Neural Comput & Applic
DOI 10.1007/s00521-011-0526-5

ORIGINAL ARTICLE

# Data validation and missing data reconstruction using self-organizing map for water treatment

**B. Lamrini · El-K. Lakhal · M-V. Le Lann · L. Wehenkel**

Received: 19 February 2009 / Accepted: 12 January 2011


**Abstract** Applications in the water treatment domain generally rely on complex sensors located at remote sites. The processing of the corresponding measurements for generating higher-level information such as optimization of coagulation dosing must therefore account for possible sensor failures and imperfect input data. In this paper, self-organizing map (SOM)-based methods are applied to multiparameter data validation and missing data reconstruction in a drinking water treatment. The SOM is a special kind of artificial neural networks that can be used for analysis and visualization of large high-dimensional data sets. It performs both in a nonlinear mapping from a high-dimensional data space to a low-dimensional space aiming to preserve the most important topological and metric relationships of the original data elements and, thus, inherently clusters the data. Combining the SOM results with those obtained by a fuzzy technique that uses marginal adequacy concept to identify the functional states (normal or abnormal), the SOM performances of validation and reconstruction process are tested successfully on the experimental data stemming from a coagulation process involved in drinking water treatment.

**Keywords** Anomaly detection · Coagulation process · Data validation · Drinking water treatment · Missing data reconstruction · Self-organizing maps

B. Lamrini (✉)
IRCAM/Centre Georges-Pompidou,
1, Place Igor-Stravinsky, Paris, France
e-mail: Bouchra.Lamrini@ircam.fr; blamrini@yahoo.fr

B. Lamrini · El.-K. Lakhal
Faculté des Sciences Semlalia, Laboratoire d'Automatique, de l'Environnement et des Procédés de Transfert, Université Cadi Ayyad, P.O. Box: 2390', 40000 Marrakech, Morocco

M-V. Le Lann
Laboratoire d'Analyse et d'Architecture des Systèmes, LAAS-CNRS, 7, Avenue du Colonel Roche, 31077 Toulouse Cedex 4, France

L. Wehenkel
Systems and Modelling Research Unit, Institute Montefiore (B28, P32), University of Liege, Grande Traverse 10, Sart-Tilman, 4000 Liege, Belgium

## 1 Introduction

To improve drinking water quality while reducing operating costs, many drinking water utilities are investing in advanced process control and automation technologies. The use of artificial intelligence technologies, specifically artificial neural networks [7, 9, 15], is increasing in the drinking water treatment industry as they allow for the development of control tools capable to meet the requirements of these production units in order to obtain an optimal treatment and guarantee a good quality of supply. Given the strong evolution of the raw water characteristics, an important property for such system is indeed the robustness with regard to the sensors failings or to the unexpected raw water characteristics, owing to accidental pollution for example. Coagulation process is one of the critical processes performed in the drinking water treatment, involving many biological, physical, and chemical phenomena [17]. The control of a good coagulation is essential for maintenance of satisfactory treated water quality and economic plant operation. Thus, an over-dosage can lead both to an increase in the operating costs and to public health concerns. While an under-dosage can cause failure to meet the water quality targets, as the coagulation has a strong impact on the clarification step. The main objective of this work is to validate

Published online: 05 February 2011

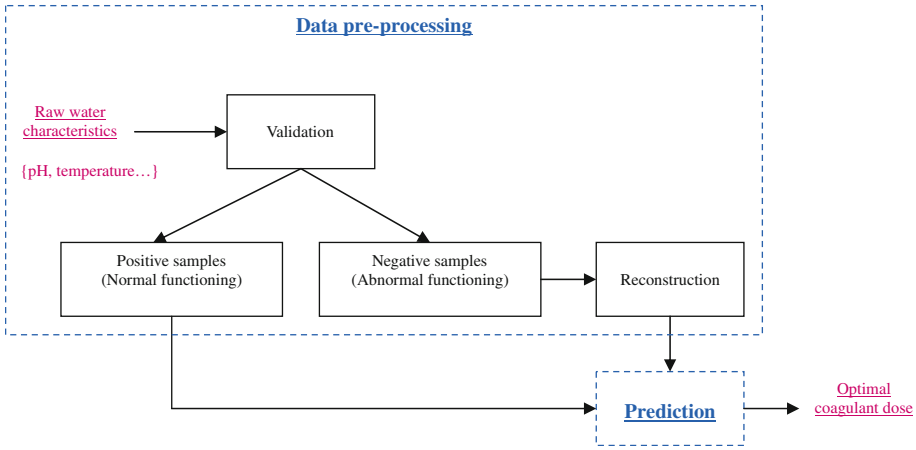


**Fig. 1** Structure of system for automatic coagulation control

and rebuild the measurements of characteristics raw water so as to provide reliable inputs to the automatic coagulation control system (Fig. 1).

In many anomaly detection applications, abnormal (negative) samples are not available at the training stage. For instance, in a computer security application, it is difficult, to have information about all possible attacks. In the machine-learning approaches, the lack of samples from the abnormal class causes difficulty in the application of supervised techniques. Therefore, the obvious machine-learning solution is to use an unsupervised algorithm. For this, we adopted an unsupervised learning approach based on the self-organizing map algorithm introduced by Kohonen [12]. The self-organizing map is one of the most popular artificial neural network models in the unsupervised learning category. It has been successfully applied in various engineering applications [13] covering, for instance, areas like data classification [26], process monitoring and control [2, 10], and fault diagnosis [21]. It has also proven to be a valuable tool in process control of water treatment [23–25].

This paper will first describe the application site chosen for data validation and missing data reconstruction. The integrated research approaches we are undertaking to pre-processing raw water characteristics are given in Sect. 3. Finally, experimental results are presented and discussed in Sect. 4.

## 2 Overview of study area

The drinking water treatment plant concerned in this study is the drinking water treatment Rocade plant located at Marrakech, Morocco. It provides water to more than 1.5 million inhabitants. The raw water is extracted from the Rocade channel. In case of resource failure (raw, pollution, etc.), the treatment plant takes raw water from the pumping plant Takerkoust. Sixty percent of the city needs are ensured by this plant, the complement is brought by underground resources (well, drilling, etc.). It has a nominal capacity to treat 1,400 l/s of water. The treated water is stored in two tanks and transported through the water supply network. The drinking water treatment plant involves physical and chemical processes. Figure 2 presents a schematic overview of the various operations needed to treat the raw water in the

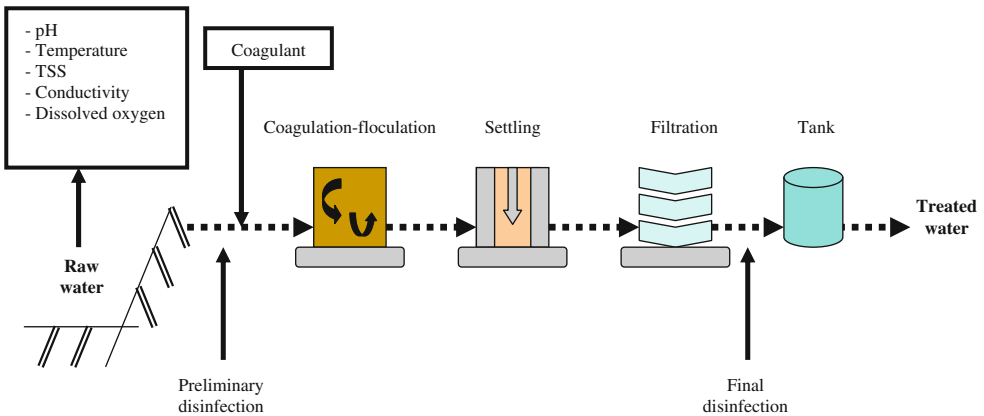


**Fig. 2** Simplified synopsis of the Rocade water treatment plant

Rocade plant. The treatment consists essentially of preliminary disinfection, coagulation–flocculation, settling, filtration, and final disinfection.

