A Spatial AMH Copula-Based Dissimilarity Measure to Cluster Variables in Panel Data

F. Marta L. Di Lascio (marta.dilascio@unibz.it)
Free University of Bozen-Bolzano https://orcid.org/0000-0002-9297-4261
Andrea Menapace
Free University of Bozen-Bolzano
Roberta Pappadà
University of Trieste

Research Article

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

Posted Date: December 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1145716/v1

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

F. Marta L. Di Lascio (Corresponding author) [1]

Faculty of Economics and Management

Free University of Bozen-Bolzano,

Piazza Università, 1, 39100 Bolzano (BZ) – Italy

ORCID: 0000-0002-9297-4261

Andrea Menapace

Faculty of Science and Technology

Free University of Bozen-Bolzano,

Piazza Università, 1, 39100 Bolzano (BZ) – Italy

ORCID: 0000-0003-0778-9721

Roberta Pappadà

Department of Economics, Business, Mathematics and Statistics "B. de Finetti"

University of Trieste,

Via A. Valerio, 4/1, 34127 Trieste (TS) – Italy

ORCID: 0000-0002-4852-0561

[1] **Email address for correspondence:** marta.dilascio@unibz.it

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

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

[1*]Faculty of Economics and Management, Free University of Bozen-Bolzano, Piazza Università, 1, Bozen-Bolzano, 39100, Italy.
[2]Faculty of Science and Technology, Free University of Bozen-Bolzano, Piazza Università, 1, Bozen-Bolzano, 39100, Italy.
[3]Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, via A. Valerio, 4/1, Trieste, 34127, Italy.

*Corresponding author(s). E-mail(s): marta.dilascio@unibz.it;
Contributing authors: andrea.menapace@unibz.it;
rpappada@units.it;

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. In this paper, we develop a clustering methodology based on a novel dissimilarity measure to analyze a high temporal resolution panel data for district heating demand in the Italian city Bozen-Bolzano. Starting from 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 proposed measure is able to better distinguish between objects with low dissimilarity than classic rank-based dissimilarity measures. Moreover, our proposal is defined in a spatial version that is able to

take into account the spatial location of the compared objects. We investigate the proposed measure through Monte Carlo studies and compare it with the corresponding spatial Kendall's correlation-based dissimilarity measure. Finally, the application to real data 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, District heating demand, Panel data, Spatial distance.

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. Renewable energy systems require a fully reshape of the traditional infrastructure and a rethink of the technologies involved. 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 [1] by incorporating renewable sources and reducing waste of energy in a flexible urban energy system [2].

Developing stochastic methods to analyze high frequency DH energy data provided by smart grids is currently a demanding challenge (see, e.g. [3], [4]). In particular, there is a need for an in-depth analysis of heating data (see, e.g. [5, 6]) to enhance the management and planning of the heating system [7]. 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.

In the hierarchical agglomerative clustering framework, 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 [8]. 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 the choice of pairwise dissimilarity measures is extensive (see [9] and references therein). 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., [10], [11], [12], [13], [14], and [15]), 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 [16]. 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., [17], [18], [19], [20], [21]).

While many contributions in the context of clustering r.v.s have focused on detecting a strong association between extreme values (see, e.g., [22] and [23]), this paper focuses on the ability to differentiate r.v.s characterized by a similar and low level of dependence, motivated by the features of the panel data of DH demand. As discussed by [24], 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 [25] to cluster r.v.s in the agglomerative hierarchical clustering context, proposing a new spatial 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., [26], [27], [28]). Extensions of these approaches, considering both temporal and cross-sectional dependence via copulas, can be found in, e.g., [29], [30], but to the best of our knowledge, there are no methodological procedures dedicated to panel data analysis, which is our focus. Hence, we develop a procedure for clustering panel data with characteristics suitable for the proposed AMH copula-based dissimilarity measure. We, then, analyze panel data of DH demand in an Italian city. While some studies exploit copulas in the field of DH demand (see, e.g., [31], [32]), 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., [33], [34]).

The remainder of the paper is organized as follows. First, Section 2 describes our motivation and provides a toy example introducing the framework of our proposal. Then, we define a new dissimilarity measure and present its theoretical advantages in Section 3. In Section 4, we compare our proposal with a classic dissimilarity measure through a 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 5. Section 6 highlights the most relevant implications and summarizes our main findings.

