COPULA-BASED FUZZY CLUSTERING OF TIME SERIES

Pierpaolo D'Urso[1] , Marta Disegna[2] and Fabrizio Durante[3]

[1] Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma, Rome, Italy, (e-mail: pierpaolo.durso@uniroma1.it)

[2] Faculty of Management, School of Tourism, Bournemouth University, UK (e-mail: disegnam@bournemouth.ac.uk)

[3] Faculty of Economics and Management, Free University of Bolzano, Bolzano, Italy, (e-mail: marta.disegna@unibz.it, fabrizio.durante@unibz.it)

ABSTRACT: Motivated by a real problem of tourism destination and regional studies, we develop a clustering algorithm to identify similar patterns of tourism time series. The algorithm joins the copula–approach to cluster analysis with the fuzzy methodology, allowing to extend usual clustering methods for time series based on Pearson's correlation and to capture the uncertainty that arises assigning units to clusters.

KEYWORDS: Copula, Fuzzy partitioning around medoids, Time series, Tourism destination.

1 Introduction

Clustering of time series is an important tool in several research and applied fields, especially in finance and economics where practitioners are often interested in identifying similarities in patterns across time. As such, several methods have been developed according to the different meanings that one can assign to the general statement that "two time series are close each other". These methods may group, for instance, time series that have similar values, functional shapes, autocorrelation structure, or good approximation by prototype objects (see Caiado *et al.* , 2015 and references therein).

Here we focus on a recent approach that consists of interpreting similarities among time series in terms of their co-movements (see, for instance, De Luca & Zuccolotto, 2011). Durante *et al.* , 2014; Durante *et al.* , 2015b; Durante *et al.* , 2015a suggest the use of copula to capture the co-movements of two or more time series. This is a rank–invariant approach whose main feature is that time series are supposed to group together when they are positively concordant, i.e. large (respectively small) values of one series at a given time tends to be associated with large (respectively small) values of the other series at the same time. In particular, in this work we will exploit the use of copulas for clustering time series by combining it with the advantages provided by a fuzzy clustering approach (D'Urso, 2015) and a Partitioning Around Medoids (PAM) procedure.

This method has been developed starting from a practical problem concerning tourism destinations, i.e. to identify agglomerations of cities/towns belonging to the same destination characterized by a common evolution of the tourist flows over time. Through this analysis, we have the opportunity to: 1) identify agglomerations of cities/towns that have experienced a common tourist evolution; 2) recognize the medoid of each agglomeration, i.e. the city/town that characterizes each agglomeration and that can be considered as the touristic attractor of a given sub-region. To demonstrate how this method works, the case study of South–Tyrol region (Northern Italy) is presented.

2 Methodology

The proposed clustering algorithm can be schematically presented as follows. Let us assume that $\mathbf{X}$ is a $(T \times n)$ data matrix, if necessary properly transformed in order to remove the seasonal parts, where $\mathbf{x}_i$ is a generic element that represents the time series of the i-th unit $(i = 1, \ldots, n)$. The Copula-based dissimilarity measure between $\mathbf{x}_i$ and $\mathbf{x}_j$ $(i, j = 1, \ldots, n$ and $i \neq j)$ can be formalized as a suitable distance between the copula C_{ij} (expressing the dependence between the units i and j) and the copula M_2 that expresses the co-monotone dependence between two time series, namely

$$D(\mathbf{x}_i, \mathbf{x}_j) = \left| C_{ij} - M_2 \right|. \tag{1}$$

In other words, if the time series are exactly co-monotone, then their dissimilarity is equal to 0, while the dissimilarity increases when we have a small deviation from such an extreme case. Usually, the distance is assumed to be the supremum norm or a Cramer-von Mises distance (L^2 norm). Here, the copula can be either estimated parametrically or replaced by its empirical version.

The Fuzzy k–Medoids (FkMdd) clustering model is hence combined with the Copula-based dissimilarity measure, leading to the novel Copula-Fuzzy clustering algorithm (C-FkMdd). The objective function of this algorithm can be formalized as follow:

$$\begin{cases} \min: & \sum\limits_{i=1}^{n} \sum\limits_{s=1}^{k} u_{is}^{m} D(\mathbf{x}_i, \widetilde{\mathbf{x}}_s) = \sum\limits_{i=1}^{n} \sum\limits_{s=1}^{k} u_{is}^{m} \left| C_{is} - M_2 \right| \\ s.t. & \sum_{s=1}^{k} u_{is} = 1, \qquad u_{is} \geq 0 \end{cases} \tag{2}$$

where $\mathbf{x}_i$ and $\widetilde{\mathbf{x}}_s$ are the time trajectories of the i-th unit and of the s-th medoid, respectively; u_{is} is the membership degree of the unit i to the cluster s; $D(\mathbf{x}_i, \widetilde{\mathbf{x}}_s)$ is the Copula-based dissimilarity measure for multivariate time series as described in (1) between the i-th unit and the s-th medoid; $m > 1$ is a parameter which controls for the fuzziness of the obtained partition.

