unibz
Fakultät für Wirtschaftswissenschaften
Facoltà di Economia
Faculty of Economics and Management

BEMPS –
Bozen Economics & Management Paper Series

NO 89/ 2021

A spatially-weighted AMH copula-based dissimilarity measure to cluster variables in panel data

F. Marta L. Di Lascio, Andrea Menapace, Roberta Pappadà

A spatially-weighted AMH copula-based dissimilarity measure to cluster variables in panel data

F. Marta L. Di Lascio [*] Andrea Menapace [†] Roberta Pappadà [‡]

Abstract

Investigating thermal energy demand is crucial for the development of sustainable cities and efficient use of renewable sources. Despite the advances made in this field, the analysis of energy data provided by smart grids is currently a demanding challenge due to their complex multivariate structure and high-dimensionality. In this paper, we develop a clustering methodology based on a novel copula-based dissimilarity measure suitable for analyzing a high temporal resolution panel data for district heating demand. Inspired by the characteristics of this data, we explore the usefulness of the Ali-Mikhail-Haq copula in defining a new dissimilarity measure to cluster variables in a hierarchical framework. We show that our proposal is particularly sensitive to small dissimilarities based on tiny differences in the dependence level. Therefore, the measure we introduce is able to better distinguish between objects with low dissimilarity than classic rank-based dissimilarity measures. Moreover, our proposal considers a weighted version of the copula-based dissimilarity that embeds the spatial location of the involved data objects. We investigate the proposed measure through Monte Carlo studies and compare it with the corresponding Kendall's correlation-based dissimilarity measure. Finally, the application to real data concerning the Italian city Bozen-Bolzano makes it possible to find clusters of buildings homogeneous with respect to their main characteristics, such as energy efficiency and heating surface, to support the design, expansion and management of district heating systems.

Keywords: Ali-Mikhail-Haq copula, Cluster analysis, Dissimilarity measure, District heating demand, Panel data, Spatial weight.

Jel Codes: C10, C33, C38

[*]Faculty of Economics and Management, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy. Email: marta.dilascio@unibz.it

[†]Faculty of Science and Technology, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy. Email: andrea.menapace@unibz.it

[‡]Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Italy. Email: rpappada@units.it

1 Introduction

Understanding thermal consumption in urban areas is a crucial need to increase the sustainability and efficiency of energy systems and reduce world climate change [Lund et al., 2014]. Renewable energy systems require a fully reshape of the traditional infrastructure and a rethink of the technologies involved [Lund et al., 2018]. District heating (DH hereafter) is one of the key technologies involved in the ongoing process aimed at developing sustainable cities and improving the efficiency of the heating sector. Indeed, DH is defined as an energy distribution system that provides heat through a network of pipes to buildings in a neighborhood or a town by incorporating renewable sources and reducing waste of energy in a flexible urban energy system [Frederiksen and Werner, 2013].

Developing stochastic methods to analyze high frequency DH energy data provided by smart grids is currently a demanding challenge (see, e.g. Sharma and Saini [2015], Ma et al. [2017]). In particular, there is a need for an in-depth analysis of heating data to enhance the management and planning of the heating system and the efficient use of renewable energy sources (see, e.g. Menapace et al. [2021]). In this context, clustering methods enable the investigation of the structure underlying the data generating process (DGP hereafter), serving as the basis for further learning, such as forecasting and anomaly detection. Specifically, the identification of DH users that are similar according to relevant characteristics contributes to efficiently plan the DH and manage heat production and distribution.

In the hierarchical agglomerative clustering framework [Everitt et al., 2011], the core idea is to construct the hierarchical relationship among the objects to be grouped starting from a set of clusters each containing a single object to a single cluster containing all the objects [Kaufman and Rousseeuw, 1990]. Hierarchical clustering requires a pairwise dissimilarity measure to compare singletons and a linkage rule to compare clusters. The most widely used linkage rules are the average, the complete, and the single. The literature on hierarchical clustering methods is extensive and applications have been successfully performed in various contexts (see, e.g., Bengtsson and Cavanaugh [2008], Nguyen [2016], Alvarez-Esteban et al. [2016], Di Lascio et al. [2018]). In clustering random variables (r.v.s hereafter), copula-based measures of association have been used in a variety of application contexts (see, e.g., Nazemi and Elshorbagy [2012], Di Lascio et al. [2017], Pappadà et al. [2018] and De Luca and Zuccolotto [2021]), as they allow describing complex dependence structures and addressing specific features of the joint distribution of r.v.s, such as asymmetries and tail dependence [Durante and Sempi, 2015]. Indeed, copula models allow us to describe the dependence structure of the DGP separately from the marginal distributions, yielding a much greater degree of flexibility in specifying and estimating the dependence relationship. For instance, the copula approach makes it possible to define pairwise dissimilarities as well as multivariate dissimilarities in terms of concordance or tail dependence measures (see, e.g., Kojadinovic [2010], Durante et al. [2015], De Luca and Zuccolotto [2017], Bonanomi et al. [2019], Fuchs et al. [2021]).

While many contributions in the context of clustering r.v.s have focused on detecting a strong association between extreme values (see, e.g., Durante et al. [2014] and Côté and Genest [2015]),

this paper focuses on the ability to differentiate r.v.s characterized by a low level of dependence and small dissimilarities according to the features of the DH demand data analysed in this work. As discussed by Kruskal [1977], cluster analysis is appropriate to extract information from small dissimilarities. Here, we analyze hourly panel data concerning the thermal energy demand of residential users in the Italian city of Bozen-Bolzano in 2016. To this aim we explore the potential of the Ali-Mikhail-Haq (AMH hereafter) copula [Ali et al., 1978] to cluster r.v.s in the agglomerative hierarchical clustering context, proposing a new AMH copula-based dissimilarity measure to investigate the theoretical and applied properties. Since the most used copula-based dissimilarity measures involve Kendall's τ correlation coefficient, we empirically compare the performance of the proposed measure with the corresponding version based on Kendall's τ through Monte Carlo studies.

As mentioned above, the theoretical contribution of this paper is applied to panel data. In the context of time series data analysis, hierarchical clustering algorithms exploiting copula-based dissimilarity measures have been used to detect the co-movements of r.v.s (see, e.g., De Luca and Zuccolotto [2011], Disegna et al. [2017], Reddy and Ganguli [2013]). Extensions of these approaches, considering both temporal and cross-sectional dependence via copulas, can be found in, e.g., Yi and Liao [2010], Rémillard et al. [2012], but to the best of our knowledge, there are no methodological procedures dedicated to panel data analysis, which is our focus. Hence, the proposed AMH copula-based dissimilarity measure is exploited in the development of a procedure for clustering panel data. While some studies use copulas in the field of DH demand (see, e.g., Di Lascio et al. [2020], Di Lascio et al. [2021]), copula-based clustering has not yet been developed – or only marginally – in relation to energy or the more general environmental sciences field (see, e.g., Luo et al. [2019], Just and Łuczak [2020]).

