CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE UNIVERSITY PRESS

PROCEEDINGS E REPORT
ISSN 2704-601X (PRINT) - ISSN 2704-5846 (ONLINE)

SCIENTIFIC PROGRAM COMMITTEE

Giovanni C. Porzio (chair) (University of Cassino and Southern Lazio - Italy)

Silvia Bianconcini (University of Bologna - Italy)
Christophe Biernacki (University of Lille - France)
Paula Brito (University of Porto - Portugal)
Francesca Marta Lilja Di Lascio (Free University of Bozen-Bolzano - Italy)
Marco Di Marzio ("Gabriele d'Annunzio" University of Chieti-Pescara - Italy)
Alessio Farcomeni ("Tor Vergata" University of Rome - Italy)
Luca Frigau (University of Cagliari - Italy)
Luis Ángel García Escudero (University of Valladolid - Spain)
Bettina Grün (Vienna University of Economics and Business - Austria)
Salvatore Ingrassia (University of Catania - Italy)
Volodymyr Melnykov (University of Alabama - USA)
Brendan Murphy (University College Dublin -Ireland)
Maria Lucia Parrella (University of Salerno - Italy)
Carla Rampichini (University of Florence - Italy)
Monia Ranalli (Sapienza University of Rome - Italy)
J. Sunil Rao (University of Miami - USA)
Marco Riani (University of di Parma - Italy)
Nicola Salvati (University of Pisa - Italy)
Laura Maria Sangalli (Polytechnic University of Milan - Italy)
Bruno Scarpa (University of Padua - Italy)
Mariangela Sciandra (University of Palermo - Italy)
Luca Scrucca (University of Perugia - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)
Mariangela Zenga (University of Milan-Bicocca - Italy)

LOCAL PROGRAM COMMITTEE

Carla Rampichini (chair) (University of Florence - Italy)

Chiara Bocci (University of Florence - Italy)
Anna Gottard (University of Florence - Italy)
Leonardo Grilli (University of Florence - Italy)
Monia Lupparelli (University of Florence - Italy)
Maria Francesca Marino (University of Florence - Italy)
Agnese Panzera (University of Florence - Italy)
Emilia Rocco (University of Florence - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)

CLADAG 2021
BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

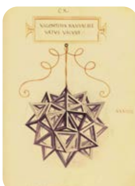
FIRENZE UNIVERSITY PRESS
2021

CLADAG 2021 BOOK OF ABSTRACTS AND SHORT PAPERS : 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021/ edited by Giovanni C. Porzio, Carla Rampichini, Chiara Bocci. — Firenze : Firenze University Press, 2021.
(Proceedings e report ; 128)

https://www.fupress.com/isbn/9788855183406

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-340-6 (PDF)
ISBN 978-88-5518-341-3 (XML)
DOI 10.36253/978-88-5518-340-6

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs
Front cover: Illustration of the statue by Giambologna, *Appennino* (1579-1580) by Anna Gottard



CLAssification and Data
Analysis Group (CLADAG)
of the Italian Statistical
Society (SIS)

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

This book is printed on acid-free paper
Printed in Italy

INDEX

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

Contributed Papers

ALI-MIKHAIL-HAQ COPULA TO DETECT LOW CORRELATIONS IN HIERARCHICAL CLUSTERING

F. Marta L. Di Lascio[1], Andrea Menapace[2], and Roberta Pappadà[3]

[1] Faculty of Economics and Management, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy, (e-mail: marta.dilascio@unibz.it)

[2] Faculty of Science and Technology, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy, (e-mail: andrea.menapace@unibz.it)

[3] Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Italy, (e-mail: rpappada@units.it)

ABSTRACT: In this work we introduce a new dissimilarity measure based on the Ali-Mikhail-Haq copula, motivated by the empirical issue of detecting low correlations and discriminating variables with very similar rank correlation. This issue arises from the analysis of panel data concerning the district heating demand of the Italian city Bozen-Bolzano. In the hierarchical clustering framework, we empirically investigate the features of the proposed measure and compare it with a classical dissimilarity measure based on Kendall's rank correlation.

