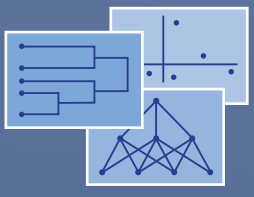
Studies in Classification, Data Analysis, and Knowledge Organization

Theodore Chadjipadelis · Berthold Lausen · Angelos Markos · Tae Rim Lee · Angela Montanari · Rebecca Nugent *Editors*

Data Analysis and Rationality in a Complex World



Springer

Studies in Classification, Data Analysis, and Knowledge Organization

More information about this series at http://www.springer.com/series/1564

Theodore Chadjipadelis · Berthold Lausen ·
Angelos Markos · Tae Rim Lee ·
Angela Montanari · Rebecca Nugent
Editors

Data Analysis and Rationality in a Complex World

Springer

Editors
Theodore Chadjipadelis
Department of Political Sciences
Aristotle University of Thessaloniki
Thessaloniki, Greece

Berthold Lausen
Department of Mathematical Sciences
University of Essex
Colchester, UK

Angelos Markos
School of Education
Democritus University of Thrace
Alexandroupolis, Greece

Tae Rim Lee
Department of Data Science and Statistics
Korea National Open University
Seoul, Korea (Republic of)

Angela Montanari
Department of Statistical Sciences
"Paolo Fortunati"
University of Bologna
Bologna, Italy

Rebecca Nugent
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA, USA

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-030-60103-4 ISBN 978-3-030-60104-1 (eBook)
https://doi.org/10.1007/978-3-030-60104-1

Mathematics Subject Classification: 62-XX, 62-06, 62-07, 62Hxx, 62H30, 62Pxx, 62Jxx

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains revised versions of the selected papers presented at the 16th Biennial Conference of the International Federation of Classification Societies (IFCS 2019) organized by the Greek Society of Data Analysis (GSDA), held in Thessaloniki, Greece on 26–29 August 2019. The theme of the conference was "Data Analysis and Rationality in a Complex World". Rationality is a critical issue, as we experience it today. The COVID-19 outbreak revealed also the complexity. Data Analysis is -not the only, but a critical tool for handling information and making decisions under uncertainty on many occasions and for many scientific areas. Rationality is about decision-making based on facts, political and social choice, and the Interest of the People. Authorities, universities, and institutions should take care in order to improve everyday life and solve major political and social problems bringing together Data Science [improve rationality], free and fair Elections [secure free choice and responsibility], and Governance [handling a complex World].

Theodore Chadjipadelis (Aristotle University of Thessaloniki) chaired the Local Organizing Committee and the Scientific Program Committee with Berthold Lausen (IFCS President) and Tae Rim Lee (Korea National Open University) as the vice-chairpersons. The conference encompassed 178 presentations in 56 sessions, including 8 plenary talks and 2 workshops. With 224 attendees from 29 countries, the conference provided a very attractive interdisciplinary international forum for discussion, mutual exchange of knowledge, and cross-disciplinary cooperation.

This volume presents 37 articles dealing with theoretical aspects, methodological advances, and practical applications in domains relating to classification and clustering. The contributions were selected in a second reviewing process after the conference. In addition to the fundamental areas of classification and clustering, the volume contains manuscripts concerning data analysis and statistical modelling in application areas such as economics and finance, computer science, political science, and education. The contributions are listed in alphabetical order with respect to the authors' names.

For the convenience of the reader, the content of this volume is briefly reviewed: *Bellanger et al.* present an agglomerative hierarchical clustering method with temporal ordering constraints. *Chadjipadelis & Teperoglou* employ hierarchical clustering and multiple correspondence analysis to analyze political competition in EU member states at the occasion of the 2019 European Parliament elections. *Champagne Gareau et al.* present a graph clustering technique to improve the efficiency of an electric vehicle planner. *Di Mari et al.* present an approach for computing the coefficient of determination for mixtures of regressions in the Gaussian framework. *Dziechciarz & Dziechciarz-Duda* present a procedure for survey data collection based on fuzzy coding. *Ferreira & Marques* study the relationships between performance measures in discrete supervised classification. *Ganczarek-Gamrot et al.* evaluate value-at-risk measures to assess the risk of price changes in the energy market. *Górecki et al.* define and evaluate measures of mutual dependence for multivariate functional data. *Iodice D'Enza et al.* present a chunk-wise version of iterative principal component analysis for single imputation of "tall" data sets. *Jimeno et al.* run a benchmarking study to evaluate the performance of different clustering methods for mixed-type data. *Kazana et al.* employ a joint dimension reduction and clustering approach to investigate entrepreneurs' attitudes toward a green infrastructure plan. Kitanishi et al. apply a topological data analysis mapper and a spatial perception method to systematically visualize the relationships among pharmaceutical data. *Koutsoupias & Mikelis* combine the use of text mining and multivariate data analysis methods to explore a set of textual documents. *Krężołek & Trzpiot* present an approach to estimate extreme risk using the Hill estimator and its modifications. *Lelu & Cadot* evaluate a series of clustering methods on text data. *Liang & Lee* present experimental results to obtain a rule-of-thumb for choosing the basis spacing for process convolution Gaussian process models. *McLachlan & Ahfock* review and present new results about using the Gaussian mixture model for partially classified data. *Menexes & Koutsos* combine correspondence analysis and ordinary kriging to display values of quantitative variables as supplementary onto factorial maps. *Moschidis & Thanopoulos* apply dimension reduction and clustering techniques to study heterogeneity in e-commerce data from official statistics. *Murugesan et al.* run a benchmarking study to highlight the advantages and drawbacks of spectral clustering, DBSCAN, and k-means on simulated and empirical data. Nakayama employs Bayesian network analysis to model trends in consumer web communication data of new products. *Nicolussi et al.* consider chain graph models for categorical variables to evaluate the level of perceived health in the EU. *Nienkemper-Swanepoel et al.* present a visualization approach to identify the missing data mechanism in incomplete multivariate categorical data. *Okada & Yokoyama* introduce a procedure for assembling one-mode three-way proximities from one-mode two-way proximities, and a method for hierarchical clustering of one-mode three-way proximities. *Panagiotidou & Chadjipadelis* explore the views and attitudes of first-time young voters about Europe and Democracy using multivariate data analysis techniques. *Pratsinakis et al.* compare hierarchical clustering approaches for binary data from molecular markers using external criteria for cluster validation. Smaga introduces