The remainder of the paper is organized as follows. We define a new dissimilarity measure and present its theoretical properties in Section 2. In Section 3, we compare our proposal with a classic dissimilarity measure through a Monte Carlo simulation study and discuss the advantages and limitations of the new dissimilarity measure. We then illustrate a clustering methodology based on the proposed dissimilarity via the application to panel data in Section 4. Section 5 highlights the most relevant implications and summarizes our main findings, relating the more technical mathematical results to the Appendix A.

2 AMH copula-based dissimilarity measure

Copulas originated in the context of probabilistic metric spaces via Sklar's theorem Sklar [1959] stating that a copula $C(\cdot)$ is a joint distribution function with uniform margins. The advantages of the copula-based approach in contexts where dependence is relevant are well known, since copulas potentially enable describing any kind of complex multivariate dependence structure of the DGP, such as non-linear and non-Gaussian relations, heavy tails, and asymmetries [Durante and Sempi, 2015]. In the literature, a myriad of copula models have been proposed, each able to describe a particular dependence pattern. Here we focus on the Ali-Mikhail-Haq copula function

that has been introduced by [Ali et al., 1978] and whose statistical properties have been studied by Kumar [2010]:

$$C^{\text{AMH}}(u_1, u_2) = \frac{u_1 u_2}{1 - \theta^{\text{AMH}}_{u_1 u_2}(1 - u_1)(1 - u_2)} \tag{1}$$

where $\theta^{\text{AMH}}_{u_1,u_2} \in [-1, 1[$ is its dependence parameter whose domain in terms of Kendall's τ coefficient is $[-0.1817, 0.3333[$. Thus, the AMH copula function can be used to describe both positive and negative correlation of r.v.s, even though it is not suitable for very high positive or negative correlations. The dependence parameter of the AMH copula can be estimated using the estimation methods available in the literature (see, e.g., Cherubini et al. [2004]).

In the hierarchical clustering context copula has been largely used to define dissimilarities in terms of measures of association (see, e.g., [Fuchs et al., 2021] and references therein). Here, the decision on which clusters should be merged is based on the dissimilarity between two objects and a linkage rule specifying the dissimilarity between two clusters of objects. Such linkage is usually a function of the pairwise dissimilarities of objects in the clusters. In the light of the empirical data features, our purpose is twofold: on the one side, we need to take into account the spatial location of objects to compare and, on the other side, define a dissimilarity measure able to differentiate objects with low and very similar dependence. Hence, we propose the following measure based on the AMH copula that takes into account the spatial information of objects:

$$d^{\text{AMH}}_{jj'} = c_{jj'} \sqrt{2(1 - \theta^{\text{AMH}}_{jj'})} \tag{2}$$

where $c_{jj'} = \exp(g_{jj'} / \max(G)) - \delta_{jj'}$, with $\delta_{jj'} = 0 \, \forall j \neq j'$, and 1 otherwise, and $G = (g_{jj'})$ is the spatial weights matrix that can be calculated starting from the geographic distance (based on longitude and latitude information) of all the pairs (j, j') with $j, j' = 1, \ldots, p$. Such a weighting scheme emphasizes the dissimilarity of objects that are further apart. Moreover, the measure in Eq. (2) is a dissimilarity measure, since it satisfies the two properties of a dissimilarity measure whose proof is trivial:

$$\begin{aligned} &\text{P1.} \quad d^{\text{AMH}}_{jj'} \geq 0 \quad \forall \, j, j', \quad \text{and} \quad d^{\text{AMH}}_{jj'} = 0 \quad \text{if and only if } j = j' \\ &\text{P2.} \quad d^{\text{AMH}}_{jj'} = d^{\text{AMH}}_{j'j} \quad \forall \, j, j'. \end{aligned}$$

The proposed dissimilarity measure takes values in $[0, 2\exp(1)]$ and it considers minimum dissimilarity only between variables with maximum comonotone (positive) dependence. In addition, d^{AMH} is decreasingly monotone with respect to θ^{AMH}, and this property means that the dissimilarity degree tends to vanish as soon as approaching the comonotonic (positive) case.

It is worth stressing that: (i) when $c_{jj'} = 1$, the proposed dissimilarity only depends on the association of the considered pair (j, j') and Eq. (2) maps dissimilarity values from $[1.1547, 1.5373]$ to $[0, 2]$ in light of the relationship between θ^{AMH} and Kendall's τ (see Eq. (A.1) in the Appendix A); (ii) Eq. (2) is not intended to measure spatial dependence and, thus, it is not related to dissimilarities based on spatial association or heterogeneity (see, e.g., Anselin and Rey [2010], Anselin [1995]), but it only takes into account the spatial location of the r.v.s to cluster;

(iii) the use of spatial information $c_{jj'}$ has an important effect on the dissimilarity between two objects and differentiates the AMH copula-based dissimilarity measure from the corresponding dissimilarity based on Kendall's τ correlation coefficient $d^{\tau} = c_{jj'}\sqrt{2(1-\tau_{jj'})}$ as explained in detail below.

To stress the effect of the spatial weight and the different behaviour of d^{AMH} and d^{τ}, we assume that the AMH copula model is the true model and express d^{τ} as $d^{f\left(\theta^{\text{AMH}}\right)} = c_{jj'}\sqrt{2(1-f\left(\theta^{\text{AMH}}\right))}$, where $f\left(\theta^{\text{AMH}}\right)$ is given in Eq. (A.1), Appendix A. We then mathematically analyse the two measures, and show the different behaviour of the two. In particular, we compute the difference between the two measures, and between the partial derivatives of order 1 and 2 of the two measures with respect to θ^{AMH}. The resulting mathematical expressions are in Eq.s (A.2–A.4) in Appendix A, while the shape of the three equations by varying $\theta^{\text{AMH}} \in [-1, 1[$ and $c_{jj'} \in [1, \exp(1)]$ is shown in Fig. 1. The difference between the two dissimilarities (Fig. 1, left) shows a monotonically decreasing (increasing) behaviour for negative (positive) values of dependence by varying $c_{jj'}$. The slope as well as the curvature of the plane changes with the dependence and the spatial weight. Both partial derivatives differences are monotonically increasing in θ^{AMH} and $c_{jj'}$. Especially in the difference of slopes (Fig. 1, middle), the impact of the spatial weight is different for the two dissimilarity measures; indeed, as the spatial weight increases, the difference in the slope of the two dissimilarity measures increases too. The differential increments of the two dissimilarity measures is greater than zero for all $c_{jj'} \in [1, \exp(1)]$. Hence the difference between the two considered measures is monotonically increasing and convex in $c_{jj'}$.

