, and F. Herrera (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology* 62(7), 1382–1402.
- Dias, S. and P. Brito (2015). Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining* 8(2), 75–113.
- Diday, E. (2016). Thinking by classes in data science: The symbolic data analysis paradigm. *WIREs Computational Statistics* 8(5), 172–205.
- Giordano, G. and P. Brito (2014). Social networks as symbolic data. In *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*, pp. 133–141. Springer.
- Irpino, A. (2024). *HistDAWass: Histogram-Valued Data Analysis*. R package version 1.0.8.

- Irpino, A. and R. Verde (2015). Basic statistics for distributional symbolic variables: A new metric-based approach. *Advances in Data Analysis and Classification* 9(2), 143–175.
- Irpino, A., R. Verde, and F. d. A. T. de Carvalho (2014). Dynamic clustering of histogram data based on adaptive squared wasserstein distances. *Expert Systems with Applications* 41(7), 3351–3366.
- Lancichinetti, A. and S. Fortunato (2012). Consensus clustering in complex networks. *Scientific Reports* 2, 336.

Session IV

Representations and Dimension Reduction

Pairwise Logratio Functions: A Symbolic Data Analysis Perspective for Density-Valued Data

Rui Nunes^{1,*}, Karel Hron², Jitka Machalová², Peter Filzmoser³

1. Faculty of Sciences, University of Porto, Porto, Portugal & LIAAD-INESC TEC, Portugal
 2. Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, Czech Republic
 3. Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria
- *Contact author: up201400313@up.pt

Keywords: Pairwise Logratio Function, Symbolic Data Analysis, Density-Valued Data, Bayes Spaces, Compositional Data Analysis

In the context of Symbolic Data Analysis (SDA) and in many modern datasets, we no longer record one value for each observational unit; instead, an observation may be a distribution. This is the case, for example, when we analyze the range of wages in a region or the concentration profile of a pollutant. The challenge is that a probability density is positive and integrates to one, so standard Euclidean methods can produce misleading results or depend on an arbitrary normalization. This makes Bayes spaces $\mathcal{B}^2(I)$ a natural framework for density-valued data, in the same way that Aitchison geometry (Aitchison, 1986) provides an appropriate framework for compositional data. Since Bayes space is a separable Hilbert space (Egozcue et al., 2006), there exists an isometric isomorphism with the L^2 space, called the centered log-ratio (*clr*) transformation (van den Boogaart et al., 2014). Let $f \in \mathcal{B}^2(I)$ be a strictly positive density on a bounded interval I , with length $\eta = |I|$.

$$clr(f)(t) = \log f(t) - \frac{1}{\eta} \int_I \log f(u) d\lambda(u), \quad \forall t \in I.$$

clr maps Bayes-space elements to the zero-integral subspace of $L^2(I)$ and consequently permits the use of standard Functional Data Analysis (FDA) tools; nevertheless, the centering term $\frac{1}{\eta} \int_I \log f(u) d\lambda(u)$ depends on the complete domain. If attention is restricted to a sub-interval, the centered log-ratio representation must be recalculated rather than simply restricted.

In this work we propose the pairwise log-ratio (PLR) transformation, we argue that the PLR function is a natural bridge between the two traditions: it preserves the full distributional observation while expressing its information through directly interpretable comparisons between parts of the support. The proposed *plr* function is a mapping $plr(\cdot, \cdot) : I \times I \rightarrow \mathbb{R}$, defined as

$$plr(f)(x, y) = \log \frac{f(x)}{f(y)}.$$

In Figure 1 we can observe two representations of the *plr* function.

A typical symbolic-data workflow may represent each unit by a density and then compare units, cluster them, or regress their distributional outcomes on explanatory variables. The *plr* function complements this workflow in three ways. First, it provides a distribution-preserving representation. SDA analysis can retain the pairwise structure of the entire support. This is meaningful even when two units have similar means but different shapes. Second, *plr* is naturally compatible with FDA, all its techniques can be applied after representing *plr* functions in a finite basis. Compositional splines are particularly useful for this purpose because they respect positivity and the relative

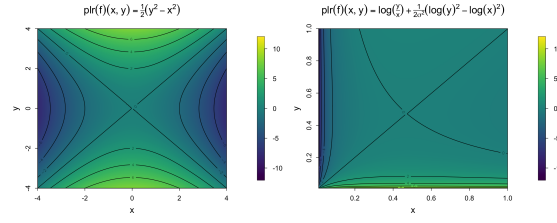


Figure 1: Pairwise logratio functions for $X \sim N(0, 1; I = [-4, 4])$ (left) and for $X \sim LN(0, 0.75; I = [0, 1])$ (right).

geometry inherited from compositional data analysis. Third, plr is stable under domain restriction. If attention is restricted from I to a subdomain $I_1 \subset I$, the reduced representation is simply the restriction of $plr(f)$ to $I_1 \times I_1$. This is valuable in SDA when analysts focus on a scientifically relevant range.

Two examples illustrate the proposed approach. In the first, zinc concentration distributions from two Czech districts are compared. The plr surface identifies where the contaminated district has locally greater relative concentration than the reference district and remains interpretable when the analysis is restricted to a selected concentration range. In the second application, income distributions across U.S. states are represented by plr functions and analyzed through function-on-scalar regression. Covariates such as gross domestic product and unemployment rate then explain variation in pairwise distributional dominance, rather than only variation in a state-level average. These examples suggest a practical SDA pipeline: estimate one density per unit; compute its plr function; represent plr function using a compositional-spline basis; and perform functional regression, dimension reduction, or clustering on the coefficients. The fitted effects can then be mapped back to pairs (x, y) and read as changes in relative density. This makes model output easier to communicate to domain experts than coefficients tied to an arbitrary global centering convention.

The pairwise log-ratio function offers a concise conceptual bridge between Bayes space, FDA, and SDA. It treats each distribution as a genuine observation, preserves relative information, and gives every modeled value a transparent interpretation as a comparison between two support locations. By combining plr with compositional splines and functional methods, researchers can move from descriptive symbolic summaries to principled distributional modeling without sacrificing interpretability. The main limitation is computational and statistical: plr functions live on a product domain, and estimates can be unstable where a density is close to zero. Smoothing, positivity-preserving density estimation, and regularized basis representations are therefore important. Future work should develop inferential procedures for symbolic samples, robust versions for sparse histograms, and extensions to multivariate distributions.

Acknowledgements This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (<https://doi.org/10.54499/UID/50014/2025>).

References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.
- Egozcue, J. J., J. L. Díaz-Barrero, and V. Pawłowsky-Glahn (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica* 22(4), 1175–1182.

van den Boogaart, K. G., J. Egozcue, and V. Pawlowsky-Glahn (2014). Bayes Hilbert spaces.
Australian & New Zealand Journal of Statistics 54(2), 171–194.

Masking Without Misleading: Learning Privacy-Preserving Representations against Reconstruction Attacks

Jiayu Shang¹, Eufrásio de A. Lima Neto², Marcelo R.P. Ferreira², Telmo Silva Filho^{1,*}

1. University of Bristol, United Kingdom

2. Universidade Federal da Paraíba, Brazil

*Contact author: telmo.silvafilho@bristol.ac.uk

Keywords: Privacy-preserving Data Publication, Reconstruction Attack, Utility-Privacy Trade-Off, Stochastic Data Anonymisation, Representation Learning.

The public release of labelled datasets underpins reproducible machine learning benchmarking, yet exposing raw features invites reconstruction attacks (Dwork and Roth, 2014), and conventional anonymisation techniques tend to degrade the data so heavily that they no longer support a faithful comparison among methods. We argue that benchmarking depends only on the relative ordering of classifiers, not on their absolute performance, so we study transformations that protect raw features while preserving this ordering.

We introduce a two-stage framework that first maps the data onto a two-dimensional nonlinear manifold with t-SNE (van der Maaten and Hinton, 2008) or UMAP (McInnes et al., 2018) and then perturbs the projected coordinates with calibrated uniform noise. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the Standard-Scaled (Pedregosa et al., 2011) original feature matrix. The first step produces an embedding $\mathbf{Z} \in \mathbb{R}^{n \times d}$, which is then modified by an interval-valued (Bock, 2000) mechanism that produces a noisy embedding $\mathbf{W}$ by perturbing each coordinate z_{ij} independently:

$$w_{ij} = z_{ij} + \delta_{ij}^+ - \delta_{ij}^-, \quad \delta_{ij}^\pm \sim \mathcal{U}(0.0001, u_k), \quad (6)$$

where u_k is the noise upper limit for level $k \in \{1, 2, 3\}$ and $\mathcal{U}(a, b)$ denotes the continuous uniform distribution on $[a, b]$. The asymmetric two-sided perturbation prevents trivial cancellation while ensuring the expected displacement is zero.

