Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 19 (2006) 877–888
Neural Networks
www.elsevier.com/locate/neunet

2006 Special Issue

Unfolding preprocessing for meaningful time series clustering

Geoffroy Simon[a,*,1], John A. Lee[a,2], Michel Verleysen[a,b,3]

[a] *Machine Learning Group - DICE, Université Catholique de Louvain, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium*
[b] *SAMOS-MATISSE, UMR CNRS 8595, Université Paris I - Panthéon Sorbonne, Rue de Tolbiac 90, F-75634 Paris Cedex 13, France*

Abstract

Clustering methods are commonly applied to time series, either as a preprocessing stage for other methods or in their own right. In this paper it is explained why time series clustering may sometimes be considered as meaningless. This problematic situation is illustrated for various raw time series. The *unfolding preprocessing* methodology is then introduced. The usefulness of unfolding preprocessing is illustrated for various time series. The experimental results show the meaningfulness of the clustering when applied on adequately unfolded time series.

Keywords: Clustering; Time series; Self-organizing maps

1. Introduction

Time series analysis is a task encountered in many fields, for example mathematics, statistics, econometrics, system identification, physics and machine learning to cite only a few. In all these fields there exist various methods of analyzing, describing, extracting features or even forecasting time series. Among all these data analysis tools the partitioning, vector quantization and similarity search methods can be grouped under the general heading of clustering methods.

Clustering is usually applied to time series in two different ways. In the context of time series prediction, for example, clustering can be applied to time series either as preprocessing for another time series prediction model (Cottrell, Girard, & Rousset, 1997; Vesanto, 1997) or (Dablemont et al., 2003) for example) or as a prediction model in its own right ((Martinetz, Berkovich, & Schulten, 1993; Simon, Lendasse, Cottrell, Fort, & Verleysen, 2004; Walter, Ritter, & Schulten, 1990) to cite only a few references). In practice most of the clustering methods work with subsequences of the initial time series. Such subsequences are vectors of successive past values that can be extracted from the series using a sliding window. In the remainder of the text vectors of successive past values are called regressors.

Despite the frequent use of these clustering methods on time series very few works try to answer fundamental questions such as the meaningfulness of time series clustering. Recently it has been proved that in some cases clustering on time series is meaningless (Keogh, Lin, & Truppel, 2003). Some authors used this proof to conclude that subsequence or regressor clustering is *always* meaningless (Bagnall, Janakec, & Zhang, 2003; Hetland, 2003; Mahoney & Chan, 2005). It is shown in Struzik (2003) that time series clustering is not meaningless but rather difficult due to the auto-similarity of the series. It is also proved in Denton (2004) that kernel-based clustering methods are able to provide meaningful clustering for time series. In this paper it will be shown that clustering is indeed meaningful once the regressors are *unfolded*, i.e. when some adequate preprocessing is performed.

In order to show the meaningfulness of clustering the unfolding preprocessing is introduced. This methodology is based on embedding (Sauer, Yorke, & Casdagli, 1991). The main idea of embedding is to use delayed coordinates in the regressors. Selection of the optimal delay can be done using either the autocorrelation function (Abarbanel, 1997; Kantz & Schreiber, 1997; Kaplan & Glass, 1995) or mutual information (Abarbanel, 1997; Fraser & Swinney, 1986; Kantz & Schreiber, 1997). The limit of these two common approaches is that the first lag is usually selected with great care but the influence

* Corresponding author.
E-mail address: simon@dice.ucl.ac.be (G. Simon).
URL: http://www.dice.ucl.ac.be/~gsimon/ (G. Simon).
[1] G. Simon is funded by the Belgian F.R.I.A.
[2] J.A. Lee is a Scientific Research Worker of the Belgian F.N.R.S.
[3] M. Verleysen is a Research Director of the Belgian F.N.R.S.

0893-6080/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2006.05.020

of the other delayed components (selected using multiples of this lag) is not taken into account. In this paper a measure is proposed to determine whether the selected delay is sufficient to unfold efficiently the series in a p-dimensional space of regressors. The idea of the proposed criterion is to observe the mean distance of a data set to the principal diagonal of its space. The whole methodology allows creating unfolded regressors with much more informative content than those obtained using the simple sliding window technique. The meaningfulness of time series clustering on these unfolded time series will then be shown on artificial and real time series using a Self-Organizing Map (Kohonen, 1995). Self-Organizing Maps are used here as a clustering tool thanks to their vector quantization property. The meaningfulness of time series clustering will be observed using criteria that measure the relative differences between regressor distributions.

The paper is organized as follows. Notations and definitions are given in Section 2, which ends with a brief description of Self-Organizing Maps. The criterion to measure the meaningfulness of clustering is detailed in Section 3. The so-called 'meaninglessness' of time series clustering is shown in the same section using graphical representations in 2- and 3-dimensional spaces. The *unfolding preprocessing* methodology is presented and illustrated in Section 4. Experimental results are provided in Section 5. Section 6 concludes the paper.

2. Regressor clustering using Self-Organizing Maps

2.1. Time series and regressors: Definitions and notations

A time series S is a series of values x_t, with $1 \leq t \leq n$, measured from a time varying process. The x_t values are ordered according to the time index and sampled with constant frequency. Note that, for reasons of simplicity, x_t is considered as scalar here though it could be a multidimensional vector.

In the context of time series prediction one has to build a model of the series. For example regression models have at least one output and some inputs corresponding either to past values x_t of the time series itself or to past values u_t of other related processes. These inputs are grouped in a vector describing the state of the process at a given time t and defined as:

$$x_{t-p+1}^{t} = \{x_t, x_{t-1}, \ldots, x_{t-p+1}, u_t, u_{t-1}, \ldots, u_{t-q+1}\}. \quad (1)$$

The $p+q$-dimensional vector x_{t-p+1}^{t} is called the *regressor*. For reasons of simplicity the regressors will be limited in this paper to p-dimensional vectors of past values:

$$x_{t-p+1}^{t} = \{x_t, x_{t-1}, \ldots, x_{t-p+1}\}. \quad (2)$$

Intuitively, the regressors can be seen as subsequences obtained from the initial time series using a sliding window.

The question regarding how to choose the length p of the regressor is decisive for the model. This question can be answered elegantly for linear models using various criteria (AIC, BIC, MDL, etc. (Ljung, 1999)) which may be extended in a nonlinear context (Kantz & Schreiber, 1997; Ljung, 1999). Model structure selection strategies using statistical resampling methods (like cross-validation, k-fold cross-validation, bootstrap (Efron & Tibshirani, 1993)) are also often used in order to find a correct value for the model complexity p. Another approach to choose p is to estimate the dimension of the time series using the correlation dimension, as proposed in Ding, Grebogi, Ott, Sauer, and Yorke (1993), Grassberger and Procaccia (1983) and Sauer et al. (1991) and illustrated in Babloyantz, Nicolis, and Salazar (1985) and Camastra and Colla (1999) to cite only a few. However, this question of choosing an adequate value of p is quite independent from the unfolding goal of this work; p will therefore be deemed to be fixed a priori throughout the rest of this paper.

2.2. Self-Organizing Maps

The Self-Organizing Map (SOM) algorithm is an unsupervised classification algorithm proposed in the 1980s (Kohonen, 1995). Since its introduction it has been applied in many areas, from classification to robotics to cite only two (Kohonen, 1995). The intuitive graphical representations that can be obtained from the SOM have made this tool a very popular nonlinear clustering method. Theoretical aspects of the SOM have also been studied in detail (Cottrell, Fort, & Pagès, 1998).

The main concepts of the SOM algorithm can be summarized as follows. During the learning stage, the SOM units, or *prototypes*, are moved within the data distribution. Each prototype is linked to neighbouring ones according to a 1- or 2-dimensional grid whose shape and size are chosen a priori. The learning consists in presenting each datum and selecting the closest prototype with respect to some distance measure. This winning prototype is moved towards the data while its neighbours are also moved according to constraints given by the grid. At the end of the learning stage, the final prototype positions represent a discrete and rough approximation of the data density that is therefore partitioned into *clusters*. Each cluster is associated with one of the prototypes.

The two main properties of the SOM are vector quantization and topology preservation. In this paper, the SOM is used as a clustering tool. Data to be clustered are the p-dimensional regressors obtained from the time series.