2 Motivation

Consider a set of n-dimensional data objects where each object is a realization of a r.v. representing a time series of the phenomenon of interest (for instance, a time series of residuals obtained after removing any trend over time and autocorrelation effects). The correlation between any pair of r.v.s can be used to determine the dissimilarity measure for clustering time series based on similar behavior over time. In particular, for two r.v.s X_j and $X_{j'}$, the dissimilarity can be expressed in terms of Kendall's rank correlation coefficient, $\tau_{jj'}$, by adopting, for instance, the function $\sqrt{2(1-\tau_{jj'})}$ as in [11] that highlights the differences in the dissimilarity values and ranges in $[0, 2]$.

As a toy example, we consider five variables extracted from the data we analyze in Section 5. Assume that the Kendall's correlations computed for five variables denoted with X, Y, Z, W, V form the following symmetric matrix:

$$\mathbf{R} = \begin{array}{c} \begin{array}{ccccc} X & Y & Z & W & V \end{array} \\ \begin{pmatrix} 1.000 & -0.103 & -0.077 & 0.102 & -0.028 \\ -0.103 & 1.000 & 0.079 & -0.001 & 0.170 \\ -0.077 & 0.079 & 1.000 & 0.117 & -0.041 \\ 0.102 & -0.001 & 0.117 & 1.000 & 0.139 \\ -0.028 & 0.170 & -0.041 & 0.139 & 1.000 \end{pmatrix} \end{array} \begin{array}{c} X \\ Y \\ Z \\ W \\ V \end{array}$$

As can be noted, the values of the pairwise rank correlations are overall low and homogenous, with the highest value 0.139 associated with the pair (V, W). In this situation, the aforementioned dissimilarity will be characterized by limited variability (in the range $[1.289, 1.485]$), due to tiny differences in the rank correlations. The heat maps of distances for the five considered variables are shown in Fig. 1 (middle). The heat map of the Kendall-based dissimilarities allows us to visualize the strong homogeneity within the represented objects, providing information on the partition of the data based on the hierarchical agglomerative method with complete linkage through the dendrograms of rows and columns. Moreover, to assess the overall quality of the dendrogram, we consider the agglomerative coefficient [8] (AC hereafter) that is given by the normalized average of all the "1-dissimilarities" to the first cluster that the singleton is merged with, divided by the dissimilarity of the merger in the final step of the algorithm. The AC value is equal to 0.095, showing the very poor quality of the obtained dendrogram.

Since each variable in the data we analyze in Section 5 is associated with a spatial location, it is reasonable to include spatial information in the computation of the dissimilarity. Hence, we assess the clustering via the expression $d^{\tau}_{jj'} = c_{jj'}\sqrt{2(1-\tau_{jj'})}$, where $c_{jj'} = \left(\frac{1}{w_{jj'}} - \delta_{jj'}\right)$ with $w_{jj'}$ denoting the spatial weight associated with pair (j, j') and $\delta_{jj'}$ being the usual Kronecker δ [35]. The spatial weight can be computed through the standard exponential function of spatial distance [35], i.e. $w_{jj'} = \exp(-g_{jj'}/\max_{jj'}\{g_{jj'}\})$, where $g_{jj'}$ is the geographic distance between j and j' on the World Geodetic System ellipsoid [36]. We thus use a spatial Kendall-based dissimilarity measure $d^{\tau}_{jj'}$ where the strength of the dissimilarity increases with the spatial distance. The corresponding heat map shown in Fig. 1 (right) clearly indicates the usefulness of spatial weights in better discerning very similar objects, e.g., pairs (X, Y)

and (Y, W). However, this does not completely overcome the issue of discerning the r.v.s that are weakly correlated and not distant in space. For instance, pairs (X, Y) and (Y, Z) result in cells with similar color intensity in both cases (see the middle and right panels of Fig. 1), and, although AC has increased to the more reasonable value of 0.514, the overall quality of the dendrogram is still poor. From an engineering point of view, this weakness is a bottleneck for