We evaluate the approach across multiple classification datasets, nine classifiers, and twelve configurations spanning three noise levels. Utility is measured by the Spearman rank correlation (ρ_s) between the classifier rankings obtained before and after transformation, and privacy by a k-nearest neighbour reconstruction attack quantified through mean squared error and the coefficient of determination. This work makes the following contributions:

- **Structural privacy of non-linear projections.** We demonstrate that, even before any noise is added, t-SNE and UMAP yield a mean reconstruction MSE that is $1.74\times$ and $2.17\times$ higher, respectively, than the value obtained by interval-based perturbation alone.
- **Noise-level interaction effect.** We show that adding *medium* interval noise to a t-SNE projection simultaneously *improves* classifier-ranking utility (ρ_s increases by 0.09 relative to bare t-SNE) and substantially increases reconstruction resistance (MSE increases by $2.24\times$), a counterintuitive win-win that does not hold at high noise levels.
- **Classifier-type robustness analysis.** Tree-based classifiers (CART, XGBoost, Random Forest) retain ranking stability under all transformations, whereas linear models (LDA, LR) degrade most severely, a contrast that offers guidance to practitioners when selecting evaluation metrics.

- **Practical configuration guidelines.** We derive dataset-adaptive upper limits and provide concrete recommendations for practitioners: t-SNE with medium interval noise attains the best utility–privacy balance, whereas UMAP with high interval noise is preferable when maximum privacy is required at the cost of ranking preservation.

References

- Bock, H. H. (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin, Heidelberg: Springer-Verlag.
- Dwork, C. and A. Roth (2014). *The Algorithmic Foundations of Differential Privacy*. Foundations and trends in theoretical computer science. Now.
- McInnes, L., J. Healy, N. Saul, and L. Großberger (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software* 3(29), 861.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.
- van der Maaten, L. and G. Hinton (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(86), 2579–2605.

Biplots in Symbolic Data Analysis

Sugnet Lubbe¹, Paula Brito^{2,3*}

1. MuViSU (Centre for Multi-dimensional Data Visualisation), Department of Statistics and Actuarial Science, Stellenbosch University, South Africa

2. School of Economics and Management, University of Porto, Porto, Portugal

3. LIAAD – INESC TEC, Porto, Portugal

*Contact author: mpbrito@fep.up.pt

Keywords: Principal Component Analysis, Biplots, Interval-Valued Data, Histogram-Valued Data.

Principal Component Analysis (PCA) biplots (see, e.g. [Gower et al. \(2011\)](#)) provide a powerful and elegant framework for the low-dimensional visualisation of multivariate data $X_{n \times p}$. Using the singular value decomposition, the data matrix is approximated by a low-rank representation that is optimal in the least-squares sense, thereby capturing the main patterns of variation in the data. In a PCA biplot, the units are represented as points, typically in a two-dimensional space while the variables may be represented as calibrated axes. This dual representation provides a rich graphical summary, allowing relationships among observations and variables to be interpreted simultaneously.

We extend the classical PCA biplot paradigm to the broader framework of Symbolic Data Analysis (SDA) allowing for interval and histogram-valued data, which represent intrinsic variability within each unit. The observations are represented by the corresponding quantile functions, assuming a uniform distribution within each interval or histogram bin. In this work, we focus on the L_2 Wasserstein distance and build on the descriptive framework of [Irpino and Verde \(2015\)](#). We construct a matrix of quantile functions and develop a decomposition analogous to the singular value decomposition of a conventional data matrix. Based on this decomposition, we define linear combinations of quantile functions. These combinations play a role directly comparable to that of linear combinations of variables in classical PCA, but they operate in a space of distributions rather than on scalar observations. Consequently, when units are projected onto the plane spanned by the first two principal components, their positions reflect variation not only in location but also in the spread and shape of the underlying distributions.

Rather than being represented by a single point, each unit is visualised as a two-dimensional density. This representation captures the distributional nature of the data and conveys information about both variability and shape, providing a richer and more informative visualisation than traditional point-based displays. In parallel with classical PCA biplots, we construct p biplot axes corresponding to the original distributional variables. These axes are calibrated in the original units of measurement, thereby preserving interpretability and maintaining continuity with classical PCA biplots. The resulting display offers a dual perspective: it enables the simultaneous examination of the relative positions and distributional characteristics of the units, together with the contributions and interpretations of the variables.

The proposed methodology is implemented in the R package *DDbiplotEZ*, available at <https://github.com/sugnet/DDbiplotEZ>. It extends the package *biplotEZ*, published on CRAN [Lubbe et al. \(2025\)](#), and provides tools for constructing distributional data objects and visualising them using distributional PCA biplots. Illustrative examples demonstrate the practical usefulness

of the approach and show how it can reveal structure in complex data.

Acknowledgements

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (<https://doi.org/10.54499/UID/50014/2025>).

References

- Gower, J. C., S. Lubbe, and N. J. Le Roux (2011). *Understanding Biplots*. John Wiley & Sons.
- Irpino, A. and R. Verde (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification* 9(2), 143–175.
- Lubbe, S., N. J. le Roux, J. Nienkemper-Swanepoel, R. Ganey, R. Buys, Z. Adams, and P. Manefeldt (2025). biplotEZ: EZ-to-Use Biplots, R package version 2.3.

Session V

Robust Statistics

The L1-Wasserstein Distance for Interval-Valued Data: Barycenter Computation and Robust Clustering

M. Rosário Oliveira^{1,2,*}, Sara Suarez-Fernandez³, Diogo Pinheiro², Adán Rodríguez-Martínez³, Agustina Bouchet³

1. Department of Mathematics, Instituto Superior Técnico, Lisbon, Portugal

2. CEMAT, Instituto Superior Técnico, Lisbon, Portugal

3. Department of Statistics and O.R. and M.T., University of Oviedo, Oviedo, Spain

*Contact author: rosario.oliveira@tecnico.ulisboa.pt

Keywords: Interval-Valued Data, L1 Wasserstein Distance, Mallows Distance, Robustness, Clustering

Interval-valued data preserve the internal variability of an observation that is otherwise lost when a phenomenon is summarized by a single value, making them a central object of study in Symbolic Data Analysis. While the Mallows (L2-Wasserstein) distance is the standard metric for comparing interval-valued observations by modelling their underlying microdata distributions (for details see [Oliveira et al. \(2022\)](#)), the L1 counterpart remains comparatively unexplored, despite the well-known robustness advantages of L1-type criteria over their squared, L2 analogues. In this work, we study the L1-Wasserstein distance for interval-valued data under a latent-variable model linking each interval to its underlying microdata distribution ([Oliveira et al., 2024](#)). We derive an explicit, closed-form expression for the distance in terms of the centres and ranges of the intervals and of the assumed latent distribution.

Building on this distance, we characterize its associated barycenter (the natural L1 analogue of the median for interval-valued data) under a continuous uniform latent distribution. Leveraging both the distance and its barycenter, we propose a dynamic clustering algorithm for interval-valued data, and prove that it guarantees monotone convergence to a local optimum of the associated adequacy criterion. We benchmark the proposed approach against Partitioning Around Medoids (PAM) ([Kaufman and Rousseeuw, 1990](#)) using both the L1-Wasserstein and Mallows' distances, as well as k-means variants, across synthetic scenarios in which cluster structure is driven by interval centres, ranges, or both. The L1-Wasserstein-based methods consistently match or outperform their L2 counterparts, and are markedly more robust in several scenarios. These robustness properties, established through simulation under both contaminated and uncontaminated settings, are further validated on real-world data, whose complexity extends beyond what simulated scenarios can capture.

References

- Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.
- Oliveira, M. R., M. Azeitona, A. Pacheco, and R. Valadas (2022). Association measures for interval variables. *Advances in Data Analysis and Classification* 16(3), 491–520.
- Oliveira, M. R., D. Pinheiro, and L. Oliveira (2024). Location and association measures for interval-valued data based on Mallows' distance. *Advances in Data Analysis and Classification*, arXiv:2407.05105.