permutation and bootstrap tests for the repeated measures analysis of variance for functional data. *Sokołowski & Markowska* present an algorithm for creating a robust distance matrix between observations with outliers. *Srakar & Vecco* present a clustering algorithm for polygonal data. *Stalidis et al.* evaluate the performance of multiple correspondence analysis and hierarchical clustering, as well as a two-layer shallow neural network for personalized supermarket offer recommendations. *Szilágyi & Lengyel* present the results of an empirical study on what motivates the participants of the sharing economy in Hungary using structural equation modeling. *Tai & Frisoli* run a benchmark comparison of minimax linkage to other hierarchical clustering methods using multiple performance metrics on data sets with known clustering structure. *Trejos-Zelaya et al.* implement and evaluate clustering algorithms based on combinatorial optimization metaheuristics. *Tsimperidis et al.* employ keystroke dynamics and machine learning models to classify unknown Internet users according to age, handedness, and educational level. *Varga & Fodor* use hierarchical clustering to derive a typology of critical raw materials with regard to technological innovation. *Vicente-Villardón et al.* extend redundancy analysis to binary data using logistic regression. *Warrens & Ebert* study the predictive power of cluster solutions based on normal mixture models when relevant outcomes are involved in the estimation procedure, using a real-world data set on school motivation.

We would like to express our gratitude to all members of the scientific program committee, for their ability in attracting interesting contributions. A special thanks is due to the local organizing committee for a well-organized conference. We also thank the session organizers for supporting the spread of information about the conference, and for inviting speakers, the reviewers for their timely reports, and Veronika Rosteck and Boopalan Renu of Springer Nature for their support and dedication to the production of this volume. Last but not least, we would like to thank all participants of the conference for their interest and various activities which made the IFCS 2019 conference and this volume an interdisciplinary possibility for scientific discussion.

Colchester, UK — Berthold Lausen
Thessaloniki, Greece — Theodore Chadjipadelis
Alexandroupolis, Greece — Angelos Markos
Seoul, Korea (Republic of) — Tae Rim Lee
Bologna, Italy — Angela Montanari
Pittsburgh, USA — Rebecca Nugent
June 2020

Contents

Contributors

Daniel Ahfock University of Queensland, Brisbane, QLD, Australia

Luis Eduardo Amaya-Briceño Guanacaste Campus, University of Costa Rica, Liberia, Costa Rica

Avi Arampatzis Democritus University of Thrace, Xanthi, Greece

Éric Beaudry Université du Québec á Montréal, QC, Montréal, Canada

Lise Bellanger Université de Nantes Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, Nantes Cedex 03, France

Martine Cadot LORIA Nancy France, Vandoeuvre-lès-Nancy, France

Manuela Cazzaro University of Milano-Bicocca, Milano MI, Italy

Theodore Chadjipadelis School of Political Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece

Jaël Champagne Gareau Université du Québec á Montréal, QC, Montréal, Canada

Irene Cho Department of Mathematics and Statistics, San José State University, San José, CA, USA

Efthimia Chrisanthidou International Hellenic University, Drama, Greece

Arthur Coulon CNRS/Université de Tours, UMR 7324 CITERES, Laboratoire Archéologie et Territoires, Tours, France

Marina Delianidi International Hellenic University, Thermi, Greece

Agnese Maria Di Brisco University of Milano-Bicocca, Milano MI, Italy

Roberto Di Mari Department of Economics and Business, University of Catania, Catania, Italy

Konstantinos Diamantaras International Hellenic University, Thermi, Greece

Józef Dziechciarz Wrocław University of Economics, Wroclaw, Poland

Marta Dziechciarz-Duda Wrocław University of Economics, Wroclaw, Poland

W. Miro Ebert University of Groningen, Faculty of Behavioural and Social Sciences, Groningen, TS, The Netherlands

Ilias Eleftherohorinos Aristotle University of Thessaloniki, Thessaloniki, Greece

Ana Sousa Ferreira Faculdade de Psicologia, Universidade de Lisboa, Business Research Unit (BRU-IUL), Lisboa, Portugal

Kitti Fodor University of Miskolc, Institute of Economic Theory and Methodology, Miskolc-Egyetemváros, Hungary

Kayla Frisoli Carnegie Mellon University, Pittsburgh, PA, USA

Alicja Ganczarek-Gamrot University of Economics in Katowice, Katowice, Poland

Sugnet Gardner-Lubbe Stellenbosch University, Stellenbosch, South Africa

Tomasz Górecki Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

Philippe Husi CNRS/Université de Tours, UMR 7324 CITERES, Laboratoire Archéologie et Territoires, Tours, France

Masaya Iizuka Okayama University, Okayama, Japan

Salvatore Ingrassia Department of Economics and Business, University of Catania, Catania, Italy

Alfonso Iodice D'Enza Universitá degli Studi di Napoli Federico II, Napoli, Italy

Fumio Ishioka Okayama University, Okayama, Japan

Jarrett Jimeno Department of Mathematics and Statistics, San José State University, San José, CA, USA

Alejandra Jiménez-Romero School of Mathematics, Costa Rica Institute of Technology, Cartago, Costa Rica

Pantelis I. Kaplanoglou International Hellenic University, Thermi, Greece

Lefkothea Karapetsi Centre for Research and Technology, Thessaloniki, Greece

Iphigenia Karaveli International Hellenic University, Thermi, Greece

Alkiviadis Katsalis International Hellenic University, Thermi, Greece

Stella Kazakli International Hellenic University, Drama, Greece

Angelos Kazaklis OLYMPOS Non Profit Integrated Centre for Environmental Management, Drama, Greece

Vassiliki Kazana Department of Forestry and Natural Environment 1st km Drama-Mikrohori, International Hellenic University, Drama, Greece

Yoshitake Kitanishi Okayama University, Okayama, Japan

Thomas Koutsos School of Agriculture Faculty of Agriculture Forestry and Natural Environment Hellas, Aristotle University of Thessaloniki, Thessaloniki, Greece

Nikos Koutsoupias University of Macedonia, Thessaloniki, Greece

Dominik Krężołek Department of Demographics and Economic Statistics, University of Economics in Katowice, Katowice, Poland

