CLUSTERING TIME SERIES: AN APPLICATION TO COVID-19 DATA

Margherita Gerolimetto, Stefano Magrini

Introduction

Clustering time series has recently received a lot of attention from the literature. Similarly to cross-sectional clustering, several algorithms have been developed to carry out time series clustering and the choice of which one is more adapt depends on both the aim of the analysis itself and the typology of data at hand (for good reviews see Liao, 2005; Fu, 2011). Among all clustering algorithms, those that have been mostly used in the time series literature are: hierarchical, partitioning and model-based. Given so, it is possible to broadly identity two approaches for time series clustering: i) one that modifies cross-sectional data algorithms so that they can be employed also for time series data; ii) another that converts time series data into a cross-sectional object treatable with traditional clustering methods.

Within the first approach, a crucial issue is represented by the capability of identifying dissimilarities between time series. The usual Euclidean distance is a rather improper measure since it does not consider the correlation structure of the time series itself. Consistently with this, several proposals have been presented in literature to measure dissimilarity between pairs of time series, some of which are described in the second section of this paper.

As for the second approach, the fundamental element is the choice of the features to extract. In this vein, Wang et al. (2006) present a method for clustering time series that concentrates on their structural characteristics, whose pattern similarities are identified using Self Organizing Maps (Kohonen, 2001), an unsupervised neural network algorithm. The structural features are obtained from the time series by applying operations that best capture the underlying characteristics, for example, trend, seasonality, kurtosis, etc. In this work, we resort to the spline literature (De Boor, 1978) and consider, as a particular feature to extract, the coefficients of the p basis functions into which a series is decomposed when it is smoothed via a spline.

We apply some clustering time series methods, selected from both approaches, to analyze the daily time series of Covid-19 deaths for Italian regions between February 2020 and February 2022. Results show that there are patterns of regions that tend to

stick together across the various groupings obtained with the considered methods of clustering.

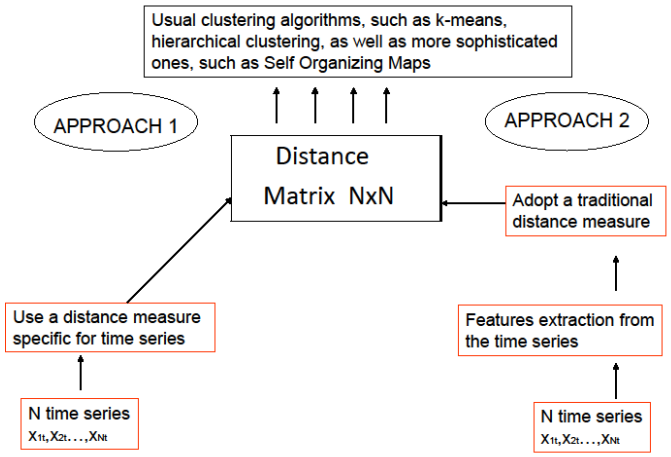
The structure of the paper is as follows. In the second section, we will present an overview on clustering time series. In the third section, we focus on clustering spline decomposition coefficients. In the fourth section, we will present our empirical analysis and some conclusions.

Time series clustering

The aim of clustering is to identify patterns by organizing data into homogenous groups where the within group object similarity is minimized and the between group objects dissimilarity is maximized. Clustering has been originally conceived for static data and, among the best-known algorithms, we mention k-means, where each cluster is represented by the mean value of the objects in the cluster, and hierarchical clustering, where data are grouped into a tree of clusters adopting agglomerative and divisive algorithms. Just like static data clustering, time series clustering requires the choice of the algorithm to form the clusters. In addition, time series data require a preliminary phase where the dynamic nature is accounted for. As anticipated in the Introduction, this can be done by choosing between two approaches.

Practically, the first approach, a.k.a *raw-data-based,* works with raw data and modifies the concept of distance measure so that it becomess compatible with the time series objects. The second approach, a.k.a *feature-based* or *model-based,* instead transforms the raw time series data into a features vector of lower dimension (or, alternatively, into a set of model parameters) and then carries out the grouping using conventional clustering methods. For a data set of N time series, the entire framework is displayed in Figure 1.

