

# Spatial weighted robust clustering of multivariate time series based on quantile dependence with an application to mobility during COVID-19 pandemic

Ángel López-Oriona<sup>1</sup>, Pierpaolo D'Urso<sup>2</sup>, José A. Vilar<sup>1</sup> and Borja Lafuente-Rego<sup>1</sup>

<sup>1</sup>*Department of Mathematics, Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña, Spain*

<sup>2</sup>*Department of Social Sciences and Economics, Sapienza University of Rome, Italy*

## Abstract

In this paper, a fuzzy clustering model for multivariate time series based on the quantile cross-spectral density and principal component analysis is improved. The extension consists of (i) a weighting system which assigns a weight to each principal component in accordance with its importance concerning the underlying clustering structure and (ii) a penalization term allowing to take into account the spatial information. The iterative solutions of the new model, which employs the exponential distance in order to gain robustness against outlying series, are derived. A simulation study shows that the introduction of the weighting system substantially enhances the effectiveness of the former approach. The behaviour of the extended model in terms of the spatial penalization term is also analysed. An application involving multivariate time series of mobility indicators concerning COVID-19 pandemic highlights the usefulness of the proposed technique.

## Keywords

clustering, multivariate time series, principal component analysis, quantile cross-spectral density, weighting system, spatial statistics, COVID-19

## 1. Introduction and related work

Clustering of time series is a central problem in data mining with applications in many fields [1, 2]. The objective is to split a large set of unlabelled time series into homogeneous groups so that similar series are placed together in the same group and dissimilar series are located in different groups. The majority of the proposed approaches concern univariate time series (UTS) [3, 4, 5], while clustering of multivariate time series (MTS) has received much less attention [6].

This paper improves a fuzzy clustering model for MTS proposed in our previous work [7]. This model is based on the so-called Quantile Cross-spectral Density (QCD) and the classical Principal Component Analysis (PCA). Although highly successful, the model in [7] suffers from two major drawbacks. First, each principal component contributes equally to the objective function of the clustering algorithm. This overlooks the fact that some principal components could have more importance than others concerning the underlying clustering structure. Second, no spatial information is considered in the clustering model, so it can not handle time series datasets containing geographical information in a proper way. This paper proposes a new fuzzy clustering model aimed at overcoming both previously mentioned issues.

---


*The 13th International Workshop on Fuzzy Logic and Applications*

✉ oriona38@hotmail.com (Á. López-Oriona); pierpaolo.durso@uniroma1.it (P. D'Urso); jose.vilar@udc.es (J. A. Vilar); borja.lafuente@udc.es (B. Lafuente-Rego)

🆔 0000-0003-1456-7342 (Á. López-Oriona); 0000-0002-7406-6411 (P. D'Urso); 0000-0001-5494-171X (J. A. Vilar); 0000-0002-0877-7063 (B. Lafuente-Rego)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In Section 2, the procedure presented in [7] is concisely revised and its weighted version is introduced. A brief numerical experiment is carried out in order to show why introducing a weighting system in the principal components space is advantageous. Section 3 describes the new proposed model, in which spatial information is taken into account. A simple simulation study is performed to assess the behaviour of the procedure. In Section 4, the technique is applied to a dataset of mobility concerning the COVID-19 pandemic. Section 5 concludes. Finally, the Appendix contains the derivation of the iterative solutions of the proposed algorithm.

## 2. Fuzzy clustering models based on the quantile cross-spectral density

In this section, we first introduce a fuzzy clustering model based on QCD and PCA. Then, we present a modification of this model in which a different weight is given to each principal component. A numerical experiment shows the usefulness of the latter approach.

### 2.1. A fuzzy clustering model based on QCD and PCA

Let  $\{\mathbf{X}_t, t \in \mathbb{Z}\} = \{(X_{t,1}, \dots, X_{t,d}), t \in \mathbb{Z}\}$  be a  $d$ -variate real-valued strictly stationary stochastic process. Denote by  $F_j$  the marginal distribution function of  $X_{t,j}$ ,  $j = 1, \dots, d$ , and by  $q_j(\tau) = F_j^{-1}(\tau)$ ,  $\tau \in [0, 1]$ , the corresponding quantile function. Fixed  $l \in \mathbb{Z}$  and an arbitrary couple of quantile levels  $(\tau, \tau') \in [0, 1]^2$ , consider the cross-covariance of the indicator functions  $I\{X_{t,j_1} \leq q_{j_1}(\tau)\}$  and  $I\{X_{t+l,j_2} \leq q_{j_2}(\tau')\}$  given by

$$\gamma_{j_1, j_2}(l, \tau, \tau') = \text{Cov}(I\{X_{t,j_1} \leq q_{j_1}(\tau)\}, I\{X_{t+l,j_2} \leq q_{j_2}(\tau')\}), \quad (1)$$

for  $1 \leq j_1, j_2 \leq d$ . Taking  $j_1 = j_2 = j$ , the function  $\gamma_{j,j}(l, \tau, \tau')$ , with  $(\tau, \tau') \in [0, 1]^2$ , so-called quantile autocovariance function (QAF) of lag  $l$ , generalizes the traditional autocovariance function.

