Reconstruction of missing data in multidimensional time series by fuzzy similarity

Piero Baraldi, F. Di Maio, D. Genini, Enrico Zio

RECONSTRUCTION OF MISSING DATA IN MULTIDIMENSIONAL TIME SERIES BY FUZZY SIMILARITY

P. Baraldi[1], F. Di Maio[1], D.Genini[1], E. Zio[1,2]

[1]*Energy Department, Politecnico di Milano, Via Ponzio 34/3, 20133 Milano, Italy*

[2] *Chair on System Science and the Energetic Challenge, European Foundation for New Energy – Paris and Supelec, Paris, France*

piero.baraldi@polimi.it

ABSTRACT

The present work addresses the problem of missing data in multidimensional time series such as those collected during operational transients in industrial plants. We propose a novel method for missing data reconstruction based on three main steps: (1) computing a fuzzy similarity measure between a segment of the time series containing the missing data and segments of reference time series; (2) assigning a weight to each reference segment; (3) reconstructing the missing values as a weighted average of the reference segments. The performance of the proposed method is compared with that of an Auto Associative Kernel Regression (AAKR) method on an artificial case study and a real industrial application regarding shut-down transients of a Nuclear Power Plant (NPP) turbine.

Keywords – time series, missing data, fuzzy similarity, Auto-Associative Kernel Regression (AAKR), operational transients in industrial plants, Nuclear Power Plant (NPP)

.

NOTATION

M	number of reference trajectories
m	index of the reference trajectories
J	number of dimensions of a trajectory
j	index of the signal
$\bar{\bar{X}}_m^{tr}$	m-th reference trajectory
T	time length of a reference trajectory
k	time index
$x_m^{tr}(k, j)$	value of signal j of trajectory m at time k
$\bar{\bar{X}}$	test trajectory
j_{miss}	signal with missing data
φ	length of the time window with missing data
t	present time in the test trajectory
$x(k, j)$	value of signal j at time $k \le t$ in the test trajectory
$\hat{x}(k, j_{miss})$	reconstruction of a missing datum
L_t	time length used for the similarity computation
$\bar{x}_t(j)$	segment of the most recent L_t measurements of signal j in the test trajectory
$\bar{x}_m^{tr}(k, j)$	segment of length L_t of signal j which ends at time k in reference trajectory m
$\delta_m^2(k, j)$	squared Euclidean distance between the monodimensional segment $\bar{x}_m^{tr}(k, j)$ and $\bar{x}_t(j)$
$\delta_m^2(k)$	squared Euclidean distance between $\bar{\bar{X}}$ and the k-th segment of the m-th reference trajectory
$\mu_m(k)$	membership function value of the "approximately zero" fuzzy set computed in $\delta_m^2(k)$
α	parameter of $\mu_m(k)$
β	parameter of $\mu_m(k)$
$d_m(k)$	distance score between $\bar{\bar{X}}$ and the k-th segment of the m-th reference trajectory
$w_m(k)$	weight given to the k-th segment of the m-th reference trajectory

1. Introduction

The problem of missing data afflicts a variety of application areas in fields such as engineering, economics and finance. The datasets available to build models are often characterized by missing values, due to various causes such as sensor faults, problems of not reacting experiments, not recovering work situations, transferring data to digital systems [Qiao et al., 2005]. Missing data can lead to problems during model development.

In this work, we consider the problem of missing data in the context of on-line condition monitoring of industrial components by empirical, data-driven models [Saxena et al., 2007; Antory, 2007]. On-line condition monitoring aims at informing on the health state of industrial components. When using empirical models, the condition monitoring performance is highly dependent on the availability and quality of the measurements used to establish (train) the model [Reifman, 1997]. Furthermore, during the use of condition monitoring models, if, for example, a sensor fails to provide an input value, the condition monitoring model may not be capable of inferring the health state of the component. Therefore, it is important to restore the missing sensor readings to provide a set of complete input data to the condition monitoring model, for its training and during its use.

Three different approaches to the missing data problem are typically proposed. The first consists in removing from the dataset all the patterns with a missing datum in at least one signal [Nelwamondo et al., 2008]. The main drawback of such approach is the loss of valuable information that may be contained in the discharged signal measurements [Almeida et al, 2010]. The effect of this is particularly relevant when the dataset is small. The second approach consists in substituting the missing data by statistical values, e.g. the mean value of the available historical data [Schafer et al., 2002]. The goodness of the substituting values depends on how close they actually are on the true (unknown) values [Timm et al., 2002]. Finally, the third approach consists in developing model to reconstruct the missing values based on the relationships between the signals [Nelwamondo et al., 2008]. In this context, empirical, auto-associative, modelling techniques such as Auto-Associative Artificial Neural Networks (AANN) [Hines et al., 1996], Principal Component Analysis (PCA) [Hines et al., 2008; Luh et al., 2011] and Auto-Associative Kernel Regression Methods (AAKR) [Garvey et al., 2006; Baraldi et al., 2012] have been applied with success to the reconstruction of missing data. AANN have been shown to work well for reconstruction, although they may require high computational costs in large multidimensional problems [Hashemian et al., 2008]. PCA and AAKR methods are “leaner” and, thus, more suitable for application to missing data reconstruction

in large datasets, due to their lower computational burden. Other more complex approaches applied to the reconstruction of missing data are based on the use of similarity measures computed taking into account the correlation [Kim et al., 2005] and entropy [Brock et al., 2008] between the data. However, the performance of these techniques, which has been proven to be very satisfactory in non time dependent problems such as those encountered in genes expression data, tends to decrease when they are applied to datasets containing non-stationary, time-varying and nonlinear signal behaviours [Borgan et al., 2011].

In this work, we address the problem of missing data in multidimensional time series, e.g. process signals monitored during turbine start-up transients in nuclear power plants or daily flow values of some rivers in a drainage basin. We assume to have available several examples of the time series, e.g. collection of the signal values measured during several, different turbine start-up transients or the daily measurements of the river flows during different years. A difficulty comes from the fact that the reconstruction of a datum missing at a given time does not depend only from the values of the other signals at that time, but also from previous values on the time series.

In this paper, we present a missing data reconstruction method that we have developed based on a Fuzzy Similarity (FS) method [Zio et al., 2010a]. A measure of similarity is computed between a set of reference multidimensional time-series segments of given length and the multidimensional segment containing the missing data; then, the missing values are reconstructed as average of the reference segments weighted by the similarity with the segment containing the missing data.
The method is tested on two case studies: a simulated four-dimensional time series and 27 signals measured during shut-down transients of nuclear power plants (NPP) steam turbines. The performance of the proposed method is compared with that of an AAKR method of literature [Baraldi et al., 2010].

The remainder of the paper is organized as follows: Section 2 introduces the problem that we want to address; Section 3 describes the proposed FS-based method and discusses the metrics that are

used to evaluate the missing data reconstruction performance; Sections 4 and 5 illustrate the application of the method to the artificial case study and to the industrial case study, respectively; finally, Section 6 draws the conclusion of the work.