Figure 1 – *Conceptual map of time series clustering.*



Raw-data-based clustering

In the *raw-data-based* approach, the dynamic nature of the objects to cluster is handled by defining an appropriate measure of distance/similarity, bearing in mind that the Euclidean distance, typically adopted in static data clustering, is not adequate for time series data because it does not take into consideration the time dependence structure. Here, given two time series X_t and Y_t, where t =1,…,T, we will present an overview of some measures of distance proposed in the literature.

A very interesting measure of distance that overcomes the limits of the Euclidean distance is represented by the Dynamic Time Warping (DTW) distance (Keogh and Ratanamahatana, 2004). DTW algorithms come from the engineering literature and their aim is comparing discrete sequences with continuous sequences. In the present context, the logic of DTW is to firstly align the time series (intuitively, they must be stretched and compressed locally so they resemble each other as much as possible) and then some measure of distance is calculated between observations that match. The alignment of the time series is the core of the DTW and it is implemented through the so-called warping function, $\phi(k)$ that remaps the time index of X_t and Y_t

$$\phi(k) = \left(\phi_x(k), \phi_y(k)\right) \tag{1}$$

where $\phi_x(k) \in \{1, \dots, T\}$ and $\phi_y(k) \in \{1, \dots, T\}$. The average cumulated distortion between the warped time series is given by

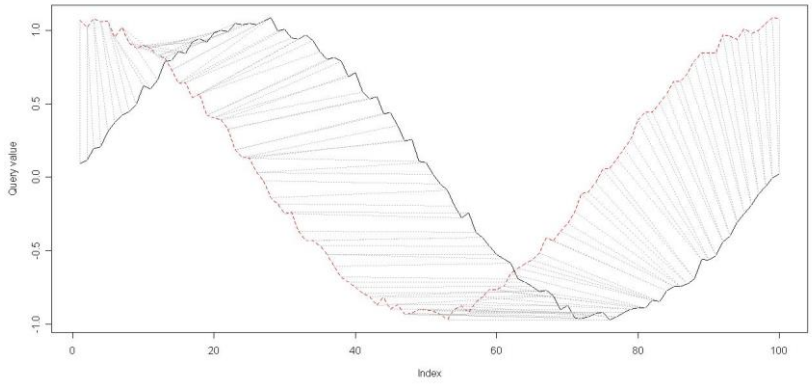
$$d_\phi(x, y) = \sum_{k=1}^{T} d\left(\phi_x(k), \phi_y(k)\right) \frac{m_\phi(k)}{M_\phi} \tag{2}$$

and the optimal alignment is the warping path $\phi(k)$ such that distortion is minimized. Moreover, on the warping function it is imposed monotonicity to ensure reasonable paths

$$\begin{aligned} \phi_x(k+1) &\geq \phi_x(k) \\ \phi_y(k+1) &\geq \phi_y(k) \end{aligned} \tag{3}$$

Eventually, the DTW distance between X_t and Y_t is the Euclidean distance between observations aligned via the warping function, as in the example in Figure 2.

Figure 2 – *Example of alignment of the time indexes of two time series (simulated data).*



Note: DTW time index matching of two time series (simulated from sinusoidal waves)

Another way of overcoming the limit of the Euclidean distance is the proposal by Galeano and Peña (2000) who present a method to assess similarity between time series focused on the comparison of their autocorrelation functions (ACF).[1] In particular, they define a metric based on the distance between the estimated

[1] Some developments of this method are in D'Urso and Maharaj (2009) and Alonso and Peña (2019).