In the case of the multivariate process  $\{\mathbf{X}_t, t \in \mathbb{Z}\}$ , we can consider the  $d \times d$  matrix

$$\mathbf{\Gamma}(l, \tau, \tau') = (\gamma_{j_1, j_2}(l, \tau, \tau'))_{1 \leq j_1, j_2 \leq d}, \quad (2)$$

which jointly provides information about both the cross-dependence (when  $j_1 \neq j_2$ ) and the serial dependence (because the lag  $l$  is considered).

In the same way as the spectral density is the representation in the frequency domain of the autocovariance function, the spectral counterpart for the cross-covariances  $\gamma_{j_1, j_2}(l, \tau, \tau')$  can be introduced. Under suitable summability conditions (mixing conditions), the Fourier transform of the cross-covariances is well-defined and the *quantile cross-spectral density* is given by

$$f_{j_1, j_2}(\omega, \tau, \tau') = (1/2\pi) \sum_{l=-\infty}^{\infty} \gamma_{j_1, j_2}(l, \tau, \tau') e^{-il\omega}, \quad (3)$$

for  $1 \leq j_1, j_2 \leq d$ ,  $\omega \in \mathbb{R}$  and  $\tau, \tau' \in [0, 1]$ . Note that  $f_{j_1, j_2}(\omega, \tau, \tau')$  is complex-valued.

The quantile cross-spectral density contains information about the general dependence structure of a given stochastic process. For a specific realization of the process, this quantity can be consistently estimated by means of the so-called smoothed CCR-periodogram,  $\hat{G}_{T,R}^{j_1, j_2}(\omega, \tau, \tau')$ , proposed by [8].

Based on previous remarks, a simple dissimilarity measure between two realizations of the  $d$ -variate process (MTS) can be defined as follows. Given the  $i$ -th MTS,  $\mathbf{X}_t^{(i)}$ , consider the set  $G^{(i)} = \{\hat{G}_{T,R}^{j_1, j_2}(\omega, \tau, \tau'), j_1, j_2 = 1, \dots, d, \omega \in \Omega, \tau, \tau' \in \mathcal{T}\}$ , where  $\Omega$  is the set of Fourier frequencies and  $\mathcal{T} = \{0.1, 0.5, 0.9\}$ . Let  $\Psi^{(i)}$  be the vector formed by concatenating separately the real and imaginary parts of the elements of the set  $G^{(i)}$ . A dissimilarity measure between the series  $\mathbf{X}_t^{(1)}$  and  $\mathbf{X}_t^{(2)}$  is defined as the squared Euclidean distance between the vectors  $\Psi^{(1)}$  and  $\Psi^{(2)}$ . We call this dissimilarity  $d_{QCD}$ .

The distance measure  $d_{QCD}$  can be used as input to the traditional fuzzy  $C$ -medoids algorithm to develop a clustering procedure for MTS. However, the numerical simulations carried out in [7] have shown that performing PCA as a preprocessing step frequently improves the effectiveness of the metric  $d_{QCD}$  in a clustering context. More

precisely, given a set of  $n$  MTS,  $\{\mathbf{X}_t^{(1)}, \dots, \mathbf{X}_t^{(n)}\}$ , the previously mentioned QCD-based features are extracted from each series, thus providing the set  $\Psi = \{\Psi^{(1)}, \dots, \Psi^{(n)}\}$ . These vectors are transformed by means of PCA, obtaining the set of score vectors  $\Psi_{PCA} = \{\Psi_{PCA}^{(1)}, \dots, \Psi_{PCA}^{(n)}\}$ . The corresponding distance used in the clustering algorithm is  $d_{QCD_{PCA}}(\mathbf{X}_t^{(1)}, \mathbf{X}_t^{(2)}) = \|\Psi_{PCA}^{(1)} - \Psi_{PCA}^{(2)}\|^2$ . From now on, we assume that this is the distance being considered, although the subscript  $PCA$  is removed for the sake of simplicity. It is worth remarking that the use of PCA has been shown to improve the accuracy of clustering algorithms in several contexts [9, 10, 11, 12].

The metric  $d_{QCD}$  is used to develop the so-called *QCD-based Fuzzy C-Medoids clustering model* (QCD-FCMd), whose goal is to find the subset of  $\Psi$  of size  $C$ ,  $\tilde{\Psi} = \{\tilde{\Psi}^{(1)}, \dots, \tilde{\Psi}^{(C)}\}$ , and the  $n \times C$  matrix of fuzzy coefficients,  $U$ , solving the minimization problem

$$\min_{\tilde{\Psi}, U} \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \|\Psi^{(i)} - \tilde{\Psi}^{(c)}\|^2 \text{ with respect to } \sum_{c=1}^C u_{ic} = 1 \forall i \text{ and } u_{ic} \geq 0, \quad (4)$$

where  $u_{ic} \in [0, 1]$  represents the membership degree of the  $i$ -th series in the  $c$ -th cluster,  $\tilde{\Psi}^{(c)}$  is the vector of QCD-based features with regards to the medoid series for the cluster  $c$ , and  $m > 1$  is a parameter controlling the fuzziness of the partition, usually referred to as fuzziness parameter.