Preliminary disinfection (chlorination) is usually a necessary pre-treatment step that destroys disease-causing bacteria, parasites and other organisms generating tastes and undesirable odors. The second stage is coagulation that involves the addition of a chemical coagulant, typically aluminum sulfate, used for destabilization (charge neutralization). A bulky precipitate is formed, which electrochemically attracts solids and colloidal particles. The solid precipitate is removed by allowing it to settle at the bottom of the tank and then periodically removing it as sludge. Then, the flocculation combines small particles into larger ones that settle out of the water as sediment. Synthetic organic polymers are generally used to promote coagulation settling or sedimentation occurring naturally as flocculated particles settled out of the water. The next stage is the filtration process, where the particles passing through the previous stages are removed. The filtered water is also treated by a final disinfection to eliminate the last micropollutants. The water is then stored in a tank and ready to be transported through the water supply network.

## 3 Anomaly detection approaches

The anomaly detection problem can be stated generally as a two-class classification problem: given an element of the space, classify it as normal or abnormal. Different terminologies can be used, such as novelty or surprise detection [5], fault detection [6], and outlier detection [20]. Accordingly, many approaches have been proposed, which include statistical [22], machine learning, and immunological inspired techniques [8].

In a preliminary survey [16], an identification of different functional states (normal or abnormal) describing the behavior coagulation process has been carried out. The identification idea is the evaluation of the significant system measurements (pH, temperature, total suspends solid, conductivity, dissolved oxygen), to recognize the normal and abnormal functional states. The identification of functional states is based on the iterative application of LAMDA (learning algorithm for multivariate data analysis) classification technique. The LAMDA methodology allows the aggregation and exploitation all information stemming from the environment process as well as expert knowledge. Raw data associated with normal state are perfectly valid (positive samples), and in the other case, raw data are declared outliers (negatives samples). Although this simple approaches proves to be sufficient in most cases, the detection of inconsistencies in the data involving more than one parameter and as well as their reconstruction requires the use of more sophisticated techniques such as Kohonen maps. Thus, the main purpose of this study is to analyze the SOM performances on the validation and reconstruction of the raw water characteristics, and this is in combination with classification results already obtained through LAMDA methodology.

### 3.1 Fuzzy technique for anomaly detection

The LAMDA (learning algorithm for multivariate data analysis) methodology is a classification technique introduced by Aguilar-Martin et al. [1] and developed by Piera-Carreté et al. [19]. More recent studies [11, 16, 18, 29] have described in detail the methodology as well as the algorithms and functions used. LAMDA is a fuzzy methodology of conceptual clustering and classification. It allows the representation of classes or concepts by means of the logic connection of all marginal information available. The formation and the recognition of classes are based on the attribution of each object to a class according to the heuristic rule of maximal adequacy. An object is then most likely to belong to the class that presents the greater adequacy degree (GAD). It models the total 'indistinguishability' (chaotic homogeneity) or homogeneity inside the description space from which the information is extracted. This is done by means of a special class called the non-informative class (NIC). This class accepts all items with the same adequacy; therefore, it introduces naturally a classification threshold. According to Fig. 3, LAMDA has two fundamental steps: learning and recognition.

#### *3.1.1 Learning*

At the first stage of learning step (self-learning or unsupervised learning), no previous information is given and LAMDA generates clusters or classes. In this case, it allows obtaining different classifications with the same data set, by changing LAMDA parameters. Using this strategy on a known data set, the expert proceeds to a knowledge-based interpretation of such classes. He modifies the LAMDA parameters in order to improve the quality of the final classification. The classes and updated learning parameters are the output of this initial learning stage. On the second stage (supervised learning), this learning allows performing a different number of choices, like learning from an initial set of classes, which can be modified by adding new classes or by updating their parameters or both.

#### *3.1.2 Recognition*

It has two alternatives, either the user allows unclassified individuals, meaning that an individual has not been recognized in any class (its adequacy degree is lower than the

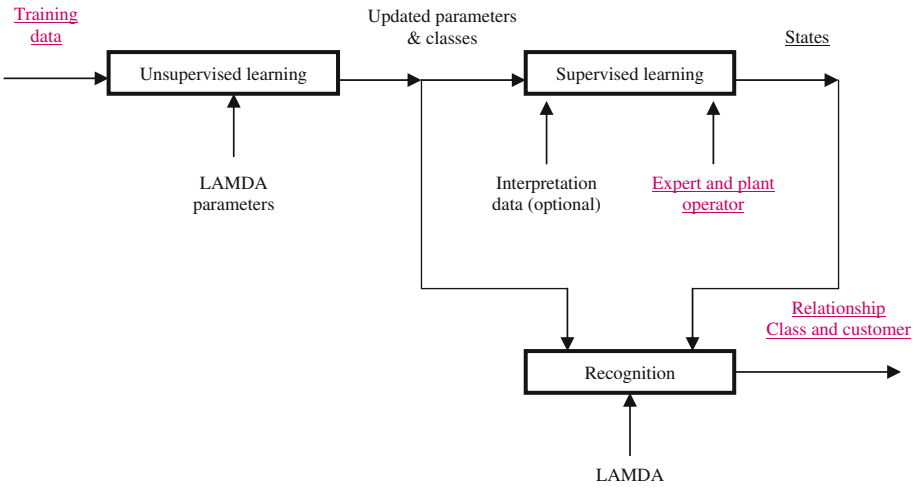


**Fig. 3** Detailed functionality of LAMDA classification tool

minimum threshold) and has been placed in the NIC class, or force every individual to be assigned to a class, in this last case the non informative class is not taken into account for recognition.

The MAD concept is a term related to how similar is one object descriptor to the same descriptor of a given class, and GAD is defined as a membership degree of one object to a given class. Classification process is performed according to a similarity criteria computed in two stages (Fig. 4). First MAD to each existing class is computed for each object descriptor. Second, these partial results will be aggregated in order to get a GAD of an individual to a class. Given that MAD depends on the nature of each descriptor, the algorithm uses general possibility functions. For quantitative descriptors, there are several options introduced in [29] to compute the MAD. One possibility function applied is a fuzzy extension of the binomial probability function, which gives as result the following expression:

$$\mathrm{MAD}\left[x_i\middle|\rho_{j,i}\right] = \rho_{j,i}^{1-d_{j,i}} - \left(1-\rho_{j,i}\right)^{d_{j,i}} \quad \text{with} \quad d_{j,i} = \left|x_i - c_{j,i}\right| \tag{1}$$

where $\rho_{j,i}$ is the possibility of the observed element to belong to a class $C_j$; $x$ is the normalized value of the quantitative descriptor for a particular element; and $c_{j,i}$ is center of $C_j$.

GAD computation is performed as an interpolation between T-Norm and T-Conorm by means of the $\alpha$ parameter. $\alpha = 1$ represents the intersection and $\alpha = 0$ means the union. In [30], some connectors used for GAD computation are presented.

$$\mathrm{GAD}_\alpha(\mathrm{MAD}_1,\ldots,\mathrm{MAD}_n) = \alpha T(\mathrm{MAD}_1,\ldots,\mathrm{MAD}_n) + (1-\alpha)S(\mathrm{MAD}_1,\ldots,\mathrm{MAD}_n). \tag{2}$$

