A spatially-weighted AMH copula-based dissimilarity measure for clustering variables: An application to urban thermal efficiency

F. Marta L. Di Lascio[1] | **Andrea Menapace**[2] | **Roberta Pappadà**[3]

[1]Faculty of Economics and Management, Free University of Bozen-Bolzano, Bolzano, Italy

[2]Faculty of Engineering, Free University of Bozen-Bolzano, Bolzano, Italy

[3]Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Trieste, Italy

Correspondence
F. Marta L. Di Lascio, Faculty of Economics and Management, Free University of Bozen-Bolzano, Piazza Università 1, 39100, Bolzano, Italy.
Email: marta.dilascio@unibz.it

Funding information
Italian Ministry of University and Research, Grant/Award Number: 2017TA7TYC; Free University of Bozen-Bolzano

Abstract

Investigating thermal energy demand is crucial for developing sustainable cities and the 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 article, we propose a novel copula-based dissimilarity measure suitable for analyzing district heating demand and introduce a procedure to apply it to high-temporal resolution panel data. Inspired by the characteristics of the considered data, we explore the usefulness of the Ali-Mikhail-Haq copula in defining a new dissimilarity measure to cluster variables in the hierarchical framework. We show that our proposal is particularly sensitive to small dissimilarities based on tiny differences in the strength of the dependence between the involved random variables. Therefore, the measure we introduce is able to distinguish between objects with low dissimilarity better than standard rank-based dissimilarity measures. Moreover, our proposal considers a weighted version of the copula-based dissimilarity that embeds the spatial location of the involved objects. We investigate the proposed measure through Monte Carlo studies and compare it with an analogous dissimilarity measure based on Kendall's correlation. 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. In turn, our findings may 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

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 full reshaping 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 & Werner, 2013).

Developing stochastic methods to analyze high-frequency DH energy data provided by smart grids is currently a demanding challenge (see, e.g., Ma et al., 2017; Sharma & Saini, 2015). 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 & Rousseeuw, 1990). Hierarchical clustering requires a pairwise dissimilarity measure to compare singletons and a linkage rule to compare clusters. The literature on hierarchical clustering methods is extensive and applications have been successfully performed in various contexts (see, e.g., Alvarez-Esteban et al., 2016; Bengtsson & Cavanaugh, 2008; Di Lascio et al., 2018; Nguyen, 2016). In clustering random variables (r.v.s hereafter), copula-based measures of association have been used to define dissimilarity indices in a variety of applied contexts (see, e.g., De Luca & Zuccolotto, 2021; Di Lascio et al., 2017; Nazemi & Elshorbagy, 2012; Pappadà et al., 2018), 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 & 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., Bonanomi et al., 2019; De Luca & Zuccolotto 2017; Durante et al., 2015; Fuchs et al., 2021; Kojadinovic, 2010).

While many contributions in the context of clustering r.v.s have focused on detecting a strong association between extreme values (see, e.g., Côté & Genest, 2015; Durante et al., 2014), this article focuses on the ability to capture low dependence and small dissimilarities between the involved r.v.s. As discussed by Kruskal (1977), cluster analysis is appropriate to extract information from small dissimilarities. 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. We then propose a new AMH copula-based dissimilarity measure and investigate its properties at both the theoretical and applied level. Since the most used copula-based dissimilarity measures involve Kendall's τ correlation coefficient, we compare the performance of the proposed measure with the corresponding version based on Kendall's τ through Monte Carlo studies.

The theoretical contribution of this article is general, but it has been motivated by the features of the panel data concerning the thermal energy demand of residential users in the Italian city of Bozen-Bolzano in 2016 that have been analyzed in this work. 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 & Zuccolotto, 2011; Disegna et al., 2017; Reddy & Ganguli, 2013). Extensions of these approaches, considering both temporal and cross-sectional dependence via copulas, can be found in, for example, Yi and Liao (2010) and Rémillard et al. (2012), but to the best of our knowledge, there is a lack of procedures dedicated to panel data analysis, which is our focus. Hence, the proposed AMH copula-based dissimilarity measure is exploited to develop a clustering procedure for panel data. While some studies use copulas in the field of DH demand (see, e.g., Di Lascio et al., 2020, 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., Just & Łuczak, 2020; Luo et al., 2019).

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

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 & 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 correlations 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. The most widely used linkage rules are the average, the complete, and the single (Everitt et al., 2011). In light of the empirical data features (see Section 4 for details), 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 distinguish objects, say r.v.s, with low and very similar dependence. Given p objects $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_j, \dots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_j$ is a n-dimensional vector of values from a r.v. X_j, we propose the following measure based on the AMH copula that also takes into account the spatial distance between any pair of r.v.s:

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

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 $j, j' = 1, \dots, p$. Hence, the weight is null if and only if $j = j'$, otherwise it is an increasing function of the geographic distance, taking values in $]1, \exp(1)]$. Such a weighting scheme emphasizes the dissimilarity of objects that are further apart. Moreover, the measure in Equation (2) is a dissimilarity measure, since it satisfies the three properties of a dissimilarity measure (see Kaufman & Rousseeuw, 1990, Chap. 1) whose proof is trivial:

P1. $d^{\text{AMH}}_{jj'} \geq 0 \quad \forall\, j, j'$

P2. $d^{\text{AMH}}_{jj'} = 0 \quad$ if and only if $j = j'$

P3. $d^{\text{AMH}}_{jj'} = d^{\text{AMH}}_{j'j} \quad \forall\, j, j'$.

As noted by a referee, the above-mentioned properties are satisfied only under some conditions on the spatial weights. Specifically, multiplicative spatial weights have to satisfy all the three above properties, that is, $c_{jj'} \geq 0$ and $c_{jj'} = c_{j'j},\ \forall j, j'$, and $c_{jj'} = 0$ if and only if $j = j'$, to ensure $d^{\text{AMH}}_{jj'}$ is a proper dissimilarity measure. In our proposal, where a geographic distance is used to define $c_{jj'}$, all the required properties are satisfied by definition.