AIDA: An R Package for the Analysis of Interval Data

Catarina P. Loureiro^{1,2,*}, M. Rosário Oliveira^{1,2}, Paula Brito^{3,4}, Lina Oliveira^{2,5}

1. CEMAT, Instituto Superior Técnico, Lisbon, Portugal
2. Departamento de Matemática, Instituto Superior Técnico, Lisbon, Portugal
3. LIAAD-INESC TEC
4. Faculdade de Economia e Gestão, Universidade do Porto, Porto, Portugal
5. CAMGSD, Instituto Superior Técnico, Lisbon, Portugal

*Contact author: catarinapadrela@tecnico.ulisboa.pt

Keywords: Mallows' Distance, Interval Minimum Covariance Determinant Estimator, Robust Interval-Mahalanobis Distance, Explainable Outlier Detection, Shapley Value, Visualization

AIDA (Loureiro, 2026) is an *R* package that provides tools for the analysis of interval-valued data, including construction, visualization, parameter estimation, robust estimation, and explainable outlier detection. The package is built around the `intData` class for representing interval-valued data while accounting for the model linking the microdata and macrodata (Oliveira et al., 2022). This class is complemented by functions for aggregating microdata and estimating parameters of its underlying latent distribution. Package functions also allow estimation of the sample barycenter and symbolic covariance matrix based on the Mallows distance (Oliveira et al., 2025).

AIDA additionally provides robust tools for settings in which interval-valued datasets contain outlying observations. In particular, it implements the Interval Minimum Covariance Determinant estimator (Loureiro et al., 2026b), which provides robust estimates of the barycenter and symbolic covariance matrix. These estimates support an outlier detection rule based on the squared robust interval-Mahalanobis distance (Loureiro et al., 2026b). Explainable outlier detection is further supported through a Shapley value-based decomposition of this distance, allowing assessment of the contributions of individual variables to an observation's outlyingness (Loureiro et al., 2026a).

The *R* package includes visualization tools for representing interval-valued data and aiding in the interpretation of outlier detection and Shapley value-based explainability results. These visualization tools can also be used in a broader context, including for conventional Shapley value-based outlier explanations obtained with the *R* package `ShapleyOutlier` (Mayrhofer and Filzmoser, 2022). In addition, a collection of interval-valued datasets is provided with the package. AIDA is, therefore, designed to assist both methodological research and applied workflows involving interval-valued data, with particular emphasis on robust estimation, outlier detection, and their interpretation.

Acknowledgments: This work was supported by FCT — Fundação para a Ciência e Tecnologia, I.P. and the Recovery and Resilience Plan through the grant [UI/BD/153720/2021](#), and through the projects [UID/04621/2025](#), [UID/50014/2025](#), [UID/04459/2025](#), and [UID/PRR/04459/2025](#).

References

- Loureiro, C. P. (2026). *AIDA: Analysis of Interval Data*. R package version 0.2.1. <https://cran.r-project.org/package=AIDA>.
- Loureiro, C. P., M. R. Oliveira, P. Brito, and L. Oliveira (2026a). Explainable Outlier Detection for Interval-valued Data. [arXiv:2606.26307](#).

- Loureiro, C. P., M. R. Oliveira, P. Brito, and L. Oliveira (2026b). Minimum Covariance Determinant Estimator and Outlier Detection for Interval-valued Data. [arXiv:2604.26769](https://arxiv.org/abs/2604.26769).
- Mayrhofer, M. and P. Filzmoser (2022). *ShapleyOutlier: Multivariate Outlier Explanations using Shapley Values and Mahalanobis Distances*. R package version 0.1.2. <https://CRAN.R-project.org/package=ShapleyOutlier>.
- Oliveira, M. R., M. Azeitona, A. Pacheco, and R. Valadas (2022, Sep). Association measures for interval variables. *Advances in Data Analysis and Classification* 16(3), 491–520.
- Oliveira, M. R., D. Pinheiro, and L. Oliveira (2025). Location and association measures for interval data based on Mallows’ distance. [arXiv:2407.05105](https://arxiv.org/abs/2407.05105).

Robust Interval Distributional regression model

Sónia Dias^{1,*}, Eufrásio de A. Lima Neto², Paula Brito³

1. ESTG - Polytechnic University of Viana do Castelo & LIAAD INESC-TEC, Portugal

2. Federal University of Paraíba, Brazil

3. FEP School of Economics and Management, University of Porto & LIAAD - INESC TEC, Porto, Portugal

*Contact author: sdias@estg.ipv.pt

Keywords: Robust Linear Regression Model, Interval Data, Quantil Function, Gaussian Kernels

Over the years, several linear regression models have been developed within the context of Symbolic Data Analysis to deal with symbolic variables, namely interval-valued variables and, more recently, histogram-valued variables. In this context, the Interval Distributional (ID) regression model, proposed by Dias and Brito (2017), constitutes an innovative approach, as it moves beyond representing intervals exclusively through their centers and half ranges, instead considering interval-valued information globally. Thus, the model allows each interval to be treated as a whole, while preserving the information concerning its internal distribution.

For an interval-valued variable X , the observation for each unit i is an interval $I_{X(i)}$ which may be represented by its center, $c_{X(i)}$ and half-range, $r_{X(i)}$, as $I_{X(i)} = [c_{X(i)} - r_{X(i)}, c_{X(i)} + r_{X(i)}]$. To overcome the difficulties inherent to performing arithmetic operations on intervals, we represent them by the corresponding quantile functions $\Psi_{X(i)}(t)$ that is, the inverse of their cumulative distribution functions, under specific assumptions of the within-interval distribution.

Since the space of the quantile functions is only a semi-vector space (the product of a quantile function by a negative real number is not a non-decreasing function), the Interval Distributional (ID) regression model, proposed by Dias and Brito (2017) includes, for each explicative variable, both the quantile function $\Psi_{X_j(i)}(t)$, which represents the observation of the interval-valued variable X_j and $-\Psi_{X_j(i)}(1-t)$, which represents the respective symmetric interval.

For the response variable Y , each quantile function $\Psi_{Y(i)}$ for each unit i may be expressed as $\Psi_{Y(i)}(t) = \Psi_{\hat{Y}(i)}(t) + e_i(t)$ where $\Psi_{\hat{Y}(i)}(t)$ is the predicted quantile function for unit i , given by

$$\Psi_{\hat{Y}(i)}(t) = v + \sum_{j=1}^p a_j \Psi_{X_j(i)}(t) - \sum_{j=1}^p b_j \Psi_{X_j(i)}(1-t) \quad (7)$$

with $t \in [0, 1]$; $a_j, b_j \geq 0$, $j \in \{1, 2, \dots, p\}$, and $v \in \mathbb{R}$.

When a Uniform distribution is assumed within the observed intervals, the corresponding quantile functions that represents the observations of the interval-valued variable X_j are linear and non-decreasing with domain $[0, 1]$, given by $\Psi_{X_j(i)}(t) = c_{X_j(i)} + (2t - 1)r_{X_j(i)}$. Consequently, the respective symmetric interval is represented by $-\Psi_{X_j(i)}(1-t) = -c_{X_j(i)} + (2t - 1)r_{X_j(i)}$, $0 \leq t \leq 1$.

Therefore, the predicted quantile function $\Psi_{\hat{Y}(i)}$ in expression (7) can be rewritten as

$$\Psi_{\hat{Y}(i)}(t) = \sum_{j=1}^p (a_j - b_j) c_{X_j(i)} + v + \sum_{j=1}^p (a_j + b_j) r_{X_j(i)} (2t - 1) \quad (8)$$

with $t \in [0, 1]$; $a_j, b_j \geq 0$, $j \in \{1, 2, \dots, p\}$, and $v \in \mathbb{R}$.

Other distributions to model the interval-valued observations have also been considered, particularly the Symmetric Triangular Distribution (Dias and Brito, 2017).

The non-negative parameters of the ID model are estimated by solving a quadratic optimization problem subject to non-negativity constraints on the unknown parameters. The Mallows distance is used to quantify the dissimilarity between the predicted and the observed quantile functions. This distance also provides the basis for defining a goodness-of-fit measure for the ID model.

The ID model can be regarded as a generalization of the classical linear regression model since, when all observations consist of degenerate intervals (intervals with null range), the ID model reduces to the classical linear regression model. Furthermore, the approach can be extended to handle histogram-valued variables Dias and Brito (2015), making it a general framework applicable to different types of variables.

The presence of outliers is, however, a common problem in regression models, as atypical observations may exert influence on parameter estimation and, consequently, impair the predictive performance of the model. It is therefore essential to develop a robust version of the ID regression model capable of limiting the influence of such observations and producing more stable estimates in the presence of contaminated data.

In this work, we propose the Robust Interval Distributional Regression (RiDR) model, which incorporates robustness principles from the robust regression model Exponential-type Kernel Robust Regression, proposed by Lima Neto and de Carvalho (2018). The RiDR is based on an objective function that employs Gaussian kernels to quantify the similarity between observed and predicted values. Through an iterative reweighing procedure, the model identifies potentially atypical observations and progressively reduces their influence on parameter estimation by assigning them weights close to zero. The proposed approach therefore limits the impact of outliers on parameter estimation, contributing to greater stability of the estimates and increased robustness of interval predictions.