2. Problem Statement

We consider a training dataset containing *M* different *J*-dimensional realizations of a time series, hereafter called trajectories, and indicated by $\overline{\overline{X}}_m^{tr}, m = 1,..., M$. These reference trajectories are all complete, i.e. they do not suffer of any missing data. For simplicity of illustration, all the reference trajectories are assumed to have the same time length *T*. The generic element $x_m^{tr}(k, j)$ of $\overline{\overline{X}}_m^{tr}$ indicates the value of signal *j* of trajectory *m* at time *k.*

The objective of this work is the reconstruction of missing data in a (test) trajectory, $\overline{\overline{X}}$, that we are measuring. The length of $\overline{\overline{X}}$ can be shorter than *T*. We consider that the values of only one signal, hereafter referred to as j_{miss}, are missing in a single time window from time $t - \varphi$ until the present time *t*. The generic element of the test trajectory, $x(k, j)$, indicates the value of signal *j* at time $k \le t$. The obtained reconstruction of a missing datum of signal j_{miss} will be indicated by $\hat{x}(k, j_{miss}),\ t - \varphi \le k \le t$.

3. The Fuzzy Similarity-based reconstruction method

The proposed method for missing data reconstruction is based on three main steps:

(1) compute a measure of fuzzy similarity between multidimensional segments of the reference trajectories in a time window of length L_t and the most recent segment of length L_t of the test trajectory containing the missing datum;

(2) assign a weight to each reference segment; the weight is chosen proportional to the similarity of the reference segments to the test trajectory;

(3) reconstruct the missing datum as a weighted average of the reference segments.

At the present time *t*, the segment of test trajectory containing only the most recent L_t measurements of signal *j* is denoted as $\bar{x}_t(j) = [x(t - L_t + 1, j); x(t - L_t + 2, j), ..., x(t, j)]$ and the generic segment of length L_t of signal *j* in reference trajectory *m* which ends at time *k* is denoted as $\bar{x}_m^{tr}(k, j) = \bar{x}_m^{tr}(k - L_t + 1 : k, j)$.

With respect to the time window length, L_t, two situations may arise:

CASE A) it is longer than the time duration of the missing data segment, i.e. $L_t > \varphi$; in this case, the values $[x(t - L_t + 1, j_{miss}), ..., x(t - \varphi - 1, j_{miss})]$ are available and only the values in $[x(t - \varphi_t, j_{miss}), ..., x(t, j_{miss})]$ are missing;

CASE B) it is equal or shorter than the time duration of the missing data segment, i.e. $L_t \leq \varphi$; in this case, all the values of $\bar{x}_t(j_{miss}) = [x(t - L_t + 1, j_{miss}); x(t - L_t + 2, j_{miss}), ..., x(t, j_{miss})]$ are missing.

For example, Figures 1 and 2 show the evolution of 4 signals in a time series in a case in which we are unable to measure signal j_{miss}=1 from time *t*=21. To perform the reconstruction of the missing values, we consider a time window of length L_t =5. Thus, at time *t=21* (Figure 1), only *φ=1* measurements are missing for signal j_{miss} (case A), whereas as time passes, the number of missing measurements increases until, at time *t*=26, we reach the situation of case B where $\varphi = L_t$.

For ease of explanation, the case of missing values in a single signal is treated, but generalization to a multi-dimensional problem is straightforward.

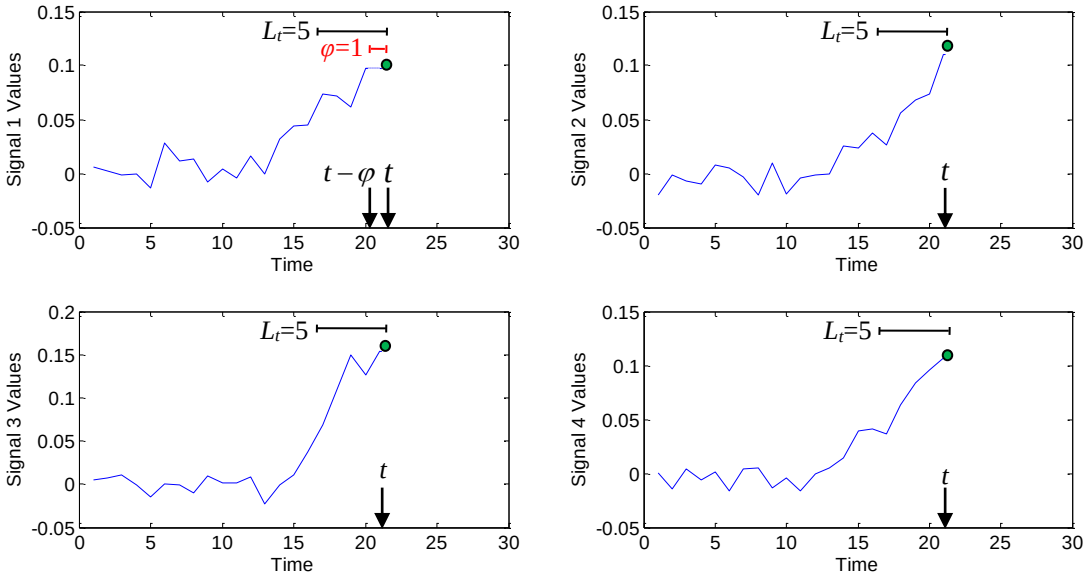

Figure 1. A 4-dimensional trajectory (present time *t*=21, missing data from *t=20)*. The time window considered for signal reconstruction has a length L_t=5.

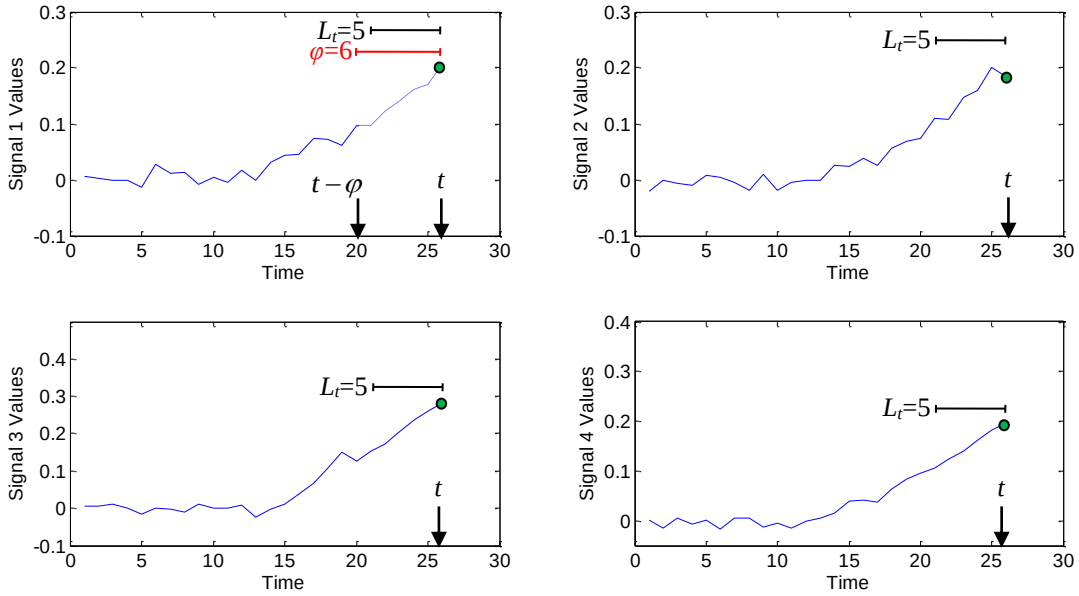

Figure 2. Same trajectory of Figure 1 (present time *t*=26).