Several hyperparameters have to be selected in the model QCD-FCMd. In particular, the number of clusters,  $C$ , the fuzziness parameter,  $m$  and the number of retained principal components,  $p$ , need to be set in advance. We propose to perform hyperparameter selection through a grid search by considering the four internal clustering quality indexes given in [13]. It is worth mentioning that the model QCD-FCMd was compared in [7] with several alternative procedures, clearly outperforming all of them.

Despite its good behaviour, The QCD-FCMd model suffers from a major limitation: all the selected principal components receive the same importance concerning the distances in (4), thus ignoring that some principal components often contain more information than others about the underlying clustering structure. Solving the limitations of the QCD-FCMd model is one of the main motivations of the present work.

## 2.2. A weighted fuzzy clustering model based on QCD and PCA

Consider the set of MTS  $\{\mathbf{X}_t^{(1)}, \dots, \mathbf{X}_t^{(n)}\}$  and denote by  $\Psi = \{\Psi^{(1)}, \dots, \Psi^{(n)}\}$  the set of QCD-based features. Assume that the subset of  $p$  first principal components was selected. In this way, each  $\Psi^{(i)}$  is a  $p$ -dimensional vector,  $\Psi^{(i)} = (\Psi_1^{(i)}, \dots, \Psi_p^{(i)})$ ,  $i = 1, \dots, n$ . We introduce now the *Weighted QCD-based Fuzzy C-Medoids clustering model* (W-QCD-FCMd), which intends to get around the limitations of the QCD-FCMd model by considering the minimization problem

$$\begin{aligned} & \min_{\tilde{\Psi}, w, U} \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \sum_{k=1}^p \left( w_k \left( \Psi_k^{(i)} - \tilde{\Psi}_k^{(c)} \right) \right)^2 \\ & \text{subject to } \sum_{c=1}^C u_{ic} = 1 \text{ and } u_{ic} \geq 0, \sum_{k=1}^p w_k = 1, w_k \geq 0, \text{ for } k = 1, \dots, p, \end{aligned} \quad (5)$$

where  $w = \{w_1, \dots, w_p\}$  is a set of  $p$  weights and the remaining terms are the same as in (4). The W-QCD-FCMd model associates the weight  $w_k$  to the  $k$ -th principal component when computing the distance in (5). Since the weights are variables in the objective function, the set  $w$  gets properly estimated. Basically, W-QCD-FCMd allows for an appropriate tuning of the influence of the various principal components when computing the dissimilarity between MTS. Hence, the algorithm can give more importance to the principal components containing the most valuable information about the structure of the dataset. The iterative solutions of (5) can be easily obtained through the Lagrange multipliers method and the procedure is quite similar to the one shown in the Appendix for computing the solutions of the model presented in Section 3.

In order to show the advantages of the W-QCD-FCMd model over QCD-FCMd model, we constructed a simple simulation study as follows. We considered a scenario involving four different generating processes. 10 series were simulated from each process and the corresponding set of 40 MTS was subject to clustering by means of both models QCD-FCMd and W-QCD-FCMd. The generating processes were bivariate vector moving average processes

with vectorized matrices of coefficients  $(-0.3, 0, 0, 0)$ ,  $(0.5, 0, 0, 0)$ ,  $(0, 0, 0.3, 0)$  and  $(0, 0, 0, 0.7)$ , and Gaussian innovations. The number of clusters was set to  $C = 4$ . The series length was set to  $T = 300$ . Several values for the fuzziness parameter,  $m$ , and the retained number of principal components,  $p$ , were taken into account. Specifically, we considered  $m = 1.4, 1.6, 1.8, 2$  and  $p = 1, 2, 3, 4, 5$ .

According to our clustering purpose, the partition defined by the generating processes was assumed to be the true partition (the ground truth) and the assessment of both approaches was carried out with the fuzzy extension of the Adjusted Rand Index [14], denoted by FARI. The simulation procedure was repeated 200 trials and average values of FARI were calculated for each combination of  $m$  and  $p$ . Table 1 contains the results. It is clear that, overall, the model W-QCD-FCMd substantially outperformed QCD-FCMd in most of the settings. The only exception was when  $p = 1$ , when both models are the same (only one weight is present in W-QCD-FCMd for the first principal component). We have performed additional numerical experiments, achieving similar results. Therefore, we conclude that the W-QCD-FCMd model substantially improves the QCD-FCMd model.

		$m = 1.4$	$m = 1.6$	$m = 1.8$	$m = 2$
$p = 1$	QCD-FCMd	<b>0.58</b>	0.53	<b>0.50</b>	0.40
	W-QCD-FCMd	0.55	<b>0.55</b>	0.49	<b>0.45</b>
$p = 2$	QCD-FCMd	0.70	0.53	0.36	0.26
	W-QCD-FCMd	<b>0.86</b>	<b>0.82</b>	<b>0.72</b>	<b>0.55</b>
$p = 3$	QCD-FCMd	0.78	0.53	0.38	0.31
	W-QCD-FCMd	<b>0.91</b>	<b>0.84</b>	<b>0.65</b>	<b>0.52</b>
$p = 4$	QCD-FCMd	0.45	0.33	0.23	0.20
	W-QCD-FCMd	<b>0.60</b>	<b>0.48</b>	<b>0.38</b>	<b>0.30</b>
$p = 5$	QCD-FCMd	0.32	0.23	0.17	0.13
	W-QCD-FCMd	<b>0.28</b>	<b>0.32</b>	<b>0.25</b>	<b>0.18</b>

**Table 1**

Average FARI obtained by QCD-FCMd and W-QCD-FCMd in the simulation scenario. For each pair  $(p, m)$ , the best result is shown in bold.