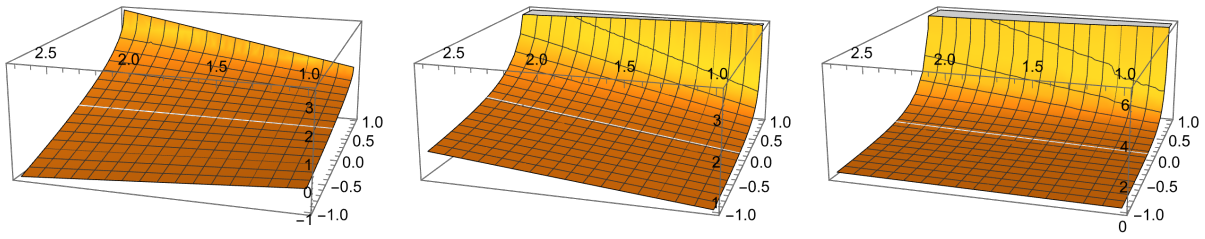

Figure 1: Comparison between d^{AMH} and $d^{f\left(\theta^{\text{AMH}}\right)}$: difference between $d^{f\left(\theta^{\text{AMH}}\right)}$ and d^{AMH} in Eq. (A.2) (left), and between first partial derivative in Eq. (A.3) (middle), and second partial derivatives in Eq. (A.4) (right) of d^{AMH} and $d^{f\left(\theta^{\text{AMH}}\right)}$ (z-axis) versus $\theta^{\text{AMH}} \in [-1, 1[$ (y-axis), and $c_{jj'} \in [1, \exp(1)]$ (x-axis).

Finally, since the parametric space of the dependence parameter θ^{AMH} tends to amplify the difference between low-rank correlations, allowing us to distinguish objects with tiny differences in dissimilarity values, the proposed measure is particularly useful when variables exhibit low

dependence and the dissimilarity values show homogeneity. As will be clear in the empirical application in Section 4, this also results in a dendrogram that is less flattened and dense, i.e., with a wide distance between clusters so that a later fusion takes place at a higher level of dissimilarity than the previous one. Hence, the hierarchy of clusters is better highlighted, improving the interpretation and cutting of the dendrogram.

3 Monte Carlo study

Here, we provide a simulation study to assess the goodness of the proposed dissimilarity measure in Eq. (2) with respect to the weighted Kendall-based dissimilarity measure d^{τ}. In particular, we want to investigate the ability of d^{AMH} to discriminate objects with small and close correlation values, also taking into account the spatial information. To this end, we consider five different scenarios of a three-dimensional DGP based on copulas differing from the AMH (see Table 1) and we generate $K = 3$ independent samples, each representing a cluster generated from a specific copula model. Inspired by the case study analysed in Sect. 4 we set all the dependence parameters to small values and we generate $n = 150$ realizations of $p = 41$ r.v.s (which can be interpreted, for instance, as serially uncorrelated time series). The cluster size p_k (with $k = 1, \ldots, K$) is randomly chosen from 2 to $(41 - (K + 1))$ to ensure that each cluster has at least 2 elements and the size of the whole clustering is p. The five considered scenarios are simulated by using different settings for spatial information. In particular, the weights are computed by using the exponential form described in Section 2, where $g_{jj'}$ is the distance between the two points j and j' computed according to the generated coordinates. We consider two different settings for the geographic position of points. In one case, we generate points in the plane in such a way that one cluster of points is clearly distant from the other two that conversely show some overlap: we use the following cluster centers $(100, 100)$, $(500, 300)$, and $(600, 200)$ to generate points by adding a random noise distributed as $\mathcal{N}(0, 100)$ and each cluster size is chosen randomly as described above. Here, $g_{jj'}$ is the Euclidean distance between the simulated plane coordinates. In the other case, we compute the weights starting from the geographic positions on the WGS ellipsoid of the points observed in the panel data application described in the section 4 adding a uniform random noise. We therefore simulate 10 different scenarios, and for each, perform 500 Monte Carlo replications.

Table 1: Simulated scenarios used in the Monte Carlo study.

Scenario	Cluster 1	Cluster 2	Cluster 3
1	Clayton, $\tau = 0.05$	Clayton, $\tau = 0.15$	Clayton, $\tau = 0.25$
2	Gumbel, $\tau = 0.25$	Frank, $\tau = 0.1$	Clayton, $\tau = 0.2$
3	Gumbel, $\tau = 0.2$	Frank, $\tau = 0.2$	Clayton, $\tau = 0.2$
4	Clayton, $\tau = 0.2$	Clayton, $\tau = 0.2$	Clayton, $\tau = 0.2$
5	Gumbel, $\tau = 0.2$	Gumbel, $\tau = 0.2$	Gumbel, $\tau = 0.2$

To measure the performance of d^{AMH} and d^{τ}, we compute (i) the Adjusted Rand Index [Hubert and Arabie, 1985] (ARI hereafter) to assess the agreement between the partitions obtained using the two compared measures given the true number of clusters and (ii) the agglomerative coefficient [Kaufman and Rousseeuw, 1990] (AC hereafter) as the average width of the banner [Rousseeuw, 1986] describing the strength of the clustering structure to assess the overall quality of the dendrogram. The distribution of ARI for each simulated scenario is shown in Fig. 2. Here, the partition is obtained by cutting the dendrogram so that three clusters are identified. It is evident that the proposed spatially-weighted AMH copula-based dissimilarity measure provides partitions very different from those obtained using the spatially-weighted Kendall-based dissimilarity, irrespective of the scenario, the linkage rule, and the spatial weights. It is interesting to note that the role of the spatial weights is crucial and negatively affects the agreement between the two measures when the weights are empirically computed, and therefore not related to the simulated within-cluster dependence. The resulting ARI values support the already theoretically discussed differences between d^{AMH} and d^{τ} justifying the use of the AMH copula dependence parameter as an alternative to the Kendall's coefficient. The AC distribution for each simulated

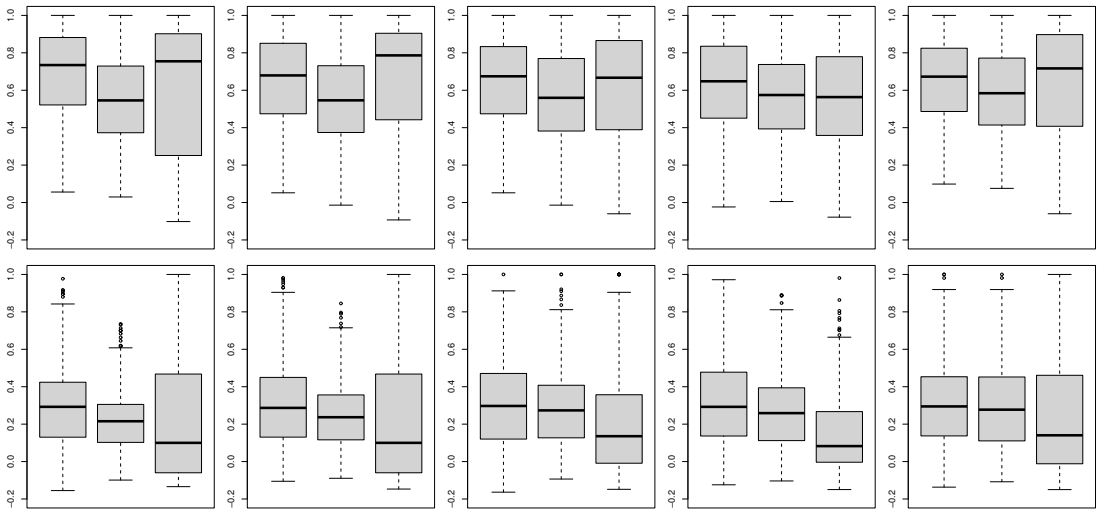

Figure 2: Boxplots of ARI (y-axis) comparing the partitions obtained through d^{AMH} and d^{τ} with $K = 3$ by varying i) the linkage method between the average, the complete (maximum), and the single (minimum) (x-axis), ii) the scenario among the five given in Table 1 (panels by columns), and iii) the spatial settings among random weights and empirical weights plus a random noise (panels by rows) - see text for details. Sample size is $n = 150 \times p = 41$. The number of Monte Carlo replications is 500.