The signal reconstruction method is based on the following steps:

Step 1: Trajectory pointwise difference computation

In this step, we compute the squared Euclidean distance difference between the segment of the test trajectory containing the missing data and the segments obtained from the reference trajectories.

In CASE A, characterized by $\varphi < L_t$, the squared Euclidean distance is computed taking into account all the available L_t measurements for the signals $j \neq j_{miss}$ without missing data in the test trajectory and only the $L_t - \varphi$ available measurements for signal j_{miss}. In practice, the test segments $\bar{x}_t(j)$ of the test trajectory $\bar{\bar{X}}$ at time instant t and the k-th segment of the m-th reference trajectory are:

$$\begin{aligned} &\bar{x}_t(j) = [x(t-L_t+1, j); x(t-L_t+2, j); ...; x(t-\varphi, j)] && \text{if } j = j_{miss} \\ &\bar{x}_t(j) = [x(t-L_t+1, j); x(t-L_t+2, j); ...; x(t, j)] && \text{if } j \neq j_{miss} \end{aligned} \tag{2}$$

$$\begin{aligned} &\bar{x}_m^{tr}(k, j) = [x_m^{tr}(k-L_t+1, j); x_m^{tr}(k-L_t+2, j); ...; x_m^{tr}(k-\varphi, j)] && \text{if } j = j_{miss} \\ &\bar{x}_m^{tr}(k, j) = [x_m^{tr}(k-L_t+1, j); x_m^{tr}(k-L_t+2, j); ...; x_m^{tr}(k, j)] && \text{if } j \neq j_{miss} \end{aligned} \tag{3}$$

The squared Euclidean distance between the L_t elements of the k-th segment of the m-th reference trajectory $\bar{x}_m^{tr}(k, j)$ and the elements of the test time segment $\bar{x}_t(j)$ of the j-th signal, is given by:

$$\delta_m^2(k, j) = \left| \bar{x}_m^{tr}(k, j) - \bar{x}_t(j) \right|^2 \tag{4}$$

The distance is finally reduced to [Zio et al., 2010b]:

$$\delta_m^2(k) = \frac{\sum_{j=1}^{J} \delta_m^2(k, j)}{J} \tag{5}$$

In CASE B), characterized by all missing values of signal j_{miss} in L_t, i.e. $\varphi \geq L_t$, only the remaining signals $j \neq j_{miss}$ are considered for the computation of the squared Euclidean

distance. In practice, the test segments $\bar{x}_t(j)$ of the test trajectory $\bar{\bar{X}}$ and the *k*-th segment of the *m*-th reference trajectory, are:

$$\bar{x}_t(j) = [x(t - L_t + 1, j); x(t - L_t + 2, j); ...; x(t, j)]$$
$$\bar{x}_m^{tr}(k, j) = [x_m^{tr}(k - L_t + 1, j); x_m^{tr}(k - L_t + 2, j); ...; x_m^{tr}(k, j)] \qquad j \neq j_{miss} \qquad (6)$$

The squared Euclidean distance $\delta_m^2(k, j)$ between the L_t components of the *k*-th time segment of the *m*-th reference trajectory $\bar{x}_m^{tr}(k, j)$ and the L_t measurements of the test time segment $\bar{x}_t(j)$ of the *j*-th signal, $j \neq j_{miss}$, is calculated as in eq. (4).

The distance is finally divided by the number of signals without missing values $J - 1$:

$$\delta_m^2(k) = \frac{\sum_{j=1}^{J-1} \delta_m^2(k, j)}{J - 1} \qquad (7)$$

Step 2: Trajectory pointwise similarity and distance score computation

To account for a gradual transition between 'similar' and 'non-similar' we introduce an "approximately zero" fuzzy set [Zio et al., 2010c] taken, in this work, as a bell-shaped function, whose membership function value computed in $\delta_m^2(k)$ is equal to:

$$\mu_m(k) = e^{-\left(-\frac{\ln(\alpha)}{\beta^2}\delta_m^2(k)\right)} \qquad (8)$$

The parameters α and β are set by the analyst: the larger the value of the ratio $\ln(\alpha)/\beta^2$, the narrower the fuzzy set and the stronger the definition of similarity [Zio et al., 2010a].

The distance score $d_m(k)$ between two segments is, then, computed as:

$$d_m(k) = 1 - \mu_m(k) \qquad k = L_t, ..., T \ , \ m = 1, ..., M \tag{9}$$

Step 3: weights computation

The basic idea behind the weighted reconstruction is that all the reference training trajectories carry useful information for the reconstruction of the missing data in the currently developing trajectory [Zio et al., 2010a]. To this aim, weights are computed with a decreasing monotone function [Zio et al., 2010a], such that the smaller the distance $d_m(k)$ the larger the weight given to the *k*-th segment of the *m*-th reference trajectory:

$$w_m(k) = (1 - d_m(k)) e^{-\frac{1}{\beta} d_m(k)} \qquad k = L_t, ..., T \ , \ m = 1, ..., M \tag{10}$$

The same value of β used in eq.(8) is here employed in order to reduce the number of parameters to be set.

Step 4: Missing Values Reconstruction

The reconstruction of the missing datum, $\hat{x}(t, j_{miss})$, in the test trajectory is the weighted sum of the last elements $x_m^{tr}(k, j_{miss})$ of the reference training segments [Baraldi et al., 2010]:

$$\hat{x}(t, j_{miss}) = \frac{\sum_{m=1}^{M} \sum_{k=L_t}^{T} w_m(k) x_m^{tr}(k, j_{miss})}{\sum_{m=1}^{M} \sum_{k=L_t}^{T} w_m(k)} \tag{11}$$

4. Application to an artificial case study

We have considered an artificial case study built by simulating *M*=114 trajectories of length *T* =100 time steps (in arbitrary units) in a *J*=4 dimensional signal space, whose projections into the 4-

dimensional space are plotted in Figures 3-6. The trajectories are simulated by choosing for each signal one of the three following functional behaviors:

$$x_j = 2\alpha_1 a\left[1+erf\left(\frac{(t-e)-u}{\sqrt{2}}\right)\right]+10^{-3\omega} \tag{12}$$

$$x_j = \alpha_2\left(c^{d^{(t-e)}}-c\right)+10^{-3\omega} \tag{13}$$

$$x_j = \alpha_3 b(t-e)+10^{-3\omega} \tag{14}$$

where j=1, 2, 3, 4 is the signal index and a, b, c, d, e, μ, $\omega, \alpha_1, \alpha_2$ and α_3 are values randomly sampled from the probability distributions listed in Table I.

Parameter	Distribution
a	Uniform [0.45,0.55]
b	Uniform [0.3,0.4]
c	Uniform [1.1,1.3]
d	Uniform [1.2,1.3]
e	Uniform [0,2]
α_1	Uniform [1,1.5]
α_2	Uniform [1,1.2]
α_3	Uniform [0.9,1.1]
u	Uniform [2.2,2.7]
ω	Normal (0,1)

TABLE I. Distribution of the random parameters *a, b, c, d, e, u, ω, α1, α2 and α3*

One hundred trajectories have been simulated using eq.(12) for signals 1, 2 and 4, and eq.(14) for signal 3 (Table II). These trajectories are intended to reproduce the nominal behavior of a system, e.g. the operation of a plant when no anomalies or faults occur. Hereafter, they will be referred to as 'normal condition trajectories'. To artificially reproduce possible anomalous behaviors of the system, 14 trajectories characterized by signal functional behaviors different from those of the

corresponding signals in the 100 trajectories have been simulated according to the eqs. reported in Table II.