Mirosław Krzyśko Interfaculty Institute of Mathematics and Statistics, Calisia University, Kalisz, Poland

Koji Kurihara Okayama University, Okayama, Japan

Niël Le Roux Stellenbosch University, Stellenbosch, South Africa

Herbert K. H. Lee University of California, Santa Cruz, USA

Alain Lelu Université de Franche-Comté (rtd), Besançon, France

Levente Lengyel Institute of Economic Theory and Methodology, University of Miskolc, Miskolc-Egyetemváros, Hungary

Waley W. J. Liang University of California, Santa Cruz, USA

Panagiotis Madesis Centre for Research and Technology, Thessaloniki, Greece

Vladimir Makarenkov Université du Québec á Montréal, QC, Montréal, Canada

Angelos Markos Democritus University of Thrace, Alexandroupoli, Greece

Małgorzata Markowska Wrocław University of Economics and Business, Wrocław, Poland

Anabela Marques Escola Superior de Tecnologia do Barreiro, IPS, CIIAS, Barreiro, Portugal

Geoffrey McLachlan University of Queensland, Brisbane, QLD, Australia

George Menexes School of Agriculture Faculty of Agriculture Forestry and Natural Environment Hellas, Aristotle University of Thessaloniki, Thessaloniki, Greece

Kyriakos Mikelis University of Macedonia, Thessaloniki, Greece

Stratos Moschidis Hellenic Statistical Authority, Piraeus, Greece

Alex Murillo-Fernández Atlantic Campus, University of Costa Rica, Turrialba, Costa Rica

Nivedha Murugesan Department of Mathematics and Statistics, San José State University, San José, CA, USA

Atsuho Nakayama Tokyo Metropolitan University, Hachioji-shi, Japan

Federica Nicolussi University of Milan, Milano MI, Italy

Johané Nienkemper-Swanepoel Stellenbosch University, Stellenbosch, South Africa

Symela Ntoanidou Aristotle University of Thessaloniki, Thessaloniki, Greece

Akinori Okada Rikkyo University 3-18-1 Ozenji Higashi Asao-ku Kawasaki-shi, Kanagawa-ken, Japan

Francesco Palumbo Universitá degli Studi di Napoli Federico II, Napoli, Italy

Georgia Panagiotidou School of Political Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece

Georgios Peikos Democritus University of Thrace, Xanthi, Greece

Eduardo Piza-Volio CIMPA & School of Mathematics, Faculty of Science, University of Costa Rica, San José, Costa Rica

Emmanouil D. Pratsinakis Aristotle University of Thessaloniki, Thessaloniki, Greece

Antonio Punzo Department of Economics and Business, University of Catania, Catania, Italy

Dimitrios Raptis International Hellenic University, Drama, Greece

Madhumita Roy Department of Mathematics and Statistics, San José State University, San José, CA, USA

Theodosios Siomos International Hellenic University, Thermi, Greece

Łukasz Smaga Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

Andrzej Sokołowski Cracow University of Economics, Cracow, Poland

Andrej Srakar Institute for Economic Research (IER), Ljubljana and Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia

George Stalidis International Hellenic University, Thermi, Greece

Roland Szilágyi Institute of Economic Theory and Methodology, University of Miskolc, Miskolc-Egyetemváros, Hungary

Xiao Hui Tai University of California, Berkeley, CA, USA

Eftichia Teperoglou School of Political Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece

Athanasios Thanopoulos Hellenic Statistical Authority, Piraeus, Greece

Cristina Tortora Department of Mathematics and Statistics, San José State University, San José, CA, USA

Javier Trejos-Zelaya CIMPA & School of Mathematics, Faculty of Science, University of Costa Rica, San José, Costa Rica

Grażyna Trzpiot Department of Demographics and Economic Statistics, University of Economics in Katowice, Katowice, Poland

Ioannis Tsimperidis Democritus University of Thrace, Xanthi, Greece

Beatrix Varga University of Miskolc, Institute of Economic Theory and Methodology, Miskolc-Egyetemváros, Hungary

Marilena Vecco Burgundy School of Business—Universite Bourgogne Franche-Comte, Bourgogne Franche-Comte, France

Laura Vicente-Gonzalez Departamento de Estadística, Universidad de Salamanca, Salamanca, Spain

Jose L. Vicente-Villardon Departamento de Estadística, Universidad de Salamanca, Salamanca, Spain

Mario Villalobos-Arias CIMPA & School of Mathematics, Faculty of Science, University of Costa Rica, San José, Costa Rica

Matthijs J. Warrens University of Groningen, Faculty of Behavioural and Social Sciences, Groningen Institute for Educational Research, Groningen, TG, The Netherlands

Waldemar Wołyński Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

Satoru Yokoyama Aoyama Gakuin University 4-4-25 Shibuya, Tokyo, Japan

Nefeli Zagourgini International Hellenic University, Drama, Greece

PerioClust: A Simple Hierarchical Agglomerative Clustering Approach Including Constraints

Lise Bellanger, Arthur Coulon, and Philippe Husi

Abstract PerioClust is a hierarchical agglomerative clustering (HAC) method including temporal (resp. spatial) ordering constraints. This new semi-supervised learning algorithm is designed to consider two potentially error-prone sources of information associated with the same observations. One reflects dissimilarities in the "feature space" and the other the temporal (resp. spatial) constraint structure between the observations. A distance-based approach is adopted to modify the distance measure in the classical HAC algorithm using a convex combination to take into account the two initial dissimilarity matrices. The choice of the mixing parameter is, therefore, the key point. We define a criterion based on cophenetic distances, as well as a resampling procedure to ensure the good robustness of the proposed clustering method. The dendrogram associated with this HAC can be interpreted as the result of a compromise between each source of information analysed separately. We illustrate our clustering method on two real data sets: (i) an archaeological one containing temporal information, (ii) a socio-economical one containing geographical information.