scenario is shown in Fig. 3. According to Kaufman and Rousseeuw [1990] an AC close to 1 indicates that tight clusters that are far away from each others, i.e. a very clear clustering structure, have been identified. Instead, when the AC is close to zero, "the data set does not contain very natural clusters which would have been formed sooner [...] as all dissimilarities between objects

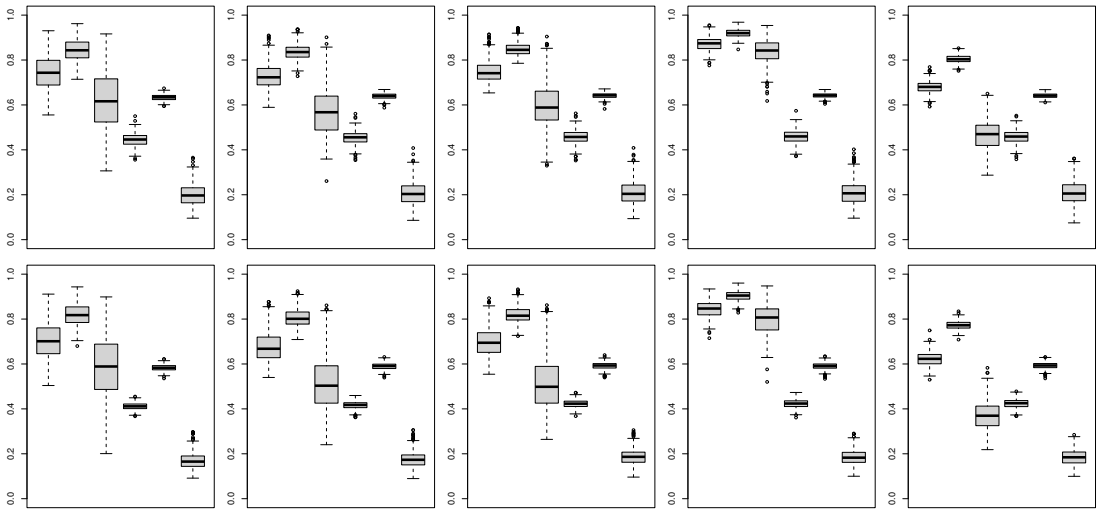

Figure 3: Boxplots of AC (y-axis) by varying i) the pairwise dissimilarity measure between d^{AMH} and d^{τ}, ii) the linkage method between the average, the complete (maximum), and the single (minimum) (x-axis starting with the average linkage and d^{AMH}, continuing with the complete linkage and d^{AMH}, and ending with the single linkage and d^{τ}), iii) the scenario among the five provided in Table 1 (panels by columns), and iv) the spatial settings among random weights, and empirical weights plus a random noise (panels by rows) - see text for details. Sample size is $n = 150 \times p = 41$. The number of Monte Carlo replications is 500.

are of the same order of magnitude" [Kaufman and Rousseeuw, 1990]. Here, it is evident that the proposed dissimilarity measure outperforms the measure based on Kendall's τ irrespective of the scenario, the linkage rule, and the setting of spatial weights. The complete linkage appears to be better than the average and the single linkages, and the use of spatial information appears to have a positive but mild effect on the performance of the proposed measure.

4 Application to panel data

4.1 District heating system and thermal energy demand

In this section, we describe the data concerning the thermal consumption of the residential users connected to the DH of the Italian city Bozen-Bolzano. The heating demand of Bozen-Bolzano is partially supplied by a DH system that is in constant expansion to sustain the municipality's climate actions [Menapace et al., 2020]. The Bozen-Bolzano DH concerns a network of about $20\ Km$ pipes, a centralized production center mainly based on a waste-to-energy plant, 220 $MW\ h$ thermal storage, and more than 200 heat exchanger substitutions [Menapace et al., 2019]. Each substation is endowed with a smart heat meter that provides high frequency and accurate resolution data used by operators to monitor the system.

Here we use time series of the thermal energy demand (TED hereafter, in $kW\ h$) of 41 residential users (i.e., one or more buildings with homogeneous characteristics fed by one or more DH substations) connected to the Bozen-Bolzano DH during the winter week from 08/01/2016 to 14/01/2016 (see Fig. 4). We also use the time series of meteorological data, such as outdoor temperature (TEMP hereafter, in $°C$) and solar radiation (RAD hereafter, in W/m^2) provided by the S. Maurizio weather station. The meteorological data, indeed, present significant dependence on heating demand and can help the proper modelling of the TED panel data Soutullo et al. [2016]. The observed time series have been pre-processed to remove outliers due to meter or transmission system failures, and then aggregated to obtain hourly observations.

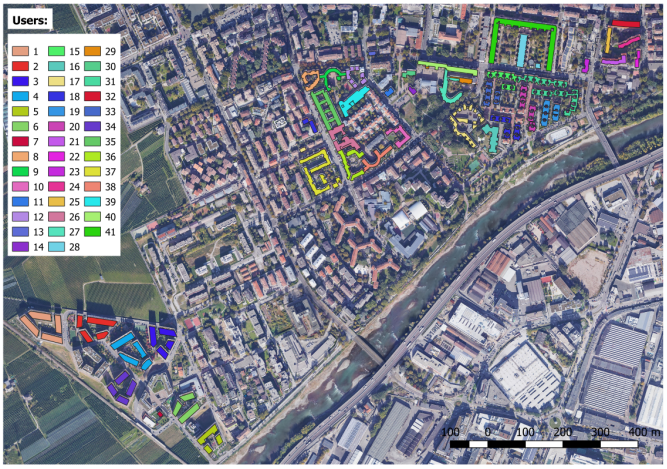

Figure 4: Map of the sample of users in the different districts fed by the Bozen-Bolzano DH.

The final aim of this application is to identify and characterize clusters of homogeneous buildings with respect to the behavior of TED. Therefore, the aim of the cluster analysis is to provide useful information to improve the efficiency and sustainability of the DH of Bozen-Bolzano through a proper schedule of the heat production and management of the network and the thermal reservoir. For instance, consider two users with clearly different behaviors: Fig. 5 (top) represents the typical heating profile of a new or renovated building with continuous operation control that maintains the indoor temperature constant throughout the entire day with morning and evening peaks; Fig. 5 (bottom) corresponds to a typical non-renovated building with a night setback control that leads to null demand during the night and a sharp peak in the early morning. To verify and assess the quality of the clustering results, the following additional information is used: heating surface (in dam^2), energy class (in $kW\ h/m^2/year$), age category (ranges from class 1, the oldest for buildings before 1918, to class 9, the newest for buildings after 2005), and mean yearly heat consumption (in $MW\ h/year$).