The performance of the proposed model for missing data reconstruction will be evaluated in terms of its accuracy, i.e. the ability of providing correct reconstructions of missing data. The metric used is the Mean Square Error (MSE) between the reconstructions provided by the model and the true values [Baraldi et al., 2010], averaged over a number of different test trajectories.

Trajectories	Signal 1	Signal 2	Signal 3	Signal 4
1-100	Eq. (12)	Eq. (12)	Eq. (14)	Eq. (12)
101-105	Eq. (12)	Eq. (13)	Eq. (14)	Eq. (12)
106-109	Eq. (13)	Eq. (14)	Eq. (12)	Eq. (12)
110-112	Eq. (13)	Eq. (14)	Eq. (13)	Eq. (12)
113-114	Eq. (14)	Eq. (14)	Eq. (14)	Eq. (12)

Table II Equations used to simulate the signal evolution in the 114 trajectories of the case study

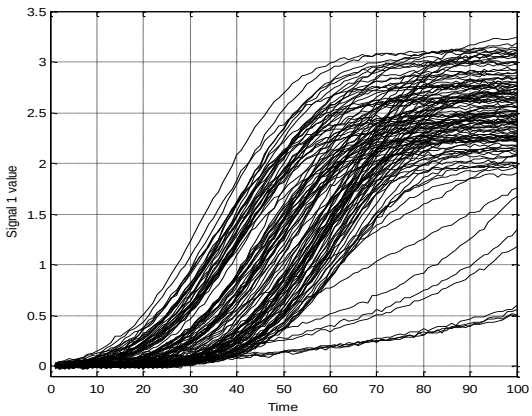

Figure 3 Projection on the signal 1 axis of the 114 simulated trajectories

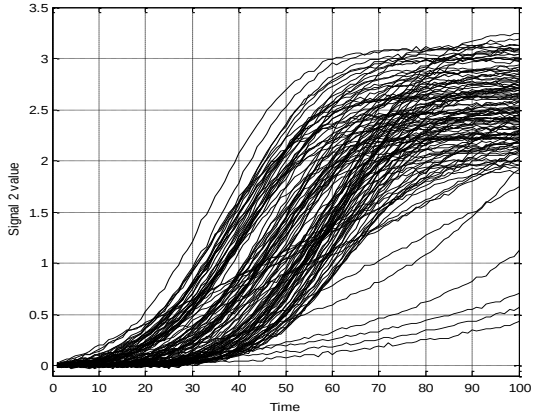

Figure 4 Projection on the signal 2 axis of the 114 simulated trajectories

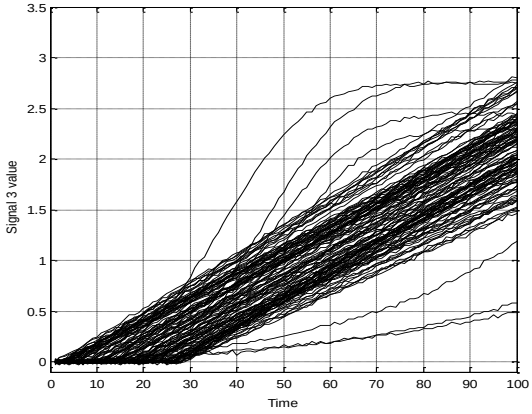

Figure 5 Projection on the signal 3 axis of the 114 simulated trajectories

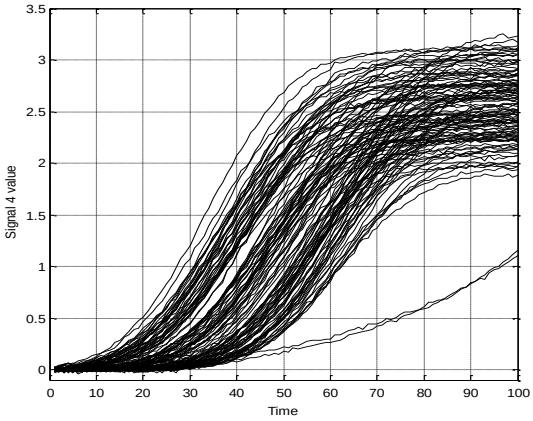

Figure 6 Projection on the signal 4 axis of the 114 simulated trajectories

4.1 Results

The application of the proposed signal reconstruction method requires to properly set the values of the length of the time window for the segment of values used for the reconstruction, L_t, and of the parameters α and β in eqs.(8) and (10). Table III reports the optimal values of α and β found with a systematic procedure based on the computation of the MSE on a validation set. In practice, we have considered all the possible combinations of L_t, α and β with L_t=2,5,10,20 and α and β with discrete values taken from the intervals $[10^{-6}, 10^{-1}]$ and $[10^{-3}, 10^{-1}]$, respectively. Finally, the parameter setting with minimum MSE is selected. Notice that the larger L_t, the smaller the ratio $\ln(\alpha)/\beta^2$ (in eq. (8)): this means that the the larger L_t, the larger the fuzzy set that defines the similarity between reference trajectories and test time segment. This is expected because the similarity μ in eq. (8) has to accommodate larger the Euclidean distance between the segments (eq. (7)) due to larger L_t, i.e., longer reference trajectories and time segments.

L_t	α	β
2	0.8	0.01
5	$1x10^{-5}$	0.05
10	$1x10^{-5}$	0.05
20	1x10-5	0.05

TABLE III. Optimal values of the parameters α and β in correspondence of different time segment lengths L_t

Figure 7 shows the accuracy metric, MSE, as a function of the length of the time segment, L_t, obtained in the reconstruction of $\varphi = 20$ data in signal 1 (upper left), 2 (upper right), 3 (lower left) and 4 (lower right). In this case, the best results are obtained with $L_t = 10$.