The proposed dissimilarity measure only takes into account the comonotone (positive) dependence, and d^{AMH} is decreasingly monotone with respect to θ^{AMH}, meaning that the dissimilarity degree tends to vanish as soon as approaching the comonotonic (positive) case.

It is worth stressing that: (*i*) when spatial weights are omitted, that is, the spatial information is ignored, the proposed dissimilarity only depends on the association of the considered pair (j, j') and, assuming the AMH copula as the true model, Equation (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 Equation (A1) in the Appendix A); (*ii*) Equation (2) is not intended to measure spatial dependence and,

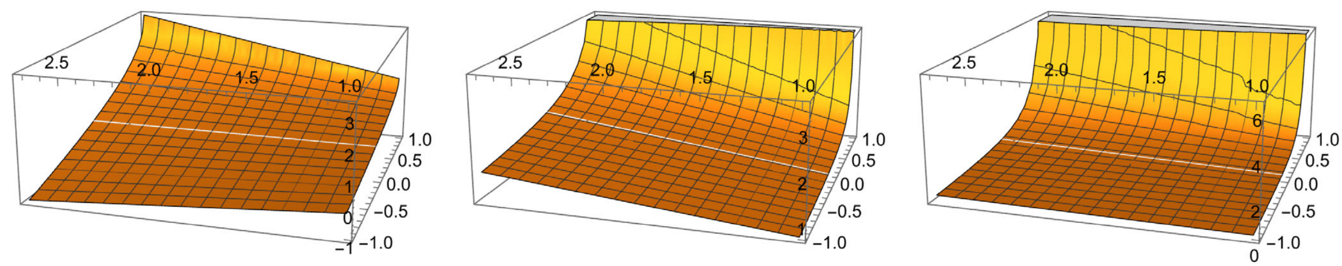

FIGURE 1 Comparison between d^{AMH} and $d^{f(\theta^{\mathrm{AMH}})}$: Difference between $d^{f(\theta^{\mathrm{AMH}})}$ and d^{AMH} in Equation (A2) (left), and between first partial derivative in Equation (A3) (middle), and second partial derivatives in Equation (A4) (right) of d^{AMH} and $d^{f(\theta^{\mathrm{AMH}})}$ (z-axes) versus $\theta^{\mathrm{AMH}} \in [-1, 1[$ (y-axes), and $c_{jj'} \in]1, \exp(1)]$, $\forall\, j \neq j'$ (x-axes).

thus, it is not related to dissimilarities based on spatial association or heterogeneity (see, e.g., Anselin, 1995; Anselin & Rey, 2010), 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 behavior of d^{AMH} and d^{τ}, we assume that the AMH copula model is the true model and express d^{τ} as $d^{f(\theta^{\mathrm{AMH}})} = c_{jj'}\sqrt{2\left(1-f\left(\theta^{\mathrm{AMH}}\right)\right)}$, where $f\left(\theta^{\mathrm{AMH}}\right)$ is given in Equation (A1), Appendix A. We then mathematically analyze the two measures and show the different behavior 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 Equations (A2)–(A4) in Appendix A, while the shape of the three equations by varying $\theta^{\mathrm{AMH}} \in [-1, 1[$ and $c_{jj'} \in]1, \exp(1)]$, $\forall\, j \neq j'$ is shown in Figure 1.

The difference between the two dissimilarities (Figure 1, left) shows a monotonically decreasing (increasing) behavior 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 (Figure 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 are greater than zero for all $c_{jj'} \in]1, \exp(1)]$, $\forall\, j \neq j'$. Hence the difference between the two considered measures is monotonically increasing and convex in $c_{jj'}$.

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, implying that jumps between subsequent fusions are better spotted from the dendrogram. Hence, the hierarchy of clusters is better highlighted, improving the cutting of the dendrogram and the interpretation of the resulting partition.

3 | MONTE CARLO STUDY

Here, we provide a simulation study to assess the goodness of the proposed dissimilarity measure in Equation (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 spatial information. To this end, we consider a p-dimensional DGP consisting of K copula models differing from the AMH and independent of each other. The dimension p_k of each copula is randomly chosen from 2 to $(p-(K+1))$ to ensure that $p_k \geq 2$ with $k = 1, \ldots, K$, and $\sum_{k=1}^{K} p_k = p$. We consider five different scenarios of the described DGP that are summarized in Table 1. Inspired by the case study analyzed in Section 4 we set all the copula dependence parameters to small values (see Table 1), $p = 41$, $K = 3$, and we generate an independent sample of size $n = 150$ from each of the three copulas involved in the considered scenario. Since each copula model in the DGP represents the dependence structure underlying a cluster, we are thus simulating a partition consisting of $K = 3$ clusters containing at least two elements, that is, r.v.s to be clustered, and such that the size of the whole clustering is $p = 41$.