KEYWORDS: Ali-Mikhail-Haq copula; cluster analysis; dissimilarity measure; low correlation.

1 Introduction

Copula-based measures of association have been employed in clustering procedures in a variety of applied contexts, since they allow to describe complex multivariate dependence structures and address specific features of the joint distribution of random variables, such as asymmetries and tail dependence (Durante & Sempi, 2015). For instance, the copula approach made it possible to define pairwise dissimilarities in terms of concordance or tail dependence measures (see, e.g., Fuchs *et al.*, 2021, and the references therein).

While many contributions in this context have focused on detecting high association between extremely low/high values, in this paper we focus on modeling weak correlation and the ability to discriminate objects with low and very similar degree of dependence. This issue comes from the features of the district heating (DH hereafter) demand from residential users of the Italian city of Bozen-Bolzano. We thus propose a new dissimilarity measure based on the

Ali-Mikhail-Haq (AMH hereafter) copula, and empirically compare it with a classical dissimilarity measure based on Kendall's τ coefficient.

The contribution is organized as follows. First, we introduce the copula-based dissimilarity measures (Sect. 2). Second, we present the cluster analysis performed to compare the proposed AMH-based dissimilarity with the one based on Kendall's τ (Sect. 3). Finally, Sect. 4 summarizes the main findings.

2 Kendall's τ- and AMH-based dissimilarity

Here, we want to perform an agglomerative hierarchical clustering (AHC hereafter) of m continuous random variables $(X_1,\ldots,X_m)$ defined on the same probability space by taking into account their stochastic dependence. A typical dissimilarity measure used in the AHC algorithm can be defined in terms of Kendall's τ coefficient as follows

$$d^{\tau}_{jj'} = \sqrt{2(1-\tau_{jj'})} \in [0,2] \tag{1}$$

where $\tau_{jj'},\ j,j' \in \{1,\ldots,m\}$, is computed from n observations of the pair $(X_j,X_{j'})$. From a different perspective, one can assume a specific copula function, motivated by its ability to capture some features of the joint behaviour observed from the data. Here we focus on the AMH copula function $C(u_1,u_2) = (u_1u_2)/(1-\theta(1-u_1)(1-u_2))$, where $\theta \in [-1,1]$. The AMH copula is very suitable for modeling low degree of association since the corresponding range for τ is $[-0.1817, 0.3333]$. Hence, we introduce a new dissimilarity measure

$$d^{\text{AMH}}_{jj'} = \sqrt{2(1-\theta_{jj'})} \in [0,2] \tag{2}$$

where $\theta_{jj'}$ is the dependence parameter of the AMH copula that can be estimated via one of the methods in the literature (see, e.g., Gunky *et al.*, 2007).

3 Application to district heating demand

We analyse time series data concerning the heat demand (in kWh) of $m = 41$ residential users connected to the DH of Bozen-Bolzano, which has been identified as a key technology for the development of sustainable cities. We consider $n = 150$ hourly observations in the period Jan 1–Jan 14, 2016. We first tackle serial dependence in the original time series by adopting a dynamic panel regression model (Wooldridge, 2002), that takes into account the relationships between DH demand and meteorological variables, such as temperature and solar radiation. Then, the residual time series are used to estimate the 41×41 dissimilarity matrices based on Eqs. (1) and (2) to use in the AHC algorithm. The crucial point is that all pairs of users have a quite low Kendall's τ (the minimum is -0.2, the highest value is 0.39). Thus, in principle, d^{AMH} should be able to better distinguish objects with low and very similar degree of association. On the basis of both the informativeness of the final clusters and the separation index by Akhanli & Hennig, 2020, we decide to adopt the complete linkage method and cut the dendrogram at $k = 3$ for both the dissimilarities.

