

Detecting cycles in Complex Time Series Databases

Francesco Giordano, Maria Lucia Parrella and Marialuisa Restaino¹

Dept. of Economics and Statistics, University of Salerno

e-mail: giordano@unisa.it, mparrella@unisa.it, mlrestaino@unisa.it

Abstract: A large part of the data stored in financial, medical and scientific databases consists in time series realizations of unknown processes. The automatic statistical modelling of such data may be a very hard problem when the time series show "complex" features, such as nonlinearity, local nonstationarity, high frequency, long memory and periodic components. In such a context, the aim of this paper is to analyze the problem of detecting automatically the different periodic components in the data, with particular attention to the short term components (weakly, daily and intra-daily cycles). We focus on the analysis of real time series from a large database provided by an Italian electric company. This database may be considered complex in the previous sense.

Keywords: time series data mining, classification, cycles.

1. Introduction

Time series data mining has attracted great attention in the statistical community in recent years. The automatic statistical modelling of large databases of time series may be a very hard problem when the time series show complex features, such as nonlinearity, nonstationarity, high frequency, long memory and periodic components. Classification and clustering of such "complex objects" may be particularly beneficial for the areas of model selection, intrusion detection and pattern recognition. As a consequence, time series clustering represents a first mandatory step in temporal data mining research.

In this paper we consider the case when the time series show periodic components of different frequency. Most of the data generated in the financial, medical, biological and other fields present such features. A clustering procedure useful in such a context has been proposed in Giordano, La Rocca and Parrella (2008). It is based on the detection of the dominant frequencies explaining large portions of variation in the data through a spectral analysis of the time series. Like every clustering methods, also the procedure proposed in Giordano, La Rocca and Parrella (2008) depends on some sort of *tuning parameters*, which directly influence the results of the classification. The purpose of this paper is to perform a sensitivity analysis on such parameters, and to control if there is a particular behaviour in the results which can help in the determination of the optimal partition and in the dynamical detection of cycles in the database. In the third section we propose a method for the selection of the best partition for the clustering procedure.

¹Address of correspondence: Dept. of Economics and Statistics, University of Salerno, Via Ponte Don Melillo, 84084, Fisciano (SA)

2. The clustering algorithm and the tuning parameters

Consider a stationary process $\{X_t; t = 1, 2, \dots\}$. Basing on the Wold decomposition of the process, the spectral density function $f_X(\omega_j)$, $-\pi \leq \omega_j \leq \pi$, may be viewed as the sum of a discrete function $f_V(\omega_j)$ and a continuous function $f_Z(\omega_j)$. The first is associated with the harmonic component of the process and identifies the discrete periodic components (cycles). The second is associated with the linear process of the process. The proposed clustering method points at detecting and classifying the discrete periodic components of the process, so it focus on $f_V(\omega_j)$. The algorithm is based on the following main steps (for the details, see Giordano, La Rocca and Parrella, 2008):

1. For each time series (user u), estimate the spectral density function $f_X^u(\omega_j)$, $\forall j$.
2. Use the *Whittle test* with a Bartlett window to test the hypothesis $H_0: f_V^u(\omega_j) = 0$, $\forall j$. Derive the *relevant discrete frequencies* for the user u , as those frequencies ω_j for which H_0 is rejected. Denote these frequencies with $\hat{\omega}_j^u$, $j = 1, 2, \dots$
3. For each user u , extract the most important frequency $\hat{\omega}_{(1)}^u$, called the *dominant frequency*, which is converted into the correspondent period $\hat{P}$ (hours, days, etc...).
4. Derive the distribution of the *dominant periods* P_i . Denote with δ_i the percentage of users which have P_i as estimated dominant period. Let $\delta_1 \geq \delta_2 \geq \dots \geq \delta_r$.
5. For a fixed integer h_1 , consider the first *most observed* dominant periods, and denote with $\Delta_{h_1} = \sum_{i=1}^{h_1} \delta_i$ the percentage of dominant periods globally covered by the first h_1 positions. The parameter h_1 is the **first tuning parameter**.
6. Define the binary matrix $\mathbf{D} = (T \times h_1)$, whose generic element $d_{u,s}$ is equal to one if the period P_s appears in the first h_2 relevant discrete periods for the user u (derived in step 2); otherwise $d_{u,s}$ is equal to zero. The matrix $\mathbf{D}$ acts as a dissimilarity matrix, and h_2 is the **second tuning parameter**.
7. By considering the different combinations of the relevant periods P_i , $i = 1, \dots, h_1$, derive the $k = 2^{h_1}$ clusters of users by associating the rows (=users) of the matrix $\mathbf{D}$ with the same sequence of zeroes/ones. Note that some clusters could be empty, due to possible exclusive disjunction relationships between some of the periodic components. Denote with $k^* \leq k$ the number of *positive clusters*.
8. Derive the “best partition” of clusters by exploring different values for h_1 and h_2 .