autocorrelation coefficients of time series X_t and Y_t, denoted by, respectively $\hat{\rho}_x$ and $\hat{\rho}_y$

$$d_{ACF} = \sqrt{(\hat{\rho}_x - \hat{\rho}_y)\Omega(\hat{\rho}_x - \hat{\rho}_y)} \tag{4}$$

where Ω is some matrix of weights. In the same vein, but on the frequency domain side, Caiado et al. (2006) propose a metric built on the logarithm of the normalized periodogram of series X_t and Y_t, at frequencies $w_j=2\pi j/T$, $j=1,\ldots,[T/2]$, denoted by, respectively, $\log NP_x(w_j)$ and $NP_y(w_j)$

$$d_{LNP} = \sqrt{\sum_{j=1}^{[T/2]} \left(\log NP_x(w_j) - \log NP_y(w_j)\right)^2} \tag{5}$$

that is in fact the Euclidean distance between $\log NP_x(w_j)$ and $NP_y(w_j)$. Moreover, the authors propose a measure of distance based on the Kullback-Leibler information metric, still calculated in the frequency domain

$$d_{KLFD} = \sum_{j=1}^{[T/2]} \left[\frac{NP_x(w_j)}{NP_y(w_j)} - \log \frac{NP_x(w_j)}{NP_y(w_j)} - 1 \right] \tag{6}$$

Those reported in this section are just some examples of measures of distance that we considered of interest and hence adopted in the following empirical analysis. By no means, this must be intended as an exhaustive presentation. For example, given their different logic, we did not discuss the proposal by Piccolo (1990), who developed a measure of distance between ARIMA models based on their AR(∞) parametrization, nor the time series clustering method based on forecast densities by Alonso et al. (2006).

Feature-based clustering

Raw-data-based clustering implies working with high dimensional spaces and this can sometimes be a serious issue also because of the amount of noise typical of data collected at fast sampling rates. In such cases, feature-based clustering can address this concern.

The idea behind the feature-based approach is dimension reduction. This implies that distance/similarity is evaluated among features extracted from each time series instead of the original time series themselves. This allows the use of simpler measures of distance, such as Euclidean, because the extracted features resemble a static object as in traditional clustering. Once the distance between the features is

calculated, the usual clustering algorithm can be adopted. As we said before, these can be k-means or hierarchical clustering, but Wang et al. (2006) also proposed the use of more sophisticated methods, such as Kohonen Self Organizing Maps.

It is important to remark that while most feature extraction methods are generic in nature, the extracted features are instead application dependent. Put it differently, one set of features that work well in one application might be not relevant in another.

Wang et al. (2006) proposed to treat the time series data to obtain a set of global measures of, for example, trend, seasonality, serial correlation, non-linear autoregressive structure, skewness, kurtosis. These (and other more specialized time series features) concisely represent the relevant characteristics of each time series, thus providing a finite set of inputs to a clustering algorithm that can then assess similarity and differences between time series. In other words, every time series is seen as a single object and from the complexity of a matrix of *N* time series, each of extension *T* (the time series length) as in approach 1, here we have *N* objects whose extension is *p*, much smaller than *T*.

Feature-based clustering with splines

With the idea that the set of features to be extracted can always be updated and extended, here we resort to the spline literature and propose to extract the coefficients of the *p* basis functions into which a series is decomposed when it is smoothed via a spline. The objective is to concisely represent the time series, capturing not just the absolute value of the series but also its0 shape. This is done via B-Splines, which are a way to approximate non-linear functions by using a smooth piece-wise combination of polynomials (De Boor, 1978) and the positions where the pieces meet are known as knots. An example of B-spline with 4 knots and degree 3 is in Figure 3.

Specifically, B-Splines have two components, a basis function and the coefficients. The basis determines the hyperparameters, *i.e.* how many knots and what degree of polynomial to use in each model. The coefficients are then multiplied by this basis to approximate the original data. The idea is that by combining *p* polynomials using different weights, or coefficients, it is possible to obtain a non-linear estimate of the original time series. Least squares estimates can be used to get the best fitting coefficients.

Once the *p* coefficients of the basis functions are estimated for each time series of the data set, every time series is seen as a single object whose dimensionality is *p*. A Euclidean distance matrix can easily be calculated and any clustering algorithm can be adopted.

Figure 3 – *Example of cubic B-spline (simulated data).*

Note: Red, pink, black and green curves are the basis functions into which the simulated mexican hat-shaped function is decomposed via B-spline.