To assess the behavior and performance of the proposed model, simulation studies have been conducted under different scenarios and results compared with those obtained from other robust regression models. Real-data applications allow analyzing and comparing the performance of the Interval Distributional regression model in its non-robust and robust versions. This comparison makes it possible to assess the ability of the RiDR model to reduce the influence of outliers and determine the extent to which the incorporation of the robustness procedure leads to improvements in estimation and predictive performance in interval regression.

Acknowledgements: This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (<https://doi.org/10.54499/UID/50014/2025>).

References

- Dias, S. and P. Brito (2015). Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8 (2), 75–113.
- Dias, S. and P. Brito (2017). Off the beaten track: A new linear model for interval data. *European Journal of Operational Research* 258 (3), 1118–1130.
- Lima Neto, E. A. and F. A. de Carvalho (2018). An exponential-type kernel robust regression model for interval-valued variables. *Information Sciences* 454-455, 419–442.

Robust Principal Component Analysis of Symbolic Matrix-valued Data

Marcus Mayrhofer¹, A. Pedro Duarte Silva^{2,*}, Paula Brito³, Peter Filzmoser¹

1. Institute of Statistics and Mathematical Methods in Economics, TU Wien, Austria

2. Universidade Católica Portuguesa, Católica Porto School of Business and Economics & CEGE, Portugal

3. Fac. Economia e Gestão, Universidade do Porto & LIAAD-INESC TEC, Portugal

*Contact author: psilva@ucp.pt

Keywords: Histogram Data, Matrix-Valued Data, Robust Covariance Estimation.

We consider numerical distributional data, where for each unit $S_i, i = 1, \dots, n$ and each variable $Y_j, j = 1, \dots, p$, a distribution is recorded, representing the corresponding observed variability. In our model, each distribution is represented by a set of indicators: a central statistic C , and the logarithm transformation of inter-quantile ranges, for a chosen set of quantiles $\phi_1, \dots, \phi_q$, denoted by $R_h^*, h = 1, \dots, m$, where $m = q - 1$ is the number of considered intervals. Typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles Brito and Duarte Silva (2025); interval-valued data are represented by midpoints and log-ranges ($m = 1$) Brito and Duarte Silva (2012). Furthermore, we consider alternative structures of the variance-covariance matrix. In the most general formulation we allow for non-zero correlations among all central statistics and log-ranges; for distributional variables there are however other cases of interest, whether the variables, the central statistics, and the different log-ranges, are or are not correlated between or among themselves, leading to five different configurations in the distributional data case, and four configurations for interval-valued data. In this work we consider these data as matrix-valued data, represented as a tensor of dimension $n \times p \times q$, following Mayrhofer et al. (2025), $X \sim ME(M, \Sigma_{var}, \Sigma_{ind}, g)$, where

- Σ_{var} is $p \times p$ and gathers variances and covariances between the variables Y_j
- Σ_{ind} is $q \times q$ and gathers variances and covariances between the considered $q = m + 1$ indicators, $C, R_1^*, \dots, R_m^*$
- $g(z) = \frac{\exp(-z/2)}{2\pi^{pq/2}}$

In matrix-variate elliptical distributions, $X \sim ME(M, \Sigma_{row}, \Sigma_{col}, g)$ means that $X \in \mathbb{R}^{p \times q}$ is a random matrix with density

$$|\Sigma_{row}|^{-q/2} |\Sigma_{col}|^{-p/2} g(\text{tr} [\Sigma_{row}^{-1} (X - M) \Sigma_{col}^{-1} (X - M)^t])$$

where M specifies the distribution center, Σ_{row} and Σ_{col} describe the variances and covariances amongst rows and columns, and $g()$ is an univariate function that controls the shape of the distribution in its tails. In our work, we consider the particular case of the Gaussian matrix-variate distribution with variables (*var*) and indicators (*ind*) used respectively as rows (*row*) and columns (*col*) of the matrix X .

In this model, the global covariance matrix Σ is written as $\Sigma = \Sigma_{ind} \otimes \Sigma_{var}$. This implies a more parsimonious model that reduces the number of parameters to be estimated. The different covariance configurations correspond to setting Σ_{ind} and/or Σ_{var} as block-diagonal matrices. The Matrix Minimum Covariance Determinant (MMCD) method Mayrhofer et al. (2025) accounts for

the matrix-variate data structure and robustly estimates the mean matrix, as well as the row-wise Σ_{var} and column-wise Σ_{ind} covariance matrices.

Given n samples $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^{p \times q}$, let $\hat{\Sigma}_{var} \in \mathbb{R}^{p \times p}$ and $\hat{\Sigma}_{ind} \in \mathbb{R}^{q \times q}$ denote mode-wise estimators of the row and column covariance matrices, obtained via either the matrix maximum likelihood estimator [Dutilleul \(1999\)](#) or the matrix minimum covariance determinant method [Mayrhofer et al. \(2025\)](#). PCA is then performed independently on each estimator via eigendecomposition,

$$\hat{\Sigma}_{var} \mathbf{U} = \mathbf{U} \mathbf{\Lambda}_{var}, \quad \hat{\Sigma}_{ind} \mathbf{V} = \mathbf{V} \mathbf{\Lambda}_{ind}, \quad (9)$$

yielding row-space loadings $\mathbf{U} \in \mathbb{R}^{p \times k}$ and column-space loadings $\mathbf{V} \in \mathbb{R}^{q \times \ell}$. The matrix PCA scores are then

$$\mathbf{Z}_i = \mathbf{U}^T (\mathbf{X}_i - \hat{\mathbf{M}}) \mathbf{V} \in \mathbb{R}^{k \times \ell}, \quad (10)$$

where $\hat{\mathbf{M}} \in \mathbb{R}^{p \times q}$ is an estimate of the mean matrix, providing a bilinear low-rank representation that captures variance along both matrix dimensions.

This framework was applied to a sample of 1544 distributional data units of Spotify track data, collected from the Spotify's public API, and aggregated by artist. The original microdata consisted of 170663 track observations on 19 music attributes. For the purpose of this analysis we have considered nine numerical attributes (valence, acousticness, danceability, duration, energy, liveness, loudness, speechiness, and tempo) and nine distributional indicators (the median, and the log-ranges of the intervals defined by the *min*, *max* and the $q_{0.125}$, $q_{0.25}$, $q_{0.375}$, $q_{0.5}$, $q_{0.75}$, and $q_{0.875}$ quantiles) for each distributional data unit. The results lead to low-rank representations with interpretable and interesting components for both matrix dimensions. Furthermore, the comparison between the PCA results based on maximum likelihood and MMCD covariance estimates highlighted the impact of unit outliers, and the importance of robustly estimating covariance matrices in symbolic matrix-valued data.

Acknowledgments

Financial support from the Fundação para a Ciência e Tecnologia through project UID/00731/2025: Research Centre in Management and Economics (DOI <https://doi.org/10.54499/UID/00731/2025>), and project UID/50014/2025 (<https://doi.org/10.54499/UID/50014/2025>) is gratefully acknowledged.

References

- Brito, P. and A. P. Duarte Silva (2012). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics* 39(1), 3–20.
- Brito, P. and A. P. Duarte Silva (2025). Parametric models for distributional data. *Advances in Data Analysis and Classification* 19(4), 1119–1146.
- Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* 64(2), 105–123.
- Mayrhofer, M., U. Radojićić, and P. Filzmoser (2025). Robust covariance estimation and explainable outlier detection for matrix-valued data. *Technometrics* 67(3), 516–530.

Session VI

Networks

Monte Carlo Directed Interval Louvain (MC-DIL): Consensus Community Detection in Directed Networks with Interval-Valued Weights

Hélder Alves^{1,*}, Paula Brito², Pedro Campos²

¹ ISSSP, Instituto Superior de Serviço Social do Porto & LIAAD-INESC TEC, Porto, Portugal

² FEP, Faculdade de Economia e Gestão, Universidade do Porto & LIAAD-INESC TEC, Porto, Portugal

*Contact author: helder.alves@issssp.pt

Keywords: Community Detection, Directed Networks, Interval-Valued Networks, Louvain Algorithm, Monte Carlo Simulation.

