

# Self-Organizing Map for erroneous data processing in time series analysis

Bassam ABDEL LATIF and Grégoire MERCIER  
GET / ENST Bretagne / dpt. ITI  
Technopole Brest-Iroise,  
CS 83818, F-29238 Brest Cedex 3, France  
Email: Bassam.alatif@enst-bretagne.fr

**Abstract**—In the context of this study, changes in the vegetation coverage are to be detected in the Brittany region in France, based on seasonal and regional scales. Thus, many low resolution images from MODIS satellite are considered per season. Unfortunately, the Brittany region in the west of France is often covered by clouds, and few dates per season with clear coverage are available.

A nonparametric regression procedure, depending on the presence of clear small zones during the season, for predicting the value of data contaminated by weather conditions is presented. The problem is the erroneous data (pixels contaminated by clouds or shadows), which may be seen as a problem of an incomplete data as soon as an outlier detection technique is properly chosen and applied. The idea is to use the Kohonen's Self Organizing Map (SOM) onto clear zones (free of clouds and shadow) for training. Then, incomplete data may be recovered by projecting the data onto this trained SOM.

To detect pixels contaminated by clouds and shadows, the Box and Whisker method has been used with an overall accuracy of 97.7636% and a Kappa coefficient of 0.9652.

## I. INTRODUCTION

In our study, we are trying to detect changes in the vegetation coverage in the Brittany region in France on seasonal and regional bases. To be able to make the study on the regional base, we need some images with low resolution pixels, that will minimize the cost and the time of processing. For the study to be on a seasonal base, we need many images per season. The Brittany region in the north-west of France is often covered by clouds, and there are few dates per season with clear coverage. MODIS satellite has a moderate resolution of 250m for its first two bands, Red and Near-infrared. Moreover, its time resolution is very high, it can capture an image every day. The information source in MODIS bands related to the vegetation cover is its two bands; band 1 (red, 620-670 nm), band 2 (near-infrared (NIR), 841-876 nm). Although the number of bands is limited,

the two bands are in the most important spectral regions for remote sensing of vegetation. To compensate the shortage of bands available, we are trying to extract information from a time series of MODIS images. It requires a set of images that cover all the season, and obviously it has to be clear to carry out the study.

Unfortunately, it is difficult to find coverages that are all clear because of clouds that may appear randomly. So some means of prediction or regression of zones covered by clouds are necessary. In this study we choose a nonparametric regression procedure depending on the presence of some clear zones during the season. The method is adapted from the Kohonen's Self-Organizing Map (SOM) [1]. The idea is to construct the SOM, using the clear zones as training patterns, to characterize the *normal* temporal behavior of the landscape. Then, the rest of the data may be projected onto this SOM to recover masked areas. The SOM has proved a great capability to deal with the incomplete data [2], [3] which seems to be a strong feature in dealing with erroneous data.

## II. THE SOM FOR MISSING VALUES

### A. Principle of the SOM

There exist many versions of the SOM, the basic philosophy, however, is very simple and already effective. SOM defines a mapping from the input data space  $\mathcal{X}^n$  onto a regular one or two-dimensional array of  $M$  nodes (see fig. 1). With every node  $m$ , a reference vector  $C_m \in \mathcal{X}^n$  is associated. An input vector  $x \in \mathcal{X}^n$  is compared with the  $C_m$ , and the best match is defined as "response": the input is thus mapped onto this location.