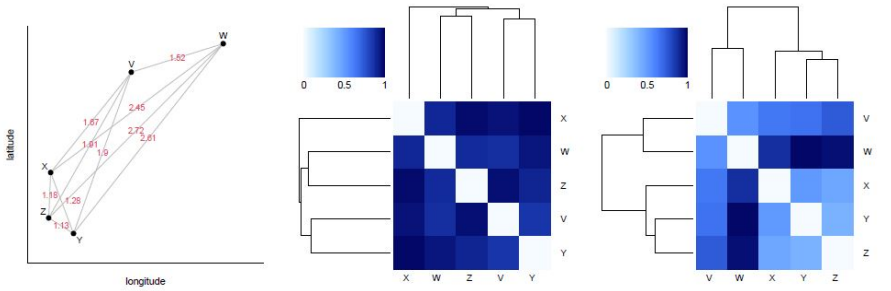

Fig. 1 From left to right: Locations (longitude and latitude coordinates) of the time series objects used for the motivating example (spatial weights are reported in red for each pair of variables) and the heat maps of the 5×5 dissimilarity matrix based on Kendall's τ rank correlation ignoring (middle) and including (right) the spatial information. The dissimilarities are normalized to $[0, 1]$ for comparison purposes.

several DH system operations since an accurate clustering of DH users would allow us to better schedule heat production, thermal storage management, and integration of renewable sources, e.g. solar and geothermal energy [5].

Motivated by the empirical issue illustrated above, in what follows we propose a new dissimilarity measure that appears to be a valid approach whenever spatial correlation-based clustering of r.v.s is of interest and the data present small but non-negligible correlations.

3 A spatial AMH copula-based dissimilarity measure

Copulas originated in the context of probabilistic metric spaces via Sklar's noted theorem [37] 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 [16]. 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 [25]:

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

where $\theta^{\mathrm{AMH}}_{u_1,u_2} \in [-1, 1[$ is its dependence parameter. Although the AMH copula function can be used to describe both positive and negative correlation of r.v.s, it is not suitable for very high positive or negative correlations. The domain of the AMH copula dependence parameter in terms of Kendall's τ coefficient is $[-0.1817, 0.3333]$. For details on the statistical properties of the AMH copula, see [38]. The dependence parameter of the AMH copula can be estimated using the estimation methods available in the literature (see, e.g., [39]).

Our interest in the AMH copula rests on the possibility of defining a new dissimilarity measure able to differentiate among objects with low and very similar dependence. Hence, we propose a measure that exploits the above discussed characteristics of the AMH copula and is also able to take into account

the spatial information between objects:

$$d_{jj'}^{\mathrm{AMH}} = c_{jj'}\sqrt{2(1-\theta_{jj'}^{\mathrm{AMH}})} \tag{2}$$

where $c_{jj'} = \left(\frac{1}{w_{jj'}} - \delta_{jj'}\right)$, $w_{jj'}$ is the spatial weight associated with the two objects j and j' and $\delta_{jj'}$ is the usual Kronecker δ [35]. Note that when $j = j'$, $w_{jj'} = \delta_{jj'} = 1$, so that $c_{jj'} = 0$. The spatial weight can be calculated starting from the geographic distance (based on longitude and latitude information) of the two objects in such a way that it decreases with the geographic distance. A possibility for defining the spatial weight is the function introduced in Section 2: $w_{jj'} = \exp(-g_{jj'}/\max_{jj'}\{g_{jj'}\})$, where $g_{jj'}$ is the geographic distance between j and j' on the World Geodetic System ellipsoid [36]. As a result, two objects j and j' are more dissimilar the further apart they are. Hence, $0 \leq d_{jj'}^{\mathrm{AMH}} \leq 2\max_{jj'}\{c_{jj'}\}$. The measure in Eq. (2) is a dissimilarity, since it satisfies the two properties of a dissimilarity measure whose proofs are trivial:

P1. $d_{jj'}^{\mathrm{AMH}} \geq 0 \quad \forall j, j', \quad$ and $\quad d_{jj'}^{\mathrm{AMH}} = 0 \quad$ if and only if $j = j'$

P2. $d_{jj'}^{\mathrm{AMH}} = d_{j'j}^{\mathrm{AMH}} \quad \forall j, j'$

The proposed dissimilarity measure considers minimum dissimilarity only between variables with maximum comonotone (positive) dependence. Moreover, 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. Since the parametric space of the dependence parameter θ^{AMH} tends to amplify the difference between low-rank correlations,

which allows 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 5, 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.