Community detection in weighted networks typically represents edge weights as fixed values, without accounting for their variability.. However, some empirical networks are more naturally represented by interval-valued weights, particularly when aggregation preserves the range of values observed at a finer level of analysis (Alves et al., 2023). In a Symbolic Data Analysis setting, such intervals preserve within-entity variability that would be lost by replacing them with a single representative value, such as the midpoint. We introduce **Monte Carlo Directed Interval Louvain (MC-DIL)**, a simulation-based framework that propagates interval-valued edge variability through directed modularity optimisation and summarises the resulting partition variability through pairwise coassociation, consensus communities, and vertex-level stability.

Method

Let $L = (L_{ij})$ and $U = (U_{ij})$ denote the lower- and upper-bound adjacency matrices of a directed interval-weighted network, where $L_{ij}, U_{ij} \in \mathbb{Z}_{\geq 0}$ and $L_{ij} \leq U_{ij}$. Thus, each edge weight is represented by an inclusive interval of non-negative integers $[L_{ij}, U_{ij}]$. At Monte Carlo replication $s = 1, \dots, S$, each edge weight is sampled independently from the discrete uniform distribution over its interval as

$$A_{ij}^{(s)} \sim \text{DU}\{L_{ij}, \dots, U_{ij}\}, \quad \Pr(A_{ij}^{(s)} = a) = \frac{1}{U_{ij} - L_{ij} + 1}.$$

Equiprobability within intervals and independence across edges are modelling assumptions rather than properties of interval-valued data. Each sampled matrix defines a point-valued directed weighted network whose community structure is evaluated using resolution-adjusted directed modularity (Leicht and Newman, 2008),

$$Q(\mathbf{c}; A^{(s)}, \gamma) = \frac{1}{m^{(s)}} \sum_{i,j} \left[A_{ij}^{(s)} - \gamma \frac{k_i^{\text{out},(s)} k_j^{\text{in},(s)}}{m^{(s)}} \right] \delta(c_i, c_j).$$

Here, $m^{(s)}$ is the total edge weight, $k_i^{\text{out},(s)}$ and $k_j^{\text{in},(s)}$ are the out- and in-strengths, and $\delta(c_i, c_j)$ indicates community co-membership. For each sampled network, the directed Louvain heuristic (Blondel et al., 2008) is run R times with independent random vertex-order initialisations, and the partition attaining the highest final directed modularity is retained. Since community labels are arbitrary across Monte Carlo replications, the S retained partitions are summarised by the pairwise coassociation matrix

$$\hat{p}_{ij} = \frac{1}{S} \sum_{s=1}^S \mathbf{1}(c_i^{(s)} = c_j^{(s)}).$$

Let $K^{(s)}$ denote the number of communities in the retained partition s , and let $K^* = \text{mode}\{K^{(1)}, \dots, K^{(S)}\}$. The coassociation matrix is converted into the dissimilarity matrix $d_{ij} = 1 - \hat{p}_{ij}$, to which average-linkage hierarchical clustering is applied. Cutting the resulting dendrogram into K^* clusters yields the consensus partition, following the general rationale of consensus clustering (Lancichinetti and Fortunato, 2012). For

vertex i , internal stability I_i is its mean coassociation with its consensus community, while external ambiguity E_i is the largest mean coassociation with an alternative community. The stability margin $M_i = I_i - E_i$ measures separation from the strongest alternative.

Empirical application and results

MC-DIL was applied to municipality-level directed mobility flows for mainland Portugal. After retaining municipality-level edges with weights of at least 20, municipalities were grouped by NUTS3 region. For each directed pair of distinct regions $i \rightarrow j$, $[L_{ij}, U_{ij}]$ was the minimum–maximum range of the retained municipality-level edge weights, preserving within-pair variability rather than replacing it by an aggregated point value. This produced a directed interval-weighted network with 24 NUTS3 regions and 187 directed edges; self-loops were fixed at $[0, 0]$.

MC-DIL used $S = 100$ Monte Carlo replications, $R = 5$ random vertex-order Louvain runs per sampled network, $\gamma = 1$, and discrete-uniform edge-weight sampling, for a total of 500 Louvain optimisations. The retained partitions contained three to five communities: $K = 4$ occurred in 67% of replications, compared with $K = 3$ in 27% and $K = 5$ in 6%. Accordingly, $K^* = 4$, and the consensus partition comprised four communities containing 9, 5, 6, and 4 NUTS3 regions, respectively. All 24 regions had positive stability margins ($M_i = 0.268\text{--}0.813$), so coassociation with the assigned consensus community exceeded that with the strongest alternative. PT1C3 (Alto Alentejo) had the smallest margin ($M_i = 0.268$) and thus the least clearly separated assignment.

Interpretation

Let $T_R(A, \Omega; \gamma)$ denote the highest-modularity partition retained from R stochastic Louvain runs under optimisation randomness Ω . The protocol-conditional pairwise estimand is

$$p_{ij}^{(R, \gamma)} = \Pr[T_R(A, \Omega; \gamma)_i = T_R(A, \Omega; \gamma)_j \mid L, U],$$

and the coassociation estimate $\hat{p}_{ij}$ is its Monte Carlo estimator. Thus, $\hat{p}_{ij}$ is not a distribution-free probability of membership in a uniquely defined “true” community: changing the within-interval sampling model, R , γ , or the Louvain protocol changes the estimand. MC-DIL combines interval-based edge variability with Louvain stochasticity; multiple runs reduce, but do not eliminate, algorithmic sensitivity.

Rather than collapsing each interval to a single value, MC-DIL propagates interval-valued edge variability through directed community detection and characterises persistence and ambiguity across sampled networks and optimisation paths. The Portuguese NUTS3 application illustrates that a dominant four-community structure can coexist with variation in community counts and heterogeneous vertex-level stability. Applications should assess Monte Carlo stability and sensitivity to S , R , γ , random seeds, and alternative within-interval sampling models.

Acknowledgements

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (<https://doi.org/10.54499/UID/50014/2025>).

References

- Alves, H., P. Brito, and P. Campos (2023). Community detection in interval-weighted networks. *Data Mining and Knowledge Discovery* 38(2), 653–698.
- Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10), P10008.
- Lancichinetti, A. and S. Fortunato (2012). Consensus clustering in complex networks. *Scientific Reports* 2, 336.
- Leicht, E. A. and M. E. J. Newman (2008). Community structure in directed networks. *Physical Review Letters* 100(11), 118703.

Event Sequences and Networks:

Part 1. Relational Events

Vladimir Batagelj^{1,2,3,*}

1. IMFM Ljubljana

2. University of Primorska IAM Koper

3. University of Ljubljana FMF Ljubljana

*Contact author: vladimir.batagelj@fmf.uni-lj.si

ORCID: 0000-0002-0240-9446

Keywords: Event Sequence, Network, Distribution, Dissimilarity.

Traditional data analysis uses single numbers or distinct categories. SDA handles aggregated or variable data that contains internal variation and/or structure.

Sequence analysis is a subfield of longitudinal data analysis closely related to time series analysis, but with an emphasis on **categorical** values represented as states. There is [The Sequence Analysis Association \(SAA\)](#). For recent overviews, see [Liao et al. \(2022\)](#); [Yeshchenko and Mendling \(2023\)](#); [van der Linden et al. \(2023\)](#); [Raab and Struffolino \(2022\)](#).

Usually we analyze a finite set of event sequences $\mathbf{S}$. A raw event sequence S is a list of events $S = (e_1, e_2, \dots, e_n)$, n is the length of the sequence S . Each event is a list of named attributes

$$e_i = (ind = i, t = t_i, a_1 = a_{1i}, a_2 = a_{2i}, \dots, a_m = a_{mi})$$

where ind is the event's index, t is the event's temporal information, and a_i s are data attributes describing different event characteristics. Their values can be structured/symbolic objects. From S we usually obtain the set $\mathbf{S} = \{S_x\}$ by partitioning S by selected categorical attribute a : $S_x = \{e_i \in S : a_i = x\}$.

In an event, some attributes can be missing. Each event has at least one data attribute defined (has a value). An attribute a_i is called general if it is defined for all events in sequences from $\mathbf{S}$. The attribute ind and, usually, the time t are general attributes.

The temporal information can be described in different ways

- implicitly – the event's index determines the time. It expresses the ordering of events,
- $t_i \in \mathbb{N}$ – it expresses the ordering of events or starts of events on a discrete timeline (hours, days, years, etc.),
- $t_i = (s_i, f_i)$ – it specifies the time interval of event e_i ; s_i is the start and f_i is the finish time of the event; $s_i \leq f_i$,
- $t_i = (s_i, d_i)$ – it specifies the time interval of event e_i ; s_i is the start and d_i is the duration time of the event; $d_i \geq 0$.