One might say that the SOM is a "nonlinear projection" of the probability density function of the high-dimensional input data onto the two-dimensional array. Let  $x \in \mathcal{X}^n$  be an input data vector. It may be compared with all the  $C_m$  in any metrics; in practical applications,

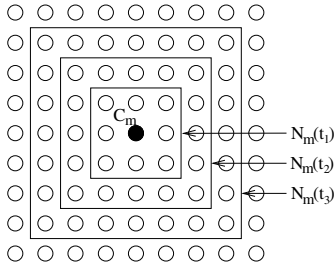


Fig. 1. The Kohonen's Self Organizing Map and topological neighborhood  $N_{m_x}(t_i)$  ( $t_1 < t_2 < t_3$ ) of the winning neuron  $C_{m_x}$ .

the smallest of the Euclidean distance  $\|\mathbf{x} - C_m\|$  is usually used to define the best-matching node, denoted by the subscript  $m_x$ :

$$\|\mathbf{x} - C_{m_x}\| = \min_{m \in \{1, \dots, M\}} \|\mathbf{x} - C_m\|. \quad (1)$$

Thus  $\mathbf{x}$  is mapped onto the node  $m_x$  relatively to the values  $C_m$ . An "optimal" mapping would be the one that maps the probability density function  $p(\mathbf{x})$  in the most "faithful" fashion, preserving at least the local structures of  $p(\mathbf{x})$ . During the learning stage, these nodes which are close to each other on the SOM will learn from the same input. The values of the  $C_m$  can be found as convergence limit of the learning process (2), where the initial values of the  $C_m(0)$  can be fixed arbitrary, e.g., randomly:

$$C_m(t+1) = C_m(t) + h_{m, m_x}(t)[\mathbf{x}(t) - C_m(t)]. \quad (2)$$

In (2),  $t$  is the time parameter (*i.e.* the number of iterations), and  $h_{m, m_x}(t)$  the so-called *neighborhood kernel*. It is a function defined over the lattice points; usually  $h_{m, m_x}(t) = h(d(m, m_x), t)$  where  $d(m, m_x)$  is the distance between the location of  $C_m$  and  $C_{m_x}$  on the SOM. While increasing  $d(m, m_x)$ , or increasing  $t$ ,  $h_{m, m_x}(t)$  decreases monotonically to 0. The average width and the form of  $h_{m, m_x}$ , defines the "stiffness" of the "elastic surface" to be fitted to the data set.

Let their index in the neighborhood of  $m_x$  be denoted by the set  $N_{m_x}(t)$ .

$$h_{m, m_x}(t) = \begin{cases} \alpha(t) \exp(-\frac{d(m, m_x)}{2\sigma^2(t)}) & \text{if } m \in N_{m_x}(t), \\ 0 & \text{if } m \notin N_{m_x}(t). \end{cases}$$

The value of  $\alpha(t)$  is then identified with a *learning-rate factor* ( $0 < \alpha(t) < 1$ ). Both  $\alpha(t)$  and the radius of  $N_{m_x}(t)$  are usually decreasing monotonically in time (during the ordering process). To obtain almost sure convergence,  $\alpha(t)$  must verify [6] :

$$\sum_{t=0}^{+\infty} \alpha(t) = +\infty, \text{ and } \sum_{t=0}^{+\infty} \alpha(t)^2 = A < +\infty,$$

$\sigma(t)$  defines the width of the neighborhood; it corresponds to the radius of  $N_{m_x}(t)$  above. In our application we used  $\alpha(t)$  and  $\sigma(t)$  which begin with their initial values and vanish with time. Typically,  $\alpha(t) = \alpha(t-1)(\frac{\text{length}-t}{t})$  and  $\sigma(t) = \sigma(t-1)(\frac{\text{length}-t}{t})$ , where 'length' is the training length.

### B. SOM algorithm with missing values

The processing of data which contains missing values is a complicated and always awkward problem [2]. In our case, we have no missing data but erroneous data in the time series. So, it is necessary to recognize, omit, and then replace them by using the SOM. The erroneous data recognition process will be discussed in section III-A, here we will explain how the SOM may be used to process the missing values.