Keywords Semi-supervised learning algorithm · Non-strict constrained approach · Hierarchical agglomerative clustering · Weighted average distance · Cophenetic matrix

L. Bellanger (✉)
Université de Nantes Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, 2 rue de la Houssinière BP 92208, 44322 Nantes Cedex 03, France
e-mail: lise.bellanger@univ-nantes.fr

A. Coulon · P. Husi
CNRS/Université de Tours, UMR 7324 CITERES, Laboratoire Archéologie et Territoires, 40 rue James Watt, ActiCampus 1, 37200 Tours, France

T. Chadjipadelis et al. (eds.), *Data Analysis and Rationality in a Complex World*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-60104-1_1

1 Introduction and Motivation

Clustering problems can be addressed with a variety of methods, all requiring dedicated techniques for the data preprocessing phase of its own. There is an abundant literature on the clustering subject, see, for example, Aggarwal and Reddy (2014), Everitt et al. (2001), Kaufman and Rousseeuw (2005). The two most widely used clustering algorithms are partitional and hierarchical clustering. In this paper, we concentrate on hierarchical clustering whose approach consists in developing a binary-tree-based data structure called the dendrogram. More specifically we propose a new constrained Hierarchical Agglomerative Clustering (HAC) method named PerioClust, which belongs to the class of semi-supervised learning algorithms. It is a non-strict constrained procedure that has been originally developed to answer chronological problems in archaeology based on artefacts data. The method is designed to consider two potentially error-prone sources of information associated with the same observations. One reflects dissimilarities in the "feature space" and the other the temporal (resp. spatial) constraint structure between the observations. A distance-based approach is adopted to modify the distance measure in the classical HAC algorithm using a convex combination of the two initial dissimilarity matrices. The choice of the mixing parameter is, therefore, the key point. We define a selection criterion for this parameter based on cophenetic distances, as well as a resampling procedure to ensure the good robustness of the proposed clustering method.

This article is organized as follows. In Sect. 2, we describe the existing methods. In Sect. 3, we present the proposed procedure. In Sect. 4, we illustrate and compare our approach using two real datasets.

2 Existing Constrained HAC Methods

Constrained clustering is a class of semi-supervised learning algorithms that differ from its unconstrained counterpart in that the only admissible clusters are those that more or less strictly respect the relationship. User-specified constraints we are interested in here are those called instance-level constraints (Davidson and Basu 2007), specifying requirements on pairs of objects. Several researchers proposed extending classical algorithms for handling instance-level constraints. Two general approaches exist: (i) constrained-based ones where the clustering algorithm is modified to integrate pairwise constraints, (ii) distance-based ones where only the distance measure is modified in the existing clustering algorithm.

In constrained-based approach, the HAC methods included in the Lance and Williams general clustering model are easily modified to incorporate the constraint of continuity. Clustering algorithms with temporal (or spatial) constraint need to state unambiguously which objects are neighbours. The most common constrained clustering solution is to use simple connecting schemes, proceeding as described in Legendre and Legendre (2012) in the case of temporal constraints. In this approach called the chronological clustering method, the constraint is imposed on the cluster-

ing activity. Another constrained-based approach, proposed by Chavent et al. (2018) and called hclustgeo, consists of proposing a Ward-like hierarchical algorithm including spatial constraints through two dissimilarity matrices and a mixing parameter. It has the potential advantage to be a non-strict constrained procedure. However, it imposes the underlying aggregation measure which leads to a Ward-like hierarchical clustering process. In general terms, HAC with constraint-based approaches have some disadvantages: (i) it can occasionally produce reversals in the dendrogram, except with complete linkage (Ferligoj and Batagelj 1982), (ii) it usually considers only the dissimilarities between linked units (e.g. chronological clustering method), which may be too restrictive in some fields such as archaeology, as we will see below, (iii) the choice and interpretation of the mixing parameter could be sensitive points (e.g. hclustgeo method).

Finally, HAC also has a long history of using spatial constraints to find specific types of clusters with the distance-based approach: the dissimilarity matrix is modified differently as, for example, a combination of geographical dissimilarities and dissimilarities on non-geographical variables. But a problem arises on how to define weight given to the geographical dissimilarities in the combination. In this work, we propose a non-strict constrained HAC approach that takes up this idea by (i) adapting it to the temporal or spatial constraints and (ii) defining weights in an objective and interpretable way.

3 The Proposed Clustering Method: PerioClust

3.1 A Distance-Based Approach

Let us consider a set of n observations. Let $\mathbf{D}_1$ be a $n \times n$ normalized[1] dissimilarity matrix (not necessary an Euclidean distance) giving dissimilarity values in the "feature space". Let $\mathbf{D}_2$ be a $n \times n$ matrix containing the normalized dissimilarities in the "constraint space". In this clustering work, we apply the HAC method to the following convex combination:

$$\mathbf{D}_\alpha = \alpha \mathbf{D}_1 + (1 - \alpha)\mathbf{D}_2 \tag{1}$$

where $\alpha \in [0; 1]$ is a fixed parameter given the importance of each dissimilarity matrix in the clustering procedure. Formula (1) defines a weighted average distance and as such makes it possible to weight each of the two sources of information calculated on the data set. When $\alpha = 0$ (resp. $\alpha = 1$), the dissimilarities obtained from dissimilarity matrix $\mathbf{D}_1$ (resp. $\mathbf{D}_2$) are not taken into account in the hierarchical clustering process. An agglomerative strategy could be chosen among those satisfying the Lance and Williams formulation. Thus, the key point here is the choice of α.

[1] Dissimilarity values are between 0 and 1.

Sokal and Rohlf (1962) developed a simple criteria, the cophenetic correlation, which provides a simple and effective method for comparing dendrograms of various sorts. The starting point is the so-called cophenetic matrix whose elements are the dissimilarity levels at which objects become members of the same cluster in the dendrogram. The correlation between the original dissimilarities and the cophenetic dissimilarities (called cophenetic correlation) is a "suitability index" of the clustering. It judges the extent to which the hierarchical structure produced by a dendrogram actually represents the data itself (see Everitt et al. 2001; Sokal and Rohlf 1962). We will base the determination of α on the optimization of an objective function based on the spirit of the cophenetic correlation.