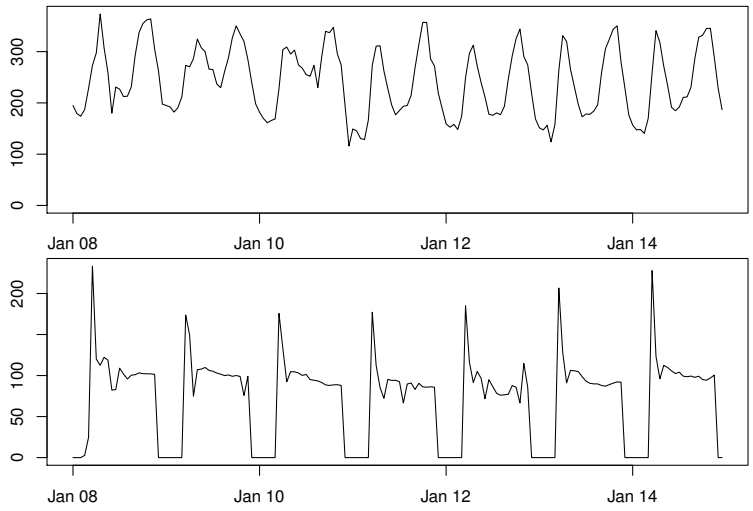

Figure 5: Time series of TED in $kW\,h$ (y-axis) of two typical users.

4.2 Clustering methodology

In this section we develop the panel data clustering procedure with the aim of finding clusters of DH residential users. The clustering methodology is based on the dependence between TED time series of each user. To consider both temporal and cross-sectional dependence, we extend the copula-based approach that has been already used for time series modeling (see, e.g., Patton [2012]) to the panel data case. To do that we first tackle serial dependence through a suitable panel regression model (see, e.g., Baltagi [1995], Wooldridge [2002]) and, next, model cross-sectional dependence between the residuals time series by applying the proposed measure in Eq. (2) in the hierarchical clustering framework. Hence, we estimate a dynamic panel regression model to the whole data set of $p = 41$ variables and $n = 150$ observations that takes into account the effect of (lagged and not) meteorological variables on TED, as well as the serial dependence of TED and individual effects μ_i, with $i = 1, \ldots, 41$. The following specified model derives from the preliminary analysis of the TED, TEMP, and RAD time series (together with their autocorrelation and partial autocorrelation functions) and a forward selection based on significant covariates:

$$\begin{aligned} \text{TED}_{it} &= \rho_1 \text{TED}_{i(t-1)} + \rho_2 \text{TED}_{i(t-24)} + \beta_1 \text{RAD}_{it} + \beta_2 \text{RAD}_{i(t-1)} + \beta_3 \text{TEMP}_{it} + \\ &\quad + \beta_4 \text{TEMP}_{i(t-3)} + u_{it} \\ &= \rho_1 \text{TED}_{i(t-1)} + \rho_2 \text{TED}_{i(t-24)} + \beta_1 \text{RAD}_{it} + \beta_2 \text{RAD}_{i(t-1)} + \\ &\quad + \beta_3 \text{TEMP}_{it} + \beta_4 \text{TEMP}_{i(t-3)} + \mu_i + \varepsilon_{it} \end{aligned} \tag{3}$$

where $i = 1, \ldots, 41$, $t = 1, \ldots, 150$, ρ_1, ρ_2, β_1, β_2, β_3, and β_4 are scalar, u_{it} is assumed to follow a one-way error component regression model with $\mu_i \sim N(0, \sigma_\mu^2)$ and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$, which are independent of each other and among themselves. Since TED_{it} is a function of μ_i, it follows that

$\text{TED}_{i(t-1)}$ is also a function of μ_i. Therefore, $\text{TED}_{i(t-1)}$ is correlated with the error term, and we use a set of instrumental variables, i.e., TED lagged from $(t-3)$ to $(t-24)$, to account for it and compute the estimation through the Arellano and Bond one-step generalized method of moments Arellano and Bond [1991].

Once the model in Eq. (3) has been estimated, the residuals of the 41 time series are extracted, and the weighted AMH copula-based dissimilarity is computed as in Eq. (2) where the 41×41 matrix of spatial weights is constructed by adopting the exponential form and the distance on the WGS ellipsoid as illustrated in Section 3. We note that the residuals show low and very similar linear Kendall's correlation ranging in $(-0.207, 0.394)$. Only two values lie outside the Kendall's τ range for the AMH copula that have been replaced with the (maximum or minimum) extreme of the range. Thus, a typical dissimilarity measure based on Kendall's τ correlation coefficient may be not able to discriminate among them, unlike d^{AMH}, which appears to be more sensitive to small correlations. Moreover, the spatial weights provide useful information about the buildings, since each district in the city is characterized by its own urban planning history. The dendrograms obtained by varying the linkage rule between average, complete, and single are shown in Fig. 6. The average and complete linkages seem to produce more balanced clusters, while the single rule exhibits the well-known chaining effect. To decide which linkage to use, we adopt the previously discussed AC where values for the average, complete, and single linkages are 0.66, 0.79, 0.41, respectively. The complete linkage is then selected, yielding the highest agglomerative coefficient that may suggest a better overall clustering structure. For completeness, we also compute the AC value for the hierarchical clustering using the weighted Kendall-based dissimilarity measure d^{τ} and the three linkages: AC is lower than that computed using the proposed measure d^{AMH} regardless of the linkage, with the highest value of 0.64 for the complete linkage. In addition, the ARI between the partitions obtained using the complete linkage and d^{AMH} or d^{τ} is equal to 0.30 confirming that the use of AMH copula leads to find a partition highly different from the one obtained using the weighted version of the Kendall's based dissimilarity measure.

As for the selection of the number of clusters to cut the dendrogram and derive the final partition, we adopt an index useful to find a compromise between within-cluster homogeneity and between-cluster separation. Specifically, we use a version of the Dunn index computed as the ratio of the minimum average dissimilarity between two clusters to the maximum average within cluster dissimilarity, which is implemented in the R package `fpc` Hennig [2020] (many other choices could have been made, see, for instance, Halkidi et al. [2001]). A large value of the computed index can be interpreted as an indication of the presence of compact and well-separated clusters. Fig. 7 (left) shows the values of the considered index for K varying between 2 and 8. Both $K = 2$ and $K = 4$ can be justified, however we select $K = 4$ since the partition into two clusters can be poorly informative. To confirm the selection we also took into account the ratio between the average distance within clusters to the average distance between clusters, leading to similar conclusions. The final partition is shown on the map in Fig. 7 (right), which underlies the important role of the spatial weights in finding clusters that take into account similar characteristics of buildings belonging to the same neighbourhood.