Quantities s_i and f_i are expressed as numbers or as date-time stamps. In some applications, an event's time can be of different types: interval (both values are defined) and instantaneous (the second value is undefined). The time can be measured on a common timeline or a relative timeline (for example, from each person's birthday). Often in our analyses, we do not need all available data attributes, and some attributes need to be recoded – we transform the raw event sequence S into a work event sequence S' . Usually, we also have some additional data about the (values of) attributes used in the event description. For example: address, age, gender, etc. of persons. Several event sequence datasets are available ([Batagelj, 2026](#)).

Sequence analysis deals with questions about the set of event sequences $\mathbf{S}$

- structure of the set S (statistics, clustering, visualization)
- models that generate sequences from the set S (transition matrices, Petri nets, Markov chains)
- frequent patterns – motifs (subsequences) in sequences from S
- networks derived from S
- etc.

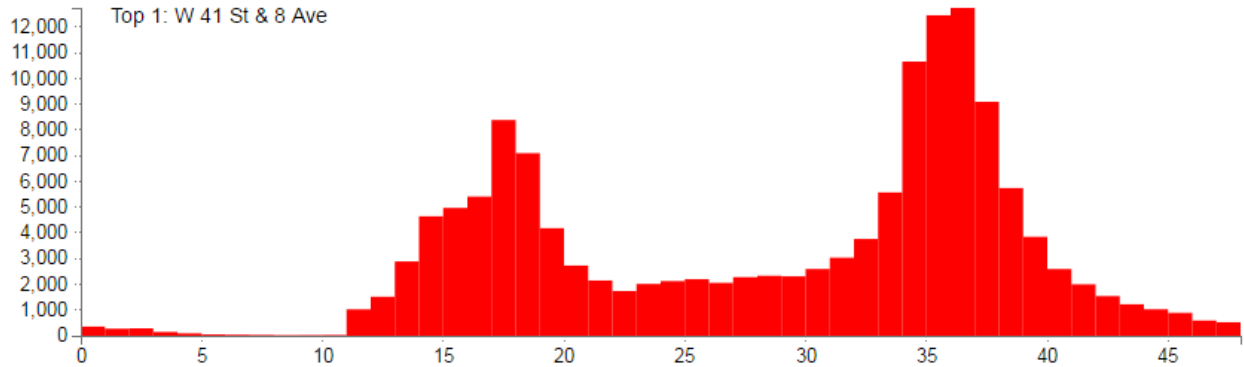


Figure 1: Distribution of departures on the most active link in the NY Citi Bike network Oct 2015–Sep 2016.

In this presentation, we will focus on using networks to analyze sets of event sequences. In some cases, events are relational – two general categorical attributes define the nodes of the network. For example, in bike rental events, there is a rental/start station and a return station. Aggregated values of parallel links can be described by symbolic objects. Sometimes, however, we can take sequences of events from S as nodes. Nodes are linked if the sequences are active at least for some time simultaneously (Oštro, 2024).

References

- Batagelj, V. (2026). Event sequences. <https://github.com/bavla/TQ/blob/master/events/README.md>.
- Liao, T. F., D. Bolano, C. Brzinsky-Fay, B. Cornwell, A. E. Fasang, S. Helske, R. Piccarreta, M. Raab, G. Ritschard, E. Struffolino, et al. (2022). Sequence analysis: Its past, present, and future. *Social science research* 107, 102772.
- Oštro (2024). Zvezoskop – an interactive visualization of career paths of active political officials. <https://ostro.si/tag/zvezoskop/>.
- Raab, M. and E. Struffolino (2022). *Sequence Analysis*, Volume 190. SAGE Publications.
- van der Linden, S., E. de Fouw, S. van den Elzen, and A. Vilanova (2023). A survey of visualization techniques for comparing event sequences. *Computers & Graphics*.
- Yeshchenko, A. and J. Mendling (2023). A survey of approaches for event sequence analysis and visualization. *Information Systems*, 102283.

Event Sequences and Networks:

Part 2. Co-Presence Networks

Anuška Ferligoj^{1,*}, Vladimir Batagelj^{2,3,4,#}

1. University of Ljubljana FDV Ljubljana

2. IMFM Ljubljana

3. University of Primorska IAM Koper

4. University of Ljubljana FMF Ljubljana

*Contact author: anuska.ferligoj@fdv.uni-lj.si

Keywords: Event Sequence, Network, Co-Presence Network, Clustering.

In this presentation, we will present an approach to transform event sequences into a network, based on the simultaneous activity of pairs of event sequences. We will demonstrate it on data from the biographies of Slovenian politicians.

A person X 's CV, $CV_X = (e_1, e_2, \dots, e_{k_x})$, consists of a sequence of events e_i . For an event $e_i = (s_i, f_i, R_i, S_i, \dots)$ we at least know its start date s_i , its end (finish) date f_i , the type R_i of the event, the state (location) S_i of the event, and maybe something more. We decided to base our analysis on the corresponding co-presence network – a weighted multi-relational temporal network $N = (V, L, w, t)$ in which the set of *nodes* V consists of studied persons. There is a *link* (edge) $\ell = (u : v; R)$ of type (relation) R between persons $u, v \in V$ iff there exist events $e_u = (s_u, f_u, R_u, S_u)$ and $e_v = (s_v, f_v, R_v, S_v)$ such that $R_u = R_v = R$ and $S_u = S_v$ and $[s, f] = [s_u, f_u] \cap [s_v, f_v] \neq \emptyset$. Parallel edges of type R can appear. They are replaced by a single edge with a weight described by a symbolic value.

The traditional sequence analysis deals mainly with the analysis of states (Ritschard and Studer, 2018). The co-presence network enables us to analyze groups of persons using network analysis methods such as cuts, cores, islands, network clustering, and blockmodeling (Batagelj et al., 2014). The proposed approaches will be illustrated on the dataset Zvezoskop of Slovenian politicians (Oštro, 2024). Figure 1 shows the co-presence network until 2022.

References

- Batagelj, V., P. Doreian, A. Ferligoj, and N. Kejžar (2014). *Understanding Large Temporal Networks and Spatial Networks*, Volume 2. John Wiley & Sons.
- Oštro (2024). Zvezoskop – an interactive visualization of career paths of active political officials. <https://ostro.si/tag/zvezoskop/>.
- Ritschard, G. and M. Studer (2018). *Sequence Analysis and Related Approaches: Innovative Methods and Applications*. Springer Nature.

Session VII

Time Series

Analysis of the Sea Level along the Portuguese Coast: A Symbolic Time Series Approach

Maria Eduarda Borges^{1*}, Paula Brito², Paulo Teles²

1. Department of Mathematics, Faculty of Sciences, University of Porto, Portugal

2. FEP School of Economics and Management, University of Porto & LIAAD - INESC TEC, Porto, Portugal

*Contact author: eduarda.varajao.borges@gmail.com

Keywords: Histogram Time Series, Interval Time Series, Portuguese Coast, Symbolic Time Series, Symbolic Time Series Clustering.

The work focuses on the analysis of sea-level dynamics along the Portuguese coast using data recorded by tide-gauge stations at a one-minute sampling frequency. Sea level exhibits pronounced intraday variability driven primarily by the tidal cycle, making it a natural candidate for symbolic representation. In this context, each day can be summarised either by the interval between its minimum and maximum recorded sea levels or by the full distribution of its minute-by-minute observations, represented as a daily histogram.

The Portuguese Hydrographic Institute (HI) provided sea-level observation datasets recorded at 21 tide-gauge stations in Portugal, covering the period from 2015/01/01 to 2025/06/14. The stations considered were Angra do Heroísmo, Aveiro, Base Naval de Lisboa, Cantareira, Faro, Figueira da Foz, Freixo, Horta, Leixões, Lisboa, Nazaré, Peniche, Ponta Delgada, Sesimbra, Setúbal Tróia, Sines, Terminal XXI, Viana do Castelo, Vila do Porto, and Vila Real de Santo António.¹

Within the theoretical framework of Symbolic Data Analysis (SDA) (Brito (2014, 2016)), extended to symbolic time series, three distance-based clustering methodologies were adapted for Interval Time Series (ITS) and Histogram Time Series (HTS). For ITS, a uniform distribution within each daily interval was assumed, whereas for HTS, a uniform distribution within each histogram sub-interval was considered. The three approaches are based on different domains: the Mallows distance for the distributional domain (Irpino and Verde (2006); Verde and Irpino (2007, 2008); Dias and Brito (2015)), the interval/histogram autocorrelation function for the time domain (Irpino and Verde (2015); Maharaj et al. (2019)), and an adapted log-normalised periodogram distance for the frequency domain (Caiado et al. (2006)).