We defined the following criterion that "balances" the weight of $\mathbf{D}_1$ and $\mathbf{D}_2$ in the final clustering:

$$CorCrit_\alpha = |Cor(\mathbf{D}_\alpha^{coph}, \mathbf{D}_1) - Cor(\mathbf{D}_\alpha^{coph}, \mathbf{D}_2)| \tag{2}$$

where $\mathbf{D}_\alpha^{coph}$ is the cophenetic matrix obtained from the HAC dendrogram with α fixed in (1). The $CorCrit_\alpha$ criterium in (2), therefore, represents the difference in absolute value between two correlations, each correlation measuring how faithfully the dendrogram with $\mathbf{D}_\alpha^{coph}$ preserves the pairwise distances between the original data points. The first correlation is associated with the comparison between the original dissimilarity matrix $\mathbf{D}_1$ and $\mathbf{D}_\alpha^{coph}$ with α fixed; while the second one compares the original dissimilarity matrix $\mathbf{D}_2$ with $\mathbf{D}_\alpha^{coph}$, α fixed. Then, in order to compromise between the information provided by $\mathbf{D}_1$ and $\mathbf{D}_2$, we determine α with $\hat{\alpha}$ such that

$$\hat{\alpha} = \text{argmin}_\alpha \, CorCrit_\alpha \tag{3}$$

In practice, we use a one-dimensional optimization procedure, combination of golden section search and successive parabolic interpolation, to obtain the value $\hat{\alpha}$ that minimizes (3). However, since this choice does not consider the potential errors in the data corpus, we also decide to create a resampling procedure adapted to obtain a percentile confidence interval for α and study its variability.

3.2 Resampling Strategy

To do this, a set of "clones" is created for each observation. A clone c of observation $i \in \{1, ..., n\}$ is defined as a copy of observation i for which dissimilarities in the "feature space" have the same values as observation i but those in the "constraint space" have been modified taking into account for a fixed i all possible profiles $i' \neq i$. Let $\mathbf{D}_1^{(c)}$ and $\mathbf{D}_2^{(c)}$ be the two $(n+1) \times (n+1)$ dissimilarity augmented matrices for clone $c \in \{1, ..., n(n-1)\}$. A HAC is then carried out using the combination defined in (1) with $\mathbf{D}_1^{(i)}$ and $\mathbf{D}_2^{(i(i'))}$. Let $CorCrit_\alpha^{(c)}$ define the same criterion as in (2) in which $\mathbf{D}_1$ and $\mathbf{D}_2$ are replaced, respectively, by $\mathbf{D}_1^{(c)}$ and $\mathbf{D}_2^{(c)}$. The adaptation of the previous reasoning to estimate α with (3) using our resampling procedure leads us to define

$$\hat{\alpha}^{(c)} = \text{argmin}_{\alpha}\, CorCrit_{\alpha}^{(c)}; c \in \{1, ..., n(n-1)\} \quad (4)$$

Based on replications, the same spirit as the Bootstrap method Efron and Tibshirani (1993), we obtain an estimated percentile confidence interval using the empirical percentiles of the distribution of $\hat{\alpha}^{(c)}$, average and standard error. A HAC can then be made using $\mathbf{D}_{\hat{\alpha}}$ or $\mathbf{D}_{\tilde{\alpha}}$, where $\tilde{\alpha}$ is in the vicinity of $\hat{\alpha}$ with respect to the confidence interval $CI_{95\%}(\alpha)$.

4 Results and Comparison on Two Real Datasets

In this section, we present the results obtained with our distance-based approach PerioClust on two real datasets: (i) Angkor data with temporal constraint, (ii) Estuary data with geographical constraint. We also compare them with those obtained with the constrained-based approach hclustgeo (Chavent et al. 2018), a comparable non-strict constrained HAC with a mixing parameter. All statistical analyses were performed using R.[2]

4.1 Archaeological Dataset: Temporal Constraints

The archaeological data come from excavations carried out in Angkor Thom (Cambodia), capital of the Khmer empire between the ninth and fifteenth centuries (Gaucher 2004). One of the major objectives here is to specify the periodization of the city, particularly from the seriation diagram (Fig. 1) otherwise called "Harris matrix" (Harris 1989) with connected sets coming from 3 disconnected archaeological sites and assemblages of pottery (quantities of different types of sherds of pottery contained in sets).

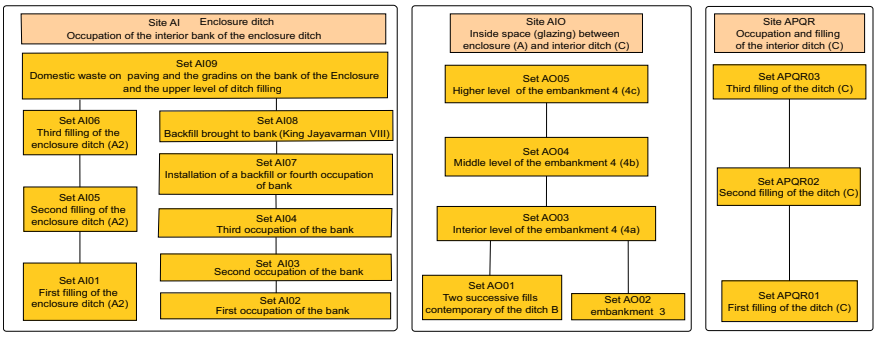


Fig. 1 Angkor: seriation diagram for three archaeological sites in relation to enclosure system

[2] http://www.r-project.org.

From the seriation diagram, it is therefore possible to construct $\mathbf{S}_2$, the symmetric adjacency matrix defined as a binary matrix of connectivity and then $\mathbf{D}_2 = \mathbf{1}_{17\times17} - \mathbf{S}_2$ associated with the 17 archaeological sets (see Sect. 3.1). Information on pottery is contained in a contingency table $\mathbf{N}$ of size 17×12 where the rows correspond to the sets and the columns to the pottery categories. As very often on this type of data Bellanger and Husi (2012), Correspondence Analysis[3] Greenacre (2016) on $\mathbf{N}$ allows to observe an arch pattern of row and column points. The first factor is related to the best chronological seriation order obtained with pottery only. Hence, the construction and overall interpretation of the chronology of the 3 sites can be enriched by combining these two sources of information using an adapted clustering method such as PerioClust. Euclidean distances between sets are calculated from all CA row components on $\mathbf{N}$. HAC on $\mathbf{D}_1$ representing pottery and $\mathbf{D}_2$ representing stratigraphy lead to the highest values of the Agglomerative Coefficient (Kaufman and Rousseeuw 2005) for Ward's criterion. It can be considered as the best aggregation strategy to adopt for this data. As a lower entanglement coefficient corresponds to a good alignment, the entanglement value of 0.39 between trees from Ward HAC with $\mathbf{D}_1$ and $\mathbf{D}_2$ indicates that they are similar, but not identical. This confirms that the information provided by pottery and stratigraphy must be considered simultaneously to solve the clustering problem.