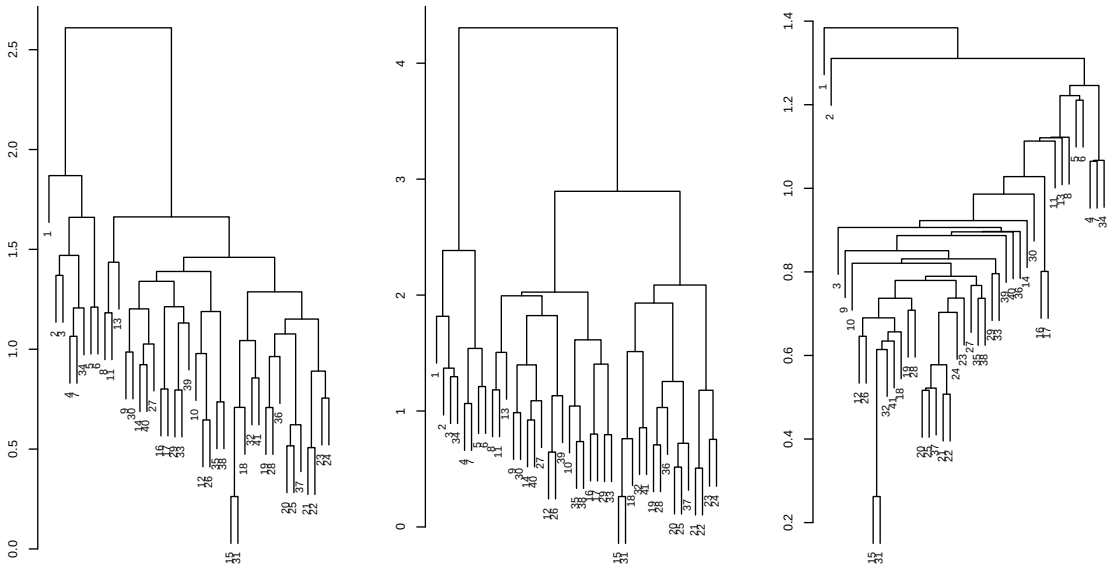

Figure 6: Dendrograms of hierarchical clustering applied to the 41 TED residual time series using the d^{AMH} dissimilarity measure and average, complete, and single linkage method (from left to right).

4.3 Clustering validation and characterization of clusters

Here we discuss the final partition obtained in the cluster analysis presented in the previous section. Figure 8 shows the clusters obtained with four time-invariant characteristics of DH users, i.e., heating surface (dam^2), age class, energy class ($kW\ h/m^2/year$), and yearly mean of heat consumption ($MW\ h/year$). The results show proper features in terms of within-cluster homogeneity and between-cluster dissimilarity. Indeed, the boxplots in Fig. 8 show low spread and low overlapping ranges. This analysis is useful to assess the quality of the final clustering obtained by analyzing the TED time series. The time-invariant characteristics highlight that the clustering methodology based on d^{AMH} groups the users well with respect to their energy performance. Indeed, worth pointing out is that the distribution of the energy class of each identified cluster differs appreciably. Specifically, clusters 1 and 2 include renovated buildings, while clusters 3 and 4 old non-renovated buildings. Cluster 2 comprises buildings that are slightly less efficient and smaller than cluster 1. Instead, cluster 3 is composed of buildings that are more efficient than those in cluster 4. The performed clustering also shows good partition in terms of age class, with a quite pronounced between-cluster dissimilarity, except for cluster 2 that includes a large variety of building ages. This is due to the inclusion in cluster 2 of quite efficient users consisting of both new and renovated buildings. Regarding the heating surface in Fig. 8, clusters 3 and 4 have medium-small sized users, while the energy-efficient buildings of clusters 1 and 2 are

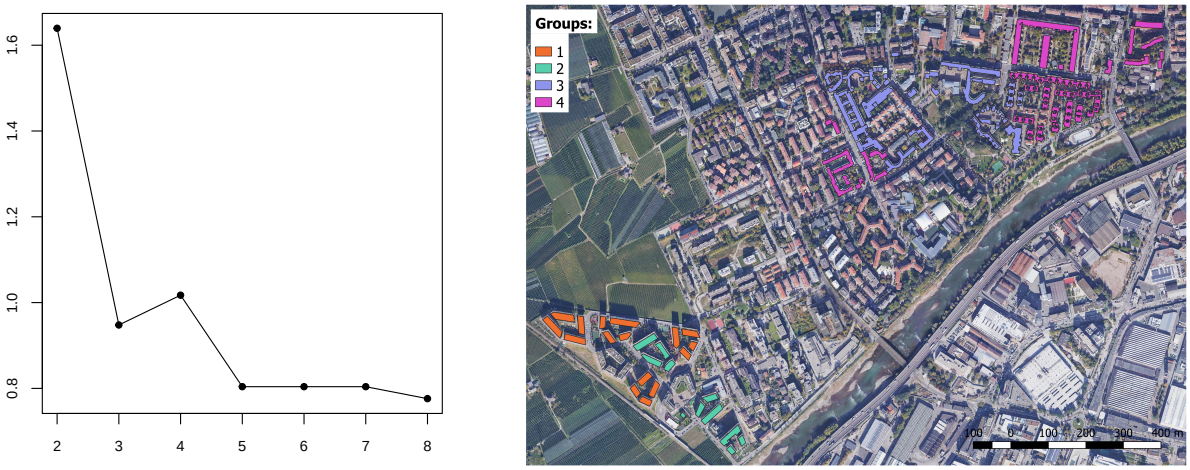

Figure 7: The Dunn-like index (y-axis) for clustering the partition into k clusters (x-axis) (left), and maps of the clusters (right) obtained applying hierarchical clustering with the d^{AMH} dissimilarity measure and the complete linkage to the 41 TED residual time series.

divided into large and medium sized users, respectively. The yearly mean of heat consumption follows analogous behavior to heating surface. The non-efficient buildings of clusters 3 and 4 have similar yearly consumption, while the buildings with high energy performance in clusters 1 and 2 are high and medium yearly consumption groups, respectively. In general, all the clusters can easily be interpreted, especially in terms of energy class and building age, by separating new and efficient users from old and inefficient ones.

In summary, the proposed dissimilarity measure allows us accurately grouping buildings according to their energy performance regardless of size using only historical heat demand information. The energy class is a crucial characteristic for any energy analysis. Indeed, the ability of d^{AMH} to identify clusters that are homogeneous in terms of energy class has several practical implications in DH, for instance, in building renovation planning, anomaly detection, forecasting heat demand, and management control.