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, $\tau = 0.2$

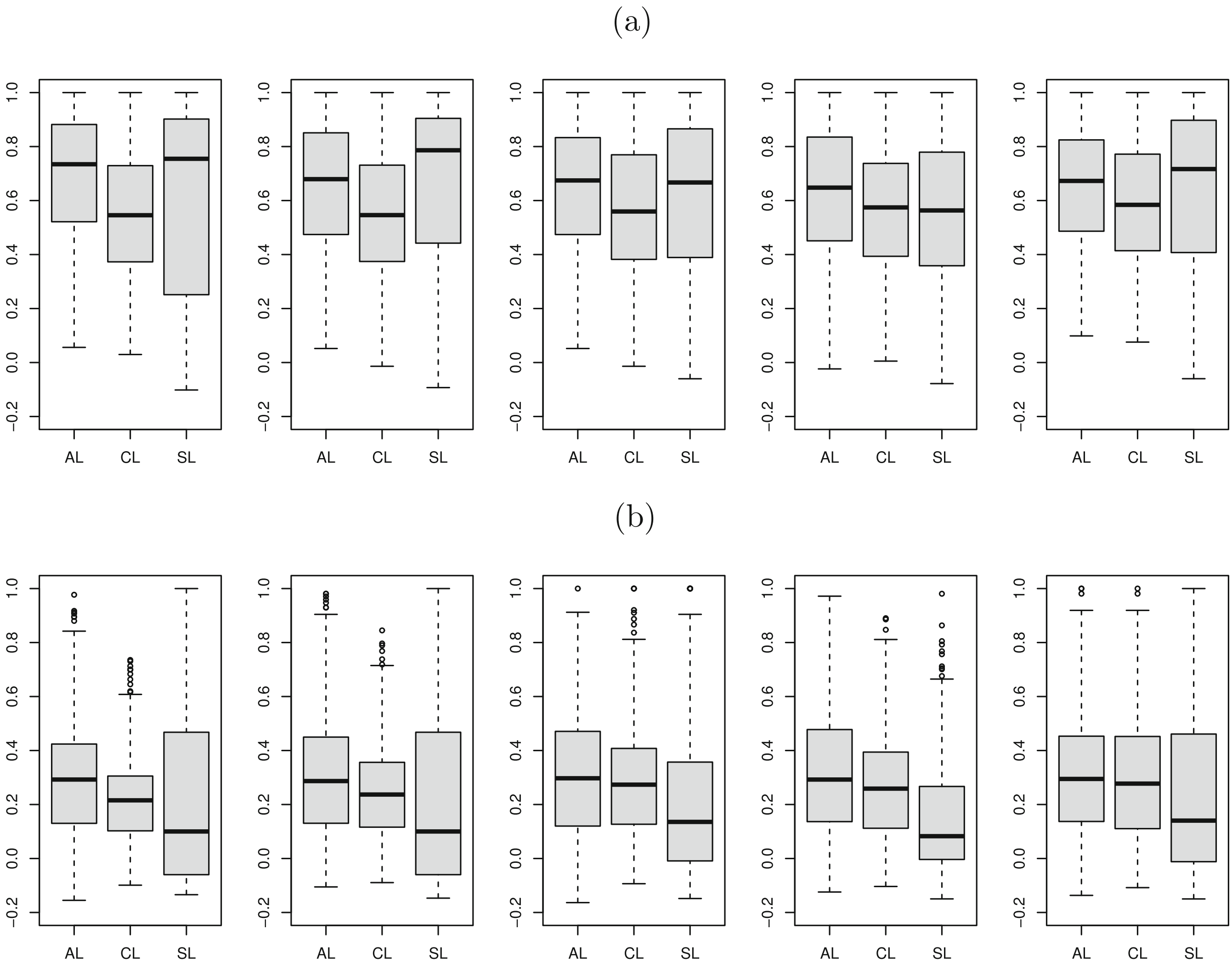

FIGURE 2 Boxplots of ARI (y-axis) comparing the partitions obtained through d^{AMH} and d^{τ} with $K = 3$ by varying (i) the linkage method between the average (AL), the complete (CL), and the single (SL) (x-axis), (ii) the scenario among the five given in Table 1 (panels by columns), and (iii) the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

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 their 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 randomly generate points by adding a Gaussian random noise distributed as $N(0,100)$ 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 set described in the Section 4 adding a uniform random noise. We therefore simulate 10 different scenarios, and for each, perform 500 Monte Carlo replications.

To measure the performance of d^{AMH} and d^{τ}, we compute (*i*) the Adjusted Rand Index (Hubert & Arabie, 1985) (ARI hereafter) to assess the agreement between the partitions obtained using the two compared measures given the true number of clusters, that is, $K = 3$, and (*ii*) the agglomerative coefficient (Kaufman & 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 Figure 2. Here, the partition is obtained by cutting the dendrogram so that three clusters are selected. 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 Kendall's coefficient. The AC distribution for each simulated scenario is shown in Figure 3. According to Kaufman and Rousseeuw (1990) an AC close to 1 indicates that