4 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 spatial Kendall-based dissimilarity measure d^{τ}. In particular, we want to investigate the ability of d^{AMH} to discriminate objects with homogeneous correlation, also taking into account the spatial information. To this end, we consider five different three-dimensional DGPs 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. Overall, we have $n = 150$ realizations of $p = 41$ r.v.s (which can be interpreted, for instance, as serially uncorrelated time series), and 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 sample size as well as the spatial weights derive from the empirical case study data set presented in Section 5. The five DGPs considered are simulated by using different settings for spatial information. In particular, when no spatial information is used $c_{jj'} = 1 \forall j, j' = 1, \ldots, p$, whereas when spatial information is used, the weights are computed using the exponential form described in Section 3, 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 next section adding a uniform random noise. We therefore simulate 15 different scenarios, and for each, perform 500 Monte Carlo replications.

Table 1 Data generating processes (DGPs) used in the Monte Carlo simulation study.

DGP	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) AC to assess the quality of the dendrogram, i.e., the whole hierarchy of clusters, and (ii) the Adjusted Rand Index [40] (ARI hereafter) to assess the quality of a specific partition, i.e., the agreement between the partition obtained given the true number of clusters and the true partition. The AC distribution for each simulated scenario is shown in Fig. 2. [8] argue that when the AC is close to zero, "the corresponding method has not found a natural structure, which can be expressed by saying that no clusters have been found, or rather that the data consists of one big cluster". Instead, an AC close to 1 indicates that a very clear clustering structure has been identified. Here, it is evident that the

proposed dissimilarity measure outperforms the measure based on Kendall's τ irrespective of the DGP, 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. On the contrary, it appears that d^{τ} is *i*) not able to discriminate among low correlated objects, and *ii*) strongly affected by the use of spatial information that helps the dissimilarity improve the overall quality of the clustering. Also, the complete linkage here shows better performance than the other two linkage rules. The distribution of ARI for each simulated scenario is shown in Fig. 3. Here, the partition is obtained by cutting the dendrogram so that three clusters are identified. It is evident that when spatial information is used, the spatial AMH copula-based dissimilarity measure outperforms the spatial Kendall-based dissimilarity measure, irrespective of the DGP and the spatial weights, but the two measures appear equivalent when unit spatial weights are used. As for d^{AMH}, when spatial information is used, the average and the complete linkage rules work better than the single rule in all the simulated scenarios. Moreover, interesting to note is that the proposed dissimilarity performance shows a mild effect in terms of the kind of DGP and level of dependence used. As for the spatial Kendall-based dissimilarity measure, neither the linkage rule nor kind of DGP appear to affect the performance of the measure when spatial information is used. Indeed, the role of the spatial weights is crucial and negatively affects the performance of d^{τ}, worsening when the weights are empirically computed, and therefore not related to the simulated within-cluster dependence.