### 3. The QCD-based exponential weighted fuzzy C-medoids clustering model with spatial component

In the context of previous sections, we propose to perform partitional fuzzy clustering on  $\{\mathbf{X}_t^{(1)}, \dots, \mathbf{X}_t^{(n)}\}$  by using the *Weighted QCD-based Exponential Fuzzy C-Medoids clustering model with Spatial component* (EW-QCD-FCMd-S), whose aim is to find the subset of  $\Psi$  of size  $C$ ,  $\tilde{\Psi} = \{\tilde{\Psi}^{(1)}, \dots, \tilde{\Psi}^{(C)}\}$ , the set of  $p$  weights  $\mathbf{w} = \{w_1, \dots, w_p\}$ , and the  $n \times C$  matrix of fuzzy coefficients,  $\mathbf{U} = (u_{ic})$ , solving the minimization problem:

$$\min_{\tilde{\Psi}, \mathbf{w}, \mathbf{U}} \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \left[ 1 - \exp \left( -\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c)})^2 \right) \right] + \frac{\gamma}{2} \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \sum_{i'=1}^n \sum_{c' \in C_c} p_{ii'} u_{i'c'}^m$$

subject to: (6)

$$\sum_{c=1}^C u_{ic} = 1, u_{ic} \geq 0, \text{ for } i = 1, \dots, n; c = 1, \dots, C, \sum_{k=1}^p w_k = 1, w_k \geq 0, \text{ for } k = 1, \dots, p,$$

where  $\beta > 0$  is a constant tuning the membership degrees to gain robustness to outlying series;  $\mathbf{P} = (p_{ii'})$  is a symmetric  $n \times n$  matrix introducing information on spatial proximity of the statistical units, with  $p_{ii'} \geq 0$  for  $i \neq i'$  and  $p_{ii} = 0$ ,  $i = 1, \dots, n$ ;  $\gamma \geq 0$  is a coefficient regulating the effect of the spatial proximity (such as described below);  $C_c = \{1, \dots, c-1, c+1, \dots, C\}$  denotes the set of all the clusters except cluster  $c$ ; and the remaining elements are the same as in (5).

The objective function in (6) is formed by two specific terms, which are described in detail below.

[1] The term  $\sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \left[ 1 - \exp \left( -\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c)})^2 \right) \right]$  corresponds, in essence, to the standard term in the classical fuzzy  $C$ -medoids clustering algorithm but including two main modifications. The Euclidean distance is

replaced by an exponential-type distance to endow the clustering procedure with higher robustness against outliers [15]. The exponential distance assigns small weights to anomalous points and larger weights to elements from well-defined and compact clusters (for more details, see [15, 16]). Moreover, this term also considers different weights for each principal component in order to improve the accuracy of the standard non-weighted version such as argued in Section 2.

[2] The second term,  $\frac{\gamma}{2} \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \sum_{i'=1}^n \sum_{c' \in C_c} p_{ii'} u_{i'c'}^m$ , is a spatial penalty term, which can be seen as a regularization element. The matrix  $\mathbf{P}$  incorporates the spatial information of the  $n$  units where the MTS were recorded and it is often restricted to indicate if two different units are or not contiguous, i.e.,  $\mathbf{P}$  is usually constructed as follows:

$$p_{ii'} = 1 \text{ if MTS } i \text{ is contiguous to MTS } i', 0 \text{ otherwise.} \quad (7)$$

The goal of the spatial regularization term is penalizing the fact that two contiguous units are located in different clusters with high membership degrees, since it is expected that neighbouring sites tend to have similar features. The constant  $\gamma$  regulates the trade-off between internal cohesion based on the set  $\Psi$  and the fact that the clusters are constituted by adjacent MTS. In this way, large values of  $\gamma$  force neighbouring spatial units to belong to the same cluster, whereas when  $\gamma = 0$ , the spatial information is not taken into account.

The optimal iterative solutions of the minimization problem (6) are given in the Appendix.

**Remark 1** (*Selection of  $\beta$* ). An appropriate choice of the hyperparameter  $\beta$  is totally crucial for a good performance of the EW-QCD-FCMd-S procedure. The value of  $\beta$  is determined in this work as the inverse of the variability in the data, which is a common choice in the literature. For more details, see [16].

**Remark 2** (*Selection of  $\gamma$* ). The selection of the optimal value for  $\gamma$  is a complex problem and frequently relies on heuristic procedures. To that aim, we adapt a procedure given in [17], which is based on maximizing the so-called within cluster spatial autocorrelation.