3. Why time series clustering may be considered as meaningless

How can the meaningfulness of time series clustering be observed? Consider two different time series. Once the regressors are formed for each series one can reasonably expect that the two sets of regressors are different as they describe two distinct processes. Applying a clustering method like the SOM algorithm on both sets of regressors reduces the information encoded in the regressors such that it is now contained in a limited number of prototypes. The meaningfulness of the clustering will then be measured from the two sets of prototypes. If the respective prototype distributions obtained by the SOM clustering tool on two distinct time series regressor sets are significantly different, it will lead to the conclusion

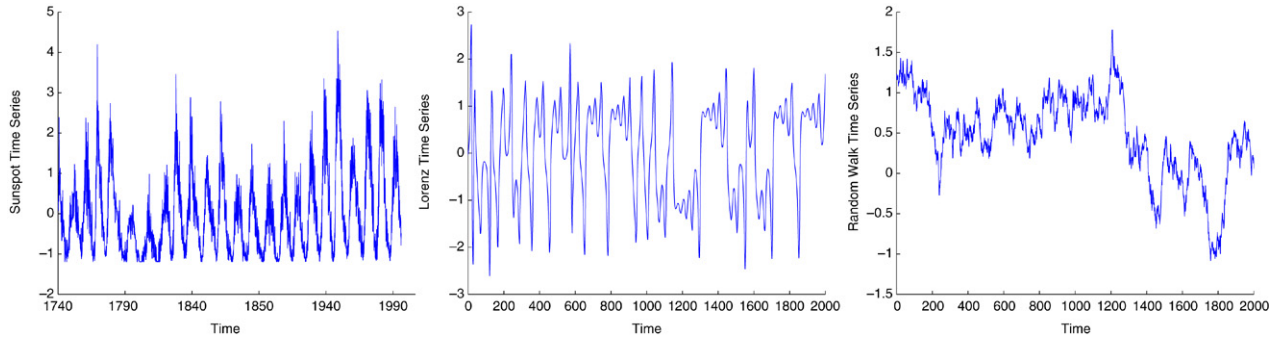

Fig. 1. Plot of the time series. From left to right: Sunspot, Lorenz and Random Walk.

that the clustering is meaningful. On the other hand, if the prototypes cannot be distinguished any more because the information contained in the regressor sets is lost by the clustering procedure, then the clustering will be considered meaningless (Simon, Lee, & Verleysen, 2005).

3.1. Criterion for meaningfulness of regressor clustering

How can significant differences be detected between two sets of prototypes? Intuitively, the prototype distributions obtained on two distinct time series should be much different from the prototype distributions obtained on a single times series in two different settings (two random initializations for example). The criterion for meaningfulness of clustering has thus to compare the prototype distributions or, more precisely, the prototype positions within the p-dimensional space of the regressors.

As time series may have different value ranges it is mandatory to scale them before computing the regressors, otherwise the prototype comparisons will obviously not be fair. In this paper, all time series are normalized according to $x_t = (x'_t - \mu_S)/\sigma_S$, where μ_S and σ_S are the time series mean and standard deviation, respectively, and x'_t are the values of the original scalar time series.

Let us consider two (normalized) time series $S_1 = \{x_t \mid 1 \leq t \leq n\}$ and $S_2 = \{y_t \mid 1 \leq t \leq m\}$, where m is not necessarily equal to n. Applying the SOM algorithm on the respective sets of regressors, both in a p-dimensional space with fixed p, leads to two prototype sets denoted $P_1 = \{\bar{x}_i \mid 1 \leq i \leq I\}$ and $P_2 = \{\bar{y}_j \mid 1 \leq j \leq I\}$. P_1 and P_2 must obviously have the same number of prototypes otherwise any comparison would not be fair.

To compare these sets of prototypes, the following measure can be used:

$$position\_difference(P_1, P_2) = \sum_{i=1}^{I} \min_j(\mathrm{dist}(\bar{x}_i, \bar{y}_j)) \quad \text{with } 1 \leq j \leq I. \tag{3}$$

This measure sums, for each prototype in P_1, the distance to its closest corresponding prototype in P_2. As defined here, the measure is not a one-to-one relation between prototypes in P_1 and P_2: some prototypes in P_2 may be selected more than once and some may not be selected. Other measures respecting a one-to-one relation could also be used. However experiments made with variants of the measure (3) lead to similar conclusions at the price of an increased computational burden. They will thus not be considered in this paper.

As already mentioned above, prototype sets obtained on the regressors from the same time series should be more or less identical. To avoid a possible influence of the initialization, the clustering is repeated K times leading to sets of prototypes sets:

$$\mathcal{P}_1 = \{P_1^k \mid 1 \leq k \leq K, \text{ with } P_1^k = \{\bar{x}_i^k \mid 1 \leq i \leq I\}\}, \text{ and} \tag{4}$$

$$\mathcal{P}_2 = \{P_2^l \mid 1 \leq l \leq K, \text{ with } P_2^l = \{\bar{y}_j^l \mid 1 \leq j \leq I\}\}. \tag{5}$$

The difference between prototype sets obtained through many runs of the SOM clustering algorithm, on a single time series on one side and on two distinct ones on the other side, can now be evaluated through the two following criteria, respectively:

$$within(\mathcal{P}_1) = \sum_{k=1}^{K}\sum_{l=1}^{K} position\_difference(P_1^k, P_1^l), \quad \text{and} \tag{6}$$

$$between(\mathcal{P}_1, \mathcal{P}_2) = \sum_{k=1}^{K}\sum_{l=1}^{K} position\_difference(P_1^k, P_2^l). \tag{7}$$

These criteria have been measured on three time series and the results are given in Fig. 2. The three times series are a real one (the sunspot series from January 1749 to February 2005 (SIDC, 2005)), an artificial one (generated using the Lorenz differential equations (Alligood, Sauer, & Yorke, 1996; Kantz & Schreiber, 1997; Kaplan & Glass, 1995)) and a random walk (generated from Gaussian white noise). Fig. 1 shows a plot of these three series. Increasing numbers of prototypes have been used in the SOM in order to observe the impact of the cluster sizes on the meaningfulness of clustering. The number of prototypes in the SOM varies between 5 and 100 by increments of 5. The prototypes and the regressors are 3-dimensional vectors; the dimension has not been optimized for prediction purposes (it is not the goal of the study) and is common to all series for comparison purposes.

The values have been computed for both criteria using successively the three series as reference. Consider for example that Sunspot is the reference time series. In that case the values *within*(Sunspot), *between*(Sunspot, Lorenz) and *between*(Sunspot, Random Walk) are plotted. Then the Lorenz and Random Walk time series are considered as references; the results are plotted in Fig. 2 from left to right.

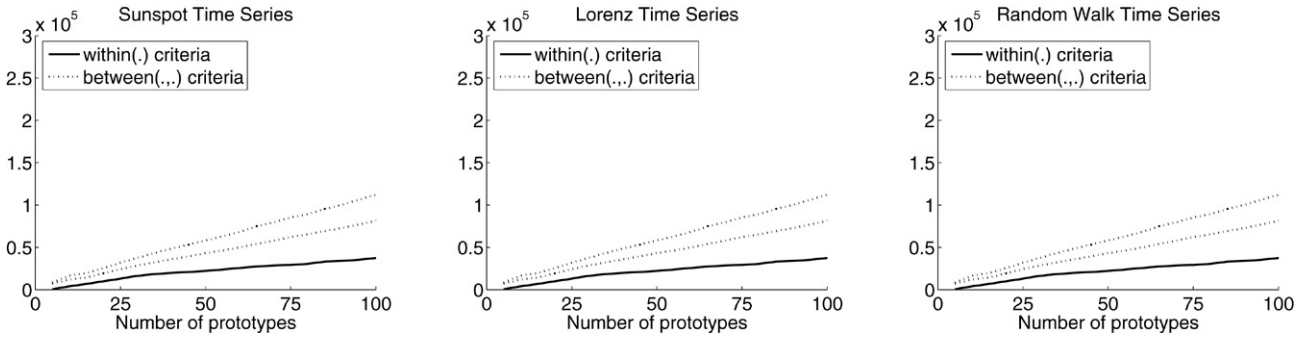

Fig. 2. Comparisons of the *within*(.) and *between*(., .) criteria on the three time series. From left to right, the Sunspot, Lorenz and Random Walk time series are successively used as reference. See text for details.