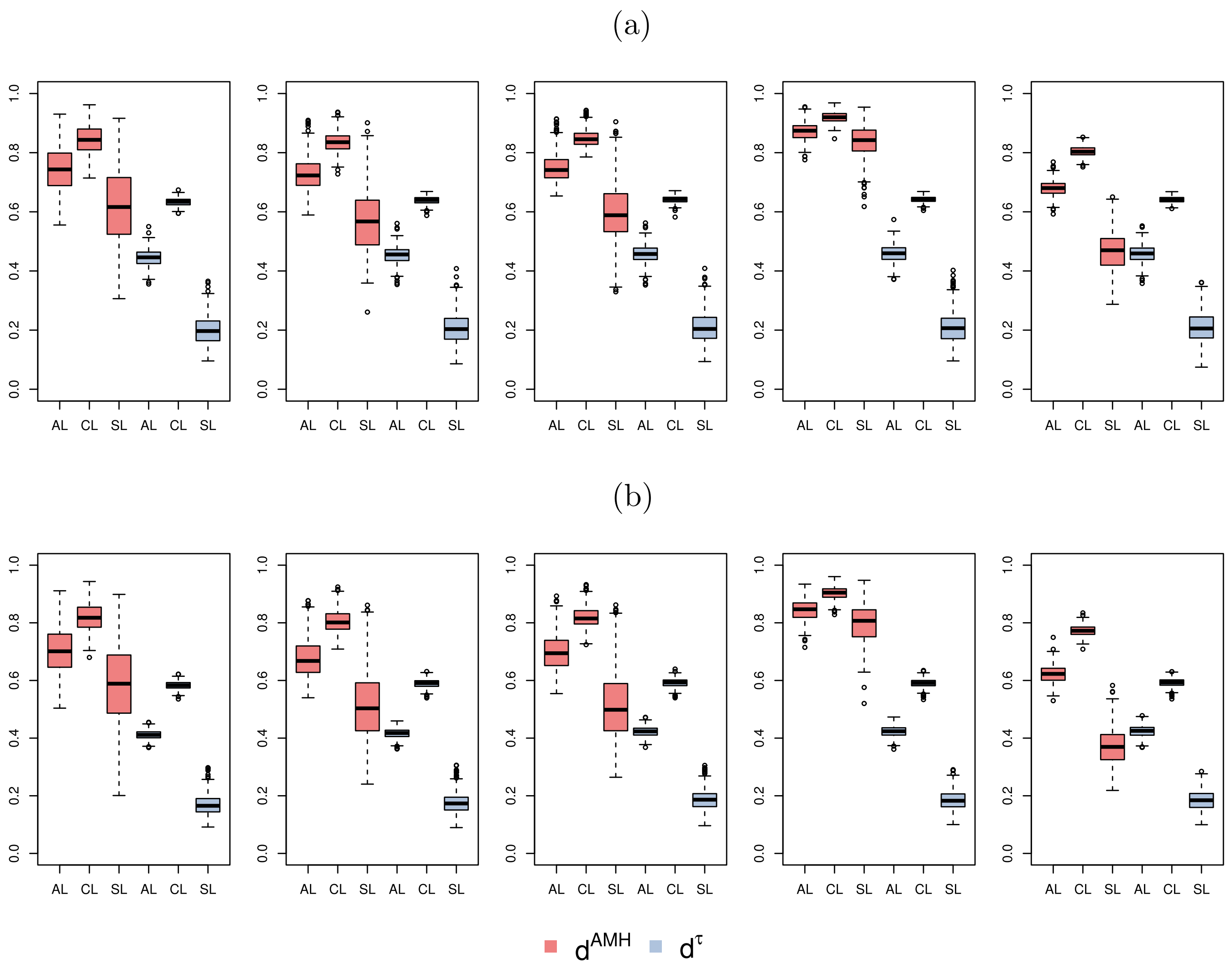

FIGURE 3 Boxplots of AC (y-axis) by varying (*i*) the pairwise dissimilarity measure between d^{AMH} and d^{τ}, (*ii*) the linkage method between the average (AL), the complete (CL), and the single (SL) (x-axis), (*iii*) the scenario among the five provided in Table 1 (panels by columns), and (*iv*) the spatial settings among (a) random weights, and (b) empirical weights plus a random noise — see text for details. The sample size is $n = 150$ and the number of r.v.s is $p = 41$. The number of Monte Carlo replications is 500.

tight clusters that are far away from each other, that is, 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 are of the same order of magnitude" (Kaufman & Rousseeuw, 1990). Here, it is evident that the proposed dissimilarity measure outperforms the measure based on Kendall's τ regardless of the scenario, the linkage rule, and the setting of spatial weights. The complete linkage turns out to be better than the average and the single linkage, 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 MWh 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 the time series of thermal energy demand (TED hereafter, in kWh) 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 January 8, 2016 to January 14, 2016 (see Figure 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 modeling of the TED panel data (Soutullo et al., 2016). The observed time series have been preprocessed to remove outliers due to meter or transmission system failures and then aggregated to obtain hourly observations.

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 for improving the efficiency and sustainability of the DH of Bozen-Bolzano through a proper schedule of the heat production and management of

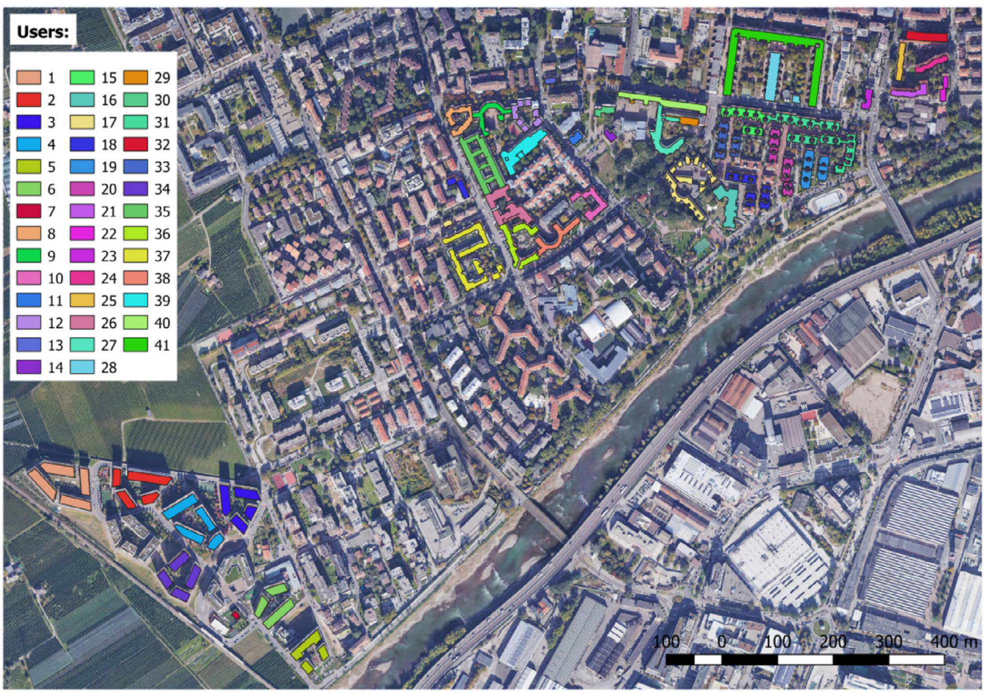

FIGURE 4 Map of the sample of users in the different districts fed by the Bozen-Bolzano DH.