3. A proposal for the selection of the optimal partition

In order to analyze the effects on the clusters of a change in the tuning parameters h_1 and h_2 , we applied the clustering procedure to a real database. The analyzed database was supplied from a large electric company. It collects the time series of energy consumption (*load curves*) observed hourly in the year 2006 for a particular category of customers. Since we look at the short term periodic components, we concentrate on the data from February to May, so the dimension of the database is 64522 users $\times$ 2880 observations. For each observed time series, we performed two different estimations of the spectral density function. The first was based on the hourly observations, and was aimed to capture the daily and intra-daily periodic components, whereas the second was based on the aggregated daily observations, and was aimed to better identify the weekly cycles. The selection of the “optimal” values for h_1 and h_2 will be made in two steps, by following different criteria: first of all we will select the optimal h_2 by considering mainly the

Daily data		h_2	1	2	3	4	5	6	7	8
$h_1 = 1$ $\Delta_1: \mathbf{0.869}$										
	Rand index (A)		—	0.87	0.94	1	1	1	1	1
	Shannon index (B)		0.85	0.93	0.96	0.96	0.96	0.96	0.96	0.96
	Number of clusters k^*	(max $k = 2$)	2	2	2	2	2	2	2	2
	Number of changes		—	4402	1806	38	0	0	0	0
	Mean(A,B)		—	0.9	0.95	0.98	0.98	0.98	0.98	0.98
$h_1 = 2$ $\Delta_2: \mathbf{0.952}$										
	Rand index (A)		—	0.87	0.94	1	1	1	1	1
	Shannon index (B)		0.51	0.56	0.57	0.58	0.58	0.58	0.58	0.58
	Number of clusters k^*	(max $k = 4$)	3	3	3	3	3	3	3	3
	Number of changes		—	4778	1942	41	0	0	0	0
	Mean(A,B)		—	0.71	0.76	0.79	0.79	0.79	0.79	0.79
$h_1 = 3$ $\Delta_3: \mathbf{0.976}$										
	Rand index (A)		—	0.86	0.94	1	1	1	1	1
	Shannon index (B)		0.36	0.4	0.41	0.41	0.41	0.41	0.41	0.41
	Number of clusters k^*	(max $k = 8$)	4	4	4	4	4	4	4	4
	Number of changes		—	4847	1951	41	0	0	0	0
	Mean(A,B)		—	0.63	0.67	0.70	0.70	0.70	0.70	0.70
Hourly data		h_2	1	2	3	4	5	6	7	8
$h_1 = 1$ $\Delta_1: \mathbf{0.780}$										
	Rand index (A)		—	0.75	0.89	0.94	0.97	0.98	0.99	0.99
	Shannon index (B)		0.94	0.75	0.62	0.55	0.51	0.49	0.47	0.46
	Number of clusters k^*	(max $k = 2$)	2	2	2	2	2	2	2	2
	Number of changes		—	8955	3702	1789	818	498	329	160
	Mean(A,B)		—	0.75	0.76	0.75	0.74	0.73	0.73	0.73
$h_1 = 2$ $\Delta_2: \mathbf{0.950}$										
	Rand index (A)		—	0.66	0.77	0.86	0.93	0.96	0.97	0.98
	Shannon index (B)		0.64	0.87	0.79	0.72	0.68	0.65	0.63	0.62
	Number of clusters k^*	(max $k = 4$)	3	4	4	4	4	4	4	4
	Number of changes		—	29896	11905	6400	2792	1779	1155	682
	Mean(A,B)		—	0.77	0.78	0.79	0.81	0.80	0.80	0.80
$h_1 = 3$ $\Delta_3: \mathbf{0.969}$										
	Rand index (A)		—	0.64	0.78	0.86	0.90	0.92	0.93	0.94
	Shannon index (B)		0.46	0.72	0.76	0.75	0.75	0.74	0.73	0.73
	Number of clusters k^*	(max $k = 8$)	4	7	8	8	8	8	8	8
	Number of changes		—	35792	18704	10998	6755	5003	3915	3252
	Mean(A,B)		—	0.68	0.77	0.80	0.83	0.83	0.83	0.83

Table 1: Performance indicators for different values of the parameters h_1 and h_2 .

movements among the clusters for different values of h_2 ; than we will select the optimal h_1 by considering also the time series aggregated in the clusters. The rationale behind this is that the parameter h_1 directly influence the number of potential clusters $k = 2^{h_1}$, and so the optimal partition must be selected by taking into account the efficiency in classifying the time series among the k clusters. On the other hand, the parameter h_2 only reflects on the number of positive clusters k^* and on the distribution of the n users among such clusters, and so for the selection of h_2 we consider mainly the stability of the clusters.