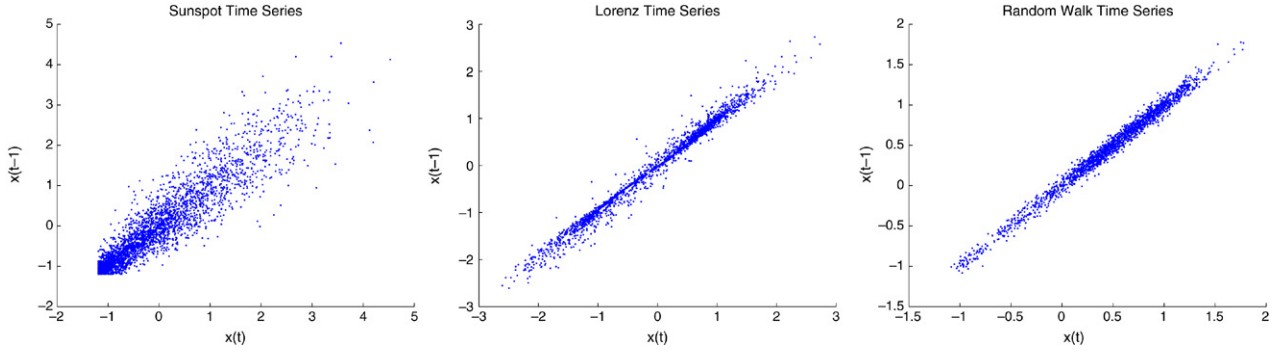

Fig. 3. Two-dimensional regressor distributions. From left to right: Sunspot, Lorenz and Random Walk.

Fig. 2 supports the meaningless character of regressor clustering. First, whatever the time series used as reference, the increase of the *within*(.) criterion with the number of prototypes remains limited. Second, the same phenomenon is noticeable for the *between*(., .) criteria. Finally, considering each plot separately, the number of prototypes does not really help in distinguishing the series. Indeed when the number of prototypes increases it is expected that the clustering should be more accurate as smaller clusters help to detect local differences. On the contrary Fig. 2 shows that this property is not verified: the *within*(.) and *between*(., .) criteria drive apart at a rate simply proportional to the number of prototypes in the SOM instead of to the (much larger) increase of respective distances between prototypes. The conclusion of meaningless clustering confirms the results in Bagnall et al. (2003), Hetland (2003), Keogh et al. (2003) and Mahoney and Chan (2005).

3.2. Illustration of the regressor clustering limitations

To understand how such a counterintuitive conclusion has to be stated from the results illustrated in Fig. 2, it is instructive to observe the regressor distributions of the three time series. Fig. 3 shows the corresponding 2-dimensional regressors.

The regressor building according to Eq. (2) is indeed problematic. If the time series is smooth the variations between successive values are small. The components of the regressors are thus highly correlated, so that the latter will concentrate around the diagonal, as shown in Fig. 3. This is clearly the case for the Sunspot and the Lorenz time series. For the Random Walk series two consecutive values are also highly correlated, by construction. It is thus not really surprising that the SOM provides a meaningless clustering as noticed from Fig. 2. In fact any method will fail to differentiate the series from such regressors because the portion of the space filled by the respective regressors is more or less the same, and very limited with respect to the available space. This is even worse in higher dimensions, as illustrated in Fig. 4 for the 3-dimensional regressors. These regressors still concentrate around a line that is, more generally, a 1-dimensional object in the p-dimensional space. The relative part of space filled by the regressor distribution will thus decrease dramatically as the regressor space dimension increases. Comparisons of regressor distributions in higher dimensions will be even more difficult than in dimension two.

4. The unfolding preprocessing

From the illustrations in Figs. 3 and 4 the goal of the unfolding preprocessing methodology is now clear: to make the regressor distribution fill a much larger part of p-dimensional space.

Intuitively the high correlation between the successive regressor values is likely to stem from too high a sampling frequency. It is therefore suggested to use a subsampling preprocessing that will make the various regressor components as independent as possible. In other words Eq. (2) should now be replaced by:

$$x^t_{t-(p-1)\tau} = \{x_t, x_{t-\tau}, x_{t-2\tau}, \ldots, x_{t-(p-1)\tau}\}, \tag{8}$$

where τ is a fixed *lag*.

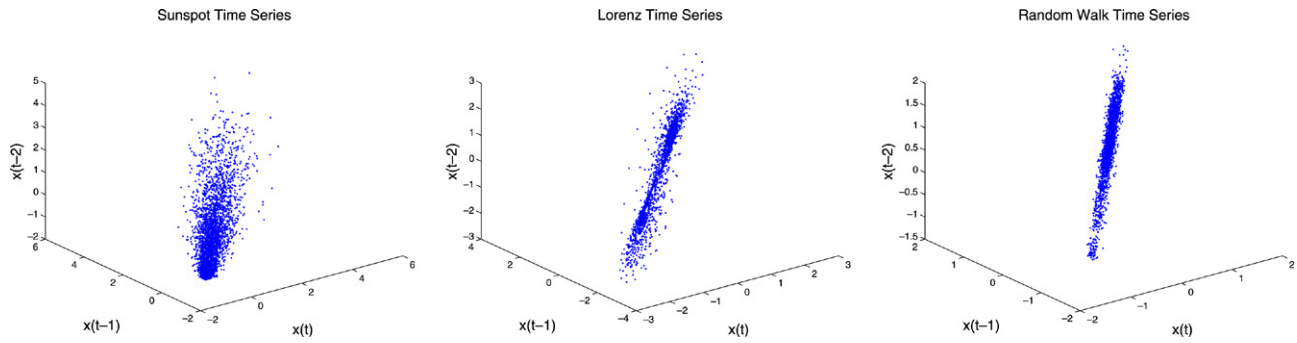

Fig. 4. Three-dimensional regressor distributions. From left to right: Sunspot, Lorenz and Random Walk.

The most difficult part of this preprocessing is the selection of an adequate value for the lag τ. In practice lag τ is selected as the lag closest to the first zero of the autocorrelation function. The autocorrelation function $C(\tau)$ is defined for real-valued time series as

$$C(\tau) = \sum_{t=1}^{n} x_t x_{t+\tau} \quad \text{for } 0 \leq \tau \leq M, \tag{9}$$

where n is the number of data in the time series and $M \leq n-1$ is the maximum time lag. Obviously, as the autocorrelation function differs from one time series to another, the selected value of τ^* will differ too.

The main limitation of the autocorrelation function is that it only computes the correlation between two variables. Though extensions to three variables exist (Gautama, Mandic, & Van Hulle, 2004), they are still limited as the goal is to use p-dimensional regressors with p possibly larger than 3. Another criterion is therefore proposed. This criterion measures the sum of the distances of the p-dimensional regressors to the diagonal in the p-dimensional space. This geometrical criterion does not measure correlation directly. Instead, it measures how much a regressor distribution differs from a straight line that represents a maximum of correlation. This sum of distances for a given time series S, denoted $\mathcal{D}(S)$, is computed as:

$$\mathcal{D}(S) = \sum_{t=1}^{n} \frac{\|v_1 - x_t\|^2 \|v_2 - v_1\|^2 - ((v_1 - x_t)^T (v_2 - v_1))^2}{\|v_2 - v_1\|^2}, \tag{10}$$

where n is the total number of p-dimensional regressors x_t that can be obtained from the original time series S; and v_1 and v_2 are two p-dimensional vectors defining the line. This formula usually presented in 2- or 3-dimensions can be generalized easily to any dimension as it is defined on vectors. In our case, as the line of concern is the diagonal of the space, v_1 and v_2 are the p-dimensional origin $(0, 0, \ldots, 0)$ and unit vector $(1, 1, \ldots, 1)$, respectively.

The complete methodology for unfolding regressors is thus the following. For the selected dimension regressors are calculated using Eq. (8). Then the graph of the distance to the diagonal criterion with respect to the lag τ can be plotted. An adequate lag corresponds to a maximum in this graph.

However, reaching the global maximum of the graph should not be considered as the unique objective. In practice, any lag that corresponds to a sufficiently high value in the graph could be considered. Local maxima or smaller lags in a plateau may be of interest too, as selecting lags that are too large may lead to difficulty of interpretation. A good rule of thumb is to limit the range of possible lags to a few times the one that would be selected traditionally, i.e. as the first zero of the autocorrelation function. Moreover, as the distance to diagonal criterion is a global criterion taking the whole regressor into account, it may happen that taking into consideration only possible maxima on this graph would lead to lags for which the autocorrelation function is not small. Therefore, in practice, the lag to be selected should correspond to a sufficiently high value of the distance to diagonal criterion; when this criterion gives several possible values, the two additional criteria should be taken into account: (1) the value of the lag should be chosen as small as possible, and (2) it should correspond to a low value of the autocorrelation function.