Let us assume that the observations may be clustered into  $M$  classes of  $n$ . When the input is an incomplete vector  $\mathbf{x}$ , we first define the set  $\mathcal{M}_x$  of the indices of the missing (absent) components.  $\mathcal{M}_x$  is a sub-set of  $\{1, 2, \dots, n\}$ . The winning code-vector  $C_{m_x}(t)$  related to  $\mathbf{x}$  is founded by using (1) at iteration  $t$ . But now, the distance  $\|\mathbf{x} - C_{m_x}(t)\|^2$  is computed with the valid components of  $\mathbf{x}$  only.

If missing data are present in the training set, the update of the code-vectors (the winning one  $C_{m_x}$  and its neighbors in  $N_{m_x}$ ) affects the valid components only. By denoting  $C_m(t) = (C_{m;1}, \dots, C_{m;k}, \dots, C_{m;n})^t$  the components of vector  $C_m(t)$  and  $\mathbf{x} = (x_1, \dots, x_n)^t$ , (2) becomes:

$$C_{m;k}(t+1) = C_{m;k}(t) + h_{m, m_x}(t)[x_k - C_{m;k}(t)] \quad (3)$$

for  $k \notin \mathcal{M}_x$  (*i.e.* for valid components). Otherwise,

$$C_{m;k}(t+1) = C_{m;k}(t). \quad (4)$$

### C. Estimation of missing values

Whatever the method used to deal with missing values, one of the most interesting properties of the algorithm is that it allows an a posteriori estimation of these missing values. Once the SOM has been trained, the missing values may simply be estimated by using:

$$\hat{x}_k = C_{m_x;k} \quad k \in \mathcal{M}_x. \quad (5)$$

When the Kohonen algorithm converges with neighbor of length 0, it is known that the code-vectors are asymptotically closed to the mean values of their classes at the end of the training phase. This estimation method therefore consists in estimating the missing values of a variable by the mean value of its class. It is obvious that

the more the compactness and the separability of the classes, the more accurate the estimation. Equation (5) may be tuned to a fuzzyfication by using membership values of the observation  $x$  to the set  $C$ . These membership values may also provide confidence intervals [2].

### III. APPLICATION TO MODIS DATA SET

This technique has been applied to a set of MODIS data dedicated to identifying bare soil in the brittany region during the winter season. Unfortunately, most of the observations are unusable due to the presence of clouds. In order to make such a monitoring available, it is necessary to process the most data as possible.

13 images are available from 25-11-2002 to 16-4-2003. Almost all of images captured in these dates are contaminated by clouds, but these dates are choosed indeed for their minimal amount of cloud coverage.

First, we tried to process band 2 (near infrared) to get clear images overall the season. Fig. 2 shows band 2 at 25-11-2002 and 4-2-2003 from a  $401 \times 401$  image in the north of Brittany region, extracted from 250m MODIS acquisition. These images are subsets from a larger images that covers the Brittany region. One can notice the cloud coverage on the two images, and see the huge amount of clouds and shadows in the latter date.

#### A. Finding outliers

The problem now is to isolate outliers (clouds or shadows) and mark them as missing data. One can use a statistical test like Grubbs' test [4], [5]. But these tests are based on the normal distribution which is nolonger the case here. In fact, cloud or shadow may be though of as *salt and pepper* noise which is not gaussian.

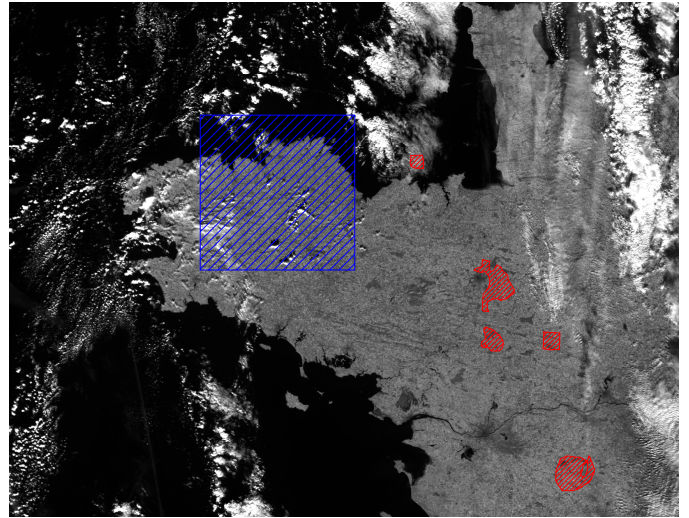
The box and whisker method has been applied instead, since it does not depends on a statistical model. The technique states that:

$x_k$  is an outlier at date  $k$  if

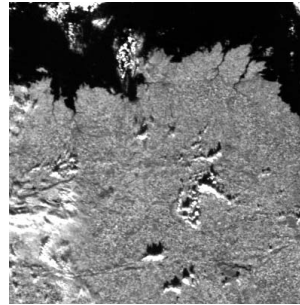
$$|x_k - 1.5(x_{3/4} - x_{1/4})| > |x_{1/2}|.$$

It is based on rank statistics:  $x_{1/2}$  is the median and  $x_{1/4}$  (resp.  $x_{3/4}$ ) the first (resp. third) quartile of the temporal signature  $x$ . One has to note that outliers are detected (and then removed) at a given date only. The temporal signature is preserved but associated to a non-empty missing component set  $\mathcal{M}_x$ .

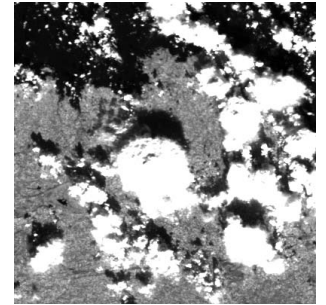
This simple method could be improved but we found an accuracy of 97.7636% with a ground truth containing three classes (Clouds, Shadows and valid data). The Kappa coefficient was 0.9652.



Training zones were selected from a  $1720 \times 1320$  MODIS image. Red hashed areas have been selected for training (these areas are weakly affected by clouds over the time). Blue hashed area shows the watershed of interest.



25-11-2002 NIR MODIS image



4-2-2003 NIR MODIS image

Fig. 2. Typical example of data affected by clouds. ©COSTEL

#### B. SOM implementation

A  $50 \times 20$  map has been implemented and trained on a specific area (shown on figure 2) where few clouds have been found. This size has been fixed in order to minimize duplicated and over trained vectors.

To ensure that the map has been fully organized and that there are no nonlinear data structures in our data set, the SOM has been represented by using the so-called " $dy, dx$ " representation [7]. This representation consists in the representation of the joint distribution of all input weights distances versus the corresponding distances in the output space. For perfectly regular maps and data structures that do not contain nonlinear relationships, the " $dy, dx$ " relation is a straight line of some slope  $a$  (where  $a$  is the weight distance between two consecutive units).

#### C. Data projection on SOM

Once the SOM has been correctly trained, the time series has to be processed. For each pixel in the original

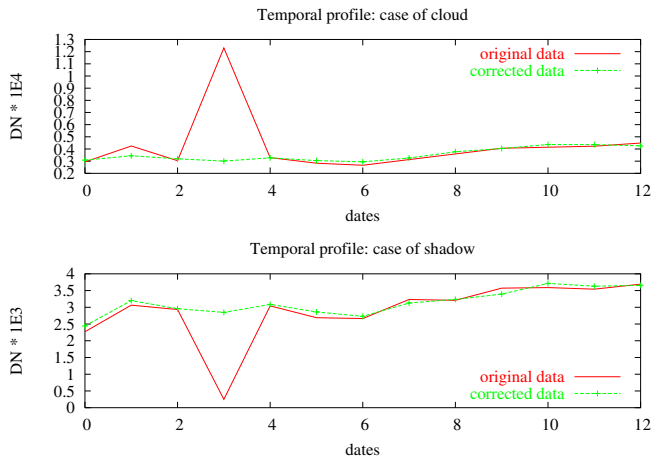


Fig. 3. Corrected temporal signatures: initial data with continued line, and its estimate line-point. The upper image shows a cloud correction while the lower a shadow.

data (*i.e.* the temporal signature through the 13 dates of red or NIR band<sup>1</sup>, the nearest neighbor in the SOM has to be found. It is worth mentioning here that we have abundant nodes in the SOM for each class, *i.e.* the variability of inter-class was taken into account.