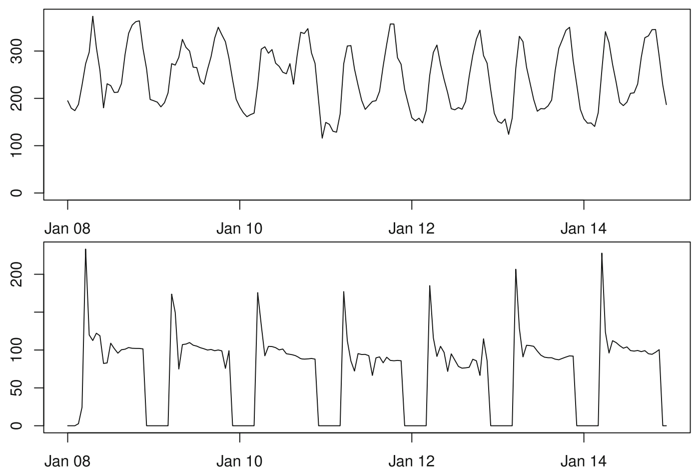

FIGURE 5 Time series of TED in $kW\,h$ (y-axis) of two typical users.

the network and the thermal reservoir. For instance, consider two users with clearly different behaviors: the top panel of Figure 5 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; the bottom panel of Figure 5 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 about the buildings is used: heating surface (in dam^2), energy consumption (in kWh/m^2/year), age class (re-coded with 9 levels, where level 1 refers to buildings aged before 1918 and level 9 to buildings aged after 2005), and mean yearly heat consumption (in MWh/year).

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 Equation (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 take into account the effect of (lagged and not) meteorological variables on TED, as well as the serial dependence of TED and individual effects μ_j, with $j = 1, \dots, p$. 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 variable selection based on significant covariates:

$$\begin{aligned}\mathrm{TED}_{jt} &= \rho_1 \mathrm{TED}_{j(t-1)} + \rho_2 \mathrm{TED}_{j(t-24)} + \beta_1 \mathrm{RAD}_{jt} + \beta_2 \mathrm{RAD}_{j(t-1)} + \beta_3 \mathrm{TEMP}_{jt} \\ &\quad + \beta_4 \mathrm{TEMP}_{j(t-3)} + \epsilon_{jt} \\ &= \rho_1 \mathrm{TED}_{j(t-1)} + \rho_2 \mathrm{TED}_{j(t-24)} + \beta_1 \mathrm{RAD}_{jt} + \beta_2 \mathrm{RAD}_{j(t-1)} \\ &\quad + \beta_3 \mathrm{TEMP}_{jt} + \beta_4 \mathrm{TEMP}_{j(t-3)} + \mu_j + \varepsilon_{jt}, \end{aligned} \tag{3}$$

where $j = 1, \dots, p$, $t = 1, \dots, n$, and ρ_1, ρ_2, β_1, β_2, β_3, β_4 are scalar; the error term $\epsilon_{jt} = \mu_j + \varepsilon_{jt}$ is assumed to follow a one-way error component regression model with $\mu_j \sim N(0, \sigma_\mu^2)$ and $\varepsilon_{jt} \sim N(0, \sigma_\varepsilon^2)$, $\forall\, j$ and $\forall\, t$, which are independent of each other and among themselves. Since TED_{jt} is a function of μ_j, it follows that $\mathrm{TED}_{j(t-1)}$ is also a function of μ_j. Therefore, $\mathrm{TED}_{j(t-1)}$ is correlated with the error term, and we use a set of instrumental variables, that is, 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 & Bond, 1991).

Once the model in Equation (3) has been estimated, the residuals of the 41 time series are extracted, and the weighted AMH copula-based dissimilarity is computed as in Equation (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 Kendall's correlations ranging in $[-0.207, 0.393]$. Only two values lie outside Kendall's τ range for the AMH copula and have been replaced with the (maximum or minimum) extreme of the range. Hence, the AMH copula is a suitable model for describing the relationships between the considered data. The dendrograms obtained by varying the linkage rule between average, complete, and single are shown in Figure 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, and 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 finding 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 obtain 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. Figure 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 Figure 7 (right). As can be noted, the spatially-weighted measure allows us to identify clusters that reflect the urban planning history of each city district without forcing in the same cluster buildings belonging to the same neighborhood.