The proposed methodology is applied to the three time series used in the previous section. First consider the Sunspot time series in 2-dimensional space. Fig. 5 presents from left to right the plots of the corresponding distance to diagonal, autocorrelation function and a zoom of the autocorrelation function for the first 400 lags. From this plot a lag of $\tau = 55$ has been selected. This lag corresponds to a large value of the distance to diagonal criterion, and is smaller than twice the lag value the autocorrelation function would give (35). Choosing the latter value would, however, lead to a poor unfolding of the series; a larger value such as $\tau = 55$ is thus preferred, despite the (small) increase of the autocorrelation (in absolute value).

The same methodology is applied to the Lorenz time series. The results are presented in Fig. 6. The distance to diagonal criterion suggests that a lag between 20 and 25 is adequate. The lag that would be selected according to the autocorrelation only would be larger (around 35), but it can be seen that the autocorrelation is not much different between 25 and 35. The lag selected in this case is the one satisfying both criterions, i.e. 25 in this case.

The methodology is applied again to the 2-dimensional regressors of the Random Walk time series. Note that by definition of a random walk there is no reason to select a lag larger than 1 for prediction purposes. However, we are not interested here in prediction, but in unfolding in the regressor space; obviously, a lag leading to a good unfolding will be much

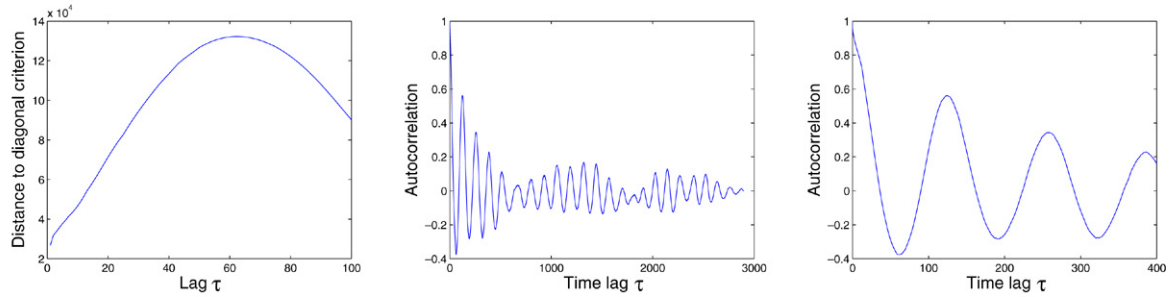

Fig. 5. From left to right: distance to diagonal, autocorrelation function and zoom of the autocorrelation function for the first 400 lags of the Sunspot time series.

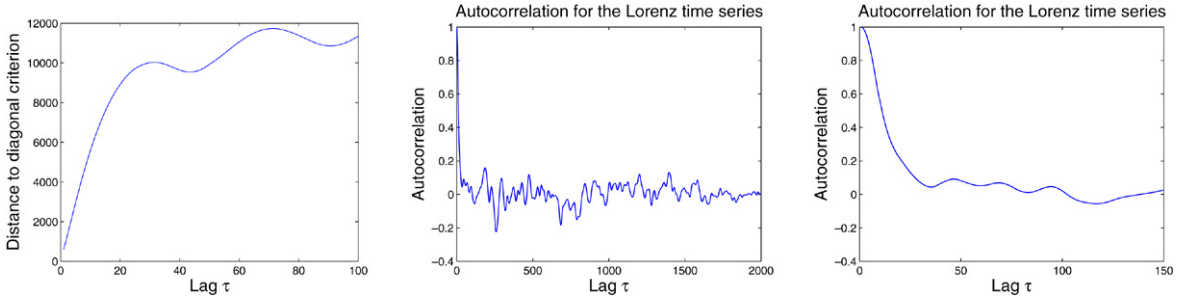

Fig. 6. From left to right: distance to diagonal, autocorrelation function and zoom of the autocorrelation function for the first 400 lags of the Lorenz time series.

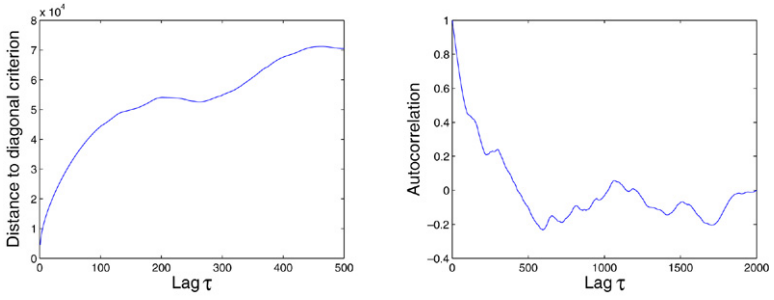

Fig. 7. Left: distance to diagonal; right: autocorrelation function of the Random Walk time series.

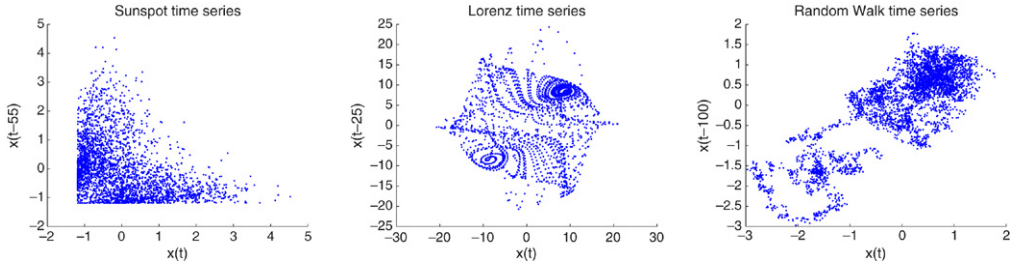

Fig. 8. Regressor distribution in a 2-dimensional space after unfolding. From left to right: Sunspot, Lorenz and random walk.

larger than 1, in order to have regressor components that are as independent as possible.

The values obtained for the two criteria are reported in Fig. 7. As a trade-off between the distance to diagonal criterion, the autocorrelation criterion and the objective to limit the value of the lag, $\tau = 100$ is selected.

Having selected these lags it is possible to observe how the regressor distributions in a 2-dimensional space are affected by the use of the unfolding preprocessing. This can be observed from Fig. 8 where the unfolded regressor distributions are plotted. The difference between Figs. 3 and 8 is clearly visible. Note that the same methodology has been applied to 3-dimensional regressors. The lags found for the Sunspot, Lorenz and Random Walk time series are respectively $\tau = 40$, $\tau = 20$ and $\tau = 100$. The corresponding unfolded regressor distributions are presented in Fig. 9 and may be compared to Fig. 4. Obviously, regressor comparisons based on SOM clustering using the *within*(.) and *between*(., .) criteria will be much more conclusive with the regressor distributions from Figs. 8 and 9 than those from Figs. 3 and 4.

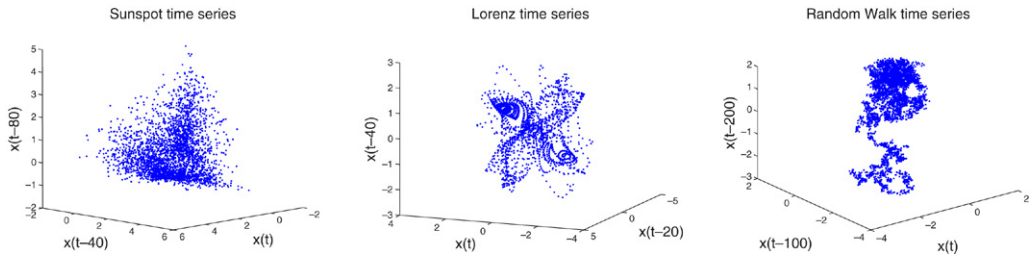

Fig. 9. Regressor distribution in a 3-dimensional space after unfolding. From left to right: Sunspot, Lorenz and random walk.

The three series used in this section were aimed to illustrate why selecting a larger lag may lead to better unfolding of the regressor distribution. Such unfolding is necessary to allow comparisons of sets of regressor prototypes, and therefore to assess the meaningful character of time series clustering. However, it should be noted that although unfolding is a necessity for using regressors in a prediction context (it is a way to make the regressors as informative as possible), unfolding is not a sufficient condition for prediction; the last example (Random Walk) clearly illustrates this fact.

Note also that in some cases, the distance to diagonal criterion gives similar lags to the autocorrelation criterion when 2-dimensional regressors are used. This is for example the case when simple AR/ARMA series are considered. However, the supplementary advantage of the distance to diagonal criterion, i.e. the possibility of using it on higher-dimensional regressors, remains.

5. Experimental results

In this section, we apply the unfolding methodology to time series before evaluating the *within*(.) and *between*(., .) criteria, in order to assess the meaningful character of time series clustering once the preprocessing is applied. Both real and artificial time series are used.