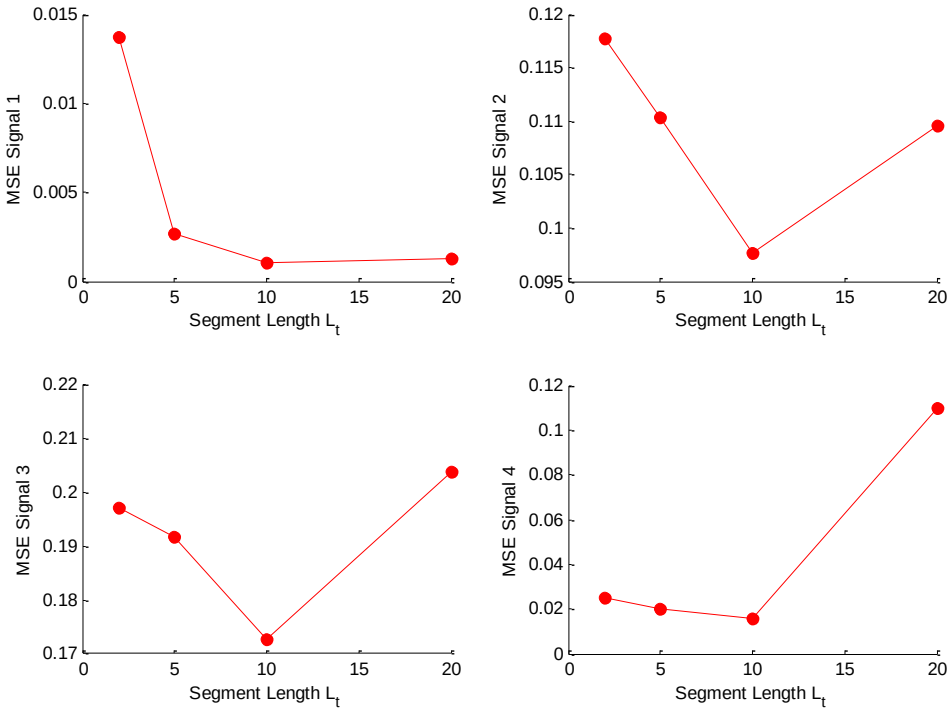

Figure 7: Reconstruction accuracy considering different values of L_t

In order to verify the performance of the signal reconstruction method, we have considered 113 trajectories to train the reconstruction model and one trajectory to test it. In the test phase, we assume to have missing data in only one signal from time t_A=21 to time t_B=40, and we generate

from a trajectory 4 different test trajectories each one containing missing data in a different signal. In order to cross-validate the accuracy computation, this procedure has been repeated 114 times, each time choosing a different test trajectory according to a leave-one-out procedure [Quan et al., 2010; Baraldi et al., 2011]. Table IV reports the accuracy obtained by the proposed method and by a literature AAKR method for comparison [Baraldi et al., 2010].

	Test on all 114 trajectories	Test on the 100 normal condition trajectories	Test on the 14 anomalous trajectories
AAKR	8.3×10^{-4}	1.7×10^{-4}	5.6×10^{-3}
FS-BASED	2.9×10^{-4}	1.5×10^{-4}	1.3×10^{-3}

Table IV Reconstruction accuracy provided by the AAKR and the proposed Fuzzy Similarity-based reconstruction method.

The two methods provide similar performances in the reconstruction of the 100 normal condition trajectories, whereas the proposed method is remarkably more accurate than the AAKR one in the reconstruction of the 14 anomalous trajectories. This is pictorially seen in Figures 8 and 9, which show the reconstructions obtained in normal condition and anomalous trajectories, respectively.

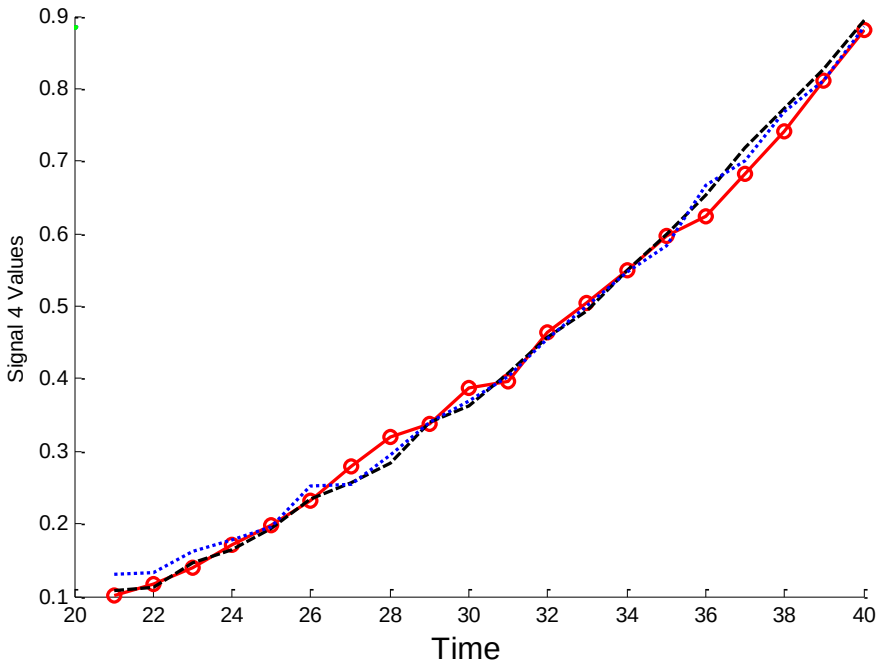

Figure 8: Missing values reconstruction for a normal condition trajectory; the true value is represented by a short dashed line, the Fuzzy Similarity-based reconstruction by a thick continuous line with circles and the AAKR reconstruction by a long dashed line

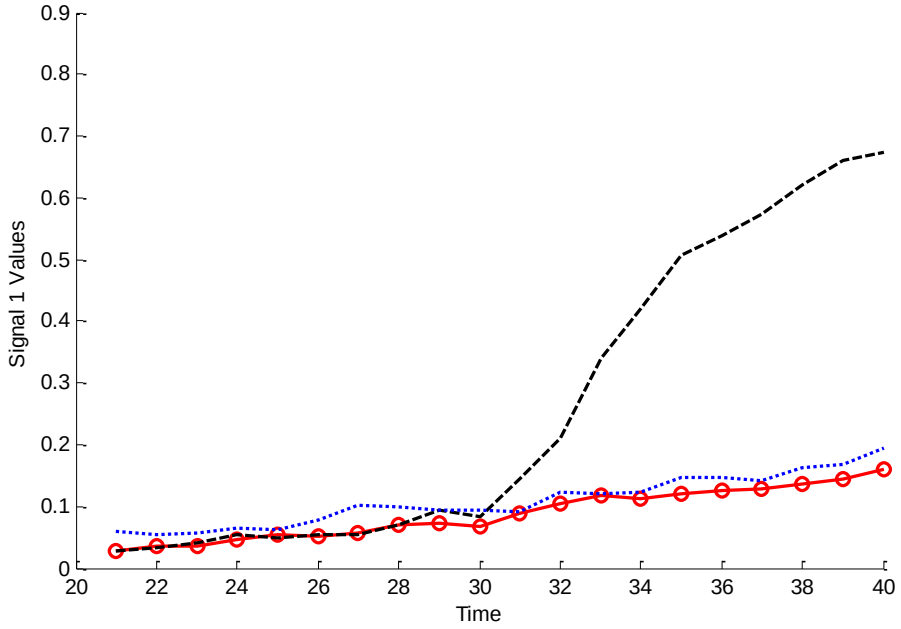

Figure 9: Missing values reconstruction of signal 1 in a test trajectory. The true value is represented by a short dashed line, the FS-based reconstruction by a thick continuous line with circles and the AAKR reconstruction by a long dashed line

It is interesting to observe that in the case of Figure 9 from *t*=32 until the end of the test trajectory, the AAKR reconstruction deviates largely from the true signal values, whereas the FS-based reconstruction is more accurate. The difference in performance between the two reconstruction methods is motivated by the different procedure used for the weights assignment. For instance, let us consider the signal 1 reconstructions provided at time *t* = 34 whose true (unknown) value $x(34,1) = 0.053$. The FS-based method identifies in the four-dimensional training input space a reference segment, $\bar{x}_{\tilde{m}}^{tr}\left(\tilde{k},1\right)$, very similar to the test segment $\bar{x}_{t=34}(1)$ (Figure 10, top, line with circles) and it assigns to this segment a very large weight (1.2×10^6). Thus, by applying eq.(11), one obtains a signal reconstruction of the missing data, $\hat{x}(34,1) = 0.080$, which is very similar to the last signal value in $\bar{x}_{\tilde{m}}^{tr}\left(\tilde{k},1\right)$. On the other hand, the AAKR is considering similarities between instantaneous signal values and not between time segments. According to this different procedure for the similarity computation, the AAKR assigns the largest weights (all in the range of [3,4]) to

several training patterns (circles in Figure 10, bottom) leading to a reconstruction $\hat{x}(34,1) = 0.55$, which is ten times larger than the signal true value $x(34,1) = 0.053$.