Empirical analysis

We now present our empirical analysis. We employ daily time series of Covid-19 deaths from 23/02/2020 to 29/03/2022 for the 19 Italian Regions and the 2 autonomous provinces of Trento and Bolzano. In particular, we consider two sets of data: i) deaths per 100,000 inhabitants and ii) deaths per 100,000 inhabitants normalized. The source of our data is "Istituto Superiore della Sanità" sticking on the official definition of Covid-19 deaths.

In our analysis with the R package Tsclust (Montero and Vilar, 2014), we adopt several clustering methods leading to as many outcomes. On the one hand, this is motivated as a form of robustness check, starting from what concerns the determination of the number of clusters. On the other hand, it is interesting to observe how results differ across clusterings. The methods we consider are selected from both the raw-data-based approach and the feature-based one. As for the first approach, we consider 4 distance matrices, calculated using the before presented distances (d_{ACF}, d_{LNP}, d_{KL}, d_{DTW}); for each of them, clustering is carried out using 5 algorithms: k-means and 4 hierarchical algorithms (Single linkage, Complete linkage, Average linkage, Ward). Eventually, the most interesting results are those based on d_{ACF} and d_{LNP}, using a k-means algorithm (only for deaths/hab) with 3 clusters.

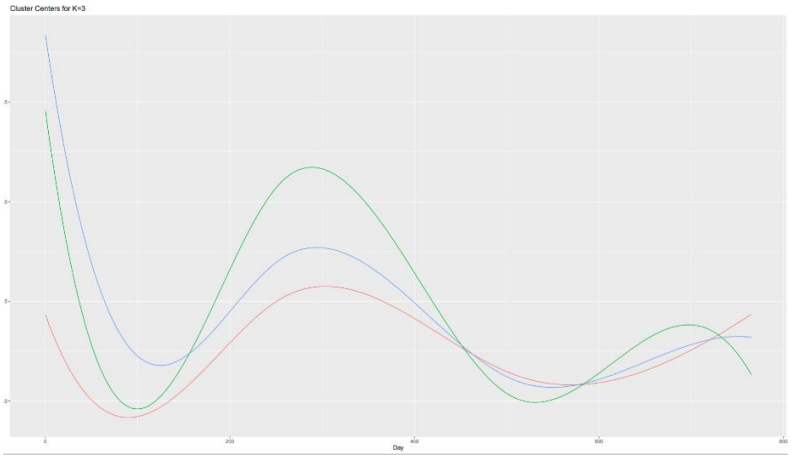
As for the second approach, the features we focus on are the coefficients of the basis functions in the spline decomposition of the time series. Specifically, we employ a cubic B-spline with 6 knots and the basis function coefficients are estimated via by least squares (for both deaths/hab and deaths/hab normalised). Given the results of the clustering analysis with the first approach, we consider here only the k-means algorithm for which the number of clusters provided as input is 3.

In the impossibility of showing the results of all clusterings (yet results for the other methods and the other data set are available upon request), we present only the outcome of the feature-based spline clustering applied to the data set on deaths per 100,000 inhabitants.

In particular, the following figures present the average spline of each of the three groups (Figure 4) in red, blue and green and the detailed composition of the groups (Figure 5). From Figure 4 it is possible to appreciate the different shape of the splines, and specifically the different slopes between peaks and troughs. From Figure 5 we observe that red group features Abruzzo, Basilicata, Calabria, Campania, Lazio, Molise, Puglia, Sardegna, Sicilia, Toscana, Umbria. The blue group contains Emilia Romagna, Liguria, Lombardia, Marche. Finally, the green group includes Bolzano, Friuli-Venezia-Giulia, Valle d'Aosta, Veneto.

Putting together the analyses conducted with the two approaches, we focus on 4 sets of results: *i)* raw-data clustering with d_{ACF} and d_{LNP}, using a k-means algorithm (only for deaths/hab) *ii*) feature-based clustering with spline for both deaths/hab and deaths/hab normalized. This leads to 4 clusterings; in each of them, the 21 time series have been divided into 3 groups.