5.1. The time series

The real time series are the following.

- The Sunspot time series (SIDC, 2005) represents the monthly number of sunspots, from January 1749 to February 2005.
- The Santa Fe A time series (Weigend & Gershenfeld, 1994) was proposed in the Santa Fe Time Series Prediction and Analysis Competition in 1991. The data were collected from a Far-Infrared-Laser in a chaotic state.
- the Polish electrical consumption time series (Osowski, Siwek, & Tran Haoi, 2001) is composed of daily measures ranging from the end of 1990 to mid 1997.

In addition to these real-world series, artificial ones have also been generated using the following linear and nonlinear equations (the three first and the four last models respectively):

- Autoregressive model, AR(3):

$$x_t = a_1 * x_{t-1} + a_2 * x_{t-2} + a_3 * x_{t-3} + \varepsilon_t. \tag{11}$$

2500 data have been generated; the first 500 data have been discarded to remove any bias due to initial conditions. The a_i coefficients, with $i = \{1, 2, 3\}$, have been chosen randomly using a normal distribution with zero mean and unit variance. ε_t has been generated using a normal distribution with zero mean and 0.1 variance.

- Autoregressive with moving average model, ARMA(4, 4):

$$\begin{aligned} x_t = {} & a_1 * x_{t-1} + a_2 * x_{t-2} + a_3 * x_{t-3} + a_4 * x_{t-4} \\ & + b_1 * \varepsilon_{t-1} + b_2 * \varepsilon_{t-2} + b_3 * \varepsilon_{t-3} + b_4 * \varepsilon_{t-4} + \varepsilon_t. \end{aligned} \tag{12}$$

Here again, 2500 data have been generated and the first 500 data have been discarded. Coefficients a_i and b_i, with $i = \{1, 2, 3, 4\}$, have been chosen randomly using a normal distribution with zero mean and unit variance. ε_t has been generated using a normal distribution with zero mean and 0.1 variance, as above.

- Autoregressive with moving average model with seasonality of 12, SARMA(3, 3):

$$\begin{aligned} x_t = {} & a_1 * x_{t-1} + a_2 * x_{t-2} + a_3 * x_{t-12} \\ & + b_1 * \varepsilon_{t-1} + b_2 * \varepsilon_{t-2} + b_3 * \varepsilon_{t-12} + \varepsilon_t. \end{aligned} \tag{13}$$

Once again, 2500 data have been generated and the first 500 data have been removed. To force the seasonality, coefficients a_i and b_i, with $i = \{1, 2\}$, have been chosen randomly using a normal distribution with zero mean and 0.02 variance while coefficients a_3 and b_3 were chosen using a normal distribution with zero mean and unit variance. ε_t has been obtained as in the previous models using a normal distribution with zero mean and 0.1 variance.

- Lorenz system (Abarbanel, 1997; Kantz & Schreiber, 1997; Kaplan & Glass, 1995):

$$\begin{aligned} \dot{x}_t &= 10(y_t - x_t), \\ \dot{y}_t &= 28x_t - y_t - x_t z_t, \\ \dot{z}_t &= x_t y_t - \frac{8}{3} z_t. \end{aligned} \tag{14}$$

4000 data have been generated using an integration step of 0.015. The first 2000 data have been discarded to remove any transient state and let the trajectory fall to the attractor.

- Rossler system (Abarbanel, 1997; Kaplan & Glass, 1995):

$$\begin{aligned} \dot{x}_t &= -(y_t + z_t), \\ \dot{y}_t &= x_t + 0.2 y_t, \\ \dot{z}_t &= 0.2 + z_t(x_t - 5.7). \end{aligned} \tag{15}$$

In this case 4000 data have been generated using an integration step of 0.075. Here again the first 2000 data have been discarded.

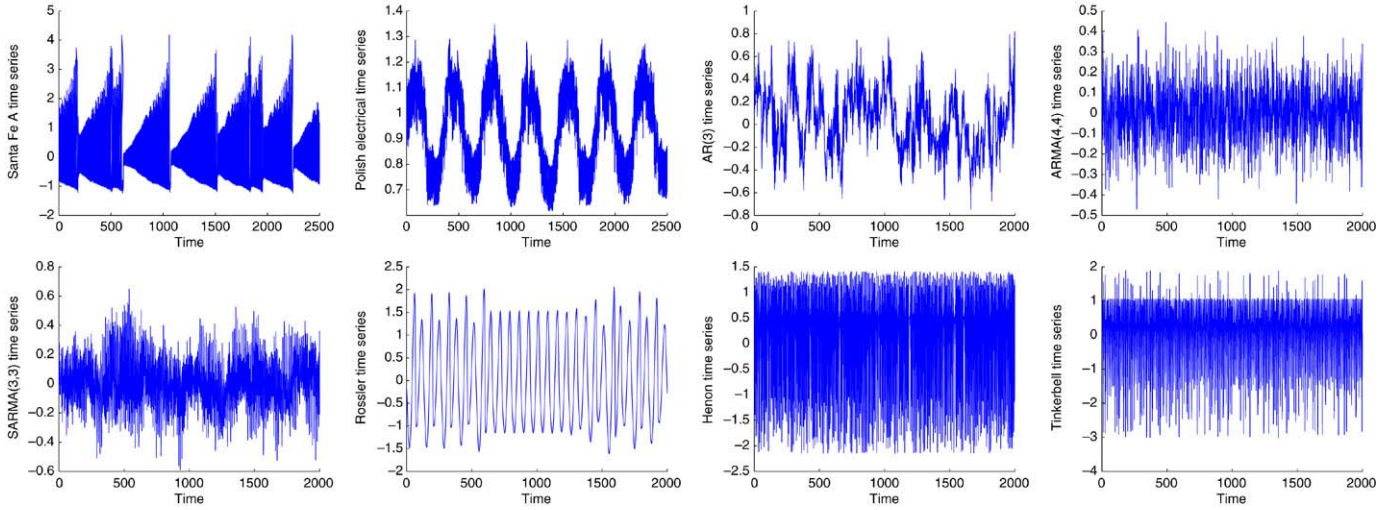

Fig. 10. From left to right, top to down: Santa Fe A, Polish electrical, AR(3), ARMA(4, 4), SARMA(3, 3) with seasonality of 12, Rossler, Henon and Tinkerbell time series.

- Henon map (Alligood et al., 1996; Kantz & Schreiber, 1997; Kaplan & Glass, 1995):

$$x_{t+1} = 1 - 1.4x_t^2 + y_t,$$
$$y_{t+1} = 0.3x_t. \tag{16}$$

  2000 data have been generated using these finite difference equations.
- Tinkerbell map (Alligood et al., 1996; Nusse & Yorke, 1998):

$$x_{t+1} = x_t^2 - y_t^2 + 0.9x_t - 0.6013y_t,$$
$$y_{t+1} = 2x_t y_t + 2x_t + 0.5y_t. \tag{17}$$

  For this system 2000 data have been generated too.
- Random walk:

$$x_{t+1} = x_t + \varepsilon_t, \tag{18}$$

  where ε_t are normally distributed random values. Here again, a series of 2000 data has been generated.

Fig. 1 shows plots of the Sunspot, the Lorenz, and the Random Walk time series, from left to right. Fig. 10 shows graphical representations of the eight other datasets, namely the Santa Fe A, the Polish electrical, the AR(3), the ARMA(4, 4), the SARMA(3, 3) with seasonality of 12, the Rossler, the Henon and the Tinkerbell time series.

5.2. Results

In this section, results obtained using time series generated using a single model are first presented. The aim is to prove that, in such a case, the *within*(.) and *between*(., .) criteria give similar values. Then, comparisons of the *within*(.) and *between*(., .) criteria values obtained from different artificial or real-world time series are presented in order to prove the meaningfulness of time series clustering.

A first experiment is performed to observe how the methodology behaves on time series generated artificially using the same model. As the series are obtained on the basis of the same equation, they are expected to be more or less indistinguishable even when unfolded: the regressor distributions should be more or less identical for each series. The example used here consists of ten Random Walk series obtained from Eq. (18). For each of these ten series, 2000 data have been generated. The lag has been chosen equal to 100 as discussed earlier. The unfolding preprocessing has been applied to all series. Fig. 12 shows the results obtained on two time series chosen from the ten generated ones; the results for the other series are similar.