5 Conclusions

In this study, we propose a new dissimilarity measure based on the Ali-Mikhail-Haq copula for the application of hierarchical clustering algorithms to spatially located time series that present both temporal and cross-sectional dependence. We validate the theoretical aspects of the proposed dissimilarity on simulated data, and exploit the presented method to analyze observed energy data. To this final aim, we develop a procedure to cluster variables in panel data having characteristics suitable for the AMH copula-based dissimilarity measure. Hence, we apply

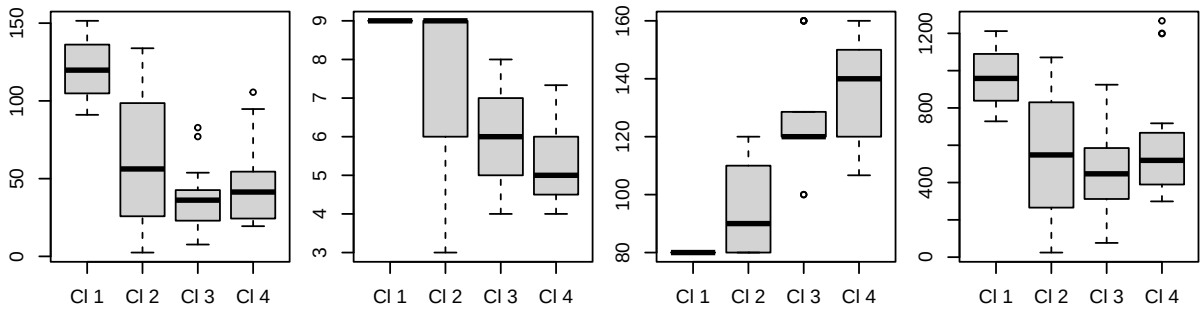

Figure 8: Time invariant characteristics of DH users (from left to right): heating surface, age class, energy class, and yearly mean of heat consumption for each cluster (from Cl 1 to Cl 4) obtained applying the hierarchical clustering with the d^{AMH} dissimilarity measure and the complete linkage to the 41 TED residual time series.

the clustering methodology to high frequency data from the DH of Bozen-Bolzano that exhibit low dependence with tiny differences in rank correlations. Our findings show the empirical usefulness of the AMH-copula based approach in identifying clusters that are well interpretable in terms of energy performance.

Our contribution responds to the current interest in the analysis of big data concerning energy demand for a sustainable and efficient planning of smart DH systems. Indeed, empirical findings are fundamental to support the optimal management of both the production and distribution of DH systems. In order to capture the interconnection between the users' energy demand, there is a need for non-standard clustering methods that are able to cope with the temporal dependence, the cross-sectional dependence, and the spatial information. Hence, considering the buildings' consumption of heating, a clustering able to take into account the above-mentioned aspects can provide crucial information when performing specific tasks for an efficient and sustainable management, such as forecasting and anomaly detection.

Despite the fact that our proposal arises from an empirical issue concerning heating demand, it can be useful in any empirical context where the interest is in the clustering of low correlated r.v.s observed at different geographic locations.

Acknowledgements

The first author acknowledges the financial support from Italian Ministry of University and Research (MIUR) under the Research Project of National Interest (PRIN) "Hi-Di NET - Econometric Analysis of High Dimensional Models with Network Structures in Macroeconomics and Finance" (grant 2017TA7TYC). The second author knowledges the Free University of Bozen-Bolzano via the research project "Techno-economic methodologies to investigate sustainable energy scenarios at urban level (TESES-URB)" - ID 2019. All the authors acknowledge Alpe-

ria S.p.A. - a company that produces and distributes energy from renewable sources - and the Bozen-Bolzano province for providing the analysed data of thermal demand.

A Appendix

Referring to the mathematical analysis presented in Sect. 2, we here provide more technical results. First, we provide the mathematical expression of the functional relationship between Kendall's τ and θ^{AMH} [Kumar, 2010]:

$$\tau = f\left(\theta^{\text{AMH}}\right) = 1 - \frac{2}{3}\frac{1}{\theta^{\text{AMH}}} - \frac{2}{3}\left(\frac{1-\theta^{\text{AMH}}}{\theta^{\text{AMH}}}\right)^2 \log\left(1-\theta^{\text{AMH}}\right) \tag{A.1}$$

Next, we report the expressions of the difference between d^{τ} and d^{AMH} in Eq. (A.2), the difference between the partial derivatives of order 1 in Eq. (A.3) and of order 2 in Eq. (A.4) of the two dissimilarity measures with respect to θ^{AMH}. Note that for simplicity we set $\theta^{\text{AMH}} = \theta$.

$$d^{f(\theta)} - d^{\text{AMH}} = c_{jj'}\left(2\sqrt{\frac{\theta + (\theta-1)^2\log(1-\theta)}{3\theta^2}} - \sqrt{2(1-\theta)}\right) \tag{A.2}$$

$$\frac{\partial d^{f(\theta)}}{\partial \theta^{\text{AMH}}} - \frac{\partial d^{\text{AMH}}}{\partial \theta} = c_{jj'}\left(\frac{\left(\frac{2(\theta-2)}{3\theta^2} + \frac{4(\theta-1)\log(1-\theta)}{3\theta^3}\right)}{2\sqrt{\frac{\theta+(\theta-1)^2\log(1-\theta)}{3\theta^2}}} + \frac{1}{\sqrt{2(1-\theta)}}\right) \tag{A.3}$$

$$\begin{aligned}\frac{\partial^2 d^{f(\theta)}}{\partial \theta^2} - \frac{\partial^2 d^{\text{AMH}}}{\partial \theta^2} = \frac{c_{jj'}}{12}\Bigg(&\frac{3\sqrt{2}}{(1-\theta)^{3/2}} - \frac{4\sqrt{3}\left(\theta(\theta-6) + (4\theta-6)\log(1-\theta)\right)}{\theta^3\sqrt{\theta + (\theta-1)^2\log(1-\theta)}} \\ &- \frac{2\sqrt{3}\left(\theta(\theta-2) + 2(\theta-1)\log(1-\theta)\right)^2}{\theta^6\left(\frac{\theta+(\theta-1)^2\log(1-\theta)}{\theta^2}\right)^{3/2}}\Bigg).\end{aligned} \tag{A.4}$$

References

M. Ali, N. Mikhail, and M. Haq. A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis*, 8(3):405–412, 1978.

P. C. Alvarez-Esteban, C. Euán, and J. Ortega. Time series clustering using the total variation distance with applications in oceanography. *Environmetrics*, 27(6):355–369, 2016.

L. Anselin. Local indicators of spatial association - LISA. *Geographical Analysis*, 27(2):93–115, 1995.

L. Anselin and S. Rey. *Perspective on spatial data analysis*. Springer-Verlag, Berlin, Germany, 2010.

M. Arellano and S. Bond. Some tests of specification for panel data : Monte carlo evidence and an application to employment equations. *Review of Economic Studies*, 58:277–297, 1991.

B. Baltagi. *Econometric Analysis of Panel Data*. John Wiley & Sons Inc., New York, 1995.

T. Bengtsson and J. E. Cavanaugh. State-space discrimination and clustering of atmospheric time series data based on kullback information measures. *Environmetrics: The official journal of the International Environmetrics Society*, 19(2):103–121, 2008.

A. Bonanomi, M. Nai Ruscone, and S. Osmetti. Dissimilarity measure for ranking data via mixture of copulae. *Statistical Analysis and Data Mining*, 12(5):412–425, 2019.

U. Cherubini, E. Luciano, and W. Vecchiato. *Copula Methods in Finance*. John Wiley & Sons Inc., Chichester, West Sussex, 2004.

M. Côté and C. Genest. A copula-based risk aggregation model. *Canadian Journal of Statistics*, 43(1):60–81, 2015.

G. De Luca and P. Zuccolotto. A tail dependence-based dissimilarity measure for financial time series clustering. *Advances in Data Analisys and Classification*, 5(4):323–340, 2011.