To apply PerioClust, we define $\mathbf{D}_\alpha$ from (1) and determine an optimal α using (3) with a confidence interval from the resampling strategy (see Sect. 3.2). We obtain $CI_{95\%}\ (\alpha)= [0.55; 0.80]$ and choose $\hat{\alpha} = 0.7$. This value indicates that for Angkor data the weight of each information source is distributed as follows: 70% for pottery and 30% for stratigraphy. This imbalance may result from the difficulty of fine interpretation of a disturbed stratigraphy, often without well-defined limits.

A Ward HAC is performed with $\mathbf{D}_{0.7}$ as defined in (1). The number of clusters to be retained was selected based on the examination of the scale of aggregation indices associated with the dendrogram (Fig. 2a) but also on the archaeologist's knowledge of the site. Indeed, the choice of 4 clusters with cluster D divided into 3 sub-clusters (Fig. 2a) seems better adapted to the chronological rhythms of the city. The fact that some clusters only include one set is archaeologically explained by the chronology:

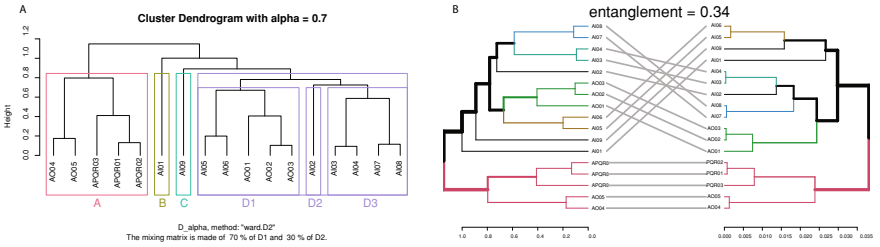


Fig. 2 **a** PerioClust with $\alpha = 0.7$, 4 clusters and 3 sub-clusters for D cluster. **b** Tanglegram between PerioClust tree (left) and hclustgeo tree (right), $\alpha = 0.7$

[3] In abbreviated form CA.

AI01 and AI02 are anterior and AI09 posterior to the most intense activity around the enclosure shown by a more rapid succession of the many other sets.

By applying hclustgeo with $\alpha = 0.7$ [4] and making 4 groups, we obtain a partition different from that of PerioClust (Fig. 2b). $\alpha = 0.7$ is also the value to choose using the quality criterion described in Chavent et al. (2018). An entanglement value of 0.34, a correlation between cophenetic matrices of 0.94 and an ARI of 0.71 indicate that PerioClust and hclustgeo give slightly different results.

From an archaeological point of view, hclustgeo's results with a very rapid grouping of the AI09, AI05 and AI06 sets are not very satisfactory (see Figs. 1 and 3). Indeed, AI09 is much more recent than the other sets and should therefore remain isolated as is the case with PerioClust. In the same way, AI01 and AI09 are very quickly grouped in the same class with hclustgeo, which poses a real problem because the oldest and most recent set are then grouped together. PerioClust presents a grouping reading more adapted to this archaeological dataset.

4.2 Estuary Dataset: Geographical Constraints

Estuary dataset is available in the ClustGeo package (Chavent et al. 2018). It is an extraction of 4 quantitative socio-economic variables for a subsample of 303 French municipalities located on the Atlantic Coast between Royan and Mimizan. The two considered dissimilarity matrices are $\mathbf{D}_1$ the Euclidean distance matrix between the municipalities performed with the 4 socio-economic variables and $\mathbf{D}_2$ the geographical distances. We set the number of classes to 5 as in Chavent et al. (2018) and compare the α values obtained with each method. For hclustgeo, $\alpha = 0.8$ was retained by the authors. If we set $\alpha = 0.8$ in PerioClust, we obtain an entanglement value of 0.84, indicating that the trees are very different.

For PerioClust, using the resampling strategy, we retain $\alpha = 0.45$ ($CI_{95\%}(\alpha)$= [0.33; 0.50]). An entanglement value of 0.8 indicates that the trees obtained with the

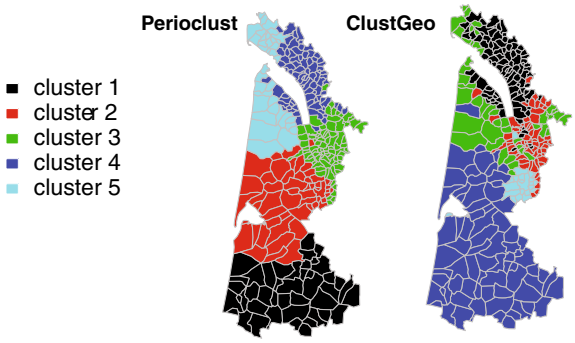


Fig. 3 Estuary: Map of the partition in 5 clusters: $\alpha = 0.45$ for PerioClust (left), $\alpha = 0.8$ hclustgeo (right)

[4] $\mathbf{D}_\alpha$ is differently defined between PerioClust and hclustgeo with $\alpha_{PerioClust} = 1 - \alpha_{hclustgeo}$. In the following, α will always refer to $\alpha_{PerioClust}$ that gives the direct importance of $\mathbf{D}_1$ in $\mathbf{D}_\alpha$.

alpha values adapted to each method are very different. Figure 3 shows that the 5 clusters are more spatially compact than those obtained for hclustgeo with $\alpha = 0.8$.

In summary, the two methods applied to these two datasets result in different dendrograms and then different partitions even if the mix parameter choice is the same.

5 Conclusions

Here, with PerioClust, we proposed a new HAC approach using temporal or spatial constraints, designed to take into consideration two sources of information. This distance-based and non-strict constrained approach is simple to implement. The modified dissimilarity matrix in the HAC is a combination of two dissimilarity matrices, so by construction all existing linkage criteria can be used. The problems of the choice and the interpretation of the mixing parameter, key points for this type of clustering methods, are solved. The mixing parameter α sets the importance of the constraint in the clustering procedure. Although PerioClust was firstly designed for archaeology, it may have a great interest in many other fields, including, for example, ecology or health (e.g. in Genome-Wide Association Studies (GWAS)).

PerioClust will be soon implemented in an R package with a shiny version also to facilitate its use.