Figure 4 – *Splines for cluster centers (data set: daily deaths per 100,000 inhabitants).*



Note: Average B-splines for each group (red, blue and green). B-spline with 6 knots and the basis function coefficients are estimated via east squares.

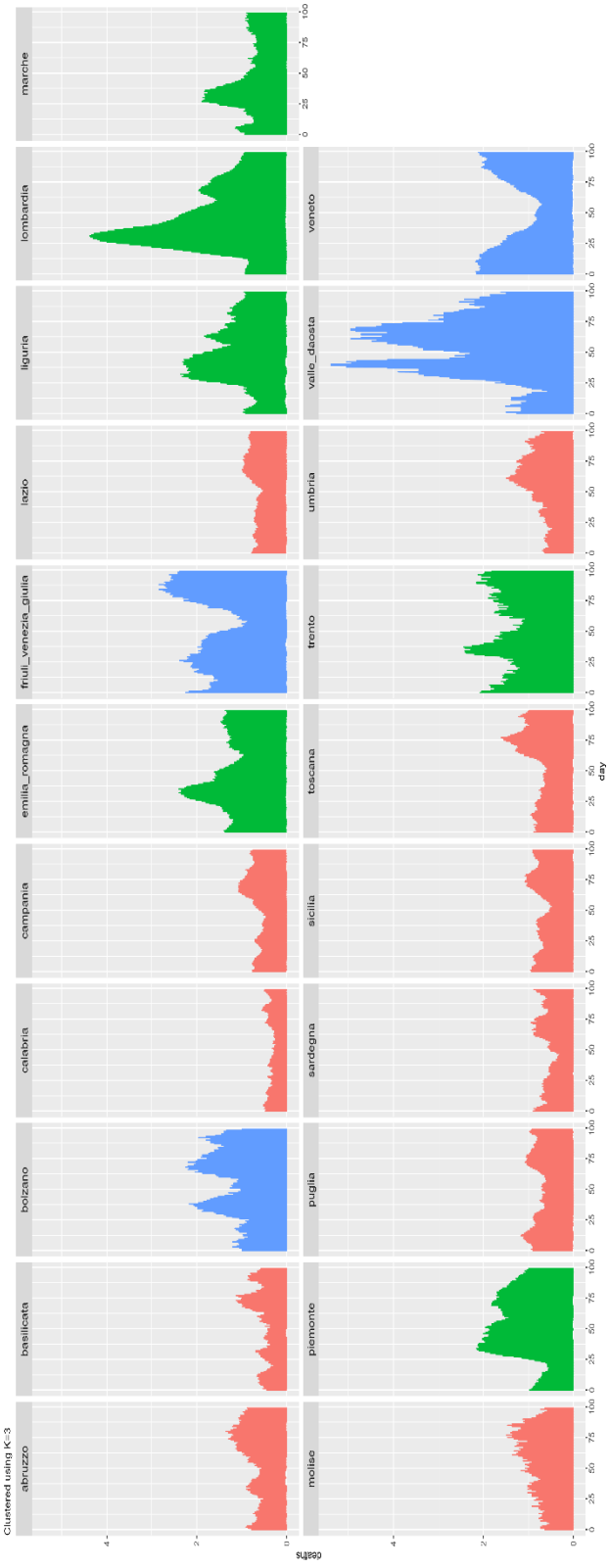


Figure 5 – *Composition of the 3 clusters- red, blue and green group.*

Note: clustering method: feature-based clustering via spline with k-means algorithm for k=3; data set: daily deaths per 100,000 inhabitants).

Figure 6 represents with different colors the patterns of associations among regions that emerge systematically across the 4 clusterings. These have been detected by building a matrix (bottom panel of Figure 6) that reports 1 every time a pair of regions sticks together over all clusterings. These pairs are then reported in the map (top panel of Figure 6) to make the geographical pattern more evident.