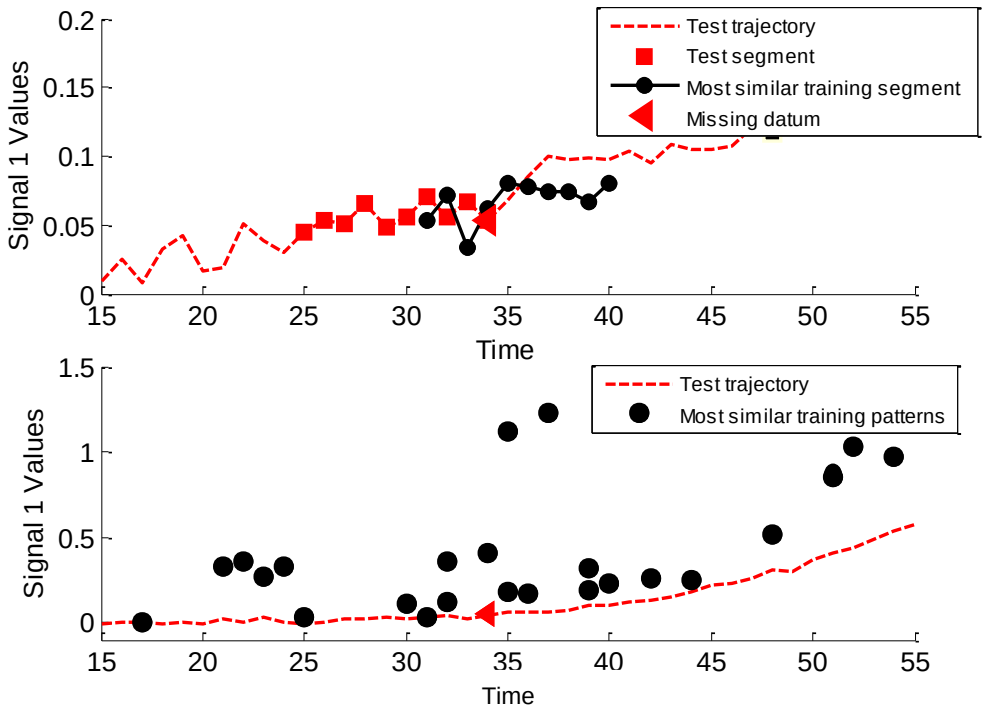

Figure 10. Upper: signal 1 evolution in the test trajectory (dotted line) including signal 1 test segment (squares) used for the reconstruction and the missing datum (triangle). The most similar segment in the training set is represented by circles. Bottom: signal 1 evolution in the test trajectories (continuous line) and missing datum (triangle). Patterns in the training set to which the largest weights are associated are represented by a circle. Notice the different scales of the signal 1 value in the two figures.

5. Application to an industrial case study

The industrial case study concerns the operation of nuclear power plant turbines during shut-down transients. We consider the values of $J = 27$ signals taken at $T = 4500$ time steps in $M = 148$ different transients. Most of the signals refer to temperatures measured in different parts of the turbines [Baraldi et al., 2010]. Figure 11 shows some examples of signal evolutions during different plant transients.

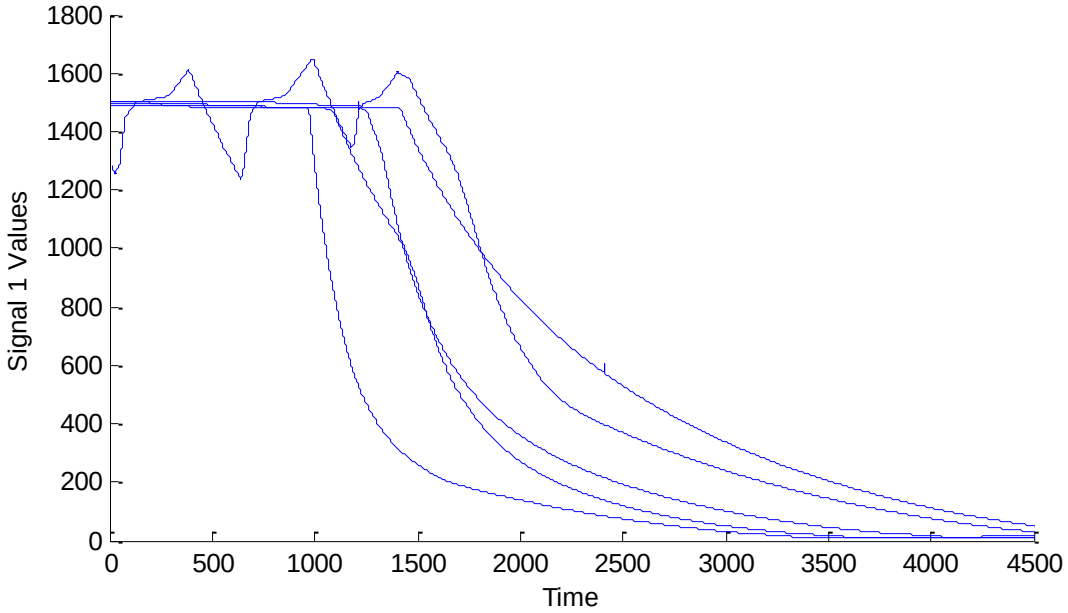

Figure 11: Evolution of a signal in different plant transients.

In order to verify the performance of the FS-based method, in this Section we perform the reconstruction of signal segments whose true values is known, but for which we assume to have missing data in time intervals of φ=20 time instants in a single signal.

The values of the FS-method parameters α and β, and of the length of the time segment have been set equal to $\alpha = 0.2$, $\beta = 0.5$ and $L_t = 10$. Figure 12 shows the overall accuracy of the FS-based and of the AAKR methods in the reconstruction of the 27 plant signals in test transients, according to a leave-one-out procedure.