Acknowledgements This research was supported in part by the ANR project ModAThom coordinated by Philippe Husi and Jacques Gaucher (EFEO). The authors wish to thank Jacques Gaucher for comments and suggestion.

References

Aggarwal, C., Reddy, C.: Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, Boca Raton (2014)

Bellanger, L., Husi, P.: Statistical tool for dating and interpreting archaeological contexts using pottery. J. Archaeol. Sci. **39**, 777–790 (2012)

Chavent, M., Kuentz-Simonet, V., Labenne, A., Saracco, J.: ClustGeo: an R package for hierarchical clustering with spatial constraints. Computat. Stat. **33**, 1799–1822 (2018)

Davidson, I., Basu, S.: A survey of clustering with instance level. ACM T Knowl. Discov. D. **77**, 1–41 (2007)

Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall/CRC, New York (1993)

Everitt, B., Landau, S., Morven, L.: Cluster Analysis, 4th edn. Oxford University Press Inc., Oxford (2001)

Ferligoj, A., Batagelj, V.: Clustering with relational constraint. Psychometrika **47**, 413–426 (1982)

Gaucher, J.: Angkor Thom, une utopie réalisée?: structuration de l'espace et modèle indien d'urbanisme dans le Cambodge ancien. Arts Asiat. **59**, 58–86 (2004)

Greenacre, M.: Correspondence Analysis in Practice. Chapman & Hall/CRC, Boca Raton (2016)

Harris, E.C.: Principles of Archaeological Stratigraphy, 2nd edn. Academic Press, London and San Diego (1989)

Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience, New York (2005)

Legendre, P., Legendre, L.: Numerical Ecology, 3rd edn. Elsevier Science BV, Amsterdam (2012)

Sokal, R.R., Rohlf, F.J.: The comparison of dendrograms by objective methods. Taxon **11**, 33–40 (1962)

What Was Really the Case? Party Competition in Europe at the Occasion of the 2019 European Parliament Elections

Theodore Chadjipadelis and Eftichia Teperoglou

Abstract The main aim of the paper is to analyse political competition in EU member states at the occasion of the 2019 European Parliament elections. At the core of our analysis are both the priorities of the national parties campaigning for the 2019 European elections and the manifestos of the transnational party groups, each consisting of national member parties from the 28 member states of the European Union. By comparing the major priorities of national actors/parties and those of the European political groups, we will be able to gauge out whether they share different or same dimensions of policy. More broadly, we will depict whether the dynamism in policy competition at the national level affects EP political groups or vice versa. The analysis is implemented through the use of correspondence analysis. Through this approach, the axes of political competition are realized.

Keywords European elections 2019 · European parliament · Policy positions · Party cohesion · Party competition

1 Introduction

Since 1979 when the first European Parliament election (EP elections) took place, national politics have dominated the electoral campaigns, party strategies and as a result, political competition at the EU level as well. Since then, in most member states of the European Union (EU), European issues have not been at stake and the EP elections have been mainly regarded as national second-order national contests (Reif and Schmitt 1980). Given the fact that the role of the European parties and transnational party groups in the EP is less clear for European citizens, at the EU level it could be argued that the so-called 'electoral connection' between the politi-

T. Chadjipadelis (✉) · E. Teperoglou
School of Political Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: chadji@polsci.auth.gr

E. Teperoglou
e-mail: efteperoglou@polsci.auth.gr

T. Chadjipadelis et al. (eds.), *Data Analysis and Rationality in a Complex World*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-60104-1_2

cal elites and the voters is weak. This aspect is considered as an important cause of legitimacy problems (e.g. Hix and Lord 1997) at the European level. In addition, in the EP elections national parties have a predominant role. This feature has somehow changed with the nomination of the Spitzenkandidaten back in the 2014 EP elections.[1] Nevertheless, despite the fact that the "nationalization" of the EU politics is a distinguishing characteristic, some scholars argue for the opposite, namely that over the past years is observed a growing policy authority of the EU level of European governance. The European Parliament has become an important actor in the EU-level policymaking process. In other words, we can observe a possible "Europeanization" of the domestic political arena in each EU member state (see among others Schmitt and Teperoglou 2018). It could be argued that the politicization of European integration has changed the content, as well as the process of decision-making (Hooghe and Marks 2009). In this framework of analysis, a central question regards the axes of political competition in EP elections.

The main objective of this article is to depict some insights into the position of national parties and European political parties/ groups competing for the 2019 EP elections on different issues. The main question of our study is whether we can identify the same or different patterns across different party families. Or in other words, to what extent is contestation at the EU level in 2019 related to the classic left-right dimension or other dimensions?

Before presenting our method and data, in the next section, we will briefly refer to the context of the 2019 EP election.

2 The Context of the 2019 European Parliament Elections

The 2019 EP election was the first one in the shadow of the Brexit referendum (for an overview about the context see Russo et al. 2019). Furthermore, the election took place in a period of an ongoing debate about the challenges for EU especially related to the immigration and refugee crises and the rise of populist parties. However, EU membership remained at quite high levels (59 % of the Europeans considered the EU membership as "something good"), but however with remarkable country variations (see the results of the post-electoral survey Eurobarometer 91.5). An important characteristic of the electoral campaign was that different initiatives have been taken in order to increase the participation especially among the younger voters.[2] Contrary to the 2014 EP elections in which the economic crisis was the main issue at stake, in the last EP election with the exception of Southern European countries, the economy was

[1] In 2014, for the first time in the history of the EP elections, there was an explicit attempt to link the results of the election with the appointment of the European Commission President. The Lisbon Treaty specified that the President of the European Commission will be elected by the European Parliament based on a proposal by the European Council, taking into account the results of the European elections (Article 17(7) TEU).

[2] Another striking feature of the last EP election was that overall turnout increased and reached 50.6%.

stabilized.[3] Finally, like in previous EP elections, the contest had in most EU member states a referendum character for some incumbent parties. The electorate wanted to send a "message" to the governmental parties and protest vote (or voting with the boot) is documented (see for the term among others Schmitt and Teperoglou 2018).

3 Operationalization: Data and Method

In our analysis, we use data from the "Your Vote Matters" platform.[4] This platform includes information about the positions of MEPs, national parties and EP political groups on 25 key issues (with three possible answers: against, in favour or abstain/ no vote for each issue]. In total, we have collected the answers of 148 national-level parties across EU and 7 groups of the European Parliament.[5] Given the fact that there was no other available data at the time of presenting and writing this article (e.g. data from Euromanifestos studies), this source is considered as a valuable one. It covers all national- and EU-level parties and a variety of issues.