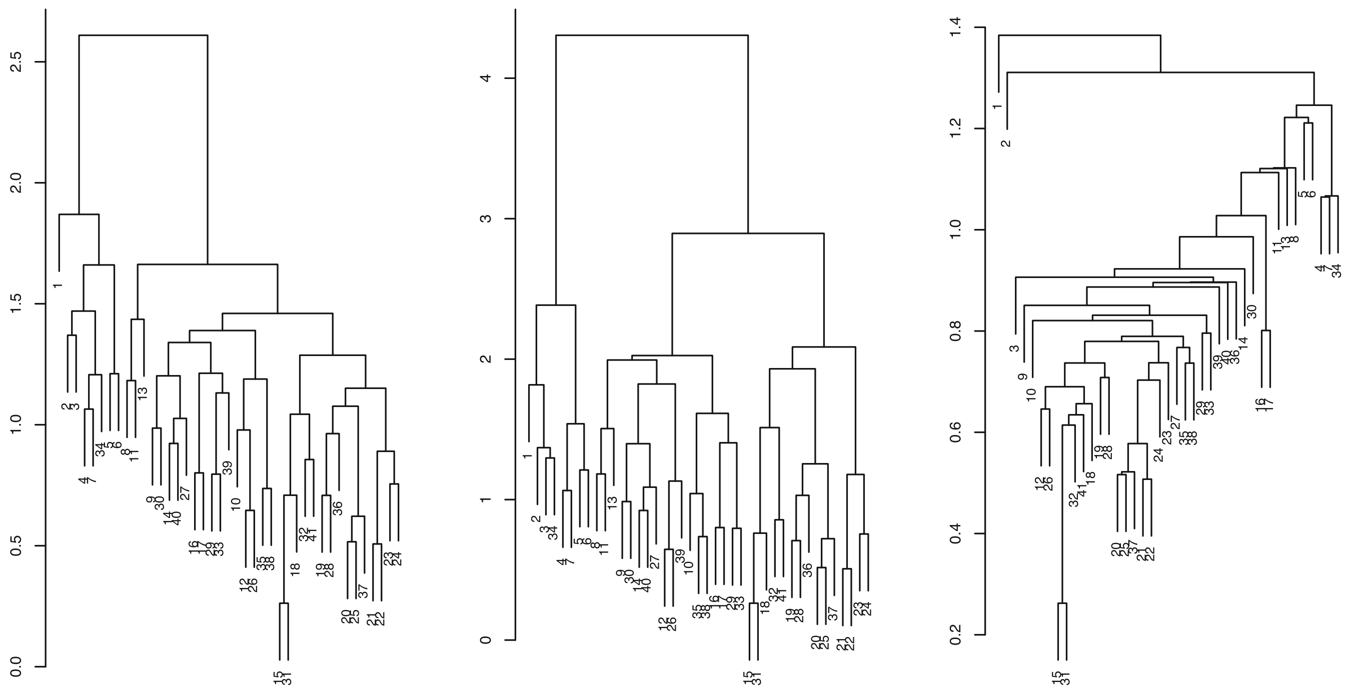

FIGURE 6 Dendrograms of hierarchical clustering applied to the 41 TED residual time series using the d^{AMH} dissimilarity measure and average, complete, and single linkage method (from left to right).

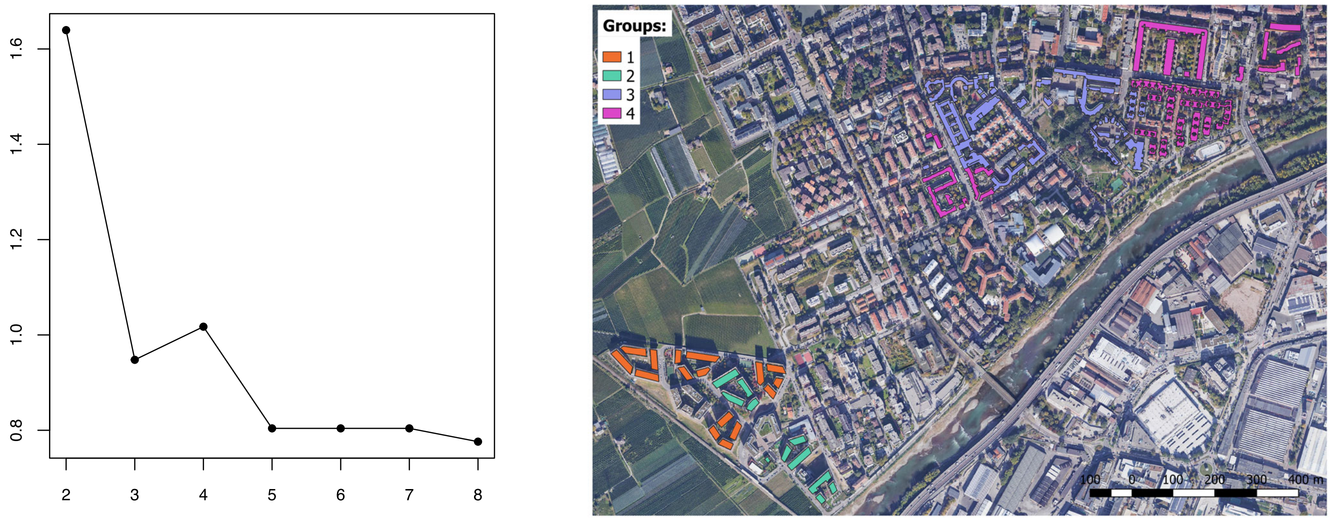

FIGURE 7 The Dunn-like index (y-axis) for clustering the partition into K clusters (x-axis) (left), and maps of the clusters (right) obtained applying hierarchical clustering with the d^{AMH} dissimilarity measure and the complete linkage to the 41 TED residual time series.

To conclude, it is worth mentioning that, following the suggestion of a referee, we further explored the comparison between the clusterings based on d^{AMH} and d^{τ}. We found that the former outperforms the latter in terms of cluster homogeneity and separation.

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, that is, heating surface, age class, energy consumption, and yearly mean of heat consumption. The results show proper features in terms of within-cluster homogeneity and between-cluster dissimilarity. Indeed, the boxplots in Figure 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 consumption 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, which includes buildings with very different 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 Figure 8, clusters 3 and 4 have medium-small sized users, while the energy-efficient buildings of clusters 1 and 2 are divided into large- and medium-sized users, respectively. The yearly mean of heat consumption follows a behavior analogous 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 consumption and building age, by separating new and efficient users from old and inefficient ones.