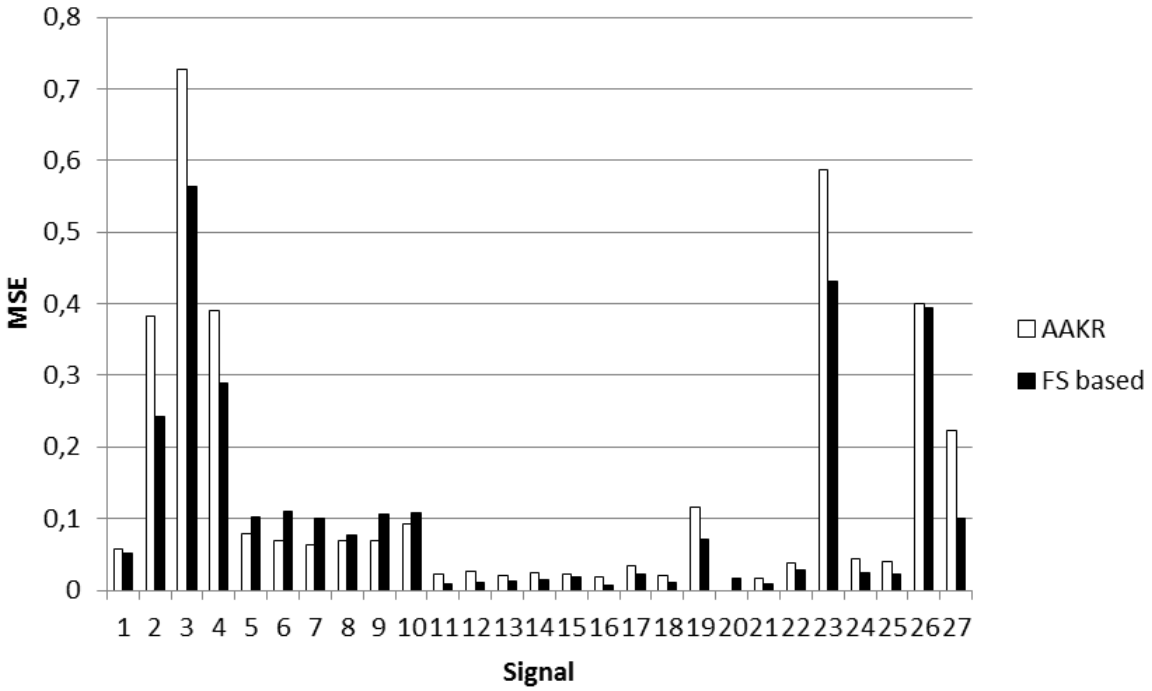

Figure 12: Average accuracy in the reconstruction of the 27 signals.

The performance of the FS-based reconstruction method is more satisfactory than that of the AAKR method. In particular, Figure 13 compares the FS-based method (continuous line with circles) and the AAKR (long dashed line) reconstructions of some signals in a transient, assuming missing data on a time window from instant t_A=81 to instant t_B=100. Notice that for several signals such as j=2, 14, 18, 22 and 27 the AAKR method provides reconstructions which remarkably deviate from the true signal values, whereas the reconstructions of the FS-based method are more accurate.

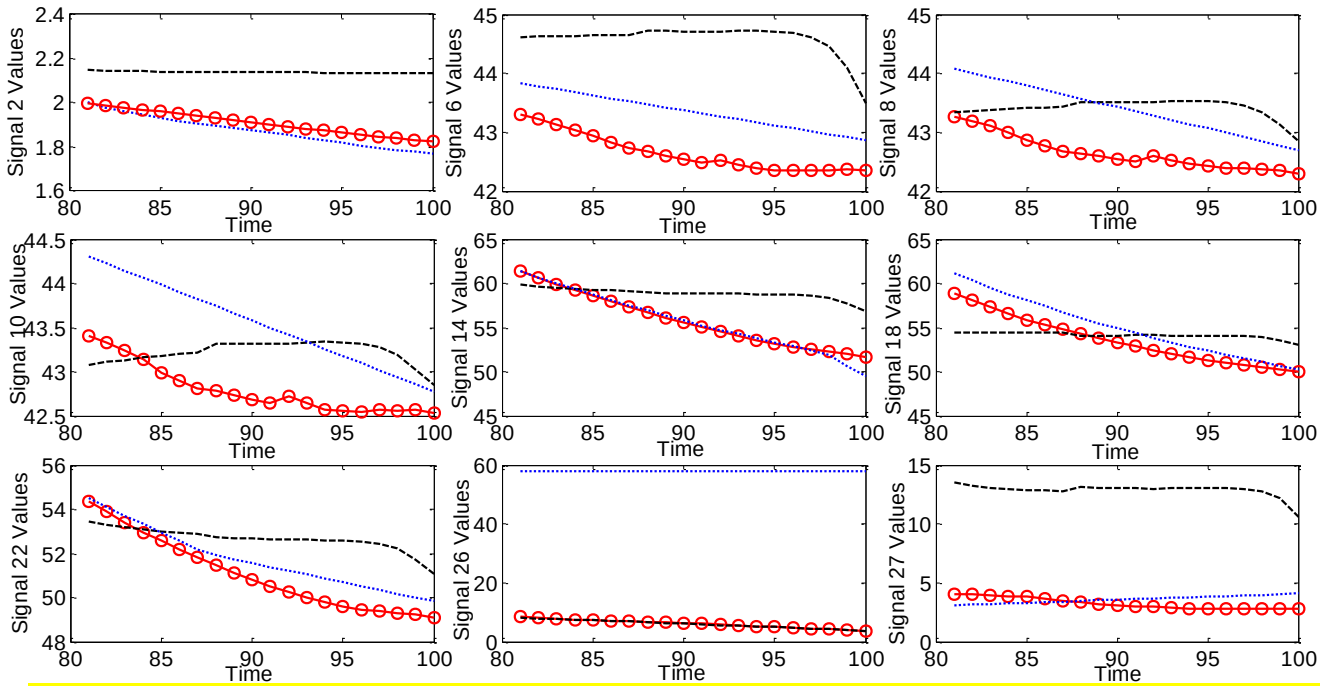

Figure 13. Reconstruction of a time window of 20 time instants in some signals. The true value is represented by a short dashed lines, the FS-based reconstruction by a continuous line with circles and the AAKR reconstruction by a long dashed line.

With respect to signal 26, both methods provide very inaccurate reconstructions. This is due to the fact that the behaviour of this transient in signal 26 is very different from the behaviour of the signal in all the considered reference trajectories as shown in Figure 14. Since the reconstruction is a weighted mean of the training values, this leads to the inaccurate signal reconstruction.

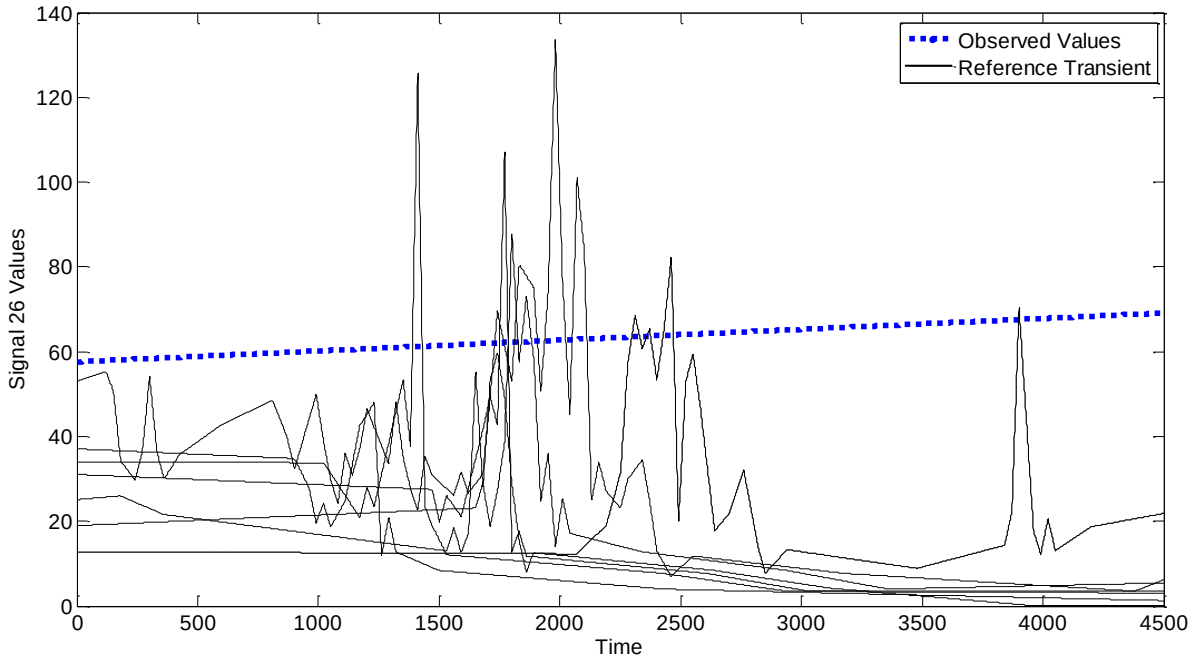

Figure 14 Signal 26 behaviour in the test (dotted line) and in the reference transients (continuous line).