From Fig. 12 two comments can be made. First, it can be observed that the *within*(.) and *between*(., .) criteria give distinct values. This is the result of ε_t in (18), which makes all series different, even though they were generated by the same model. Second, Fig. 12 shows that the differences between the *within*(.) and *between*(., .) criteria decrease after unfolding. This means that, though the initial series differ (and so their sets of regressors) because of their stochastic character and the ε_t term, the sets of regressors after unfolding are much closer to each other. In other words, the unfolding preprocessing reduces the differences between series generated from the same model; it acts as a smoother, and produces sets of regressors that better represent the series model.

Note that similar experiments performed on the Rossler and ARMA(3, 3) models, not detailed here for lack of space, lead to the same conclusions: the unfolded regressor sets are less distinguishable than the original ones.

The remainder of this section is devoted to comparison between different time series.

For 3-dimensional regressors, the lags selected using the distance to diagonal criterion are summarized in Table 1. Fig. 9 shows the regressor distributions obtained using the unfolding methodology for the Sunspot, Lorenz and the Random Walk time series, respectively, from left to right. Fig. 11 shows the corresponding unfolded regressor distributions for the eight other time series. A comparison of the values for the *within*(.) and *between*(., .) criteria is provided in Figs. 13 and 14. Note that the results for four time series only are provided due to

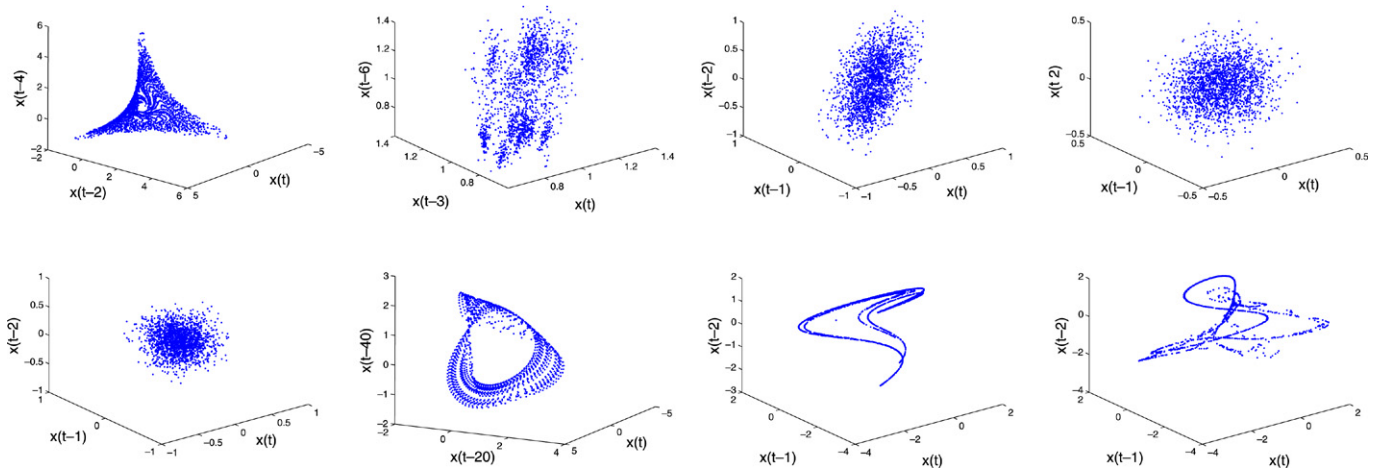

Fig. 11. 3-dimensional regressor distributions obtained using the unfolding preprocessing. From left to right, top to bottom: the Santa Fe A, Polish electrical, AR(3), ARMA(4, 4), SARMA(3, 3) with seasonality of 12, Rossler, Henon and Tinkerbell time series.

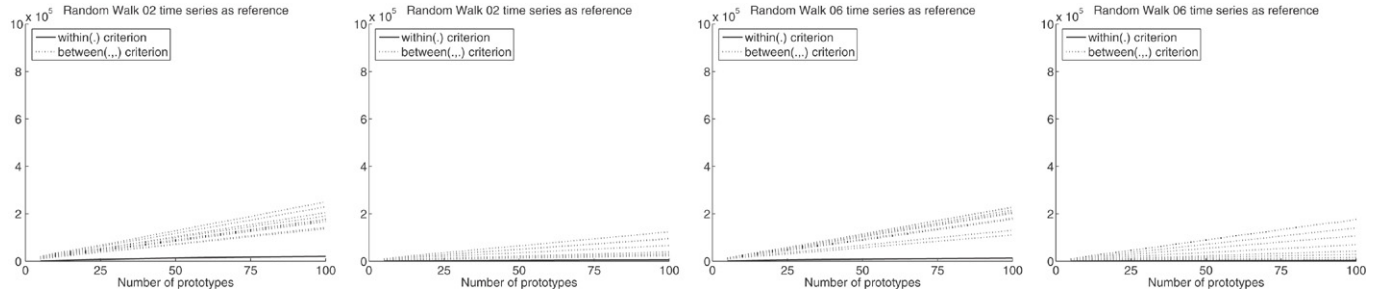

Fig. 12. *within*(.) and *between*(., .) criteria for two out of ten time series. Leftmost figures: initial regressors; rightmost figures: unfolded regressors.

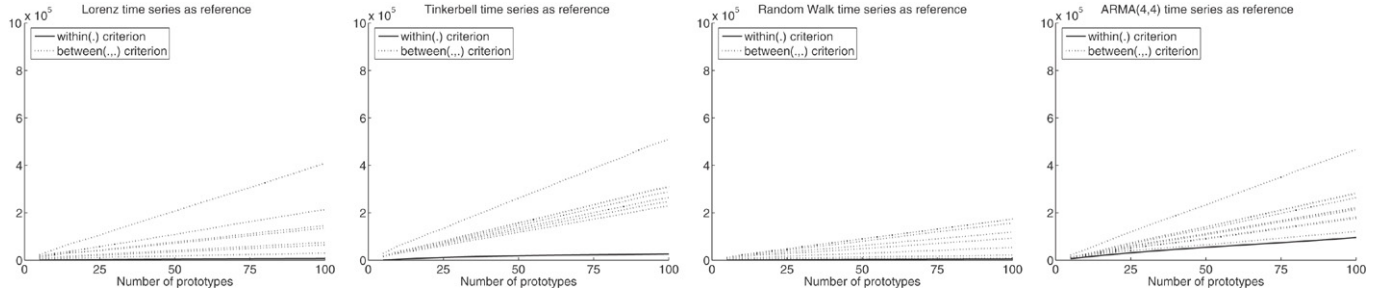

Fig. 13. *within*(.) and *between*(., .) criteria for non-preprocessed 3-dimensional regressors. From left to right: Lorenz, Tinkerbell, Random Walk and ARMA(4, 4) time series. Note that the scale is different from Fig. 2.

Table 1
Lags selected using the distance to diagonal criterion using 3-dimensional regressors

Time series	Lag	Time series	Lag
Sunspot	40	Rossler	20
Santa Fe A	2	Lorenz	20
Polish	3	Henon	1
AR(3)	1	Tinkerbell	1
ARMA(4, 4)	1	Random Walk	100
SARMA(3, 3)	1		

space limitations; the conclusions are similar for all other series. The values of the criteria in Figs. 13 and 14 were obtained using regressors from Eq. (2) and unfolded regressors from Eq. (8) respectively.

From Figs. 13 and 14, the plots of the Lorenz and Random Walk cases show a clear increase in the distance between the *within*(.) and *between*(., .) criteria when the regressors are unfolded. Concerning the Tinkerbell and the ARMA(4, 4) time series, Table 1 shows that the selected lag is 1; it is thus not surprising that no significant difference is obtained between Figs. 13 and 14 as the regressors are adequately unfolded in both cases.

Similar results can be obtained using 5-dimensional regressors. The selected lags are presented in Table 2. A comparison of the values for the *within*(.) and *between*(., .) criteria is provided in Figs. 15 and 16 for four other time series. Here again, the same conclusions may be drawn for all the other cases that are not illustrated.