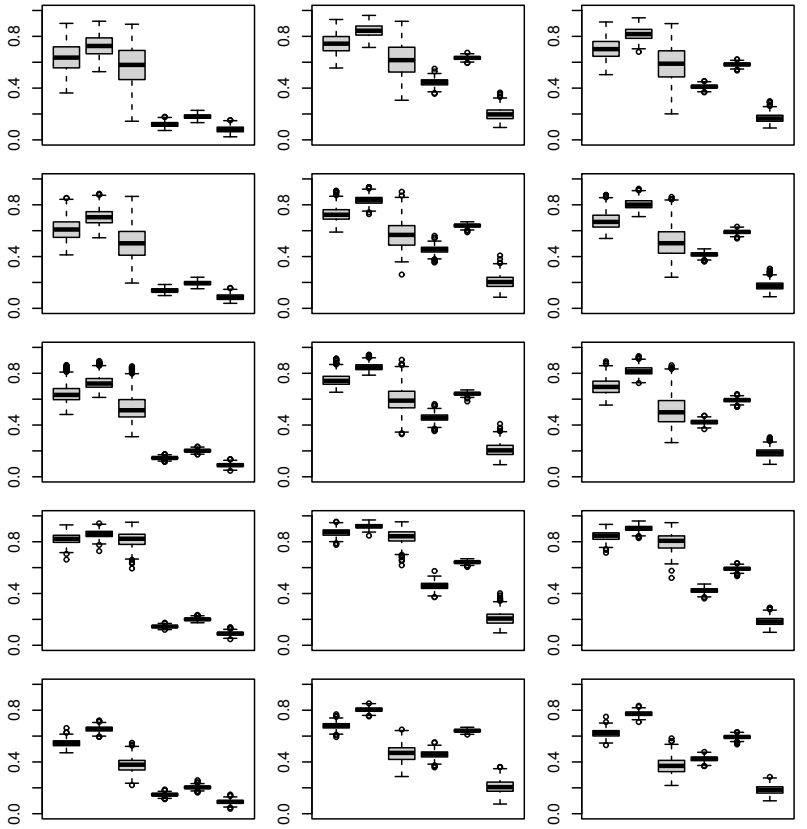

Fig. 2 Boxplots of AC (y-axis) by varying *i*) the pairwise dissimilarity measure between d^{AMH} and d^{τ}, *ii*) the linkage method between the average, the complete (maximum), and the single (minimum) (x-axis starting with the average linkage and d^{AMH}, continuing with the complete linkage and d^{AMH}, and ending with the single linkage and d^{τ}), *iii*) the DGP among the five provided in Table 1 (panels by rows), and *iv*) the spatial settings among no weights ($c_{jj'} = 1, \forall j, j'$), random weights, and empirical weights plus a random noise (panels by columns) - see text for details. Sample size is $n = 150 \times p = 41$. The number of Monte Carlo replications is 500.

5 Application to panel data

5.1 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 [41]. The

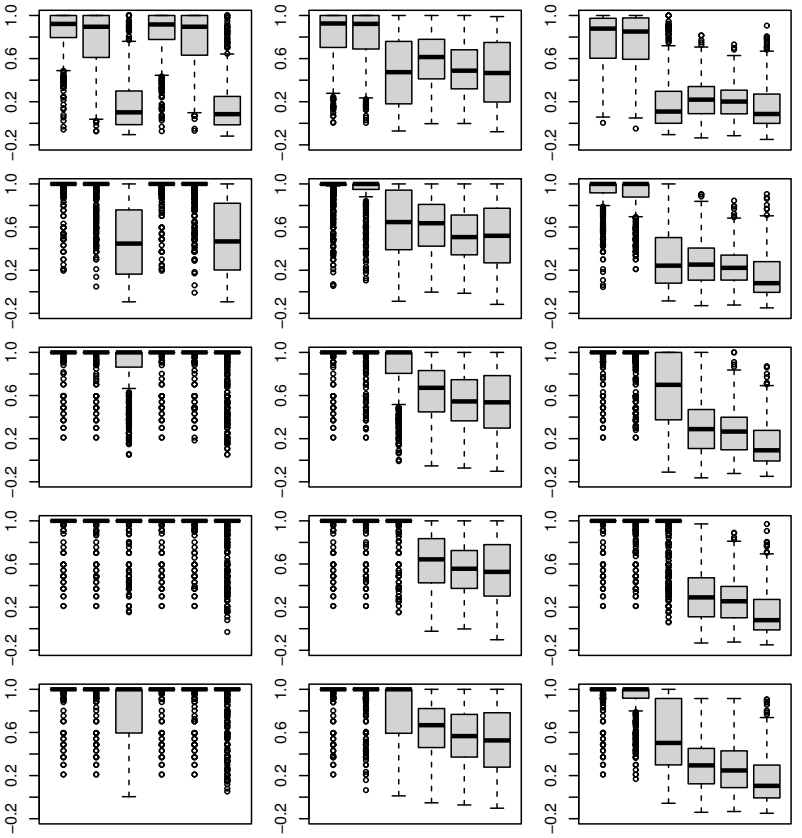

Fig. 3 Boxplots of ARI (y-axis) by varying *i*) the pairwise dissimilarity measure between d^{AMH} and d^{τ}, *ii*) the linkage method between the average, the complete (maximum), and the single (minimum) (x-axis starting with the average linkage and d^{AMH}, continuing with the complete linkage and d^{AMH}, and ending with the single linkage and d^{τ}), *iii*) the DGP among the five given in Table 1 (panels by rows), and *iv*) the spatial settings among no weights ($c_{j\,j'} = 1, \forall j, j'$), random weights, and empirical weights plus a random noise (panels by columns) - see text for details. Sample size is $n = 150 \times p = 41$. The number of Monte Carlo replications is 500.