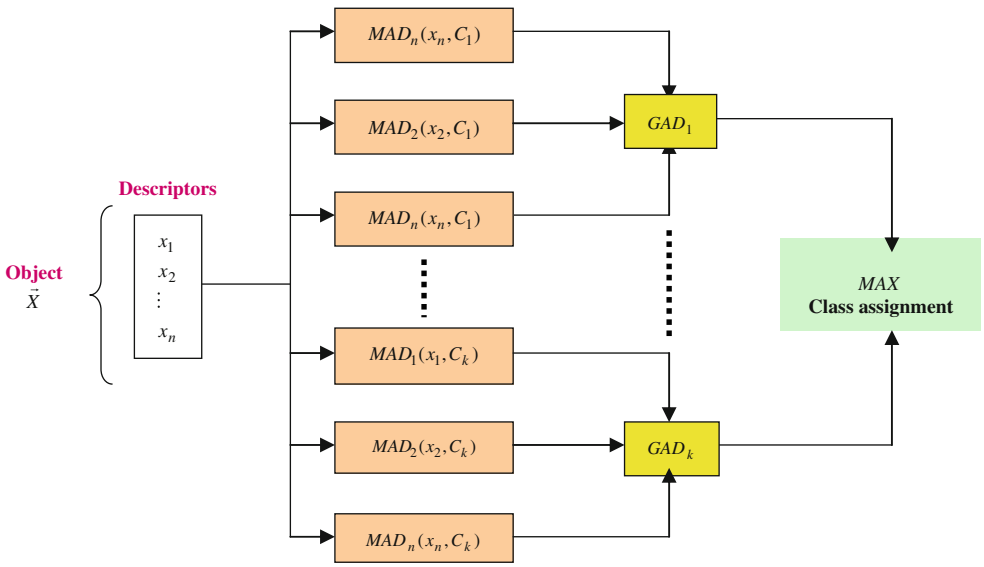


**Fig. 4** Basic LAMDA recognition methodology

### 3.2 Data validation and data reconstruction using self-organizing maps

#### 3.2.1 General considerations in SOM

The self-organizing feature maps draw some inspiration from the way we believe the human brain works. Research has shown that the cerebral cortex of the human brain is divided into functional subdivisions and that the neuron activity decreases as the distance to the region of initial activation increases [12]. There are several public domain implementations of SOM, of which we would like to highlight the SOM_PAK and Matlab SOM Toolbox, both developed by Kohonen's research group.

The Kohonen's SOM is trained using unsupervised learning to produce low-dimensional representation of the training samples while preserving the topological properties of the input space. It performs a topology preserving mapping from high-dimensional space onto map units so that relative distances between data points are preserved. The map units, or neurons, form usually a two-dimensional regular lattice. The SOM can thus serve as a clustering tool of high-dimensional data. It also has capability to generalize, i.e. the network can interpolate between previously encountered inputs. Each neuron $i$ of the SOM is represented by an $N$-dimensional weight $m_i = [m_{i1}, m_{i2}, \ldots m_{iN}]$, where $n$ is the dimensional of the input vectors. The weight vectors of the SOM form a codebook also called prototype vectors or referent vectors. The neurons of the map are connected to adjacent neurons by a neighborhood relation, which dictates the topology of the map. Usually rectangular or hexagonal topology is used. Immediate neighbors (adjacent neurons) belong to neighborhood $N_i$ of the neuron $i$. In the basic SOM algorithm, the topological relations and the number of neurons are fixed from the beginning. The number of neurons determines the granularity of the mapping, witch affects accuracy and generalization capability of the SOM. In the training phase, a given training pattern $x$ is presented to the network, and the closest unit is selected. This unit is called best matching unit (BMU), denoted here by $b$:

$$\|x - m_b\| = \min_i \{\|x - m_i\|\} \tag{3}$$

where $\|x - m_i\|$ is a distance measure, typically Euclidean.

After finding the BMU, the weight vectors of the SOM are updated. The BMU and its topological neighbors are moved closer to the input vector in the input space. The update rule [12, 27] for the weight vector of unit $i$ is:

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{bi}(t) \cdot [(x - m_i(t))] \tag{4}$$

where $\alpha(t)$ is the learning rate and $h_{bi}(t)$ is the neighborhood function how much unit $i$ is updated when unit $b$ is the winner. Both parameters decrease with time in the learning phase.

The SOM algorithm can be easily described as shown below: the first step is to define the network size, the initial learning rate and neighborhood radius. There are no theoretical results indicating the optimal values for these initial parameters. This way the user's experience plays a major role in the definition of these parameters and can be of paramount importance in the outcome of the method. The second step is the initialization of the unit's weights. These may be randomly generated, providing they have the same dimensionality as the training patterns. The next step is to initialize the training phase of the algorithm. For a number of iterations defined by the user, each pattern from the data set is selected and presented to the network. Based on Euclidean distance, the nearest unit BMU is found. The update phase consists on the update of the unit weights and depends on the distance of each unit to the BMU and to the training pattern, and on the neighborhood function and learning rate. In order for the SOM to converge to a stable solution, both the learning rate and neighborhood radius should converge to zero. Usually, these parameters decrease in a linear fashion but other functions can be used. Additionally, the update of both parameters can be done after each individual data pattern is presented to the network (iteration) or after all the data patterns have been presented (epoch). The former case is known as sequential training, and the latter is usually known as batch training.

The sequential training is usually performed in two phases. In the first phase, relatively large initial learning rate and neighborhood radius are used. In the second phase, both learning rate and neighborhood radius are small right from the beginning. This procedure corresponds to first tuning the SOM approximately to the same pace as the input data and then fine-tuning the map. After finding the BMU, the weight vectors of the SOM are updated according to equation (Eq. 4) so that the BMU is moved closer to the input vector in the input space.

The difference in batch training when compared training relies on the unit's updating process and on the non-obligation to randomly present the training patters to the network and sometimes the learning rate also be omitted. In each epoch, the input space is divided according to the distance between the map units. The division of the input space is made using Voronoi regions. These regions are polygons that include all points that are closer to a unit than to any other. The new units' weights are in this case calculated [28] according to:

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{bi}(t) \cdot x_j}{\sum_{j=1}^{n} h_{bi}(t)} \tag{5}$$

where $b$ is the BMU for the training pattern $x_i$ and $h_{bi}(t)$ is a neighborhood kernel centered on the winner unit. The new weight vectors are a weighted average of the training

patterns where the weight of each data pattern is the neighborhood function value $h_{bi}(t)$ to its BMU.

The quantization error and the topographic error are one of the several ways to evaluate the quality of a SOM after the training phase. If two prototype vectors close to each other in the input space are mapped wide apart on the grid, this is signaled by the situation where two closest best matching units of an input vector are not adjacent units. This kind of folds is considered as an indication of the topographic error in the mapping. The topographic error can be calculated as the proportion of sample vectors for which two best matching units are not adjacent (Eq. 6).

$$e_t = \frac{1}{n}\sum_{i=1}^{n} u(x_i) \tag{6}$$

where $n$ is the number of samples, $x_i$ is the $i$th sample of the data set and $u(x_i) = 1$ if the first and second best matching units of are not adjacent units, otherwise zero.

Moreover, the prototype vectors try to approximate to the data set. A consequence of this approach is the resolution error or the quantization error. To measure the resolution of the mapping, the average quantization error (Eq. 7) over the whole testing data set is usually used.

$$e_q = \frac{1}{n}\sum_{i=1}^{n} \|x_i - m_b\| \tag{7}$$

The number of map units determines the accuracy and generalization capability of the SOM. The bigger the map size the lower the quantization error, but the higher the topographic error. This is due to the neural network folds to reduce the quantization error. Moreover, the bigger the map size the higher the computational cost. Therefore, there is compromise between the increase in the topographic error and the reduction in the quantization error. A reasonable optimum solution of the compromise among the quantization error and the topographic error to determine the side lengths of the map is the heuristic formula (Eq. 8).

$$N = 5\sqrt{n} \tag{8}$$

$N$ is the number of map units and $n$ is the number of the training data samples.

### 3.2.2 Application to data validation

The invalid data have always been considered like a source of information distortion gotten from raw data. It is therefore necessary to highlight the diversity of available methods to interpret or to characterize these abnormal values, either while rejecting them in order to restore the data initial properties or while adopting methods that decrease their impact during the statistical analysis [3, 20].