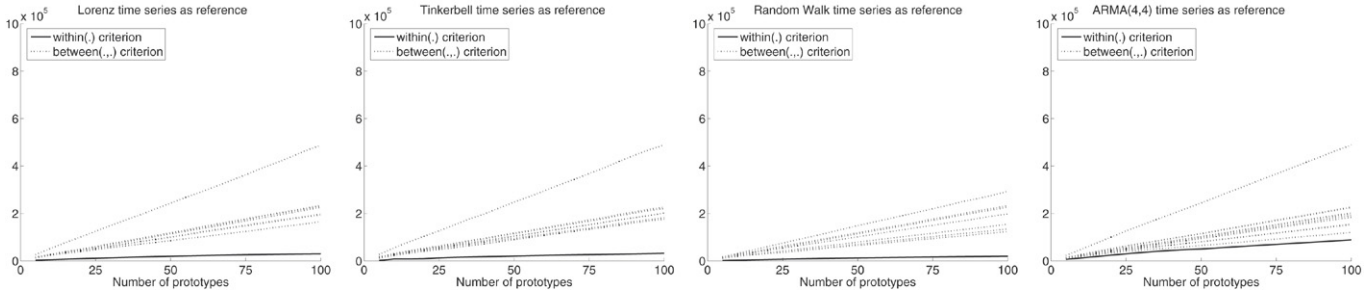

Fig. 14. *within*(.) and *between*(., .) criteria for 3-dimensional regressors preprocessed using the unfolding methodology. From left to right: Lorenz, Tinkerbell, Random Walk and ARMA(4, 4) time series.

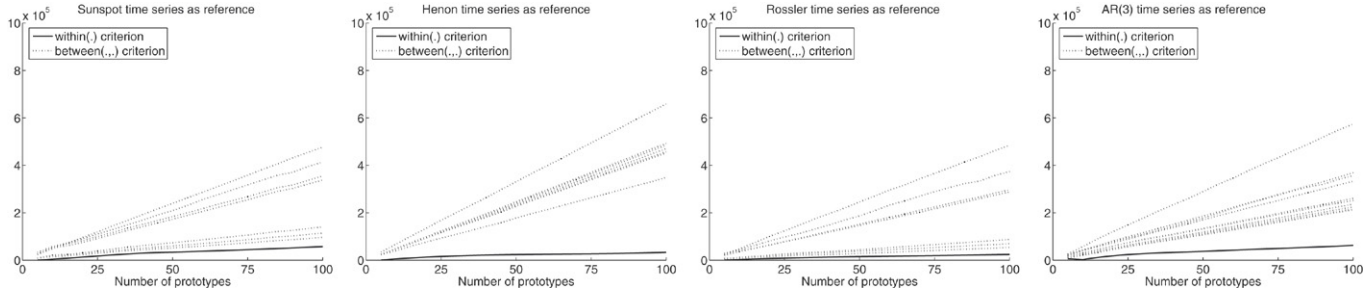

Fig. 15. Values for the *within*(.) and *between*(., .) criteria for non-preprocessed 5-dimensional regressors. From left to right: Sunspot, Henon, Rossler and AR(3) time series.

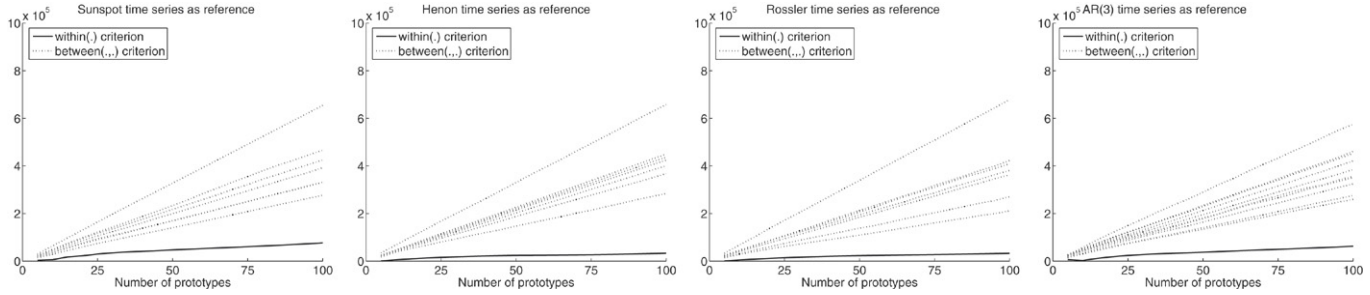

Fig. 16. Values for the *within*(.) and *between*(., .) criteria for 5-dimensional regressors preprocessed using the unfolding methodology. From left to right: Sunspot, Henon, Rossler and AR(3) time series.

Table 2
Lags selected using the distance to diagonal criterion for the used time series, for the case of 5-dimensional regressors

Time series	Lag	Time series	Lag
Sunspot	20	Rossler	10
Santa Fe A	2	Lorenz	20
Polish	2	Henon	1
AR(3)	1	Tinkerbell	1
ARMA(4, 4)	1	Random Walk	50
SARMA(3, 3)	1		

Table 3
Lags selected using the distance to diagonal criterion for the used time series, for the case of 10-dimensional regressors

Time series	Lag	Time series	Lag
Sunspot	10	Rossler	5
Santa Fe A	1	Lorenz	10
Polish	1	Henon	1
AR(3)	1	Tinkerbell	1
ARMA(4, 4)	1	Random Walk	20
SARMA(3, 3)	1		

Figs. 15 and 16 lead to the same conclusion as in the 3-dimensional case except that the *within*(.) and *between*(., .) values are even more distinct in this case; see for example the plots for the Sunspot, Rossler or AR(3) time series. Once again, approximately the same values are obtained with non-preprocessed and unfolded regressors for the Henon time series: in this case, the lag τ is also 1.

Finally, the same methodology is again applied to 10-dimensional regressors. The selected lags are presented in Table 3. Comparison of the *within*(.) and *between*(., .) values for some time series are provided in Figs. 17 and 18. As for 3-dimensional and 5-dimensional regressors, the unfolding preprocessing of regressors in a 10-dimensional space allows an easier comparison of the *within*(.) and *between*(., .) criteria.

Four supplementary conclusions may be extracted from the experimental results. First, for time series obtained from simple linear models (e.g. AR, ARMA, SARMA), the distance to diagonal criterion gives similar lags to the autocorrelation.

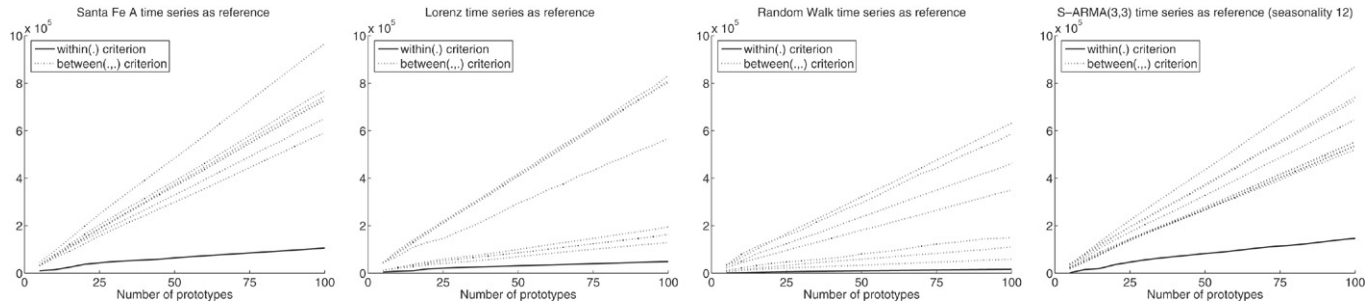

Fig. 17. Values for the *within*(.) and *between*(., .) criteria for non-preprocessed 10-dimensional regressors. From left to right: Santa Fe A, Lorenz, Random Walk and SARMA(3, 3) time series.

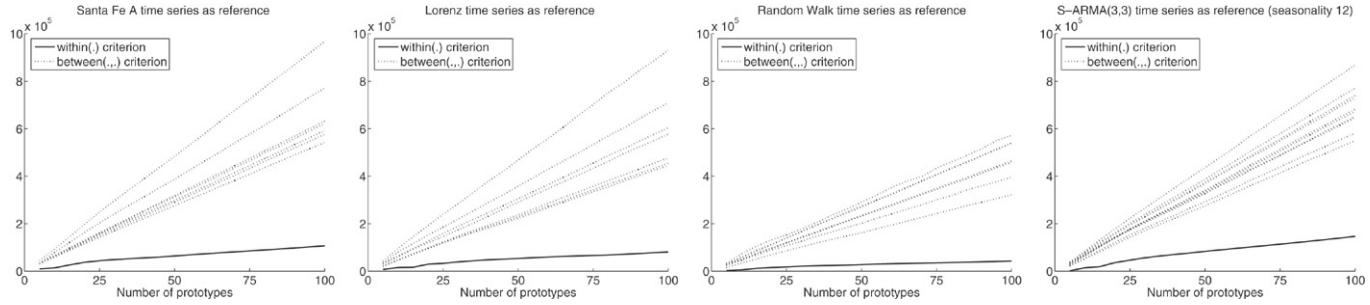

Fig. 18. Values for the *within*(.) and *between*(., .) criteria for 10-dimensional regressors preprocessed using the unfolding methodology. From left to right: Santa Fe A, Lorenz, Random Walk and SARMA(3, 3) time series.