A brief simulation study was carried out in order to assess the behaviour of the EW-QCD-FCMd-S model. The simulation mechanism involved two vector autoregressive processes of order 1, denoted by  $\text{VAR}_1$  and  $\text{VAR}_2$ , with vectorized matrices of coefficients  $(-0.2, 0.2, 0.2, 0.2)$  and  $(0.3, -0.2, 0, 0)$ , respectively, and Gaussian innovations. Ten different series were generated from each process. The first 8 MTS and MTS 19 and 20 pertained to  $\text{VAR}_1$ . On the other hand, MTS from 9 to 18 were associated with  $\text{VAR}_2$ . We encapsulated the spatial information by means of a matrix  $\mathbf{P}$  indicating that the first 10 MTS belonged to the same area, the same holding for the last 10 MTS. The matrix  $\mathbf{P}$  was constructed as

$$p_{ij} = 1, i, j = 1, \dots, 10, i \neq j, \quad p_{kl} = 1, k, l = 11, \dots, 20, k \neq l. \quad (8)$$

The remaining entries of the matrix  $\mathbf{P}$  were set to zero. The number of clusters was set to  $C = 2$ . The series length was set to  $T = 500$ . The hyperparameter  $\beta$  was chosen to be  $\beta = 1$  for illustrative purposes. Concerning the number of principal components,  $p$ , we selected the optimal value of  $p \in \{2, \dots, 6\}$  according to the FARI attained when  $\gamma = 0$  and the true partition is given by the true generating processes, which resulted  $p = 2$ .

The simulation mechanism was performed for all pairs  $(m, \gamma)$ , with  $m = 1.4, 1.6, 1.8, 2$  and  $\gamma = 0, 0.01, 0.05, 0.1, 0.2, 0.5$ . Note that, when  $\gamma = 0$ , there is no spatial penalization term, whereas when increasing the values of  $\gamma$ , more importance is given to this term, which is, the clusters are more required to be formed by series in the same area to a higher extent.

To gain insights into the behaviour of EW-QCD-FCMd-S from a spatial point of view, we recorded, for each pair  $(m, \gamma)$ , the sum of the membership degrees of the MTS 9, 10, 19 and 20 with respect to the cluster in which they must be located according to the matrix  $\mathbf{P}$ . In other words, we took the membership degrees of MTS 9 and 10 in the cluster where MTS 1-8 showed a higher membership degree. Analogously, we took the membership degrees of MTS 19 and 20 in the cluster where MTS 11-18 showed a higher membership degree. The average of the 4 previous quantities was taken into account. Note that this measure indicates to what extent the spatial requirements defined by  $\mathbf{P}$  are fulfilled. Results are given in Table 2 concerning 200 simulation trials. As expected, when  $\gamma = 0$ , the corresponding quantities are close to 0 as the grouping is made uniquely based on underlying generating processes. When the value of  $\gamma$  increases, the MTS 9, 10, 19 and 20 are more affected by the penalization term, constituting clusters along with MTS in the same areas, and the corresponding measures also increase. Indeed, for  $\gamma = 0.5$ , the

	$m = 1.4$	$m = 1.6$	$m = 1.8$	$m = 2$
$\gamma = 0$	0.05	0.07	0.08	0.08
$\gamma = 0.01$	0.06	0.08	0.11	0.14
$\gamma = 0.05$	0.17	0.23	0.27	0.29
$\gamma = 0.1$	0.44	0.45	0.44	0.43
$\gamma = 0.2$	0.78	0.69	0.64	0.58
$\gamma = 0.5$	0.90	0.86	0.81	0.76

**Table 2**

Average membership degree of MTS 9, 10, 19 and 20 with respect to the cluster they must belong according to the matrix  $\mathbf{P}$ .

spatial requirement gets so stringent that the series are located in their spatial clusters with very high membership degrees.

## 4. Application. Spatial clustering of Spanish regions based on mobility time series during COVID-19 pandemic

In this section we develop a study case related to the non-supervised classification of geographical zones in terms of their temporal records of some mobility indicators concerning the COVID-19 pandemic.

In particular, we considered mobility time series observed during COVID-19 outbreak in the 17 autonomous communities of Spain. Specifically, we recorded daily data about mobility in (i) retail and recreation places, (ii) groceries and pharmacies and (iii) transit stations. The three variables are measured as a percentage of change with respect to a baseline level associated with the pre-pandemic stage. The sample period spans from 15th February 2020 to 22nd May 2021, thus resulting serial realizations of length  $T = 1096$ . Each community is described by means of a trivariate series indicating the daily percent of change in mobility with respect to the mentioned locations. Note that, from an epidemiological point of view, it is reasonable to think that the joint behaviour of these three quantities is different depending on the geographical location of each region.

The EW-QCD-FCMd-S model was applied to the set of 17 series. Note that 5 hyperparameters had to be chosen, namely  $p$ ,  $C$ ,  $m$ ,  $\beta$  and  $\gamma$ . To choose  $p$ ,  $C$  and  $m$ , we applied a grid search based on the four internal clustering quality indexes presented in [13]. To this aim, we considered the W-QCD-FCMd clustering model, which does not require the selection of  $\beta$  and  $\gamma$ . The optimal combination was  $(p, m, C) = (2, 1.7, 4)$ . Given the selected value  $p = 2$ , the hyperparameter  $\beta$  was chosen according to the procedure indicated in [18]. The corresponding value was  $\beta = 0.432$ . The matrix  $\mathbf{P}$  was constructed so that  $p_{ij} = 1$  if communities  $i$  and  $j$  are adjacent and  $p_{ij} = 0$  otherwise. In this way, the only communities lacking neighbours are the Canary Islands and the Balearic Islands. The optimal value of the tuning parameter  $\gamma$  was chosen according to the procedure described in [17]. The optimal value of  $\gamma$  was  $\gamma = 1.76$ . The optimal weights were 0.82 and 0.18 for the first and second principal component, respectively.