Neural approaches' application to invalid data and reconstruction include generally the auto-associative neural networks (AANN) and Kohonen's SOM [25]. The AANN approach is to train a multilayer feedforward network to approximate the identity function by using target values identical to the input values. The hidden layer allows typically limiting the capacity and forces optimally the network to encode input vectors, to therefore give an

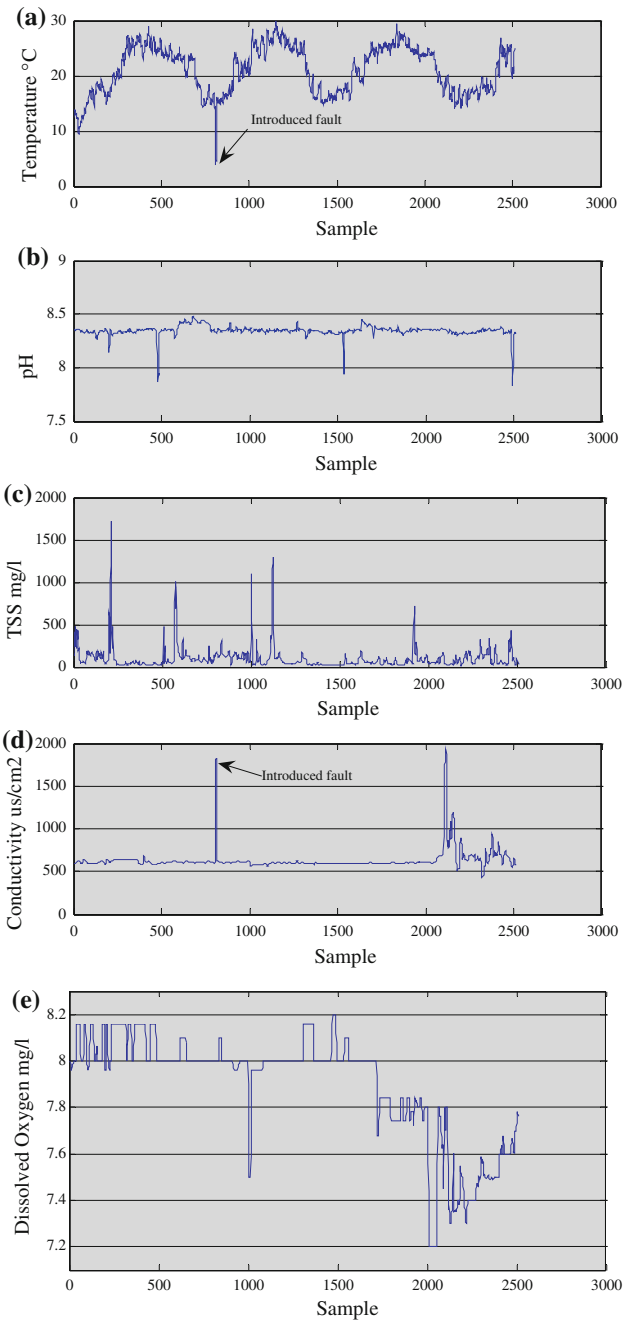


**Fig. 5** Raw water characteristics used for SOM modeling and LAMDA classification. **a** Temperature measurement with artificial fault. **b** pH measurement. **c** TSS measurement. **d** Conductivity measurement with artificial fault. **e** Dissolved oxygen measurement

information compression and dimensionality reduction. With a single hidden layer of linear units, this approach proved to be equivalent to the principal component analysis [14]. Consequently, more complex networks with nonlinearities can be seen like implementing some form of "nonlinear PCA". In Ref. [3], a multilayer perceptron with five layers has proposed for data validation and reconstruction. This network can be considered like two networks with three layers connected in series. The first network combines the redundant variables in a smaller number of variables supposed to represent the essential features of process. The second network uses the information compressed to rebuild the initial redundant measurements of the input space. This network can be used to detect invalid data, which are identified by their higher reconstruction error. However, the efficiency of such a system in the presence of incomplete input data is not fully predictable.

In this study, Kohonen's SOM is used for failure data detection and reconstruction. The SOM model combines the goals of projection and clustering algorithms and may be seen as a method for automatically arranging high-dimensional data. In our case, self-organizing maps allow not only to visualize the evolution of raw water characteristics in two dimensions, but also to detect atypical data by computing the distance between each input vector and its closest reference vector. The basic idea of data validation approach consists in determination of a confidence degree in every data sample, based on monitoring this distance. The validity of a characteristic measurement, for instance, may be put for different reasons: (1) the value is abnormally high or low; (2) the variation between two successive measurements is too important; (3) and the value is incompatible with other measurements of the same quantity obtained by an independent device, etc.

Given a $N$ prototype vectors $\{m_i, \ldots m_N\}$. Every prototype $m_k$ represents a $C_k$ class. The reference space is divided thus into $N$ classes $N(C_k)_{k=1}^{N}$. To determine the

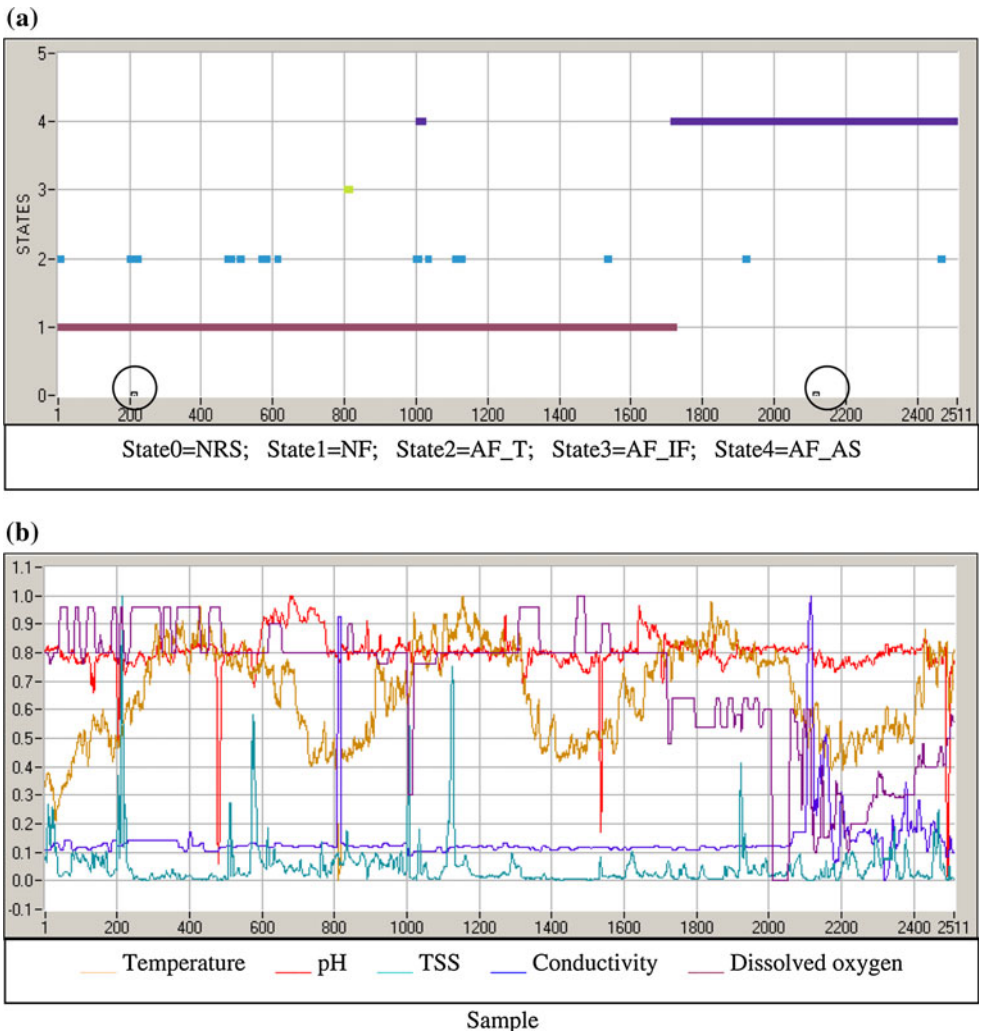


**Fig. 6** Anomaly detection results by LAMDA methodology. **a** States associated. **b** Descriptors normalized

confidence degree involves defining the activation of unit $i$ for input $x$ using a Gaussian kernel as:

$$h_i(x) = \exp\left(\frac{-1}{2\sigma_i^2}\|x - m_i\|^2\right) \quad (9)$$

where $\sigma_i^2$ is a parameter defining the size of the influence region of unit $i$. $\sigma_i^2$ may be computed as the average empirical variance of the $n$ input features, among the samples associated with unit $i$. More $\sigma_i^2$ is bigger; more the influence region of $m_i$ is bigger, and therefore more $h_i(x)$ is closer to 1.

If the activation $h_b(x)$ of the winning prototype is smaller than a specified threshold, the current sample is considered as abnormal. The contributions of each of the components of vector $x$ to the distance $\|x - m_b\|$ are then examined to determine more precisely which data should be declared as abnormal.