Each distance-based clustering method was evaluated using three clustering strategies: non-hierarchical clustering based on k -medoids, and hierarchical clustering using complete and average linkage. The resulting clusters were assessed using the Silhouette Index and the Davies–Bouldin Index.

The proposed methods were implemented in *R* to investigate whether they are able to identify meaningful groupings among the tide-gauge stations, particularly groupings consistent with their geographical location and underlying hydrodynamic characteristics.

The Mallows distance produced the most interpretable and geographically coherent partitions, clearly separating the Azorean stations, subject to an oceanic tidal regime, from Cantareira, while further isolating the riverine and estuarine stations of Base Naval Lisboa and Freixo at finer partitions. The autocorrelation function-based distance consistently grouped Base Naval Lisboa and Freixo with Faro and Horta, suggesting a shared damped or irregular tidal dynamic. For the autocorrelation function method, the choice between HTS and ITS representations mainly affects the boundary stations of Leixões, Setúbal Tróia, and Sines. In contrast, the adapted log-normalised periodogram distance yielded a weaker and less consistent cluster structure across both representations, suggesting that, as computed in this work, it is not a robust clustering criterion for this dataset. It is well known that the periodogram is not a consistent estimator since its

¹Data from the Setúbal Clube Naval station were provided but the station was excluded due to the low number of observations.

variance does not tend to zero as the sample size increases. Furthermore, the periodogram ordinates are independent even for large sample sizes. Consequently, the periodogram is unstable and tends to be very jagged even when the sample size increases. This is why the periodogram is smoothed in order to estimate the spectral density function, but the distance measure uses the unsmoothed periodogram. Moreover, much noise may still be present because of the histograms' daily frequency. Both factors may have severely affected the periodogram, distorting the results (although that did not happen to the sample autocovariance). Perhaps trying the distance measure based on the periodogram ordinates rather than the measure based on the log-normalized periodogram can provide better results. Therefore, comparing both alternatives can be insightful. We leave it for further work in time of the Workshop.

Finally, the cluster partitions obtained for both representations were compared using the Adjusted Rand Index (ARI) (Hubert and Arabie (1985)). The ARI showed that the symbolic representation chosen has a limited, moderate, and strong impact on the Mallows, autocorrelation function-based, and periodogram-based distances, respectively.

Acknowledgements

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (<https://doi.org/10.54499/UID/50014/2025>).

References

- Brito, P. (2014). Symbolic data analysis: another look at the interaction of data mining and statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4(4), 281–295.
- Brito, P. (2016). Clustering of symbolic data. *Handbook of Cluster Analysis*, 469–495.
- Caiado, J., N. Crato, and D. Peña (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis* 50(10), 2668–2684.
- Dias, S. and P. Brito (2015). Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8(2), 75–113.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2(1), 193–218.
- Irpino, A. and R. Verde (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In V. Batagelj, H.-H. Bock, and A. Ferligoj (Eds.), *Data Science and Classification*, pp. 185–192. Springer.
- Irpino, A. and R. Verde (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification* 9(2), 143–175.
- Maharaj, E. A., P. Teles, and P. Brito (2019). Clustering of interval time series. *Statistics and Computing* 29(5), 1011–1034.
- Verde, R. and A. Irpino (2007). Dynamic clustering of histogram data: using the right metric. In *Selected Contributions in Data Analysis and Classification*, pp. 123–134. Springer.
- Verde, R. and A. Irpino (2008). Comparing histogram data using a Mahalanobis-Wasserstein distance. In *COMPSTAT 2008: Proceedings in Computational Statistics*, pp. 77–89. Springer.

Interval Time Series and Spatial Contiguity in the Analysis of Air-Pollution Patterns in Brazil

Luana Lúcia Alves de Azevêdo^{1,2,*}, Paula Brito³, Paulo Teles³, Paulo Canas Rodrigues^{1,4}

1. Graduate Programme in Statistics and Data Science, Federal University of Bahia, Salvador, Brazil

2. Federal University of Rondônia, Ji-Paraná, Brazil

3. LIAAD-INESC TEC & Faculdade de Economia e Gestão, Universidade do Porto, Porto, Portugal

4. Department of Statistics, Federal University of Bahia, Salvador, Brazil

*Contact author: luana.azevedo@unir.br

Keywords: Interval Time Series, Symbolic Data Analysis, Dissimilarity Matrices, Air Pollution

Air pollution is a major environmental risk to human health and is associated with cardiovascular and respiratory diseases, stroke, asthma, and cancer ([World Health Organization, 2021](#)). In Brazil, air pollution results from different sources, including vehicular and industrial emissions, energy generation, and biomass burning, whose influence varies across regions and pollutants ([Andrade et al., 2012](#); [Kawashima et al., 2020](#)). This diversity contributes to heterogeneous concentration patterns across space and time, requiring approaches capable of jointly analysing different pollutants while preserving data variability. Symbolic Data Analysis makes it possible to represent this variability explicitly through intervals (see, e.g., [Bock and Diday, 2000](#); [Billard and Diday, 2003](#); [Brito, 2014](#)), allowing interval time series to describe the evolution of both concentration levels and their ranges.

This study analyses multivariate interval time series of the concentrations of five air pollutants: fine particulate matter (PM_{2.5}), inhalable particulate matter (PM₁₀), nitrogen dioxide (NO₂), ozone (O₃), and sulfur dioxide (SO₂). The database comprises daily municipality-level air-pollution indicators made available through the Air Pollutants Data collection ([Saldanha et al., 2025](#)). Derived from the Copernicus Atmosphere Monitoring Service global reanalysis (CAMS EAC4), the data cover 5,565 Brazilian municipalities between 1 January 2003 and 31 December 2024. For each municipality, date, and pollutant, the daily minimum and maximum concentrations define the lower and upper bounds of the interval, respectively, while the daily mean is retained as a separate descriptive measure. Exploratory analyses indicated spatial heterogeneity, asymmetric distributions, the occurrence of extreme values, and differences among pollutants in the magnitude and temporal evolution of their ranges. These results showed that municipalities with similar mean concentration levels may exhibit different variability patterns, reinforcing the relevance of the interval representation.

Dissimilarity matrices between the municipal time series were subsequently calculated. Following the literature on clustering interval-valued data and interval time series, the interval Euclidean distance, the Hausdorff distance, and the uniform and symmetric triangular Mallows distances were considered ([de Carvalho et al., 2006](#); [Maharaj et al., 2019](#)). For each measure, the dissimilarities evaluated over time were aggregated across the five pollutants.

The resulting dissimilarity matrices provide complementary representations of the differences between municipal trajectories and constitute the basis for spatially constrained clustering. In this type of approach, homogeneity among spatial units is considered jointly with geographical contiguity, so that each group corresponds to a spatially connected and internally homogeneous region ([Duque et al., 2007](#)). The spatial structure is represented by a neighbourhood matrix in which neighbouring municipalities share a territorial boundary.

Comparing the matrices makes it possible to examine how different definitions of interval proximity represent the temporal and spatial heterogeneity of air pollution. By jointly considering the daily concentration intervals of the five pollutants, the analysis expands the information used to compare municipalities and supports the identification of regions with similar multivariate dynamics. The integration of interval time

series, dissimilarity measures, and contiguity constraints thus contributes to an environmentally interpretable regionalisation of air-pollution patterns across Brazil.

Acknowledgements

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (<https://doi.org/10.54499/UID/50014/2025>).

References

- Andrade, M. d. F., R. M. de Miranda, A. Fornaro, A. Kerr, B. Oyama, P. A. de André, and P. Saldiva (2012). Vehicle emissions and PM_{2.5} mass concentrations in six Brazilian cities. *Air Quality, Atmosphere & Health* 5(1), 79–88.
- Billard, L. and E. Diday (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association* 98(462), 470–487.
- Bock, H.-H. and E. Diday (Eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin-Heidelberg: Springer-Verlag.
- Brito, P. (2014). Symbolic data analysis: Another look at the interaction of data mining and statistics. *WIREs Data Mining and Knowledge Discovery* 4(4), 281–295.
- de Carvalho, F. d. A. T., P. Brito, and H.-H. Bock (2006). Dynamic clustering for interval data based on L_2 distance. *Computational Statistics* 21(2), 231–250.
- Duque, J. C., R. Ramos, and J. Suriñach (2007). Supervised regionalization methods: A survey. *International Regional Science Review* 30(3), 195–220.
- Kawashima, A. B., L. D. Martins, S. A. A. Rafee, A. P. Rudke, M. V. de Moraes, and J. A. Martins (2020). Development of a spatialized atmospheric emission inventory for the main industrial sources in Brazil. *Environmental Science and Pollution Research* 27(29), 35941–35951.
- Maharaj, E. A., P. Teles, and P. Brito (2019). Clustering of interval time series. *Statistics and Computing* 29(5), 1011–1034.
- Saldanha, R. d. F., D. R. X. Silva, R. d. S. d. G. G. Carrijo, B. F. A. d. Oliveira, V. P. d. Matos, N. S. B. Saldanha, and M. d. M. Pedroso (2025). Air pollutants data. Zenodo data collection. <https://zenodo.org/communities/airpollutantsdata/records>. Municipality-level indicators derived from the Copernicus Atmosphere Monitoring Service global reanalysis, CAMS EAC4.
- World Health Organization (2021). *WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Geneva: World Health Organization.