The adoption of a fuzzy clustering approach allows to take into account the uncertainty that arise assigning the units to the clusters. Moreover, the membership degrees, obtained by the adoption of any fuzzy clustering algorithm, represent a measure of this kind of uncertainty. The main advantage of the FkMdd algorithm with respect to the Fuzzy k–Means (FkM) algorithm, the two most traditional fuzzy clustering algorithms, is that the prototypes obtained through the FkMdd algorithm are actually observed time series (medoids), instead of virtual time series (centroids). This allows to characterize the obtained clusters by detecting observed typical time trajectories (see, e.g., Coppi *et al.* , 2006) describing in a more realistic way the multivariate data analyzed.

3 Case study: tourism in South–Tyrol

South–Tyrol is an Italian tourist destination characterized by 116 towns and villages grouped in seven districts. The proposed clustering analysis is applied to these towns and is separately performed on the basis of two important measures that allow to study the tourist flows of a destination, i.e. the overnight stays and the arrivals. Data are monthly collected by ASTAT (the provincial institute for statistic) per each town and the period from 2008 to 2014 was taking into consideration. Moreover, the clustering analysis was conducted using the entire time series and separately distinguished between the summer and winter season.

This kind of analysis is of particular interest in the tourism economics in which the detection of the existence of agglomeration and the analysis of their trend over time-space is recognized as a key factor in promoting tourism development (Yang, 2012). In fact, a tourism agglomeration can be defined as a geographic concentrations of interconnected tourism business that cooperate but also compete creating a network of relationships that allows them to better perform certain tourism economic activities (Yang, 2012).

4 Conclusions

In this paper we suggest to combine the Copula-based dissimilarity measure with a Fuzzy Partitioning Around Medoids (FPAM) algorithm in order both to capture the rank–invariant dependence among time series and to take into account the uncertainty associated to the assignment of the elements to each cluster. This methodology is motivated by a practical application in tourism destination, assisting destination managers, planners, and politicians in the optimal planning of infrastructure development and marketing strategies as well as regional tourist policies.

References

CAIADO, J., MAHARAJ, E. A., & D'URSO, P. 2015. Time series clustering. *In:* HENNIG, C., MEILA, M., MURTAGH, F., & ROCCI, R. (eds), *Handbook of Cluster Analysis*. Chapman & Hall. in press.

COPPI, R., D'URSO, P., & GIORDANI, P. 2006. Fuzzy C-Medoids clustering models for time-varying data. *Pages 195–206 of:* BOUCHON-MEUNIER, B., COLETTI, G., & YAGER, R.R. (eds), *Modern Information Processing: From Theory to Applications*. Amsterdam: Elsevier Science.

DE LUCA, G., & ZUCCOLOTTO, P. 2011. A tail dependence-based dissimilarity measure for financial time series clustering. *Adv. Data Anal. Classif.*, **5**(4), 323–340.

DURANTE, F., & SEMPI, C. 2015. *Principles of copula theory*. London: CRC/Chapman & Hall.

DURANTE, F., PAPPADÀ, R., & TORELLI, N. 2014. Clustering of financial time series in risky scenarios. *Adv. Data Anal. Classif.*, **8**(4), 359–376.

DURANTE, F., PAPPADÀ, R., & TORELLI, N. 2015a. Clustering of time series via non–parametric tail dependence estimation. *Statist. Papers*. In press.

DURANTE, F., FERNÁNDEZ-SÁNCHEZ, J., & PAPPADÀ, R. 2015b. Copulas, diagonals and tail dependence. *Fuzzy Sets and Systems*, **264**, 22–41.

D'URSO, P. 2015. Fuzzy clustering. *In:* HENNIG, C., MEILA, M., MURTAGH, F., & ROCCI, R. (eds), *Handbook of Cluster Analysis*. Chapman & Hall. in press.

KAUFMAN, L., & ROUSSEEUW, P. J. 2005. *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: Wiley.

PORTER, M. 1998. *On competition*. Boston: Harvard Business Review Press.

YANG, Y. 2012. Agglomeration density and tourism development in China: An empirical research based on dynamic panel data model. *Tourism Management*, **33**, 1347–1359.