Table 3 shows the membership degrees concerning the 4 clustering solution given by EW-QCD-FCMd-S. For each community, the highest value for the membership degree was highlighted in bold. According to Table 3, there exist a small cluster,  $C_4$ , including the autonomous communities of Asturias (AS) and Cantabria (CA). None of the remaining communities show a high membership degree in this cluster. There are three more clusters,  $C_1$ ,  $C_2$  and  $C_3$ , each one containing 5, 4, and 6 communities, respectively. Interestingly, the community of Castile and León (CL) presents membership degrees which are spread out between the 4 clusters. Hence, this region could be considered an outlier.

In order to clarify the results in Table 3, we have depicted in Figure 1 a map of Spain where the communities were coloured according to the underlying crisp clustering partition (blue, red, yellow and green colours for Clusters  $C_1$ ,  $C_2$ ,  $C_3$  and  $C_4$ , respectively). The abbreviation of each region given in Table 3 was incorporated. It is clear from Figure 1 that the resulting partition is highly interpretable from a spatial point of view, suggesting that the climate and other geographical conditions could have influenced how people moved during the COVID-19 pandemic. Clusters  $C_3$  and  $C_4$  are formed mainly by communities located in the northern part of the country, with prototypes Galicia (GA) and AS, respectively. On the other hand, Cluster  $C_1$  is constituted by communities close to the Mediterranean Sea. The medoid of Cluster  $C_1$  is the community of Castile-La Mancha (CM). It is worth



Community	$C_1$	$C_2$	$C_3$	$C_4$
Andalusia (AN)	<b>0.71</b>	0.19	0.08	0.03
Aragon (AR)	0.25	<b>0.54</b>	0.21	0.01
Asturias (AS)	0.00	0.00	0.00	<b>1.00</b>
Balearic Islands (BI)	<b>0.77</b>	0.13	0.04	0.05
Basque Country (BC)	0.21	0.30	<b>0.44</b>	0.06
Canary Islands (CI)	0.00	<b>1.00</b>	0.00	0.00
Cantabria (CB)	0.12	0.14	0.17	<b>0.58</b>
Castile and León (CL)	0.26	0.29	<b>0.35</b>	0.10
Castile-La Mancha (CM)	<b>1.00</b>	0.00	0.00	0.00
Catalonia (CT)	<b>0.62</b>	0.18	0.10	0.10
Community of Madrid (MD)	0.09	<b>0.85</b>	0.04	0.02
Extremadura (EX)	0.04	0.17	<b>0.74</b>	0.05
Galicia (GA)	0.00	0.00	<b>1.00</b>	0.00
La Rioja (RI)	0.03	0.07	<b>0.89</b>	0.01
Navarre (NC)	0.06	0.12	<b>0.81</b>	0.01
Region of Murcia (MC)	0.23	<b>0.70</b>	0.03	0.04
Valencian Community (VC)	<b>0.76</b>	0.16	0.04	0.04

**Table 3**

Membership degrees for the 17 Spanish communities by considering the EW-QCD-FCMd-S model and a 4 cluster partition. For each community, the highest value for the membership degree was highlighted in bold.



**Figure 1:** Map of Spain highlighting the crisp clustering partition given by the model EW-QCD-FCMd-S. Each community is represented in a different colour according to its cluster (blue, red, yellow and green colour for cluster  $C_1$ ,  $C_2$ ,  $C_3$  and  $C_4$ , respectively). A darker colour was used for the medoid communities.

remarking that the only Mediterranean area not included in Cluster  $C_1$  from a crisp point of view, the region of Murcia (MC), shows a relatively high membership in this group, 0.23. Finally, Cluster  $C_2$  is the more heterogeneous cluster from a geographical perspective. This cluster is composed by a northern community, Aragon (AR), a central community, Madrid (MD), a southern zone, Murcia (MC), and the insular area of the Canary Islands (CI), which corresponds to the medoid. These islands are located far away from the Iberian Peninsula although they are shown near Andalusia (AN) in Figure 1 for the sake of simplicity. It is important to emphasize that, although the elements in Cluster  $C_2$  are not spatially connected, the strong similarity shown by the corresponding MTS offsets the spatial penalization term and provokes the formation of this group. Further investigations based on epidemiological data will be carried out so as to explain the formation of this group.

## 5. Concluding remarks

In this work, we have extended a fuzzy clustering model for multivariate time series based on QCD and PCA. The former model employs a set of features obtained through the smoothed CCR-periodogram and performs the traditional fuzzy  $C$ -medoids clustering algorithm by considering the squared Euclidean distance in the principal

components space.

The proposed extension is based on two tools. On the one hand, a system of weights is introduced in the objective function so that more importance is given to the principal components having more discriminative ability in terms of the clustering structure. On the other hand, a penalization term which permits to take into account spatial information is considered. The resulting model uses the exponential distance, thus attaining robustness against outlying series.