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 [42]. 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 one winter week from 08/01/2016 to 14/01/2016 (see Fig. 4). Moreover, we 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 observed time series have been pre-processed to remove outliers due to meter or transmission system failures, and then aggregated to obtain hourly observations. The meteorological data, presenting significant dependence on heating demand, helps the proper modelling of the TED panel data [43].

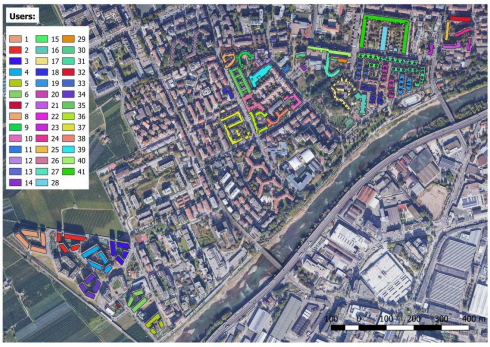

Fig. 4 Map of the sample of users in the different districts fed by the Bozen-Bolzano DH.

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,

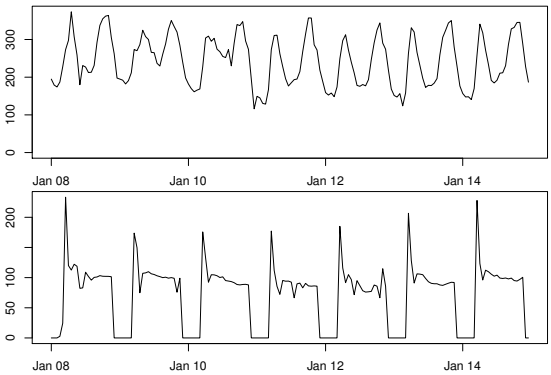

Fig. 5 Time series of TED in $kW\,h$ (y-axis) of two typical users.

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 $MW\,h/year$).

5.2 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 r.v.s representing the TED time series of each user. To consider both temporal and cross-sectional dependence, we extend the copula-based approach to time series modeling (see, e.g., [44]) to the panel data case. We first fit a suitable dynamic panel regression model (see, e.g., [45, 46]) to tackle serial dependence. Next, we model cross-sectional dependence between

the time series of residuals by exploiting copula theory and apply the proposed measure in the hierarchical clustering framework. To do that, 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:

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

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 $\mathrm{TED}_{i(t-1)}$ is also a function of μ_i. Therefore, $\mathrm{TED}_{i(t-1)}$ 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 [47].

Once the model in Eq. (3) has been estimated, the residuals of the 41 time series are extracted, and the spatial 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 4. We note that the residuals show low and very similar linear Kendall's correlation (range $(-0.207, 0.394)$). 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 0.65, 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 spatial 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. 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. We use a Dunn-like index computed as the ratio of the minimum average dissimilarity between two clusters to the maximum average within-cluster dissimilarity [48] for different values of the number of clusters K (between 2 and 8). A large value of the index can be interpreted as an indication of the presence of compact and well-separated clusters. The index plot (Fig. 7) shows that $K = 2$ and $K = 4$ can be justified. We favor the solution with $K = 4$ that can be more informative than partitioning into two clusters.