### 3.2.3 Application to data reconstruction

If vector prototypes provide a good data representation, each missing value of a given input variable can be estimated by the value of the corresponding component of the winning prototype.

Given $x$ a new vector, composed of two parts $x_o$ and $x_m$, containing, respectively, observed and missing values. The main thing is to rebuild $x_m$ from the information provided by Kohonen's card. The method proposed rest on similarity between this new vector $x = (x_o, x_m)$ and the reference vectors $m_k$. Given $X_o$ and $X_m$ the under-spaces, respectively, of $x_o$ and $x_m$ variables. $m_o$ and $m_m$ are the projections of these under-spaces. According to the activation defined by the Eq. 9, more $x_o$ is closer to $m_o$, more we will have chance that $x_m$ is closer.

$$h_i(x^o) = \exp\left(\frac{-1}{2\sigma_i^2}\left\|x^o - m_i^o\right\|\right) \quad (10)$$

The approaches of missing data estimation call for the various techniques, generally presupposing a probabilistic context. For instance, the heuristic methods (such as average and median replacing techniques) are often used and also constitute some simple and little expensive solutions. The parametric methods of maximization, as the EM (expectation maximization) algorithm [4], are extensively used and proved their efficiency, but they require the knowledge or the estimation laws of the variables probabilities. In our case, we can use a simple method that estimates missing data by the component value corresponding to winning prototype $m_b$:

$$\forall p \in M(x), \quad \hat{x}_p = m_{bp} \quad (11)$$

where $M(x)$ is the indexes set of missing values.

This method is very sensitive to the prototype change between two successive vectors $x$. To resolve this problem, we considered another method that takes in account the influence of the $k$ nearest prototypes. Each missing or

**Table 1** States associated with classes detected by LAMDA methodology

| Class name (11 classes) | State associated (8 states) | State name | Functioning type |
|---|---|---|---|
| LowSaison | Normal | NF (S1) | NF |
| HighSaison | Normal | NF (S1) | |
| Descriptor_Normal | Normal | NF (S1) | |
| Descriptor_Low | Descriptor_Alarm | AF_AS (S4) | AF |
| Descriptor_Very_Low | Descriptor_Alarm | AF_AS (S4) | |
| Descriptor_Slightly_Low | Descriptor_Alarm | AF_AS (S4) | |
| Descriptor_Elevated | Descriptor_(Slow → Stop) | AF_T (S2) | |
| Descriptor_Very_Elevated | Descriptor_(Slow → Stop) | AF_T (S2) | |
| Descriptor_Very_Elevated and artificial Fault | Descriptor_(Slow → Stop) | AF_T (S2) | |
| Artificial Fault | Artificial Fault | AF_IF (S3) | |
| NIC | Not-Recognized State | NRS (S0) | NRS |

**Table 2** SOM training parameters

| Map lattice | Map size | Neighborhood function | Neighborhood radius | Initial learning rate | Learning rate function (inv) | Epochs |
|---|---|---|---|---|---|---|
| Hexagonal | 25 × 10 | Gaussian | $\sigma_{\text{fin}} = 1$ | $\alpha_0 = 0.95$ | $\alpha(i) = \frac{\alpha_0}{(1+(100 \cdot i/T))}$ | $125 \times 10^2$ |

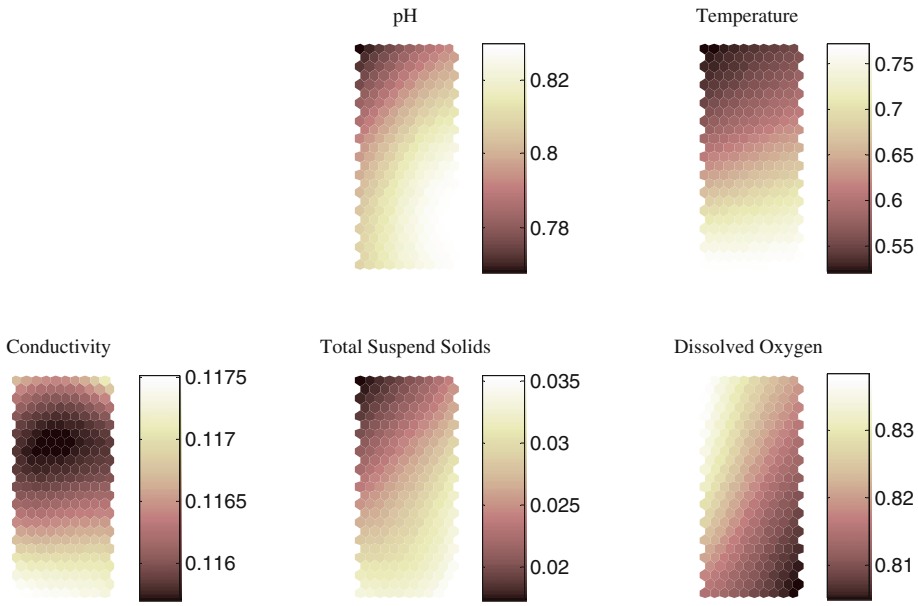


**Fig. 7** Component planes of the SOM for 5 descriptors

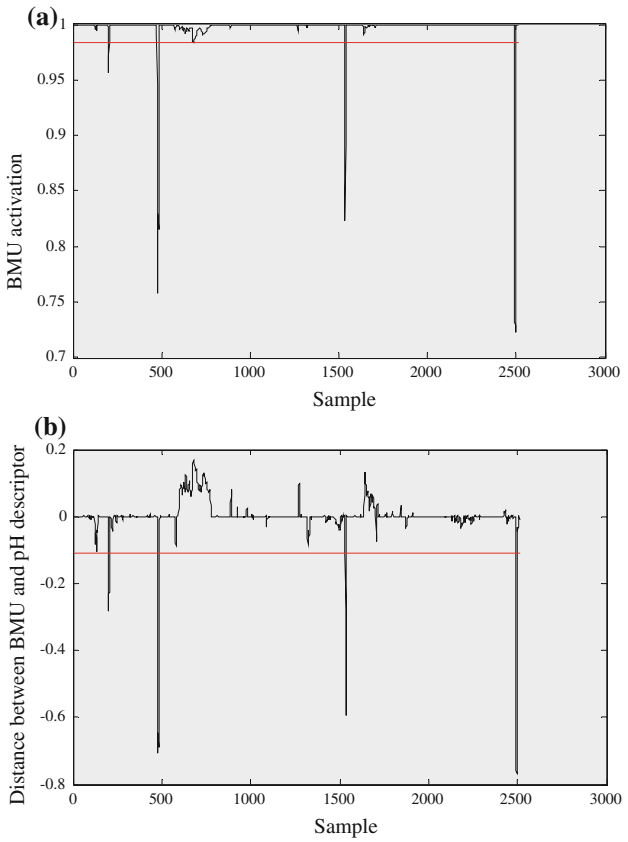


**Fig. 8** **a** Activation of the winning prototype. **b** Computed distance between winning prototype and pH descriptor

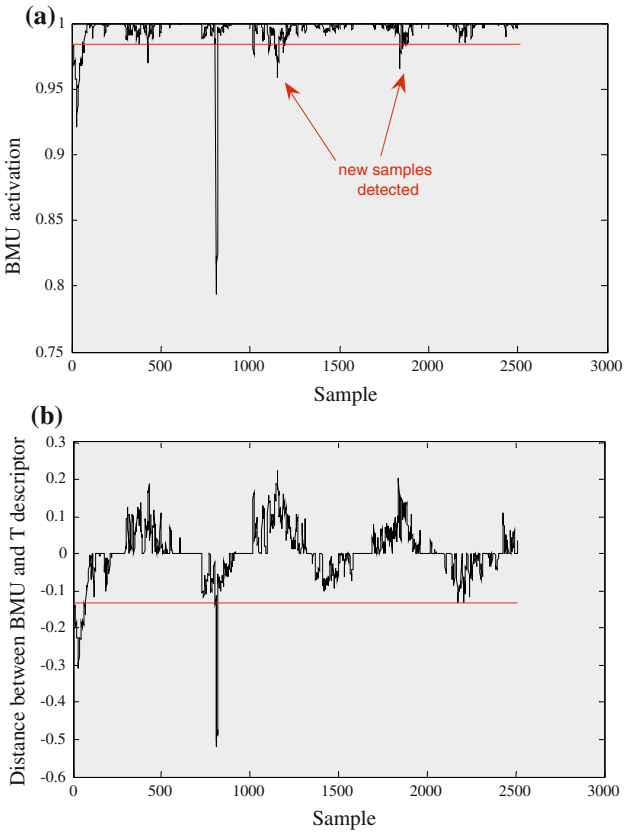


**Fig. 9** **a** Activation of the winning prototype. **b** Computed distance between winning prototype and temperature descriptor

invalid value $j$ is estimated by a combination of the corresponding component in the $k$ nearest prototypes:

$$\hat{x}(j) = \frac{\sum_{i=1}^{k} h(i) m_i(j)}{\sum_{i=1}^{k} h(i)} \quad (12)$$

where $m_i(j)$ denotes component $j$ of prototype $i$.

