HIERARCHICAL PERCENTILE CLUSTERING TO ANALYSE GREENHOUSE GAS EMISSIONS FROM AGRICOLTURE IN EUROPEAN UNION

F. Marta L. Di Lascio [2], Fabrizio Durante [1], Aurora Gatto [2]

[1] Department of Economic Sciences, University of Salento, Centro Ecotekne, 73100, Lecce, Italy, (e-mail: `fabrizio.durante@unisalento.it`)

[2] Faculty of Economics and Management, Free University of Bozen-Bolzano, Piazza Università, 1-39100, Bozen-Bolzano, Italy, (e-mail: `marta.dilascio@unibz.it`, `aurora.gatto@unibz.it`)

ABSTRACT: One of the key issues in the European Union's environmental policy concerns greenhouse gas emissions reduction, which is a challenge task to mitigate climate change. In this paper we investigate the emissions of different greenhouse gases from agriculture in the European Union countries through the innovative agglomerative hierarchical percentile clustering algorithm.

KEYWORDS: agriculture, greenhouse gas emission, hierarchical percentile clustering.

1 Introduction

Clustering methods are unsupervised techniques useful to identify structure underlying the data with the aim of gaining insights into their generating process. Durante *et al.*, 2021 proposed a method called *Agglomerative Hierachical Percentile Clustering* (AHPC) that allows us to deal with experimental errors and/or uncertainty affecting the observations. Hence, AHPC algorithm aims to cluster objects represented by repeated measurements on a set of variables.

Recently, European Union (EU) established to reduce greenhouse gas (GHG) emissions by at least 55% by 2030 compared to 1990 levels, and achieve climate neutrality by 2050. In particular, the emissions from agriculture account for about 11% of EU-27 emissions and in 2020 were about the same as in 2005. Hence, in the coming years substantial GHG emission reductions across all the sectors of the economy, including agriculture, are expected.

Our interest is to investigate the possible different behavior of EU countries in GHG emissions from agriculture through the innovative AHPC. To this aim, Sect. 2 presents the AHPC algorithm, while Sect. 3 contains the empirical analysis and a discussion of the findings. Finally, Sect. 4 concludes the paper.

2 The AHPC algorithm

The percentile clustering (PC) is a clustering method based on a dissimilarity matrix computed according to the percentile approach suggested in a seminal paper by Janowitz & Schweizer, 1989. Recently, Durante *et al.*, 2021 have developed the PC method in the hierarchical clustering framework and investigated its performance on both simulated and observed data.

Suppose to cluster d objects, i.e. statistical units, on which p different variables are observed. Also assume that each object i is associated with a set of n_i repeated observations, e.g. observations over time, and thus represented by a $(n_i \times p)$-dimensional matrix $\mathbf{X}_i$. The AHPC algorithm can be summarized as follows:

1. for each pair of objects i and j, with $i \neq j$ and $i, j = 1, \ldots, d$, i.e. for each pair of matrices $\mathbf{X}_i$ and $\mathbf{X}_j$:
   (a) compute the Euclidean distance $d_{k\ell}^{ij}$ between the k-th row of $\mathbf{X}_i$ and the ℓ-th row of $\mathbf{X}_j$ for every $k = 1, \ldots, n_i$, $\ell = 1, \ldots, n_j$;
   (b) compute the dissimilarity p_{ij} as the α-quantile of the $d_{k\ell}^{ij}$ distances computed at the step (a) by varying k and ℓ; α can obviously assume value in $[0,1]$;
2. create the $(d \times d)$-dimensional dissimilarity matrix $\mathbf{P} = (p_{ij})$ with $p_{ii} = 0$ and $p_{ij} = p_{ji}$;
3. apply the classical agglomerative hierarchical clustering algorithm (Everitt *et al.*, 2011) choosing a linkage rule among the classical ones, e.g. single, average and complete linkage, and using $\mathbf{P}$ as dissimilarity matrix.

Roughly speaking, the AHPC is a classical hierarchical clustering algorithm based on a dissimilarity matrix computed through the α-percentile of the distribution functions of Euclidean distances between each pair of considered objects. Hence, the AHPC has interesting features. Firstly, the AHPC allows us to exploit prior knowledge of each object that can be observed for a different number of times, i.e. $n_i \neq n_j$, when $i \neq j$. Secondly, the possible presence of missing values does not prevent the application of the method. Thirdly, it should be noted that the dissimilarity matrix based on the α-percentile is actually an ordered weighted aggregation function (see, e.g., Yager, 2000). However, it is important to stress that considering percentiles as dissimilarities only takes into consideration the ranks of the objects (see, e.g., Cena & Gagolewski, 2020).