Fig. 1 displays the mean daily pattern of each user (hourly heat demand over daily average heat demand (Menapace *et al.*, 2019)) by cluster, according to d^{τ} and d^{AMH}. As can be seen, a certain degree of internal homogeneity is obtained in both cases, denoting an overall good quality of the results. However, by using static features of the buildings, such as heating surface (in m^2), age class (in years), and energy class (in yearly kWh/m^2), we can highlight the diversity between the obtained partitions. The clusters based on d^{τ} are quite similar in terms of heating surface with median values in the range $(3656, 4076)$, and even though are better separated in terms of age class and energy class, they also present a source of variability. Indeed, the 75% of buildings in cluster 1 was built between 1961 and 1990, in cluster 2 almost the 70% of buildings is dated after 1981), while cluster 3 has a larger variability, and contains both recently-constructed and old energy-renovated buildings, with relatively low energy class (the third quartile is equal to 120). On the contrary, the d^{AMH} produces groups that are different in terms of heating surface (the medians are 3969, 5382, and 3102, respectively) and show within-homogeneity with respect to the energy and age class (e.g. buildings in cluster 3 are old, i.e. mostly dated before 1990, and non-efficient with first and third quartiles of energy class equal to 120 and 145, respectively).

4 Conclusions

We have introduced a new dissimilarity measure based on the Ali-Mikhail-Haq copula and empirically showed its ability to detect low correlations and discriminate among them. The application to district heating demand illustrates that the proposed measure seems to produce clusters that have a clear interpretation in terms of the relevant features, thus leading to a valuable tool to support the management and planning of a district heating system.

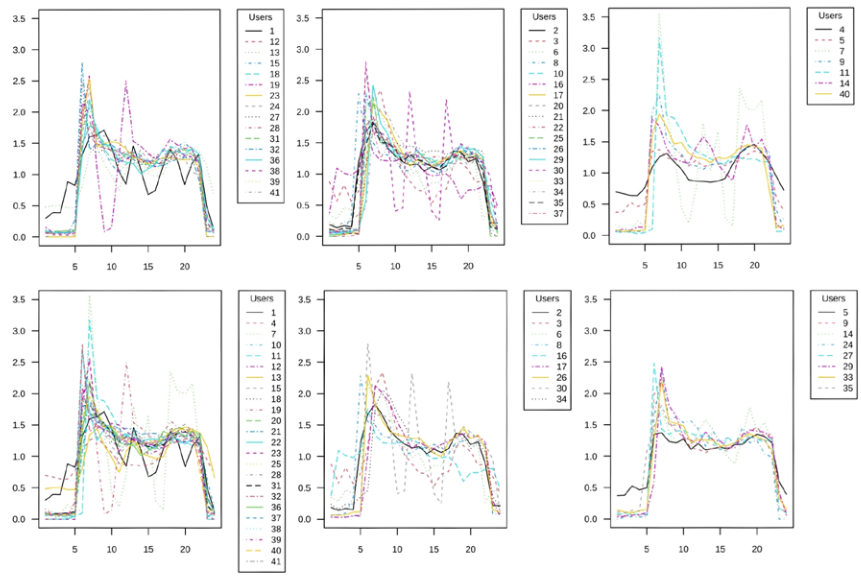


Figure 1. *Mean daily pattern (hours in x-axis) of DH users according to AHC based on* d^{τ} *and* d^{AMH} *(panels by rows) in cluster 1, 2, and 3 (panels by columns).*

References

AKHANLI, S, & HENNIG, C. 2020. Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Stat. Comput.*, **30**, 1523–1544.

DURANTE, F, & SEMPI, C. 2015. *Principles of Copula Theory*. CRC Press, Boca Raton.

FUCHS, S, DI LASCIO, F M L, & DURANTE, F. 2021. Dissimilarity functions for rank-invariant hierarchical clustering of continuous variables. *Comput. Stat. Data An.*, **159**, 107201.

GUNKY, K, SILVAPULLE, M J, & SILVAPULLE, P. 2007. Comparison of semiparametric and parametric methods for estimating copulas. *Comput. Stat. Data An.*, **51**(6), 2836–2850.

MENAPACE, A, RIGHETTI, M, SANTOPIETRO, S, GARGANO, R, & DALVIT, G. 2019. Stochastic characterisation of the district heating load pattern of residential buildings. *Euroheat and Power*, **16**(3–4), 14–19.

WOOLDRIDGE, J. 2002. *Econometrics analysis of cross section and panel data*. Cambridge: MIT Press.