## 4 Results and discussion

### 4.1 Database description

The experimental data for 4 years (2,511 samples) are used to identify the functional states (normal, abnormal, transition) by LAMDA methodology and at once to validate the detected failures before the reconstruction stage. We used 5 descriptors of raw water quality stemming from Rocade plant such as temperature (T), pH, TSS (total suspend solids), dissolved oxygen (DO), and conductivity (COND). Note that this data set covers a period of 4 years and so can be expected to account for seasonal variations of water quality. The temperature, pH and TSS parameters are strongly dependent on the seasonal phenomena (Fig. 5). According to knowledge of operator plant and our interpretations, this data set contains 963 negative samples:

- Thirty-seven negative samples associated with very low variations of pH.
- Seventy-three high measurements of TSS: the Rocade plant is in alarm state. It is normally in a slowing state, and it can change from this state to a stop state.
- One hundred and seven negative samples assigned to some very elevated measures of conductivity (caused by the presence of chlorides) with a very low variation of dissolved oxygen.
- Eight hundred and ten negative samples representing low variations of dissolved oxygen. In order to assess the robustness of the validation approach, 9 faults are introduced simultaneously in original samples (808…, 816) of temperature and conductivity descriptors. Seventy-two low variations of temperature have been also considered as the negative samples.

### 4.2 Anomaly detection using LAMDA technique

In this stage, the algorithm carefully chosen to compute the marginal adequacy degrees is $\mathrm{MAD}\left[x_i \middle| \rho_{j,i}\right] = \rho_{j,i}^{1-d_{j,i}}$

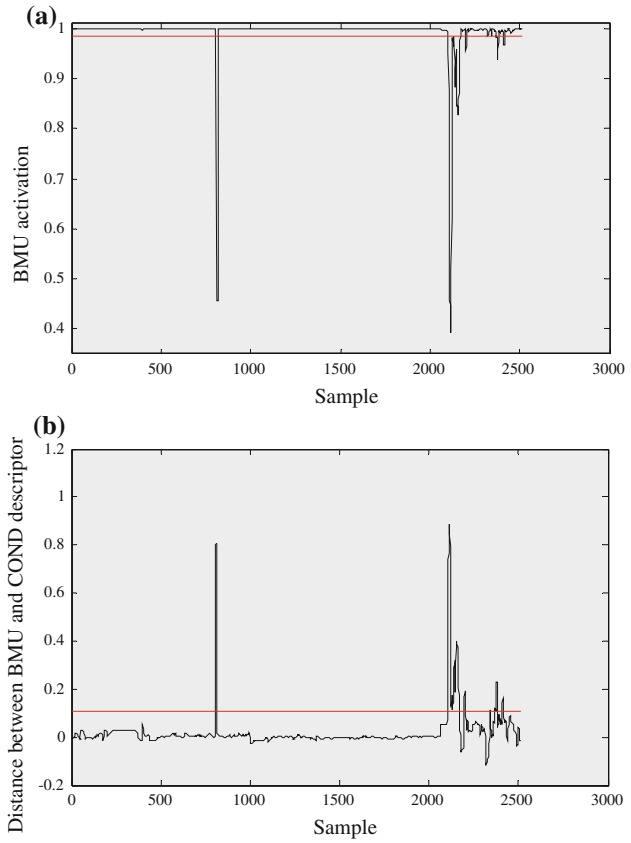


**Fig. 10** **a** Activation of the winning prototype. **b** Computed distance between winning prototype and conductivity descriptor

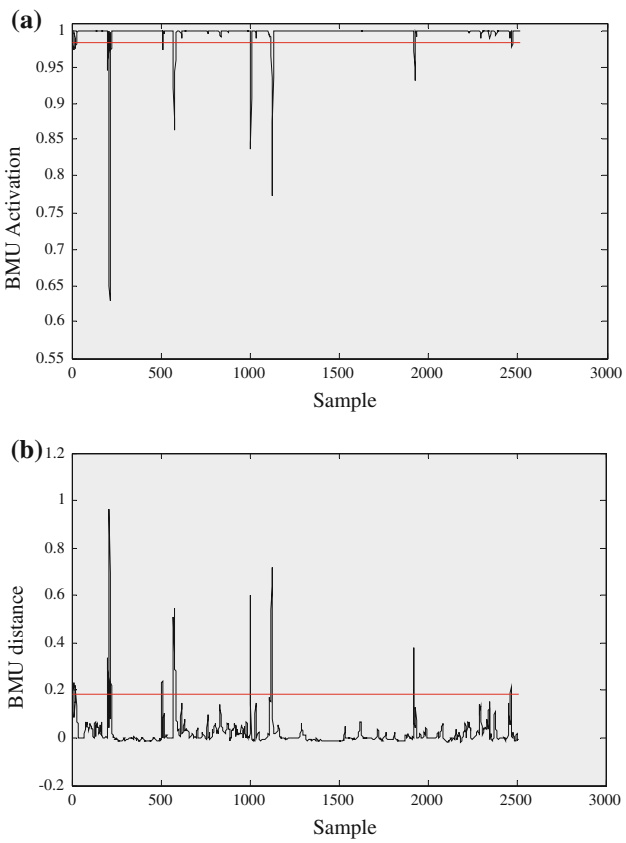


**Fig. 11** **a** Activation of the winning prototype. **b** Computed distance between winning prototype and TSS descriptor

$-\left(1-\rho_{j,i}\right)^{d_{j,i}}$. The minimum–maximum was selected as the connective family. To calculate the global adequacy degrees, we adopted an exigency level equal to $\alpha = 0.85$. Figure 6 and Table 1 show the different states obtained by the unsupervised learning on the coagulation process. These significant states are characterized from the classification information (class profile, membership matrix, etc.) [11]. While exploiting the information stemming from profile classes, e.g., normalized parameters of every class, we can note that some classes present sometimes a similar characteristics and the expert can decide to regroup these classes in a single state. Eleven classes have been identified, and according to their profile, eight functional states have been detected. This information allows us to identify significant classes and those that can be regrouped in a single state. To sum up, it was possible to identify tree types of functional states:

- *Normal Functioning "NF"*. The plant operates in the normal conditions, e.g., the describers operate with the optimal values in the high and low season (the plant operates normally in most of the time). A total of 1,610 samples have been associated with this state.
- *Abnormal Functioning "AF"*. This abnormal state includes the following: (1) the negative samples of degraded operation identified beside normal operation. The plant is in *Alarm State "AF_AS"*; (2) artificial faults that we introduced *"AF_IF"*; (3) and other negative samples that can be denoted as *Transition "AF_T"* (the descriptors may return after one time more or less long to the normal state). Eight hundred and ninety-eight samples have been associated with the functional state *"AF"*. Among 898 samples, 72 low variations of temperature have been already identified as the normal samples.
- *Not-Recognized State "NRS"*. We also see that 3 negative samples (213, 2,118, and 2,119) are not recognized by LAMDA technique. The tree samples have been placed in the NIC class.