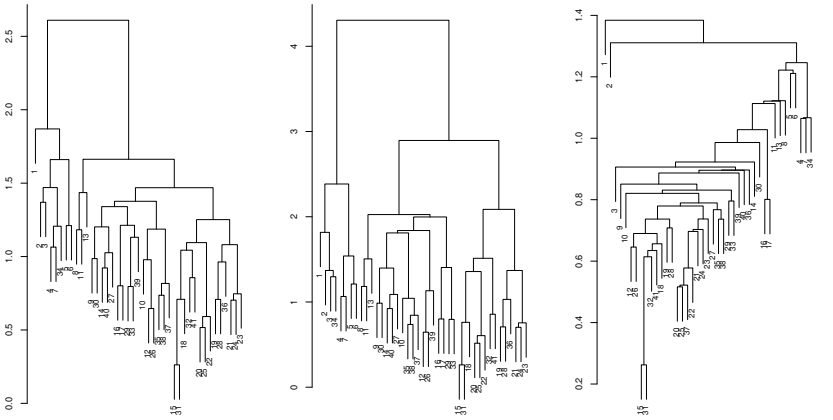

Fig. 6 Dendrograms of hierarchical clustering applied to the 41 TED residual time series using the d^{AMH} dissimilarity measure and average, complete, and single linkage method (from left to right).

The final partition is shown on the map in Fig. 7 (right) that underlies the important role of the spatial weights in finding clusters that take into account similar characteristics of buildings belonging to the same neighbourhood.

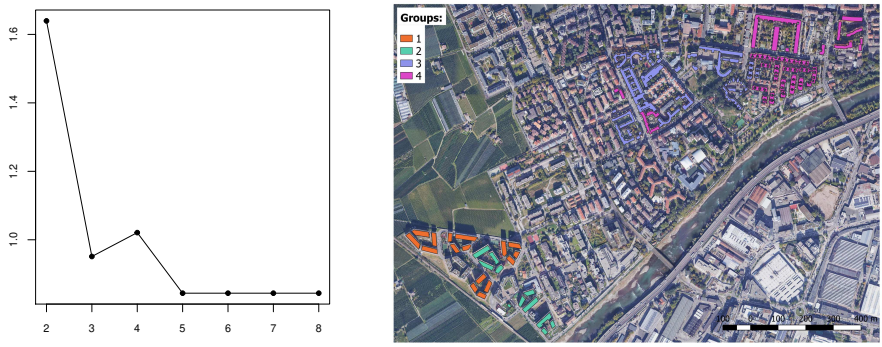

Fig. 7 The Dunn-like index (y-axis) for clustering the partition into k clusters (x-axis) (left), and maps of the clusters (right) obtained applying hierarchical clustering with the d^{AMH} dissimilarity measure and the complete linkage to the 41 TED residual time series.

5.3 Clustering validation and characterization of clusters

Here we discuss the final partition obtained in the analysis presented in the previous section. Figure 8 shows the clusters obtained with four time-invariant characteristics of DH users, i.e., heating surface, age class, energy class, 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 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 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.

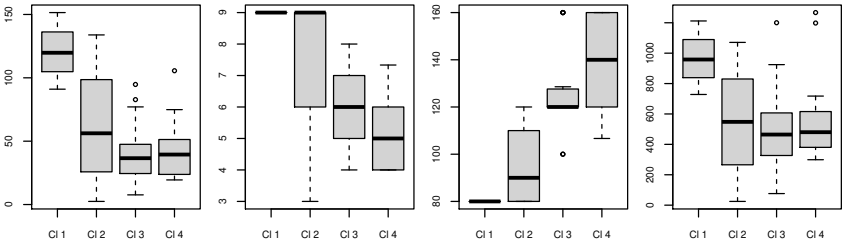

Fig. 8 Time invariant characteristics of DH users (from left to right): heating surface (dam^2), age class ($class$), energy class ($kW\,h/m^2/year$), and yearly mean of heat consumption ($MW\,h/year$) for each cluster (from Cl 1 to Cl 4) obtained applying the hierarchical clustering with the d^{AMH} dissimilarity measure and the complete linkage to the 41 TED residual time series.

In summary, the proposed dissimilarity measure allows 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.

6 Conclusions

In this study, we propose a new dissimilarity measure based on the Ali-Mikhail-Haq copula for the application of hierarchical clustering algorithms when complex temporal dependencies and spatial information are relevant. 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 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 an 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 was motivated 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 geographical sites.

Acknowledgements

The first and the second authors knowledge 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 analysed data of thermal demand.