We have showed that the consideration of the weighting system is totally advantageous in terms of clustering effectiveness. The behaviour of the model according to the spatial penalization term was also analysed by means of a brief example. Finally, the proposed technique was applied to cluster a dataset containing COVID-19 data where the MTS come from different geographic locations, leading to illuminating conclusions.

## Appendix

Here we derive the iterative solutions of the constrained minimization problem (6) via the Lagrangian multiplier method.

First, consider the corresponding Lagrangian function taking the form:

$$L(\mathbf{U}, \mathbf{w}, \boldsymbol{\lambda}, \rho) = \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \left[ 1 - \exp \left( -\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c)})^2 \right) \right] + \frac{\gamma}{2} \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \sum_{i'=1}^n \sum_{c'' \in C_c} p_{ii'} u_{i'c''}^m - \sum_{i=1}^n \lambda_i \left( \sum_{c=1}^C u_{ic} - 1 \right) - \rho \left( \sum_{k=1}^p w_k - 1 \right), \quad (9)$$

where  $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_n\}$  and  $\rho$  stand for the Lagrange multipliers concerning the constraints of the membership degrees and the weights, respectively.

By fixing  $\mathbf{w}$  and setting equal to zero the partial derivatives of  $L$  with respect to  $u_{ic}$  and  $\lambda_i$ , for arbitrary  $i \in \{1, \dots, n\}$  and  $c \in \{1, \dots, C\}$ , we obtain that

$$\frac{\partial L(\mathbf{U}, \mathbf{w}, \boldsymbol{\lambda}, \rho)}{\partial u_{ic}} = 0, \quad \frac{\partial L(\mathbf{U}, \mathbf{w}, \boldsymbol{\lambda}, \rho)}{\partial \lambda_i} = 0,$$

is equivalent to

$$m u_{ic}^{m-1} \left[ 1 - \exp \left( -\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c)})^2 \right) \right] + \gamma \sum_{i'=1}^n \sum_{c'' \in C_c} p_{ii'} u_{i'c''}^m - \lambda_i = 0, \quad \sum_{c'=1}^C u_{ic'} - 1 = 0. \quad (10)$$

From the first equation in (10) follows

$$u_{ic} = \left( \frac{\lambda_i}{m} \right)^{\frac{1}{m-1}} \left[ 1 - \exp \left( -\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c)})^2 \right) \right] + \gamma \sum_{i'=1}^n \sum_{c'' \in C_c} p_{ii'} u_{i'c''}^m \right]^{\frac{-1}{m-1}}. \quad (11)$$

Replacing (11) in the second equation of (10), we obtain

$$\left( \frac{\lambda_i}{m} \right)^{\frac{1}{m-1}} \sum_{c'=1}^C \left[ 1 - \exp \left( -\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c')})^2 \right) \right] + \gamma \sum_{i'=1}^n \sum_{c'' \in C_{c'}} p_{ii'} u_{i'c''}^m \right]^{\frac{-1}{m-1}} = 1, \quad (12)$$

which yields

$$\left( \frac{\lambda_i}{m} \right)^{\frac{1}{m-1}} = \left[ \sum_{c'=1}^C \left[ 1 - \exp \left( -\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c')})^2 \right) \right] + \gamma \sum_{i'=1}^n \sum_{c'' \in C_{c'}} p_{ii'} u_{i'c''}^m \right]^{\frac{-1}{m-1}}. \quad (13)$$

Finally, by replacing (13) in (11), we conclude that

$$u_{ic} = \left[ \sum_{c'=1}^C \left( \frac{1 - e^{(-\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c')})^2)} + S_1}{1 - e^{(-\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c')})^2)} + S_2} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad (14)$$



with  $S_1 = \gamma \sum_{i'=1}^n \sum_{c' \in C_c} p_{ii'} u_{i'c'}^m$ ,  $S_2 = \gamma \sum_{i'=1}^n \sum_{c' \in C_{c'}} p_{ii'} u_{i'c'}^m$ . Expression (14) gives the iterative solution for the membership degrees.

On the other hand, the optimal weights  $w_k$  can be determined in a similar way. Now, we fix  $u_{ic}$  and set equal to zero the partial derivatives of  $L$  with respect to  $w_k$  and the Lagrange multiplier  $\rho$  so that an equivalence occurs between

$$\frac{\partial L(\mathbf{U}, \mathbf{w}, \boldsymbol{\lambda}, \rho)}{\partial w_k} = 0, \quad \frac{\partial L(\mathbf{U}, \mathbf{w}, \boldsymbol{\lambda}, \rho)}{\partial \rho} = 0,$$

and

$$\sum_{i=1}^n \sum_{c=1}^C u_{ic}^m \exp\left(-\beta \sum_{k'=1}^p w_{k'}^2 (\Psi_{k'}^{(i)} - \tilde{\Psi}_{k'}^{(c)})^2\right) \times \left(-2w_k \beta\right) (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c)})^2 - \rho = 0, \quad \sum_{k'=1}^p w_{k'} - 1 = 0. \quad (15)$$

From the first equation in (15), we can write  $w_k$  as follows:

$$w_k = -\frac{\rho}{2\beta} \left[ \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c)})^2 \times \exp\left(-\beta \sum_{k'=1}^p w_{k'}^2 (\Psi_{k'}^{(i)} - \tilde{\Psi}_{k'}^{(c)})^2\right) \right]^{-1}. \quad (16)$$