### 4.3 Data validation and data reconstruction using SOM approach

For the SOM simulation results, we used the SOM toolbox version 2.0 beta developed at the Helsinki University of

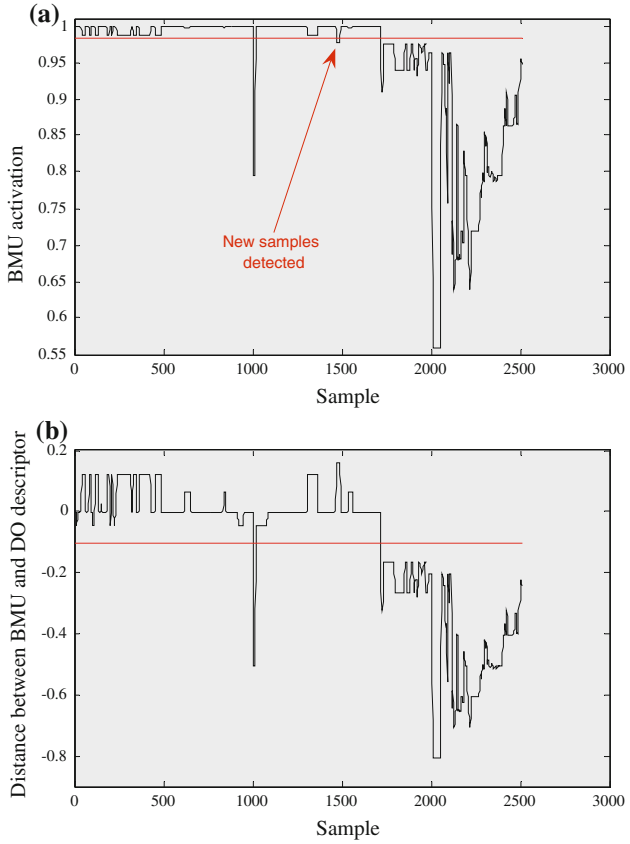


**Fig. 12** **a** Activation of the winning prototype. **b** Computed distance between winning prototype and dissolved oxygen descriptor

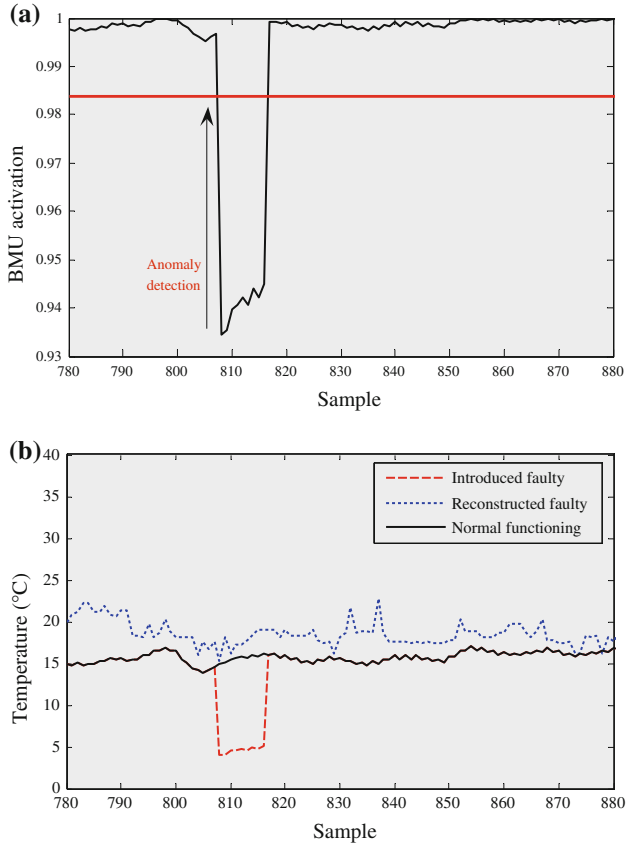


**Fig. 13** **a** Activation of the winning prototype. **b** Reconstruction of temperature faulty

Technology [28]. All descriptors are normalized in the same ranges (as by LAMDA). SOM network was trained using sequential training algorithm. Table 2 presents the training parameters adopted during this phase. The quantization error and topographic error computed are, respectively, equal to 0.26 and 0.19.

Figure 7 shows the component planes of a Kohonen map of size 10 × 25 trained on the whole data set. Each component plane shows the value of each neuron to estimate the data variable of the input space. The value is indicated with color, and the color bar on the right shows what the colors mean. The highest values correspond to dark regions and the lowest ones to light zones. It is useful to determine the several zones where the variable value is high or low and to observe any correlation or relationship among the process variables. These correlations can be detected by means of the color gradient on each component plane. Two variables with parallel gradients show a direct correlation. On the other hand, anti-parallel gradients show an inverse correlation. For instance, the temperature, pH and TSS descriptors have parallel gradients in their component planes, and therefore these tree variables are positively correlated. The dissolved oxygen is inversely (negatively) correlated to these descriptors. Such relationships between input variables are captured by the SOM and are exploited for the reconstruction of missing measures.

We have 2,511 vectors of the training set. We considered that 963 measurements (38.35%) of these data are abnormal. Afterward, we calculated the activation of winning prototype for these vectors. We sorted data in ascending order of activation. The threshold has corresponded therefore to the BMU activation of ($38.35 \times N/100$, $N = 2{,}511$) rank. We consider therefore that 38.35% of data set has a too small activation to be considered like invalid (abnormal). The threshold computed is equal to 0.984. The input vectors whose $h_b(x) < 0.984$ are then declared invalid.

The significance of the components with respect to the clustering is harder to visualize. One indication of importance is that on the borders of the clusters, values of important variables change very rapidly. The mask of the given map is used. It allows determining the quantization errors for such descriptor without the other variables contribution total. In order to visualize which descriptor should be declared precisely as faulty, the contributions of each of the components of vector x to the distance $\|x - m_b\|$ are then examined, too. These abnormal samples are then deleted to compute a new winning prototype with only normal samples. Figures 8, 9, 10, 11, and 12 show the distances computed between winning prototype and different descriptors. The variations that are abnormally very low or very high were correctly identified as being the faulty parameter. Figures 13 and 14 show the reconstruction values relating to temperature and conductivity faulty. The SOM procedure allows for the rejection of atypical samples and therefore implements some kind of "novelty detection" (Figs. 9, 12). However, this type of rejection may originate from unreliable data acquisition sources, faulty sensors, data collection errors, or merely lack of completeness of the training set. This constitutes a very conservative approach that prevents the prediction module of the system from blindly interpolating known relationships between water characteristics and coagulant dosage to previously unseen cases. It is therefore necessary to store the rejected input patterns for subsequent interpretation by the user, and possible retraining of the system in case of undue rejection of "normal" patterns. Table 3 summarizes and illustrates the various faulty data recognized by means of SOM approach. In parallel, we present the identification results recognized previously through LAMDA classification technique. We note that the neuronal approach adopted in this work almost recognized the faulty samples identified with the help of operator plant. Besides, SOM approach allows identifying the other variations that can present an alarm state for Rocade plant. The reliability and robustness of the neuronal approach are