G. De Luca and P. Zuccolotto. Dynamic tail dependence clustering of financial time series. *Statistical Papers*, 58:641–657, 2017.

G. De Luca and P. Zuccolotto. Regime dependent interconnectedness among fuzzy clusters of financial time series. *Advances in Data Analysis and Classification*, 15(2):315–336, 2021.

F. Di Lascio, F. Durante, and R. Pappadà. Copula–based clustering methods. In M. Úbeda Flores, E. de Amo, F. Durante, and J. Fernández Sánchez, editors, *Copulas and Dependence Models with Applications*, pages 49–67. Springer International Publishing, Switzerland, 2017.

F. Di Lascio, D. Giammusso, and G. Puccetti. A clustering approach and a rule of thumb for risk aggregation. *Journal of Banking & Finance*, 96:236–248, 2018.

F. Di Lascio, A. Menapace, and M. Righetti. Joint and conditional dependence modelling of peak district heating demand and outdoor temperature: a copula-based approach. *Statistical Methods and Applications*, 29(2):373–395, 2020.

F. Di Lascio, A. Menapace, and M. Righetti. Analysing the relationship between district heating demand and weather conditions through conditional mixture copula. *Environmental and Ecological Statistics*, 28(1):53–72, 2021.

M. Disegna, P. D'Urso, and F. Durante. Copula-based fuzzy clustering of spatial time series. *Spatial Statistics*, 21(A):209–225, 2017.

F. Durante and C. Sempi. *Principles of Copula Theory*. CRC Press, , Boca Raton, 2015.

F. Durante, R. Pappadà, and N. Torelli. Clustering of financial time series in risky scenarios. *Advances in Data Analisys and Classification*, 8:359–376, 2014.

F. Durante, R. Pappadà, and N. Torelli. Clustering of time series via non–parametric tail dependence estimation. *Statistical Papers*, 56(3):701–721, 2015.

B. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster Analysis*. John Wiley & Sons, Ltd, New York, 5th edition, 2011.

S. Frederiksen and S. Werner. *District heating and cooling*. Studentlitteratur AB, Lund, 2013.

S. Fuchs, F. Di Lascio, and F. Durante. Dissimilarity functions for rank-invariant hierarchical clustering of continuous variables. *Computational Statistics & Data Analysis*, 159:107201, 2021.

M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

C. Hennig. *fpc: Flexible Procedures for Clustering*, 2020. URL `https://CRAN.R-project.org/package=fpc`. R package version 2.2-9.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

M. Just and A. Łuczak. Assessment of conditional dependence structures in commodity futures markets using copula-garch models and fuzzy clustering methods. *Sustainability*, 12(6), 2020.

L. Kaufman and P. Rousseeuw. *Finding Groups in Data*. Wiley, New York, 1990.

I. Kojadinovic. Hierarchical clustering of continuous variables based on the empirical copula process and permutation linkages. *Computational Statistics & Data Analysis*, 54(1):90–108, 2010.

J. Kruskal. *The relationship between multidimensional scaling and clustering*, pages 17–44. Classification and clustering. Academic Press, New York, 1977.

P. Kumar. Probability distributions and estimation of ali-mikhail-haq copula. *Applied Mathematical Sciences*, 4(14):657–666, 2010.

H. Lund, S. Werner, R. Wiltshire, S. Svendsen, J. E. Thorsen, F. Hvelplund, and B. V. Mathiesen. 4th generation district heating (4gdh): Integrating smart thermal grids into future sustainable energy systems. *Energy*, 68:1–11, 2014.

H. Lund, P. Østeraard, M. Chang, S. Werner, S. Svendsen, P. Sorknæs, J. Thorsen, F. Hvelplund, B. Mortensen, B. Mathiesen, C. Bojesen, N. Duic, and X. Zhang. The status of 4th generation district heating: research and results. *Energy*, 164:147–159, 2018. ISSN 03605442.

B. Luo, S. Miao, C. Cheng, Y. Lei, G. Chen, and L. Gao. Long-term generation scheduling for cascade hydropower plants considering price correlation between multiple markets. *Energies*, 12(11), 2019.

Z. Ma, J. Xie, H. Li, Q. Sun, Z. Si, J. Zhang, and J. Guo. The role of data analysis in the development of intelligent energy networks. *IEEE Network*, 31(5):88–95, 2017.

A. Menapace, M. Righetti, S. Santopietro, R. Gargano, and G. Dalvit. Stochastic characterisation of the district heating load pattern of residential buildings. *Euroheat and Power (English Edition)*, 16(3-4):14–19, 2019. ISSN 1613-0200.

A. Menapace, J. Thellufsen, G. Pernigotto, F. Roberti, A. Gasparella, M. Righetti, M. Baratieri, and H. Lund. The design of 100% renewable smart urban energy systems: the case of bozen-bolzano. *Energy*, 207:118198, 2020.

A. Menapace, S. Santopietro, R. Gargano, and M. Righetti. Stochastic generation of district heat load. *Energies*, 14(17):5344, 2021.

A. Nazemi and A. Elshorbagy. Application of copula modelling to the performance assessment of reconstructed watersheds. *Stochastic Environmental Research and Risk Assessment*, 26: 189–205, 2012.

H. Nguyen. A novel similarity/dissimilarity measure for intuitionistic fuzzy sets and its application in pattern recognition. *Expert Systems with Applications*, 45:97–107, 2016.

R. Pappadà, F. Durante, G. Salvadoric, and C. De Michele. Clustering of concurrent flood risks via hazard scenarios. *Spatial Statistics*, 23:124–142, 2018.

A. Patton. A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18, 2012.

M. Reddy and P. Ganguli. Spatio-temporal analysis and derivation of copula-based intensity–area–frequency curves for droughts in western rajasthan (india). *Stochastic Environmental Research and Risk Assessment*, 27:1975–1989, 2013.

B. Rémillard, N. Papageorgiou, and F. Soustra. Copula-based semiparametric models for multivariate time series. *Journal of Multivariate Analysis*, 110:30–42, 2012.

P. Rousseeuw. *Data analysis and informatics*, volume 4, chapter A visual display for hierarchical classification, pages 743–748. North-Holland, Amsterdam, 1986.

K. Sharma and L. M. Saini. Performance analysis of smart metering for smart grid: An overview. *Renewable and Sustainable Energy Reviews*, 49:720–735, 2015.

A. Sklar. Fonctions de répartition à n dimensions et leures marges. *Publications de l'Institut de Statistique de L'Université de Paris*, 8:229–231, 1959.

S. Soutullo, L. Bujedo, J. Samaniego, D. Borge, J. Ferrer, R. Carazo, and M. Heras. Energy performance assessment of a polygeneration plant in different weather conditions through simulation tools. *Energy and Buildings*, 124:7–18, 2016.

J. Wooldridge. *Econometrics Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, 2002.

W. Yi and S. Liao. Statistical properties of parametric estimators for markov chain vectors based on copula models. *Journal of Statistical Planning and Inference*, 140(6):1465–1480, 2010.