By replacing (16) in the second equation of (15), we have

$$-\frac{\rho}{2\beta} \sum_{k'=1}^p \left[ \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m (\Psi_{k'}^{(i)} - \tilde{\Psi}_{k'}^{(c)})^2 \times \exp\left(-\beta \sum_{k''=1}^p w_{k''}^2 (\Psi_{k''}^{(i)} - \tilde{\Psi}_{k''}^{(c)})^2\right) \right]^{-1} = 1, \quad (17)$$

which yields

$$-\frac{\rho}{2\beta} = \left[ \sum_{k'=1}^p \left[ \sum_{i=1}^n \sum_{c=1}^C u_{ic}^m (\Psi_{k'}^{(i)} - \tilde{\Psi}_{k'}^{(c)})^2 \times \exp\left(-\beta \sum_{k''=1}^p w_{k''}^2 (\Psi_{k''}^{(i)} - \tilde{\Psi}_{k''}^{(c)})^2\right) \right]^{-1} \right]^{-1}. \quad (18)$$

By plugging in (16) the expression for  $-\rho/2\beta$  in (18), we obtain the iterative solution for the weights, which takes the form:

$$w_k = \left[ \sum_{k'=1}^p \frac{\sum_{i=1}^n \sum_{c=1}^C u_{ic}^m (\Psi_k^{(i)} - \tilde{\Psi}_k^{(c)})^2 \text{dexp}(i, c)}{\sum_{i=1}^n \sum_{c=1}^C u_{ic}^m (\Psi_{k'}^{(i)} - \tilde{\Psi}_{k'}^{(c)})^2 \text{dexp}(i, c)} \right]^{-1}, \quad (19)$$

where for the sake of simplicity we have used the notation  $\text{dexp}(i_1, i_2) = \exp\left(-\beta \sum_{k=1}^p w_k^2 (\Psi_k^{(i_1)} - \tilde{\Psi}_k^{(i_2)})^2\right)$ .

## References

- [1] T. W. Liao, Clustering of time series data—a survey, *Pattern recognition* 38 (2005) 1857–1874.
- [2] S. Aghabozorgi, A. S. Shirkhorshidi, T. Y. Wah, Time-series clustering—a decade review, *Information Systems* 53 (2015) 16–38.
- [3] B. Lafuente-Rego, J. A. Vilar, Clustering of time series using quantile autocovariances, *Advances in Data Analysis and classification* 10 (2016) 391–415.
- [4] J. A. Vilar, B. Lafuente-Rego, P. D'Urso, Quantile autocovariances: a powerful tool for hard and soft partitioning of time series, *Fuzzy Sets and Systems* 340 (2018) 38–72.
- [5] P. D'Urso, E. A. Maharaj, Autocorrelation-based fuzzy clustering of time series, *Fuzzy Sets and Systems* 160 (2009) 3565–3589.
- [6] P. D'Urso, E. A. Maharaj, Wavelets-based clustering of multivariate time series, *Fuzzy Sets and Systems* 193 (2012) 33–61.
- [7] Á. López-Oriona, J. A. Vilar, Quantile cross-spectral density: A novel and effective tool for clustering multivariate time series, *Expert Systems with Applications* 185 (2021) 115677.
- [8] J. Barunik, T. Kley, Quantile coherency: A general measure for dependence between cyclical economic variables, *The Econometrics Journal* 22 (2019) 131–152.
- [9] J. Xue, C. Lee, S. G. Wakeham, R. A. Armstrong, Using principal components analysis (pca) with cluster analysis to study the organic geochemistry of sinking particles in the ocean, *Organic Geochemistry* 42 (2011) 356–367.
- [10] N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, P. Patargias, Using principal component and cluster analysis in the heating evaluation of the school building sector, *Applied Energy* 87 (2010) 2079–2086.
- [11] C. Ding, X. He, K-means clustering via principal component analysis, in: *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.
- [12] S. Raychaudhuri, J. M. Stuart, R. B. Altman, Principal components analysis to summarize microarray experiments: application to sporulation time series, in: *Bioinformatics* 2000, World Scientific, 1999, pp. 455–466.
- [13] K. Zhou, C. Fu, S. Yang, Fuzziness parameter selection in fuzzy c-means: the perspective of cluster validation, *Science China Information Sciences* 57 (2014) 1–8.
- [14] R. J. Campello, A fuzzy extension of the rand index and other related indexes for clustering and classification assessment, *Pattern Recognition Letters* 28 (2007) 833–841.
- [15] K.-L. Wu, M.-S. Yang, Alternative c-means clustering algorithms, *Pattern recognition* 35 (2002) 2267–2278.
- [16] P. D'Urso, L. De Giovanni, R. Massari, Time series clustering by a robust autoregressive metric with application to air pollution, *Chemometrics and Intelligent Laboratory Systems* 141 (2015) 107–124.
- [17] R. Coppi, P. D'Urso, P. Giordani, A fuzzy clustering model for multivariate spatial time series, *Journal of Classification* 27 (2010) 54–88.
- [18] P. D'Urso, L. De Giovanni, E. A. Maharaj, R. Massari, Wavelet-based self-organizing maps for classifying multivariate time series, *Journal of Chemometrics* 28 (2014) 28–51.