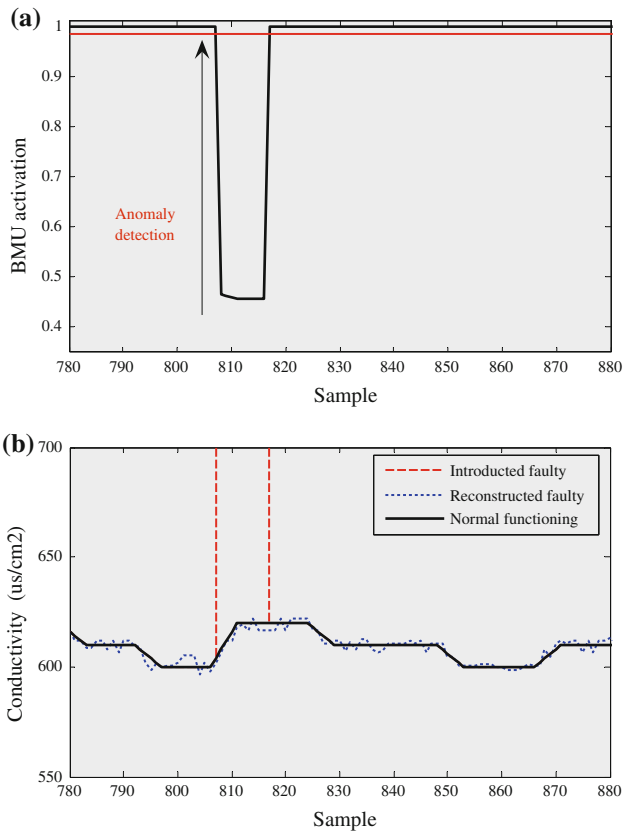


**Fig. 14** **a** Activation of the winning prototype. **b** Reconstruction of conductivity faulty

**Table 3** Results of anomaly detection obtained by means of SOM and LAMDA techniques

| Descriptors | Number of negative samples (invalid values) identified by | | | |
|---|---|---|---|---|
| | Operator | Type of variations | LAMDA | SOM |
| T | 72 | Low | No | Yes + 8 new samples (low) |
| | 9 Artificial faults | Very low | Yes | Yes |
| pH | 37 | Very low | Yes + 2 new samples (very low) | Yes |
| TSS | 73 | High | Yes | Yes + 11 new samples (slightly high) |
| COND | 107 | Very high | 49 among 107 | 49 among 107 |
| | 9 Artificial faults | Very high | Yes | Yes |
| DO | 810 | Low | 808 among 810 | Yes + 3 new samples (high) |

justified by the validation–reconstruction process of the faulty measurements and that LAMDA does not allow it in the present time.

## 5 Conclusion

In this paper, we investigated a self-organizing map approach for anomaly detection and missing data reconstruction. Experimental results using real data stemming from coagulation process involved in a drinking water treatment showed the efficiency and soundness of SOM algorithm. The results that we succeeded by this study in combination with those obtained by the fuzzy technique LAMDA show the key point of the validation–reconstruction process. It was possible to identify almost of the negative samples characterizing abnormal operation plant and in particular rebuild the faulty measurements. This approach is an environmental application that shows the utility of outlier's treatment techniques in the monitoring and the surveillance of this process type. It is clear that the final objective is to spread this neural approach to other treatment processes in order to detect at the earliest a drifts functioning or to identify a failures on an upstream unit. So, it is desirable to test new distance measures and perform additional experiments using wide variety of data sets, stemming from other processes, in order to make a fair comparison. This model will be too integrated to software neural sensor developed in a preliminary survey [15], for automatic coagulation control.

## References


1. Aguilar-Martin J, Balssa M, Lopez De Mantras R (1981) Recursive estimation of partitions: examples of learning and self teaching in RN and IN. Questiio: Quaderns d'Etadisticia, Siestems, Informatica i Investigacio Operativa, ISSN 0210-8054 5(3):150–172
2. Badran F, Thiria S, Main B (1992) Smoothing with topological map. In Proceedings of NeuroNimes92 (neural network & their applications), Nîmes, France, pp 107–115
3. Barnett V, Lewis T (1994) Outliers in statistical data (Wiley series in probability & statistics), 3rd edn. Wiley, New York
4. Biernacki C, Celeux G, Si Abdallah J-F, Govaert G, Langrognet F (2009) MIXMOD user's guide (MIXture MODelling software: high performance model-based cluster and discriminant analysis). Univ. of Franche-Comté, France. http://www-math.univ-fcomte.fr/mixmod/index.php
5. Denning D (1987) An intrusion-detection model. IEEE Trans Softw Eng 13(2):222–232
6. Fuente MJ, Vega P (1999) Neural networks applied to fault detection of a biotechnological process. Eng Appl Artif Intell 12:569–584
7. Gagnon C, Grandjean BPA, Thibault J (1997) Modelling of coagulant dosage in a water treatment plant. Artif Intell Eng 11:401–404
8. Gonzalez F, Dasgupta D (2002) Neuro-immune and self-organizing map approaches to anomaly detection: a comparison. In: Proceedings of the 1st international conference on artificial immune systems, Canterbury, UK, pp 203–211
9. Hernandez H, Le Lann M-V (2006) Development of a neural sensor for on-line prediction of coagulant dosage in a potable water treatment plant in the way of its diagnosis. In: Sichman JS et al (eds) IBERAMIA-SBIA 2006, LNAI 4140, pp 249–257
10. Kasslin M, Kangas J, Simula O (1992). Process state monitoring using self organizing maps. In: Aleksander I, Taylor J (eds) Artificial neural networks II, vol 2, Amsterdam, Netherlands, North-Holland, pp 1531–1534
11. Kempowsky T (2004). Surveillance des procédés à base de méthodes de classification: conception d'un outil d'aide pour la détection et le diagnostic des défaillances. PhD Thesis, LAAS-CNRS, Institut National des Sciences Appliquées (INSA), Toulouse, France
12. Kohonen T (1995) Self-organizing maps. Volume 30 of Springer series in information sciences. Springer, Berlin
13. Kohonen T, Oja E, Simula O, Visa A, Kangas J (1996) Engineering applications of the self-organizing map. Proc IEEE 84(10):1358–1384
14. Kramer MA (1991) Nonlinear principal component analysis using autoassociative neural networks. AIChE J 37(2):233–243
15. Lamrini B, Benhammou A, Le Lann M-V, Karama A (2005) A neural software sensor for on-line prediction of coagulant dosage: application to a drinking water treatment plant. Trans Inst Meas Control 27(3):95–213
16. Lamrini B, Benhammou A, Le Lann M-V, Lakhal El-K (2005) Detection of functional states by "LAMDA" classification technique: application to a coagulation process in drinking water treatment. Comptes Rendus Physique 6:1161–1168
17. Masschelein WJ (1997) Processus unitaires du traitement de l'eau potable. Tec & Doc Lavoisier (Ed), Paris
18. Orantes A, Kempowsky T, Le Lann MV (2006) Classification as an aid tool for the selection of sensors used for fault detection and isolation. Trans Inst Meas Control 28(5):457–480

19. Piera-Carreté N, Desroches P, Aguilar-Martin J (1989). LAMDA: an incremental conceptual clustering system. Technical report No. 89420, LAAS-CNRS, Toulouse, France
20. Planchon V (2005) Traitement des valeurs aberrantes: concepts actuels et tendances générales. Biotechnol Agron Soc Environ 9(1):19–34
21. Simula O, Alhoniemi E, Hollmen J, Vesanto J (1996). Monitoring and modelling of complex processes using hierarchical self-organizing maps. In: Proceeding of the IEEE international symposium on circuits and systems (ISCAS'96), vol supplement, pp 73–76
22. Stanimirova I, Daszykowski M, Walczak B (2007) Dealing with missing values and outliers in principal component analysis. Talanta 72:172–178
23. Trautmann T (1995) Développement d'un modèle de cartes topologiques auto-organisatrices à architecture dynamique: Application au diagnostic. PhD thesis, Univ. of Compiègne, France
24. Trautmann T, Denoeux T (1995) Comparison of dynamic feature map models for environmental monitoring. In: Proceedings of international conference on neural networks (ICNN'95), vol 1, Perth, Australia, pp 73–78
25. Valentin N, Denoeux T, Fotoohi F (1999) An hybrid neural network based system for optimization of coagulant dosing in a water treatment plant. In: Proceedings of international joint conference on neural networks (IJCNN'99), Washington
26. Vercauteren L, Sieben G, Praet M, Otte G, Vingerhoeds L, Boullart L, Lalliauw L, Roeds H (1990) The classification of brains tumours by a topological map. In: Proceedings of international conference on neural networks (ICNN'90), vol 1, Paris, pp 387–391
27. Vesanto J (1999) SOM-based visualisation methods. Intell Data Anal 3:111–126
28. Vesanto J, Alhoniemi E, Himberg J, Kiviluoto K, Parviainen J (1999). Self-organizing map for data mining in MATLAB: the SOM toolbox. Simul News Eur 25–54
29. Waissmann-Vilanova J (2000) Building a behavioural model for process supervision: application to a wastewater treatment plant. PhD thesis, LAAS-CNRS, Institut National Polytechnique (INP), Toulouse, France
30. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28