We applied two methods to find the nearest neighbor. The smallest Euclidean distance and the minimum angle between each vector in the SOM and the input vector. The smallest Euclidean distance was found to be much better in recovering the contaminated data by projection on the SOM. Fig. 4 shows the image on date 25-11-2002 recovered by using the  $50 \times 20$  SOM.

It is interesting to stress that the time series does not have to be uniformly sampled over the time. The SOM is dealing with vectors of  $13$  with no consideration to the gaps between those 13 dates.

#### D. Validation

To show the success or the failure of the proposed method in preprocessing time series data with contaminated pixels, we selected some time signatures from the study area. Fig. 3 is a typical example of the temporal signature yielded by the SOM in presence of clouds or shadows.

A deeper validation is being carried out by COSTEL laboratory in order to evaluate at what extent this method can be used in tracking the evolution of vegetation coverage.

<sup>1</sup>Two SOM have been trained for red and NIR temporal signature.

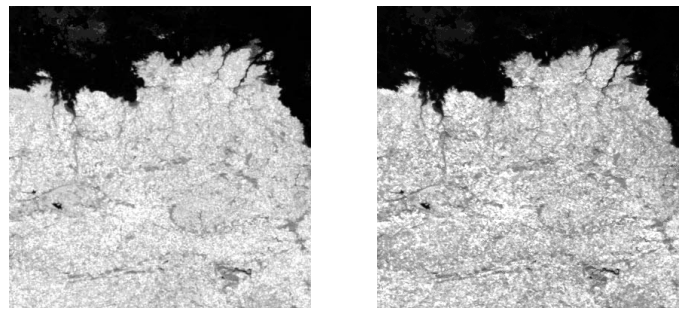


Fig. 4. The recovered data by using a  $50 \times 20$  SOM by the smallest of the Euclidean distance to project the original data onto the map.

#### IV. CONCLUSION

A technique to recover temporal series of low resolution images affected by clouds and shadows has been proposed. This method, which is based on the Kohonen's SOM, depends on training by using data taken from the same source and under the same conditions as the ones to be processed. The goal here was to predict the value of the erroneous data by means of a nonparametric algorithm, which do not require uniform sampling of the temporal series. It yields a data set free of clouds and shadows that better fit temporal analysis of the vegetation in a highly intensive agricultural area.

#### ACKNOWLEDGMENT

Authors would like to thank Rémi LECERF and Laurence HUBERT-MOY from COSTEL for supplying and preprocessing of MODIS data.

#### REFERENCES

- [1] T. Kohonen, *Self-Organizing Maps*, Second edition, Springer, 1997.
- [2] Marie Cottrell and Patrick Letrémy, Missing values: processing with Kohonen algorithm, *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, Brest, France, May, 17–20, 2005.
- [3] Ibbou, *Classification, analyse de correspondances et méthodes neuronales*. Thèse de Doctorat, Université Paris 1 Panthéon-Sorbonne, 1998.
- [4] Harvey Motulsky, Grubbs' Test for Detecting Outliers, *GraphPad Insight Issue Number 14*, Winter 1997, <http://www.graphpad.com/articles/grubbs.htm>.
- [5] NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm>.
- [6] Jeanny Hérault and Christian Jutten, *Réseaux neuronaux et traitement du signal*, Hermès, 1994.
- [7] Pierre Demartines and Jeanny Hérault, Representation of non-linear data structures through fast VQP neural network, *Proc. of Neuro-Nîmes'93*, 1993, Nîmes, France, pp. 411–424.