6. Conclusions

A fuzzy similarity-based method for missing data reconstruction has been proposed in the context of on-line condition monitoring of industrial components. The method allows performing signal reconstructions in multidimensional time series. It has been applied with success to an artificial case study and a real industrial application concerning the reconstruction of missing data in nuclear component transients, and it has been shown superior to an AAKR-based method of literature.

Given the difficulty of the signal reconstruction task in situations characterized by the presence of long segments containing missing values, we think that it would be important to associate the signal reconstruction with an estimate of its degree of confidence, which should take into account the amount and quality of the information used to perform the reconstruction. This will be object of future research activity. Indeed, future work will be devoted to estimate the degree of confidence in

the provided signal reconstruction, taking into account the amount of information available in the reference trajectories used to perform the reconstruction.

Acknowledgments

The authors are thankful to EDF R&D STEP Department for providing the data for the case study.

References

[Almeida et al, 2010] R.J. Almeida, U. Kaymak, and J.M.C. Sousa, "A new approach to dealing with missing values in data-driven fuzzy modeling", Proc. FUZZ-IEEE, pp.1-7, 2010.

[Antory, 2007] D. Antory, "Application of a data-driven monitoring technique to diagnose air leaks in an automotive diesel engine: A case study", Mechanical Systems and Signal Processing, Volume 21, Issue 2, p. 795-808, 2007.

[Baraldi et al., 2010] P. Baraldi, R. Canesi, E. Zio, R. Seraoui, R. Chevalier, Signal Grouping for Condition Monitoring of Nuclear Power Plant Components, Seventh American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technology, NPIC&HMIT 2010, Las Vegas, Nevada, 7-11 November, 2010.

[Baraldi et al., 2011] P. Baraldi, R. Razavi-Far, E. Zio, "Bagged Ensemble of FCM Classifiers for Nuclear Transient Identification", Annals of Nuclear Energy; Vol. 38 (5), pp 1161-1171, 2011.

[Baraldi et al., 2012] P. Baraldi, F. Di Maio, L. Pappaglione, E. Zio, R. Seraoui, Condition Monitoring of Electrical Power Plant Components During Operational Transients, Proceeding of the Institution of Mechanical Engineers, Part 0, Journal of Risk and Reliability, 226 (6) , pp. 568-583, 2012.

[Borgan et al., 2011] T. Borgan, T. E. Onshus, E. Lunde, Condition Monitoring Based on "black box" – and First Principal Methods, Master of Science in Engineering Cybernetics, NTNU, 2011.

[Brock et al., 2008] G. N. Brock, J.R. Shaffer, R.E. Blakesley, Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. Bioinformatics 2008;9:12.

[Garvey et al., 2006] D.R. Garvey, J. W. Hines., The Development of a Process and Equipment Monitoring (PEM) Toolbox and its Application to Sensor Calibration Monitoring, Quality and Reliability Engineering International, Vol. 22, pp. 1-13, 2006.

[Hashemian et al., 2008] H. M. Hashemian, J. Eiler, Online Monitoring, For Improving Performance of Nuclear Power Plants, Part 2: Process and Component Condition Monitoring and Diagnostics, IAEA Nuclear Energy Series No. NP-T-1.2, International Atomic Energy Agency, 2008.

[Hines et al., 1996] J. W. Hines, D. J. Wrest, R. E. Uhrig, Plant Wide Sensor Calibration Monitoring, In Proceedings of the 1996 IEEE International Symposium on Intelligent Control, Dearborn, Michigan. IEEE, Piscataway, New Jersey, pp. 378-383. September 15-18, 1996.

[Hines et al., 2008] J. W. Hines, D. Garvey, R. Seibert, A. Usynin, Technical Review of on-line monitoring techniques for performance assessment, Volume 2: Theoretical Issues, Office of Nuclear Regulatory Research, 2008.

[Kim et al., 2005] H. Kim, G.H. Golub, H.Park, Missing Value Estimation for DNA microarray gene expression data: local least squares imputation, Bioinformatics, 2005;21(2):187–98

[Luh et al., 2011] G. C. Luh, C.Y. Lin, PCA based immune networks for human face recognition Applied Soft Computing Journal, 11 (2), pp. 1743-1752, 2011.

[Nelwamondo et al., 2008] F. V. Nelwamondo, T. Marwala, Techniques for handling missing data: applications to online condition monitoring, International Journal of Innovative Computing, Information and Control, Vol. 4, 6, 2008.

[Qiao et al., 2005] W Qiao, Z. Gao, R.G. Harley, Continuous on-line identification of nonlinear plants in power systems with missing sensor measurements, Proceedings of the International Joint Conference on Neural Networks, 3, art. no. 1556141, pp. 1729-1734, 2005.

[Quan et al., 2010] T. Quan, X. Liu, X., Q. Liu Weighted least squares support vector machine local region method for nonlinear time series prediction, Applied Soft Computing Journal, 10 (2), pp. 562-566, 2010.

[Reifman, 1997] J. Reifman, Survey of artificial intelligence methods for detection and identification of component faults in nuclear power plants, Nuclear Technology, 119, pp. 76-97, 1997.

[Saxena et al., 2007] A. Saxena, A. Saad, Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems, Applied Soft Computing Journal, 7 (1), pp. 441-454, 2007.

[Schafer et al., 2002] J. L. Schafer, J.W. Graham, Missing data: our view of the state of the art, Psychological Methods, Vol. 2, pp. 147-177, 2002.

[Timm et al., 2002] H. Timm, C. Doring, R. Kruse, Fuzzy cluster analysis of partially missing datasets, in: Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and Their Implementation on Smart Adaptive Systems (EUNITE 2002), 2002.

[Zio et al., 2010a] E. Zio, F. Di Maio, A data-driven fuzzy approach for predicting the remaining useful life in dynamic failure scenarios of a nuclear system, Reliability Engineering and System Safety,Vol. 95, 2010.

[Zio et al., 2010b] E. Zio, F. Di Maio, M. Stasi, A Data Driven Approach for Predicting Failure Scenarios in Nuclear Systems, Annual of Nuclear Energy, 37, 482-491, 2010.

[Zio et al., 2010c] E. Zio, F. Di Maio, A fuzzy similiarity-based Method for Failure Detection and Recovery Time Estimation, International Journal of Performability Engineering, Vol. 6, No. 5, 2010.