As concerning the first step, table 1 reports some results useful for evaluating different values of h_2 . In particular, some values (the Shannon index and the number of clusters) refer to the set of clusters identified by a given pair (h_1, h_2) , while the other values refer to the comparison between such set of clusters and the “adjacent one”, *i.e.* the situation identified by the pair $(h_1, h_2 - 1)$ (this explain why these values are not available for $h_2 = 1$). The Rand index is a normalized index which assumes the maximum value (one) when the two situations are equal, whereas it assumes the minimum value (zero) when they are completely different (*i.e.* every two users that are classified in the same cluster in the first case are separated in the second case). It is evident that the Rand index is directly connected with the number of movements among the clusters. Another index useful for

Cycles	D_1	D_2	D_3	D_4	D_1 D_2	D_1 D_3	D_1 D_4	D_2 D_3	D_2 D_4	D_3 D_4	D_2 D_3 D_4	D_1 D_3 D_4	D_1 D_2 D_4	D_1 D_2 D_3	D_1 D_2 D_3 D_4
H_1	0.5	0.75	1	0.75	0.57	0.71	0.57	0.86	0.71	0.86	0.8	0.7	0.6	0.7	0.69
H_2	0.75	1	1	0.75	0.86	0.86	0.71	1	0.86	0.86	0.9	0.8	0.8	0.9	0.85
H_1, H_2	0.83	1	1	0.83	0.9	0.9	0.8	1	0.9	0.9	0.93	0.86	0.86	0.93	0.89

Table 2: “Goodness” index for the 45 compared partitions (normalized).

our purpose is given by the Shannon entropy, based on the relative distribution of the users among the $k = 2^{h_1}$ clusters of the partition. It assumes max value (one) when the users are equally distributed among the k clusters and minimum value (zero) when all the users are concentrated into one cluster. So this index is directly connected with the percentage of positive clusters. The results in the table show that the Rand index have a monotone behaviour while the Shannon index do not (note that the comparisons can be made only by row, relative to the same value of h_1 , because different values of h_1 determine different partitions which may not be compared, although the indexes are normalized). The mean value between the Rand index and the Shannon index will be a normalized index which takes into account both the criteria “movements among the clusters” and “percentage of positive clusters”. The max value (in bold) will identify the optimal value for h_2 .

For the selection of the parameter h_1 , a first evaluation may be made by considering the value of Δ_{h_1} . For example, we could fix an absolute threshold or we could consider a minimum relative increment for increasing values of h_1 . In our analysis, we decided to consider the first four daily cycles, denoted with D_1, D_2, D_3, D_4 and the first two hourly cycles, denoted with H_1 and H_2 . In order to explore the entire space of potential partitions, we derived all the possible combinations of one or more daily cycles with one or more hourly cycles. These amounts to 45 different partitions, and they are summarized in table 2, by crossing the rows and the columns. We introduce a normalized index (a variant of the Jaccard index) which measure the “goodness” of the classification by evaluating the agreement between the characteristics of the cluster and the structure of the aggregated time series. More precisely, for each partition we derive the aggregated time series of the clusters by summing the data observed for all the users classified in the same cluster. We expect that the aggregated time series will be characterized by the presence of the same periodic components as those which characterize the cluster. If not, it would mean that there are masking effects in the aggregation which denote that the partition is not efficient. In the table we report the results for such goodness index for the 45 partitions. Again we head for the maximum value, which identifies the “best partition”. It is interesting to note that there are 7 candidates, all of which involves the same periodic components (D_2, D_3, H_1, H_2). In order to choose one among these candidates we use again the Shannon index, to find the optimal partition with maximum heterogeneity.

References

- [1] Giordano, F., La Rocca, M., Parrella, M.L.: Clustering Complex Time Series Databases. Submitted to *SFC-CLADAG Proceedings* (2008).