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

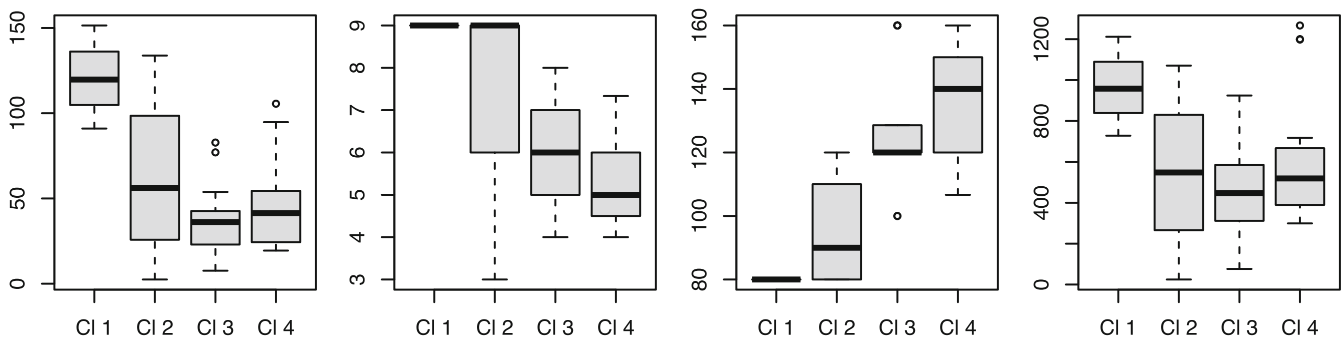

FIGURE 8 Time invariant characteristics of DH users (from left to right): Heating surface, age class, energy consumption, and yearly mean of heat consumption for each cluster (from Cl 1 to Cl 4) obtained applying the hierarchical clustering with the d^{AMH} dissimilarity measure and the complete linkage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

5 | FINAL REMARKS

5.1 | Discussion

The spatially-weighted AMH copula-based dissimilarity measure can be useful in any context where the dependence structure of the variables is suitably described by the AMH copula. From the applied viewpoint, this means that the data share characteristics of the energy data considered in this work. Despite the fact that the methodological contribution of this work is motivated by an empirical issue concerning heating demand, the d^{AMH} measure can be exploited whenever the interest is in the clustering of low correlated r.v.s observed at different geographic locations, and the relationship between variables resembles that of the AMH copula. By contrast, a clear limitation of the proposed measure is given by the underlying assumption that the data do not exhibit high-rank correlations in absolute value.

A relevant aspect to discuss concerns the spatial information introduced in the dissimilarity measure. In our proposal, we followed a standard approach to define spatial weights by exploiting exponentially transformed geographic distances (Getis & Aldstadt, 2002; Mateau & Müller, 2013). As noted by a referee, a dissimilarity measure that considers spatial information would benefit from a spatial contiguity term, which accounts for both the strength of spatial dependence and the way it is distributed in space. It is worth stressing that our proposal is not a spatial dissimilarity measure and does not intend to model nor impose a spatial dependence structure among variables, such as spatial autocorrelation, heterogeneity and so forth (see, e.g., Anselin, 1995; Anselin & Rey, 2010). However, even though d^{AMH} is not designed to model spatial dependence, it embeds the interaction between the dependence structure of the compared objects and their geographic distance. As a result, the geographic location of users is to help our AMH-based dissimilarity measure to better distinguish between different energy demand profiles. As a final remark, we emphasize that the case study presented concerns a phenomenon, that is, the energy demand, for which modeling spatial dependence is not fully appropriate; indeed, the user energy profile is not affected by spatial dynamics but only by the location of the buildings connected to the DH and their static characteristics.

5.2 | Conclusions

In this study, we propose a new dissimilarity measure based on the Ali-Mikhail-Haq copula and able to keep into account the geographic information of the objects to be compared. Our proposal has been directly motivated by an application to energy data and responds to the current interest in the analysis of big data concerning energy demand for the development of modern DH systems.

To capture the interconnection between the users' energy demand, there is a need for non-standard association measures that can cope with complex dependence structures and spatial information. Such measures can be exploited to cluster buildings in urban area based on their heat consumption with the final aim of providing crucial information supporting efficient and sustainable management of DH systems.

The proposed AMH copula-based measure has been thought to be used in the hierarchical clustering framework. After describing the theoretical aspects of the measure, we assess its performance on artificial data. In addition, we compare our proposal with the most used Kendall's τ-based dissimilarity measure (in a spatially-weighted version), both from the theoretical point of view and via Monte Carlo studies. Next, we apply the clustering methodology to high-frequency data from the DH of Bozen-Bolzano that exhibits low dependence with tiny differences in rank correlations. After removing both temporal and cross-sectional dependence through a dynamic panel regression model, we employ hierarchical clustering algorithms to the residuals time series by using the AMH copula-based dissimilarity measure. We thus empirically prove the usefulness of our procedure in identifying clusters of buildings that are well interpretable in terms of energy performance.

Our findings can provide useful insights for the optimal management of energy production and distribution systems. Indeed, efficient and sustainable DH planning would benefit from the energy characterization of users' profiles in performing relevant tasks, such as demand forecasting and anomaly detection.

ACKNOWLEDGMENTS

The first author acknowledges the financial support from the 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 acknowledges 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 Alperia S.p.A. — a company that produces and distributes energy from renewable sources — and the Bozen-Bolzano province for providing the analyzed data on thermal demand.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

REFERENCES

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

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

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

Anselin, L., & Rey, S. J. (2010). *Perspective on spatial data analysis*. Springer-Verlag.

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

Baltagi, B. (1995). *Econometric analysis of panel data*. John Wiley & Sons Inc.

Bengtsson, T., & Cavanaugh, J. E. (2008). 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.

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

Cherubini, U., Luciano, E., & Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons Inc.

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

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

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

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

Di Lascio, F. M. L., Durante, F., & Pappadà, R. (2017). *Copula-based clustering methods*. In M. Ú. Flores, E. de Amo, F. Durante, & J. F. Sánchez (Eds.), *Copulas and dependence models with applications* (pp. 49–67). Springer International Publishing.