3 Case study

In order to illustrate the usefulness of the AHPC method, we analyze the GHGs emissions from agriculture. We consider the methane-CH4, carbon dioxide-CO2 and nitrous oxide-N2O per capita emissions (expressed in CO2 equivalent) for the EU countries (excluding Malta due to lack of available CO2 data) over the period 2012-2020 (source: European Environment Agency). We standardise each variable to ease the interpretation of findings. Next, we apply the AHPC algorithm using the complete linkage and the percentile level $\alpha = 0.75$ as suggested by the Monte Carlo simulation results in Durante *et al.*, 2021. Ignoring the two-cluster solution that is low informative due to the presence of a singleton cluster, i.e. Ireland, we consider the second highest value of the Average Silhouette Width (Rousseeuw, 1987) (0.346) that suggests a partition in five clusters.

To interpret the obtained results we consider the boxplots of each analysed GHG variable according to the obtained partition (see Fig. 1). We do not represent the fifth cluster because it is only formed by Ireland whose average emissions are extremely higher than those of the other EU countries for all the considered gases (average values of CH4, CO2, and N2O are 4.632, 3.790, and 3.354, respectively). It appears that countries belonging to the second cluster are the ones with lowest GHGs emissions per capita, while the first and the third cluster only show a lower-than-average emissions per capita for two of the three considered gases. Finally, the countries in the fourth cluster are the least virtuous, with emission values for each gas considered higher than the corresponding mean value for all EU countries. It is interesting notice that the first cluster is those with the highest number of outliers, meaning that there are some countries with extreme GHG emissions per capita.

4 Conclusions

We have analysed the GHG emissions per capita in the EU, using the innovative AHPC algorithm. We have found that the GHG emissions are different across countries and kind of gas. This supports the need of adopting a common policy to reach the EU goal of reducing GHG emissions.

Acknowledgments

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). In particular,

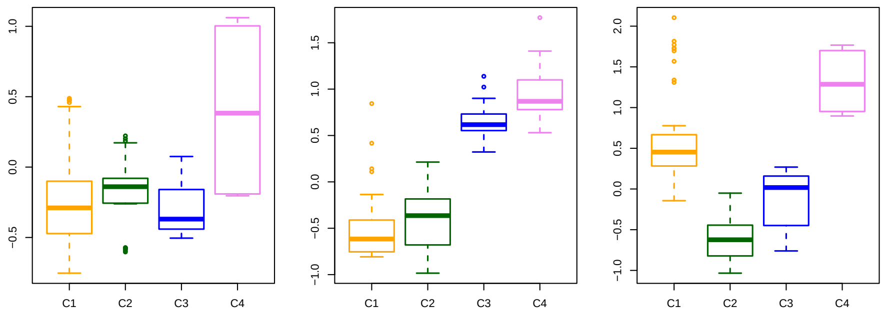

Figure 1. *Greenhouse gases emissions of CH4 (left panel), CO2 (middle panel), and N2O (right panel) by varying clusters from C1 to C4.*

our study represents an original paper related to the Spoke 4 "Multifunctional and resilient agriculture and forestry systems for the mitigation of climate change risks" and a baseline for the fulfilment of the milestones within the Task 4.2.3. "Big data analysis and decision support systems for the climate adaptation of agricultural and forestry systems". This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

FD has been supported by MIUR-PRIN 2017, Project "Stochastic Models for Complex Systems" (No. 2017JFFHSH). FD is also member of "ICSC - Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing".

References

CENA, ANNA, & GAGOLEWSKI, MAREK. 2020. Genie+OWA: Robustifying hierarchical clustering with OWA-based linkages. *Inf. Sci.*, **520**, 324–336.

DURANTE, FABRIZIO, GATTO, AURORA, & SAMINGER-PLATZ, SUSANNE. 2021. On Agglomerative Hierarchical Percentile Clustering. *Pages 616–623 of: Joint Proceedings of the 19th World Congress of IFSA, the 12th Conference of EUSFLAT, and the 11th International Summer School on AGOP*.

EVERITT, B. S., LANDAU, S., LEESE, M., & STAHL, D. 2011. *Cluster analysis*. Fifth edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

JANOWITZ, M. F., & SCHWEIZER, B. 1989. Ordinal and percentile clustering. *Math. Soc. Sci.*, **18**(2), 135–186.

ROUSSEEUW, P.J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. of Comp. and Appl. Math.*, **20**, 53–65.

YAGER, R.R. 2000. Intelligent control of the hierarchical agglomerative clustering process. *IEEE Trans. Syst. Man Cybern. Syst., Part B (Cybernetics)*, **30**(6), 835–845.