References

[1] Frederiksen, S., Werner, S.: District Heating and Cooling. Studentlitteratur, Lund (2013)

[2] 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.: The status of 4th generation district heating: research and results. Energy **164**, 147–159 (2018)

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

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

[5] Tureczek, A.M., Nielsen, P.S., Madsen, H., Brun, A.: Clustering district heat exchange stations using smart meter consumption data. Energy and buildings **182**, 144–158 (2019)

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

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

[8] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. Wiley, New York

(1990)

[9] Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edn. John Wiley & Sons, Ltd, New York (2011)

[10] Nazemi, A., Elshorbagy, A.: Application of copula modelling to the performance assessment of reconstructed watersheds. Stoch Environ Res Risk Assess (26), 189–205 (2012)

[11] Di Lascio, F.M.L., Durante, F., Pappadà, R.: Copula–based clustering methods. In: Úbeda Flores, M., de Amo, E., Durante, F., Fernández Sánchez, J. (eds.) Copulas and Dependence Models with Applications, pp. 49–67. Springer, Switzerland (2017). Chap. 4

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

[13] Nguyen-Huy, T., Deo, R.C., Mushtaq, S., Kath, J., Khan, S.: Copula statistical models for analyzing stochastic dependencies of systemic drought risk and potential adaptation strategies. Stoch Environ Res Risk Assess (33), 779–799 (2019)

[14] Shan, B., Guo, S., Wang, Y., Li, H., Guo, P.: Vine copula and cloud model-based programming approach for agricultural water allocation under uncertainty. Stoch Environ Res Risk Assess (35), 1895–1915 (2021)

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

[16] Durante, F., Sempi, C.: Principles of Copula Theory. CRC Press, , Boca Raton (2015)

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

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

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

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

[21] Fuchs, S., Di Lascio, F.M.L., Durante, F.: Dissimilarity functions for rank-invariant hierarchical clustering of continuous variables. Comput. Stat. Data An. **159**, 107201 (2021)

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

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

[24] Kruskal, J.: The relationship between multidimensional scaling and clustering. In: Ryzin, J.V. (ed.) Classification and clustering, pp. 17–44.

Academic Press, New York (1977)

[25] Ali, M., Mikhail, N.N., Haq, M.S.: A class of bivariate distributions including the bivariate logistic. J. Multivariate Anal. **8**(3), 405–412 (1978)

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

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

[28] Reddy, M.J., Ganguli, P.: Spatio-temporal analysis and derivation of copula-based intensity–area–frequency curves for droughts in western rajasthan (india). Stoch Environ Res Risk Assess (27), 1975–1989 (2013)

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

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

[31] Di Lascio, F.M.L., Menapace, A., Righetti, M.: 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)

[32] Di Lascio, F.M.L., Menapace, A., Righetti, M.: Analysing the relationship

between district heating demand and weather conditions through conditional mixture copula. Environmental and Ecological Statistics **28**(1), 53–72 (2021)

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

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

[35] Mateau, J., Müller, W.G.: Spatio-temporal Design: Advances in Efficient Data Acquisition. John Wiley & Sons Inc., Chichester, United Kingdom (2013)

[36] Karney, C.F.F.: Algorithms for geodesics. J. Geodesy (87), 43–55 (2013)

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

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

[39] Cherubini, U., Luciano, E., Vecchiato, W.: Copula Methods in Finance. John Wiley & Sons Inc., Chichester, West Sussex (2004)

[40] Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**, 193–218 (1985)

[41] Menapace, A., Thellufsen, J.Z., Pernigotto, G., Roberti, F., Gasparella,

A., Righetti, M., Baratieri, M., Lund, H.: The design of 100% renewable smart urban energy systems: the case of bozen-bolzano. Energy **207**, 118198 (2020)

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

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

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

[45] Baltagi, B.: Econometric Analysis of Panel Data. John Wiley & Sons Inc., New York (1995)

[46] Wooldridge, J.: Econometrics Analysis of Cross Section and Panel Data. MIT Press, Cambridge (2002)

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

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

Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. Methodology development was performed by F. Marta L. Di Lascio and Roberta Pappadà, and Data preparation was performed by Andrea Menapace. Sections 1, 4, 5.2 were written by F. Marta L. Di Lascio and Roberta Pappadà, Section 2 was written by Roberta Pappadà, Section 3 was written by F. Marta L. Di Lascio, Sections 5.1 and 5.3 were written by Andrea Menapace, and Section 6 was written by all the three authors. All authors read and approved the final manuscript.