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

Di Lascio, F. M. L., Menapace, A., & Righetti, M. (2020). 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.

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

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

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

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

Durante, F., & Sempi, C. (2015). *Principles of copula theory*. CRC Press.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). John Wiley & Sons, Ltd.

Frederiksen, S., & Werner, S. (2013). *District heating and cooling*. Studentlitteratur AB.

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

Getis, A., & Aldstadt, J. (2002). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, *34*(2), 130–140.

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

Hennig, C. (2020). *fpc: Flexible procedures for clustering* (R package version 2.2-9). https://CRAN.R-project.org/package=fpc

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Just, M., & Łuczak, A. (2020). Assessment of conditional dependence structures in commodity futures markets using copula-GARCH models and fuzzy clustering methods. *Sustainability*, *12*(6), 2571.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. Wiley.

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

Kruskal, J. (1977). *The relationship between multidimensional scaling and clustering*. In J. van Ryzin (Ed.), *Classification and clustering* (pp. 17–44). Academic Press.

Kumar, P. (2010). Probability distributions and estimation of Ali-Mikhail-Haq copula. *Applied Mathematical Sciences*, *4*(14), 657–666.

Lund, H., Østeraard, P. A., Chang, M., Werner, S., Svendsen, S., Sorknæs, P., Thorsen, J. E., Hvelplund, F., Mortensen, B. O. G., Mathiesen, B. V., Bojesen, C., Duic, N., & Zhang, X. (2018). The status of 4th generation district heating: Research and results. *Energy*, *164*, 147–159.

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

Luo, B., Miao, S., Cheng, C., Lei, Y., Chen, G., & Gao, L. (2019). Long-term generation scheduling for cascade hydropower plants considering price correlation between multiple markets. *Energies*, *12*(11), 2239.

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

Mateau, J., & Müller, W. G. (2013). *Spatio-temporal design: Advances in efficient data acquisition*. John Wiley & Sons Inc.

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 (English Edition)*, *16*(3-4), 14–19.

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

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

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

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

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

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

Reddy, M. J., & Ganguli, P. (2013). 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.

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

Rousseeuw, P. J. (1986). *A visual display for hierarchical classification*. In E. Diday, Y. Escoufier, L. Lebart, J. Pages, Y. Schektman, & R. Tomassone (Eds.), *Data analysis and informatics* (Vol. *4*, pp. 743–748). North-Holland.

Sharma, K., & Saini, L. M. (2015). Performance analysis of smart metering for smart grid: An overview. *Renewable and Sustainable Energy Reviews*, *49*, 720–735.
Sklar, A. (1959). Fonctions de répartition à n dimensions et leures marges. *Publications de l'Institut de Statistique de L'Université de Paris*, *8*, 229–231.
Soutullo, S., Bujedo, L. A., Samaniego, J., Borge, D., Ferrer, J. A., Carazo, R., & Heras, M. R. (2016). Energy performance assessment of a polygeneration plant in different weather conditions through simulation tools. *Energy and Buildings*, *124*, 7–18.
Wooldridge, J. (2002). *Econometrics analysis of cross section and panel data*. MIT Press.
Yi, W., & Liao, S. S. (2010). Statistical properties of parametric estimators for Markov chain vectors based on copula models. *Journal of Statistical Planning and Inference*, *140*(6), 1465–1480.

APPENDIX A.

Referring to the mathematical analysis presented in Section 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):

$$\tau = f\left(\theta^{\mathrm{AMH}}\right) = 1 - \frac{2}{3}\frac{1}{\theta^{\mathrm{AMH}}} - \frac{2}{3}\left(\frac{1-\theta^{\mathrm{AMH}}}{\theta^{\mathrm{AMH}}}\right)^2 \log\left(1-\theta^{\mathrm{AMH}}\right). \tag{A1}$$

Next, we report the expressions of the difference between d^{τ} and d^{AMH} in Equation (A2), the difference between the partial derivatives of order 1 in Equation (A3) and of order 2 in Equation (A4) of the two dissimilarity measures with respect to θ^{AMH}. Note that for simplicity we set $\theta^{\mathrm{AMH}} = \theta$.

$$d^{f(\theta)} - d^{\mathrm{AMH}} = c_{jj'}\left(2\sqrt{\frac{\theta + (\theta-1)^2\log(1-\theta)}{3\theta^2}} - \sqrt{2(1-\theta)}\right), \tag{A2}$$

$$\frac{\partial d^{f(\theta)}}{\partial \theta^{\mathrm{AMH}}} - \frac{\partial d^{\mathrm{AMH}}}{\partial \theta} = c_{jj'}\left(\frac{\left(\frac{2(\theta-2)}{3\theta^2} + \frac{4(\theta-1)\log(1-\theta)}{3\theta^3}\right)}{2\sqrt{\frac{\theta+(\theta-1)^2\log(1-\theta)}{3\theta^2}}} + \frac{1}{\sqrt{2(1-\theta)}}\right), \tag{A3}$$

$$\begin{aligned}\frac{\partial^2 d^{f(\theta)}}{\partial \theta^2} - \frac{\partial^2 d^{\mathrm{AMH}}}{\partial \theta^2} = \frac{c_{jj'}}{12}\Bigg(&\frac{3\sqrt{2}}{(1-\theta)^{3/2}} - \frac{4\sqrt{3}\left(\theta(\theta-6) + (4\theta-6)\log(1-\theta)\right)}{\theta^3\sqrt{\theta + (\theta-1)^2\log(1-\theta)}} \\ &- \frac{2\sqrt{3}\left(\theta(\theta-2) + 2(\theta-1)\log(1-\theta)\right)^2}{\theta^6\left(\frac{\theta+(\theta-1)^2\log(1-\theta)}{\theta^2}\right)^{3/2}}\Bigg).\end{aligned} \tag{A4}$$