ABSTRACT: The aim of this paper is to provide a suitable instrument able to manage spatial time data array. In particular, we suggest to adopt the Fuzzy C–Medoids algorithm with Dynamic Time Warping distance and suitable spatial penalization terms, to take into account spatial autocorrelation between geographical units. The case study of the South–Tyrolean region is presented to illustrate how this model can investigate changes of the regional tourism development.

KEYWORDS: dynamic time warping, fuzzy clustering, spatial time series, tourism destination.

1 Introduction

Clustering of spatial time data finds application in several substantive fields, such as economics, marketing, tourism, finance, medicine. Data used in this kind of applications are in the form of spatial time (three-way) data array, i.e. geographical units × quantitative features × time occasions. A spatial time data array can be algebraically formalized as \( X \equiv \{ x_{ijt} : i = 1, \ldots, I; j = 1, \ldots, J; t = 1, \ldots, T \} \), where \( i \) indicates the spatial unit, \( j \) the variable, \( t \) the time, and \( x_{ijt} \) the generic element of \( X \).

Let us also define \( X_i \equiv \{ x_{ijt} : t = 1, \ldots, T \} \), i.e. the \( i \)-th multivariate time trajectory, where \( x_{ijt} = (x_{i1t}, \ldots, x_{ijt}, \ldots, x_{iJt}) (i = 1, \ldots, I; t = 1, \ldots, T) \).

Classification of such complex data requires to take into account both their spatial and time dimension. In particular, one should consider: 1) the spatial nature of the units to be clustered; 2) the characteristics of the feature space, namely the space of multivariate time trajectories; 3) the uncertainty associated to the assignment of a geographical unit to a given cluster on the basis of the above complex features (Coppi et al., 2010). In this study, we suggest the adoption of the Fuzzy C–Medoids algorithm (FCMdd) (Krishnapuram et al.,
2001) with Dynamic Time Warping (DTW) distance (Velichko & Zagoruyko, 1970; Berndt & Clifford, 1994) to compute pairwise distances between multivariate time trajectories. To cope with the different levels of spatial contiguity (i.e., spatial autocorrelation) that could characterize geographical units, suitable spatial penalization terms are added to the objective function of the fuzzy clustering method. The case study of South–Tyrol (Northern Italy) tourism development is considered to illustrate the main features of the clustering model suggested.

2 Methodology

2.1 Time and space dimensions

To compute pairwise distances between multivariate time series, we considered the DTW distance. The DTW distance stretches or compresses two time series locally, in order to make their shape as similar as possible. The greater the “effort” to equalize two time series, the greater the distance between them. Let us consider a “test” (or query) multivariate time series, represented by a matrix \(X\) of order \((T \times J)\), and “reference” multivariate time series, \(Y\), of order \((T' \times J)\), where \(T\) is not necessarily equal to \(T'\). The “local” distance between \(X\) and \(Y\) is \(d(x_t, y_{t'}\)) \(\forall t, t' \in \{1, \ldots, T, T'\}\). The “warping path” \(p = (p_1, \ldots, p_L)\), with \(p_l = (t_l, t'_l)\), \(t_l \in \{1, \ldots, T\}, t'_l \in \{1, \ldots, T'\}\) is a sequence which defines the alignment between \(X\) and \(Y\), by assigning the element \(x_{t_l}\) of \(X\) to the element \(y_{t'_l}\) of \(Y\). The total “cost” of the warping path \(p\) is \(d_p(X, Y) = \sum_{l=1}^L d(x_{t_l}, y_{t'_l})w_p\), where \(w_p\) is a step-specific weight. DTW distance for multivariate time series is defined as:

\[
D(X, Y) = \min_p d_p(X, Y). \quad (1)
\]

As regards the spatial dimension, the contiguity between areas has to be taken into account (Coppi et al., 2010). Contiguity between areas could be defined in different manners. For instance, one could regard two areas as contiguous if they belong to the same macro-area, even if they are not adjacent. To this purpose, a penalty term for each type of “contiguity” is introduced in the objective function of the clustering algorithm.

2.2 Fuzzy clustering model

In this study we propose a Fuzzy C–Medoids (FCMdd) clustering model with DTW distance and \(K\) spatial penalty terms. The clustering algorithm can be formalized as follows (Coppi et al., 2006; Coppi et al., 2010):
\[
\begin{align*}
\min : \quad & \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic}^m D(\mathbf{X}_i, \bar{\mathbf{X}}_c) + \sum_{k=1}^{K} \beta_k \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic}^m \sum_{i' = 1}^{I} \sum_{c' \in C_c} p_{kii'}u_{i'c'}^m \\
\text{s.t.} : \quad & \sum_{c=1}^{C} u_{ic} = 1, \quad u_{ic} \geq 0
\end{align*}
\]

where \(\mathbf{X}_i\) and \(\bar{\mathbf{X}}_c\) are the multivariate time trajectories of the \(i\)-th spatial unit and of the \(c\)-th medoid, respectively; \(D(\mathbf{X}_i, \bar{\mathbf{X}}_c)\) is the DTW distance for multivariate time series (1) between the \(i\)-th area and the \(c\)-th medoid; \(m > 1\) is a parameter which controls for the fuzziness of the obtained partition; \(u_{ic}\) is the membership degree of the area \(i\) to the cluster \(c\); \(\beta_k \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic}^m \sum_{i' = 1}^{I} \sum_{c' \in C_c} p_{kii'}u_{i'c'}^m\) is the \(k\)-th spatial penalty term \((k = 1, \ldots, K)\), where \(\beta_k \geq 0\) is the coefficient of the \(k\)-th spatial penalty term, which weighs its importance within the clustering process, and \(p_{kii'}\) is the generic element of the \((I \times I)\) “contiguity” matrix \(\mathbf{P}_k\).

FCMdd is suggested since the prototypes of each clusters are observed time series (medoids), instead of virtual time series (centroids) as is the case of Fuzzy C–Means (FCM) (Kaufman & Rousseeuw, 2005; Coppi et al., 2006). Furthermore, prototypes obtained with FCM are computed as weighted average of the observed time series, and this could lead to unsatisfactory results when DTW distance is adopted (Izakian et al., 2015).

3 The case study

South–Tyrol (Northern Italy) is characterized by 116 towns grouped in seven districts (valleys). The clustering analysis is based on a set of six tourist indicators annually collected by both the provincial and the national institute for statistic from 2000 to 2012. Two levels of contiguity are considered: two towns are contiguous if they are adjacent, and/or if they belong to the same district. Then, two penalty terms in the model (2) are introduced \((K = 2)\): the generic element of \(\mathbf{P}_1\), \(p_{1ii'}\), is equal to 1 if the towns \(i\) and \(i'\) are adjacent, 0 otherwise; the generic element of \(\mathbf{P}_2\), \(p_{2ii'}\), is equal to 1 if the towns \(i\) and \(i'\) are in the same districts, 0 otherwise. It is important to underline that the use of these spatial penalty terms does not constrain nearby towns to be clustered together (Coppi et al., 2010).

4 Conclusions

In this paper we suggested the use of the FCMdd algorithm with DTW distance and spatial penalty terms, which is able to manage spatial time data ar-
ray. DTW distance is adopted to cope with the time dimension, while spatial penalty terms take into account the spatial correlation between units. To illustrate the model, we analysed the tourism evolution of the South-Tyrol region during the period 2000-2012. Indeed, a deeper understanding of the tourism evolution over time and space of the towns belonging to the same tourist destination is fundamental to establish suitable regional tourism policies, planning the regional infrastructure development, and promoting tourism development (Yang, 2012).

References