Figure 6 – *Patterns common to all clusterings.*

	ITC1	ITC2	ITC3	ITC4	ITH1	ITH2	ITH3	ITH4	ITH5	ITI1	ITI2	ITI3	ITI4	ITF1	ITF2	ITF3	ITF4	ITF5	ITF6	ITG1	ITG2
ITC1				1																	
ITC2																					
ITC3																					
ITC4	1																				
ITH1																					
ITH2																					
ITH3																					
ITH4																					
ITH5																					
ITI1																					
ITI2														1	1						
ITI3																					
ITI4																					
ITF1											1				1						
ITF2											1			1							
ITF3																				1	
ITF4																					
ITF5																			1		
ITF6																		1			
ITG1																1					
ITG2																					

Piemonte and Lombardia (pale orange in the map) stick together over all groupings. The same holds for Sicilia and Campania (yellow), for Basilicata and Calabria (light blue), for Umbria, Molise and Abruzzo (green). This means that the time series of these regions tend to share some features and this emerges in a rather robust way, given the variety of methods with which these clustering are carried out.

These results are interesting not only from the methodological perspective of comparing different methods of time series clustering, but also because it represents a preliminary step of a wider project, whose aim is to investigate possible determinants of differential Covid-deaths/hab across regions.

References

ALONSO A.M., PEÑA D. 2019. Clustering time series by dependency, *Statistics and Computing*, Vol. 29, pp. 655–676.

ALONSO A.M., BERRENDERO J.R., HERNÁNDEZ A., JUSTEL A. 2006. Time series clustering based on forecast densities, *Computational Statistics and Data Analysis*, Vol. 51, pp. 762–766.

CAIADO J., CRATO N., PEÑA D. 2006. A periodogram-based metric for time series classification, *Computational Statistics and Data Analysis*, Vol. 50, pp. 2668-2684.

DE BOOR C. 1978. *A Practical Guide to Splines*. Berlin: Springer-Verlag.

D'URSO P., MAHARAJ E.A. 2009. Autocorrelation-based fuzzy clustering of time series, *Fuzzy sets and Systems*, Vol. 160, pp. 3565-3589.

FU T.C. 2011. A review on time series data mining, *Engineering Applications of Artificial Intelligence*, Vol. 24, pp. 164-181.

GALEANO P., PEÑA D. 2000. Multivariate analysis in vector time series, *Resenhas*, Vol. 4, pp. 383-404.

KEOGH E., RATANAMAHATANA C.A. 2005. Exact indexing of dynamic time warping, *Knowledge and Information Systems*, Vol. 4, pp. 358-386.

KOHONEN T. 2001. *Self-Organizing Maps*. New York: Springer-Verlag.

LIAO T.W. 2005. Clustering of time series data: a survey, *Pattern Recognition*, Vol. 38, pp. 1857-74.

MONTERO P., VILAR J.A. 2014. TSclust: An R Package for Time Series clustering, *Journal of Statistical Software*, Vol. 62, pp. 1-43.

PICCOLO D. 1990. A distance measure for classifying ARMA models, *Journal of Time Series Analysis*, Vol. 11, pp. 153-163.

WANG X., SMITH K., HYNDMAN R. 2006. Characteristic-based clustering for time series data, *Data Mining and Knowldedge discovery*, Vol. 13, pp. 335-364.

SUMMARY

In this paper we present an attempt of clustering time series focusing on Italian data about COVID-19. From the methodological point of view, we first present a review of the most important methods existing in literature for time series clustering. Similarly to cross-sectional clustering, time series clustering moves from the choice of an opportune algorithm to produce clusters. Several algorithms have been developed to carry out time series clustering and the choice of which one is more adapt depends on both the aim of the analysis itself and the typology of data at hand. We apply some of these methods to the data set of daily time series on intensive care and deaths for COVID19 stretching from, respectively, 23/02/2020 to 15/02/2022 and from 23/02/2020 to 29/03/2022. These data refer to the 19 Italian regions and the two autonomous provinces of Trento and Bolzano.

MARGHERITA GEROLIMETTO, Università Ca' Foscari Venezia, Dipartimento di Economia, margherita.gerolimetto@unive.it
STEFANO MAGRINI, Università Ca' Foscari Venezia, Dipartimento di Economia, stefano.magrini@unive.it