Second, the distance to diagonal criterion is valid whatever the dimension of the regressor; it takes into account the whole unfolding in the regressor space, rather than the correlation or other criterion between two successive values only. Third, in general the gaps between the *within*(.) and *between*(., .) criteria are larger when using higher-dimensional regressors; this is consistent with the fact that larger regressors usually better capture the dynamics of a series. Finally, it can be noticed from Tables 1–3 that the lag selected through the distance to diagonal criterion decreases (for a given time series) when the regressor dimension increases. This can be understood by the fact that, when using a large regressor, the necessary previous values will be included even if the lag τ is too small, just because the lag is multiplied by a large p in Eq. (8).

To conclude, it can be said that using the *within*(.) and *between*(., .) criteria it has been possible to observe that, once correctly unfolded with the distance to diagonal criterion, regressors obtained from distinct time series are indeed distinguishable using a clustering technique. The latter is implemented here by self-organizing maps. Similar experiments have been performed using the K-means and Competitive Learning algorithms; in both cases, it can be shown that time series clustering is also meaningful.

6. Conclusion

In this paper, the "distance to diagonal" criterion is introduced to measure how well a time series regressor distribution is unfolded. Such unfolding is necessary to preserve in the regressors the information contained in the series. The distance to diagonal criterion measures how much all components of the regressor contribute to useful information, unlike the autocorrelation function that only takes two values into account.

Unfolding the regressor distribution is a necessary condition to keep the information in the set of regressors but does not constitute a sufficient condition, as illustrated by a random walk. However, it is expected that when prediction is possible with fixed-size regressors, choosing the lag according to the distance to diagonal criterion will help in keeping information for further processing. As having a low autocorrelation of the regressor components and keeping the lag small are also important objectives, the methodology detailed in this paper suggests using these two last criteria as subsidiary ones when the distance to diagonal criterion allows the choice of a range of lags.

While the methodology is presented here for fixed lags (equal time delays between regressor components), nothing prevents us from applying the same technique to regressors with variable lags. However, several values are to be set in this case, making the search for minima a multi-dimensional search instead of a 1-dimensional one.

Recently, a set of papers have been published with arguments concerning the meaningless character of time series clustering. While explanations have been given too, it remains that these papers have led to suspicion about the usefulness of time series clustering in all situations. By using criteria that differentiate sets of regressor prototypes, this paper shows that time series clustering is meaningful, at least when adequate preprocessing is applied. Experimental results obtained using real and artificial time series indeed show that using adequate preprocessing leads to a clear distinction of the sets of the SOM prototypes and thus to the conclusion of the meaningfulness of clustering.

Acknowledgements

We would like to thank Professor Osowsky from Warsaw Technical University for providing us with the Polish electrical data used. The authors are grateful to the reviewers for their helpful comments and in-depth reading of the manuscript.

References

Abarbanel, H. D. I. (1997). *Analysis of observed chaotic data*. New-York: Springer-Verlag.

Alligood, K. T., Sauer, T. D., & Yorke, J. A. (1996). *Chaos: An introduction to dynamical systems*. Berlin: Springer-Verlag.

Babloyantz, A., Nicolis, C., & Salazar, J. M. (1985). Evidence of chaotic dynamics of brain activity during the sleep cycle. *Physics Letters, 111 A*, 152–156.

Bagnall, A. J., Janakec, G., & Zhang, M. (2003). Clustering time series from mixture polynomial models with discretised data. Technical report CMP-C03-17. School of Computing Sciences, University of East Anglia. Available from http://www2.cmp.uea.ac.uk/˜ajb/PDF/CMP-C03-17.pdf.

Camastra, F., & Colla, A. M. (1999). Neural short-term prediction based on dynamics reconstruction. *Neural Processing Letters*, *9*(1), 45–52.

Cottrell, M., Girard, B., & Rousset, P. (1997). Long term forecasting by combining kohonen algorithm and standard prevision. In W. Gerstner, A. Germond, M. Hasler, & J. D. Nicoud (Eds.), *Lecture notes in computer science*: *Vol. 1327*. *Proc. of int. conf. on artificial neural networks* (pp. 993–998). Springer.

Cottrell, M., Fort, J. -C., & Pagès, G. (1998). Theoretical aspects of the SOM algorithm. *Neurocomputing*, *21*, 119–138.

Dablemont, S., Simon, G., Lendasse, A., Ruttiens, A., Blayo, F., & Verleysen, M. (2003). Time series forecasting with SOM and local non-linear models — Application to the DAX30 index prediction. In *Proc. of workshop on self-organizing maps* (pp. 340–345).

Denton, A. (2004). Density-based clustering of time series subsequences. In *Proc. of 3rd workshop on mining temporal and sequential data, in conjunction with 10th ACM SIGKDD int. conf. on knowledge discovery and data mining*.

Ding, M., Grebogi, C., Ott, E., Sauer, T., & Yorke, A. (1993). Estimating correlation dimension from a chaotic time series: When does plateau onset occur? *Physica D*, *69*(3–4), 404–424.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

Fraser, A. M., & Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A*, *33*, 1134–1140.

Gautama, T., Mandic, D. P., & Van Hulle, M. M. (2004). The delay vector variance method for detecting determinism and nonlinearity in time series. *Physica D*, *190*, 167–176.

Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D*, *9*, 189–208.

Hetland, M. L. (2003). Evolving sequence rules, Ph.D. thesis. Norwegian University of Computer and Information Science.

Kantz, H., & Schreiber, T. (1997). *Cambridge Nonlinear Science Series*: *Vol. 7*. *Nonlinear Time Series Analysis*. Cambridge: Cambridge University Press.

Kaplan, D. T., & Glass, L. (1995). *Understanding nonlinear dynamics*. New York: Springer-Verlag.

Keogh, E., Lin, J., & Truppel, W. (2003). Clustering of time series subsequences is meaningless: implications for past and future research. In *Proc. of the 3rd IEEE int. conf. on data mining* (pp. 115–122).

Kohonen, T. (1995). *Springer Series in Information Sciences*. *Self-organizing Maps* (2nd ed.). Berlin: Springer.

Ljung, L. (1999). *System identification — theory for the user* (2nd ed.). Prentice Hall: Upper Saddle River, NJ.

Mahoney, M. V., & Chan, P. K. (2005). Learning rules for time series anomaly detection. Computer science technical report CS-2005-04. Computer Sciences Department, Florida Institute of Technology, Melbourne, FL, available from http://www.cs.fit.edu/˜tr/cs-2005-04.pdf.

Martinetz, T., Berkovich, S., & Schulten, K. (1993). Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, *4*(4), 558–569.

Nusse, H. E., & Yorke, J. A. (1998). *Dynamics: Numerical Explorations* (2nd ed.). New York: Springer-Verlag.

Osowski, S., Siwek, K., & Tran Haoi, L. (2001). Short term load forecasting using neural networks. In: *Proc. of III Ukrainian–Polish workshop* (pp. 72–77).

Sauer, T., Yorke, J., & Casdagli, M. (1991). Embeddology. *Journal of Statistical Physics*, *65*, 579–616.

Simon, G., Lendasse, A., Cottrell, M., Fort, J. -C., & Verleysen, M. (2004). Double quantization of the regressor space for long-term time series prediction: Method and proof of stability. *Neural Networks*, *17*(8–9), 1169–1181. Elsevier.

Simon, G., Lee, J. A., & Verleysen, M. (2005). On the need of unfolding preprocessing for time series clustering. In *Proc. of workshop on self-organizing maps* (pp. 251–258).

SIDC (2005). RWC Belgium, World Data Center for the Sunspot Index, Royal Observatory of Belgium, http://sidc.oma.be/index.php3.

Struzik, Z. (2003). Time series rule discovery: tough, not meaningless. In *Lecture notes in artificial intelligence*: *Vol. 2871*. *Proc. of int. symp. on methodologies for intelligent systems* (pp. 32–39). Springer-Verlag.

Vesanto, J. (1997). Using the SOM and local models in time-series prediction. In *Proc. of workshop on self-organizing maps* (pp. 209–214).

Walter, J., Ritter, H., & Schulten, K. (1990). Non-linear prediction with self-organizing maps. In *Proc. of int. joint conf. on neural networks* (pp. 589–594).

Weigend, A., & Gershenfeld, N. (1994). *Time series prediction: forecasting the future and understanding the past*. Santa Fe Institute, MA: Addison-Wesley Publishing Company.