Clusterwise Autoregressive Modelling for Interval-Valued Time Series

G. Borrata¹, R. Mattera¹, A. Balzanella¹, R. Verde^{1*}

1. DMF - University of Campania

*Contact author: rosanna.verde@unicampania.it

Keywords: Interval-valued Time Series, Autoregressive Model, Clusterwise Model, LDQ Transformation, Clustering Time Series.

The analysis of interval-valued data has attracted increasing attention within the fields of Symbolic Data Analysis and Classification, where intervals provide a natural representation of variability, uncertainty, and aggregated information. A central issue in the statistical analysis of interval-valued data concerns the choice of representation and distance measure. Different transformations—based on bounds, center–range decompositions, quantile functions, or other functional representations—induce different geometrical structures and consequently affect clustering, classification, and prediction results.

These challenges become even more relevant in the context of interval-valued time series (ITS), where both temporal dependence and within-interval variability must be taken into account. While most existing approaches focus on modelling interval bounds or derived components separately, alternative functional representations can provide a more flexible framework for capturing complex dynamic behaviours while preserving interval coherence. Several approaches to modelling and forecasting interval-valued time series have been proposed in this direction [Maia et al. \(2008\)](#).

In this contribution, we discuss the role of transformations and metrics ([Liao \(2005\)](#), [Maharaj et al. \(2019\)](#)) in the analysis of interval-valued time series and present a general framework for autoregressive modelling based on transformed interval representations. In particular, we propose considering both the midpoint–radius decomposition and, unit-specific log-linear trend models are used to estimate the level and magnitude of ITS separately.

The proposed approach allows interval-valued trajectories to be represented according LDQ transformation on detrended lower-upper bounds, where temporal dependence structures can be estimated using suitable autoregressive techniques and subsequently mapped back to the original interval domain.

Let $X_{it} = [L_{it}, U_{it}]$ denote the interval observed for unit i at time t , where L_{it} and U_{it} are its lower and upper bounds, respectively. We define the minimum component and the LDQ component as

$$M_{it} = L_{it}, \quad D_{it} = \log(U_{it} - L_{it}).$$

Consequently, each interval-valued observation can be represented by the bivariate vector

$$\mathbf{y}_{it} = \begin{pmatrix} M_{it} \\ D_{it} \end{pmatrix} = \begin{pmatrix} L_{it} \\ \log(U_{it} - L_{it}) \end{pmatrix}.$$

The clusterwise modelling principles for interval-valued data was developed by [de Carvalho et al. \(2010\)](#). In this paper proposed clusterwise autoregressive models for ITS, where the estimation and assignment are provided as follow:

- for each unit i belonging to cluster $\mathcal{C}_k$, the cluster-specific vector autoregressive model of order P , is estimated by

$$\mathbf{y}_{it} = \boldsymbol{\alpha}_k + \sum_{j=1}^P \mathbf{A}_{k,j} \mathbf{y}_{i,t-j} + \boldsymbol{\varepsilon}_{it}, \quad i \in \mathcal{C}_k,$$

where: $\alpha_k = \begin{pmatrix} \alpha_{M,k} \\ \alpha_{D,k} \end{pmatrix}$, $\mathbf{A}_{k,j} = \begin{pmatrix} \phi_{MM,kj} & \phi_{MD,kj} \\ \phi_{DM,kj} & \phi_{DD,kj} \end{pmatrix}$, $\epsilon_{it} = \begin{pmatrix} \epsilon_{M,it} \\ \epsilon_{D,it} \end{pmatrix}$.

- Given the estimated cluster-specific parameters, each unit is assigned to the model that minimizes its total prediction error:

$$\hat{k}_i = \arg \min_{k=1,\dots,K} \sum_{t=P+1}^T \left\| \mathbf{y}_{it} - \alpha_k - \sum_{j=1}^P \mathbf{A}_{k,j} \mathbf{y}_{i,t-j} \right\|^2.$$

Model estimation and cluster assignment are alternated until the partition no longer changes or the reduction in the overall prediction error falls below a predefined threshold.

The predicted lower and upper bounds are finally obtained through the inverse transformation:

$$\hat{L}_{i,t+1} = \hat{M}_{i,t+1}, \quad \hat{U}_{i,t+1} = \hat{M}_{i,t+1} + \exp\left(\hat{D}_{i,t+1}\right).$$

The methodology is evaluated through simulation experiments examining cluster recovery, predictive accuracy, and through an application to financial data. The empirical analysis considers different companies listed on the Italian stock market, observed daily from 1 January 2024 to 1 January 2025. The companies operate in energy, utilities, aerospace, telecommunications, luxury goods, healthcare, and financial services. Daily minimum and maximum prices define the interval-valued time series, whose dynamics are summarized through the coefficients of their log-linear trends and the autocorrelations of the detrended series.

In conclusion, this contribution offers a broader perspective on the interplay between symbolic representations, distance-based methods, and dynamic modelling in the emerging field of interval-valued time-series analysis.

References

- de Carvalho, F. d. A. T., G. Saporta, and D. N. Queiroz (2010). A clusterwise center and range regression model for interval-valued data. In Y. Lechevallier and G. Saporta (Eds.), *Proceedings of COMPSTAT 2010*, Heidelberg, pp. 461–468. Physica-Verlag.
- Liao, T. W. (2005). Clustering of time series data: A survey. *Pattern Recognition* 38(11), 1857–1874.
- Maharaj, E. A., P. Teles, and P. Brito (2019). Clustering of interval time series. *Statistics and Computing* 29(5), 1011–1034.
- Maia, A. L. S., F. d. A. T. de Carvalho, and T. B. Ludermir (2008). Forecasting models for interval-valued time series. *Neurocomputing* 71(16–18), 3344–3352.

Author Index

A. Balzanella, 64
A. Pedro Duarte Silva, 50
Adelaide Freitas, 19
Adán Rodríguez-Martínez, 45
Agustina Bouchet, 45
Amedeo Napoli, 21
Ana Felizardo Henriques, 19
Andrej Srakar, 17
Angelo Leogrande, 33
Anuška Ferligoj, 57

Carlo Drago, 33
Catarina P. Loureiro, 46

David Loiret, 21
Diogo Pinheiro, 45

Egor Dudyrev, 21
Eric C. Soares, 24
Eufrásio de A. Lima Neto, 40, 48

Fernando Sebastião, 19
Francisco de A. T. de Carvalho, 26

G. Borrata, 64
Gentian Hoxhalli, 33
Getúlio, J. A. do Amaral, 24

Hélder Alves, 53

Jasminka Dobša, 29, 31
Jiayu Shang, 40
Jitka Machalová, 37
José Nataniel A. de Sá, 26
João Marôco, 19

Karel Hron, 37

Liang-Ching Lin, 13
Lina Oliveira, 46
Luana Lúcia Alves de Azevêdo, 62
Lucija Šutić Barbarić, 29
Luka Kronegger, 31

M. Rosário Oliveira, 45, 46
Maja Buhin Pandur, 29, 31
Marcelo R.P. Ferreira, 26, 40

Marcus Mayrhofer, 50
Maria Eduarda Borges, 60
Mehdi Kaytoue, 21
Miguel Couceiro, 21
Miranda Novak, 29

Oldemar Rodríguez Rojas, 15

Paula Brito, 42, 46, 48, 50, 53, 60, 62
Paulo Canas Rodrigues, 62
Paulo Teles, 60, 62
Pedro Campos, 53
Peter Filzmoser, 37, 50

R. Mattera, 64
R. Verde, 64
Renata M. C. R. de Souza, 24
Rui Nunes, 37

Sónia Dias, 48
Sara Suarez-Fernandez, 45
Sergei O. Kuznetsov, 21
Simona Korenjak-Černe, 29
Sugnet Lubbe, 42

Telmo Silva Filho, 40

Vladimir Batagelj, 55, 57