The first step in our method was to select and group the issues/ questions in specific categories. In total, we have selected items that are grouped into five different categories. These are: economic issues, law and order issues, immigration and refugee issues, environmental issues and finally, institutional issues[6] (for further information about the issues see Figs. 4, 5, 6, 7 and 8 in the Appendix). The second main step was the use of hierarchical cluster analysis for grouping the issues to categories based on the agreement, disagreement or no vote-no opinion of the party on a specific issue (see Fig. 1). The third step was the creation of our dataset with the answers of the parties to these questions per category. Finally, 18 out of the 25 items were used in the analysis.

Data analysis was based on Hierarchical Cluster Analysis (HCA) and Multiple Correspondence Analysis (MCA) in two steps (Chatzipadelis 2017). In the first step, HCA was used to assign subjects to distinct groups according to their response patterns. The main output of HCA was a group or cluster membership variable, which

[3] For example the EU28 unemployment rate was at 6.3% in June 2019 (Eurostat). It remained high in Greece (17.6% in April 2019), Spain (14.0%) and Italy (9.7%).

[4] Data from: https://yourvotematters.eu. YourVoteMatters.eu is a multilingual digital platform designed as an innovative communication tool between the 2019 European elections' candidates and their electorate. The platform is developed by a consortium of five European organizations: Riparte il Futuro (Italy), VoteWatch Europe (Belgium), European Citizen Action Service (Belgium), Vouliwatch (Greece) and Collegium Civitas (Poland) with the aim of enhancing the dialogue between all the actors involved in the elections (politicians, political parties, citizens, organizations and stakeholders).

[5] These are: European People's Party (EPP), Progressive Alliance of Socialists and Democrats (S&D), Renew Europe (Renew), Greens-European Free Alliance (Greens–EFA), Identity and Democracy (ID), European Conservatives and Reformists (ECR) and European United Left–Nordic Green Left (GUE-NGL).

[6] We have excluded from the analysis in total seven issues related mostly to external relations of the EU with US, China, Russia, etc. or specific economic measures.

reflects the partitioning of the subjects into groups. Furthermore, for each group, the contribution of each question (variable) to the group formation was investigated, in order to reveal a typology of behavioural patterns.

In the second step, the group membership variable, obtained from the first step, was jointly analysed with the existing variables via Multiple Correspondence Analysis (MCA) on the so-called Burt table (Greenacre 2017). The Burt table is a symmetric, generalized contingency table, which cross-tabulates all variables against each other. The main MCA output is a set of orthogonal axes or dimensions, which summarize the associations between variable categories into a space of lower dimensionality, with the least possible loss of the original information contained in the Burt table. HCA is then applied on the coordinates of variable categories on the factorial axes. Note that this is now a clustering of the variables, instead of the subjects. The groups of variable categories can reveal complex discourses. Bringing the two analyses together, behavioural patterns and complex discourses are used to construct a semantic map for the variables and the subjects.

ISSUES	1st GROUP	2nd GROUP	3rd GROUP	4th GROUP	5th GROUP
Economy A1		1	2		0
Economy A2		1		2	0
Law & order A4	1		0		2
Immigr A5	1		0	2	
Immigr A6	0		1	2	
Law & order A7		1		2	0
Environ.A8		1		2	0
Environ A10	1		0	2	
Economy A11		1		2	0
Economy A12		1/2			0
Economy A13	1		0	2	
Economy A14	1		0	2	
Economy A15		1		2	0
Economy A17		1		2	0
Economy A18		1		2	0
Institutional A19	1		0	2	
Institutional A20		1		2	0
Institutional A21		0		2	1

Fig. 1 Operationalization

From the hierarchical cluster analysis, as presented in Fig. 1, the main result is that all the issues are split into five subgroups. More specifically, the first and second group represents mainly an agreement with most of the issues, regardless if the issue is about economy, institutional issues, etc. On the contrary, the last group (5th) is in contrast with the second one, while the third one mainly with the first one. Finally, the fourth group is the one that includes parties that do not express in most of the cases any opinion on the various issues. In the next section of the presentation of the main findings, we try to answer the question of whether the party family plays any role for this variation.

4 Findings

In order to include party family in our analysis, we have grouped the total 148 national parties of the dataset in the respective EU party groups.

As we can see in Fig. 2, a main finding from our analysis is that the parties belonging to the centre-left and left constitute one group (node 303) which is distinguished from right-wing parties (mainly node 302) on the issues presented in Figs. 4, 5, 6, 7 and 8 in the Appendix. Moreover, the parties belonging to the EPP and RenewEurope (node 302) do not share the same views on these issues with parties that are classified to the groups of ECR and ID (node 304). On the other hand, the last group (node 305) is more difficult to interpret since it includes a mixture of left-wing, right-wing parties and those parties which do not belong to any political group. Finally, the first group is mainly composed of parties that belong to GUE/NGL and those which do not belong to any group (node 300). Overall, from this cluster analysis, we might conclude that the left-right dimension is a salient characteristic of the political competition at the occasion of the 2019 EP elections. However, as aforementioned, we have also identified two clusters that include parties belonging to different ideological party families. This aspect underlines the importance of the left-right divide at least regarding some issues. Therefore, a next step in our analysis is pivotal to link parties and items.

CLUSTER PARTIES (per party family)	300	302	303	304	305	Total
EPP		**69.2%**	5.1%	5.1%	**20.5%**	**39**
S&D	3.2%	3.2%	**87.1%**		6.5%	**31**
ECR	7.7%	7.7%	0.0%	**76.9%**	7.7%	**13**
RENEW EUROPE	7.7%	**69.2%**	7.7%	3.8%	11.5%	**26**
GUE/NGL	**16.7%**		**66.7%**		**16.7%**	**12**
GREENS/EFA	10.0%	5.0%	**80.0%**		5.0%	**20**
ID	12.5%	12.5%		**75.0%**	0.0%	**8**
NI	**33.3%**		16.7%	**33.3%**	**16.7%**	**6**
Total	7.1%	31.6%	36.1%	13.5%	11.6%	**155**

Fig. 2 Parties clustering