12

Self-Organizing maps for processing of data with missing values and outliers: application to remote sensing images

Bassam Abdel Latif and Gregoire Mercier

Abstract

This chapter presents how to recover a data set that contains missing values, errors and outlier values using the Self-Organizing Maps (SOM). It has been shown by many authors that if a data set contains missing values (missing components of some observations), then the SOM is a good candidate to recover it. The idea is as simple as to use the center of each subclass to estimate the missing values of a given observation. The virtue of the SOM regarding this problem is two folded: firstly, it is a non-parametric regression procedure that does not suppose any underlying models of the data set, and secondly it uses the information from similar observations to refine the positions of subclasses centers and hence gives better estimation.

Therefore, the SOM for missing value will be detailed first, and the modification of this algorithm will be proposed through the introduction of a new similarity measure (replacing the Euclidean distance, the widely used one with SOM applications) that is to be used to match different observations to their best matching neurons.

These algorithms will be presented by the help of a case study that recovers the missing and erroneous values in a set of remote sensing images (due to the presence of clouds and shadows). The application of the SOM algorithm to recover the missing data in a heavily cloudy region in France, monitor its superiority to other methods reported in the literature.

1. SOM for incomplete data

As we have seen in previous chapters, the SOM is used to cluster a set of observations $\mathbf{X} = \mathbf{x} : \mathbf{x} = x_1, \ldots, x_k, \ldots, x_n$ into a set of M clusters or classes of $\mathbb{R}^n$. The clustering is as good as the separability found between clusters. Normally, each cluster is represented by more than one neuron in the SOM output space to account for the natural variability in each cluster. Therefore, the number of neurons in the SOM, M, is usually greater than the natural number of clusters or classes found in the data set $\mathbf{X}$. Hence neighboring neurons share a considerable amount of information between them. This virtue of the SOM is used to overcome the reality that some observations, x, may have incomplete set of components, x_k, due to different reasons: insufficient data, sensor failure, experiments non-completed, etc.

Authors of Fessant & Midenet (2002) showed how the SOM algorithm may be used to estimate, recover, missing values in surveys. In Cottrell & Letrmy (2005), authors used the same method to estimate missing values in a socio-economical database. The basic ideas in their

work are: firstly, using valid components only in the training of the SOM and secondly using the synaptic weights of each neuron to estimate the missing components of the corresponding input observation. The following paragraphs will describe this process in more details.
Due to the huge size of the data used in many applications, we usually use two data sets in the clustering (or missing values estimation) using the SOM algorithm: the original one, $\mathbf{X}$, to be clustered and a representative subset of it, $\mathbf{X}'$. This last subset is smaller in size than the original set, this accelerates the training process. In the case that the original set, $\mathbf{X}$, contains observations with missing values, the training subset, $\mathbf{X}'$, may or may not contain observations with missing values. To proceed in the training phase, if an input observation x contains missing component, x_k, the set M_x is introduced to define the indices of these missing values. In this case, M_x is a sub-set of $1,2,\ldots,n$.
In the training phase, the winning neuron at iteration t, $\mathbf{C}_{m_x}(t)$, is selected in responding to the input $\mathbf{x}$ using the equation:

$$\|\mathbf{x} - \mathbf{C}_{m_x}\| = \min_{m \in \{1,\ldots,M\}} \|\mathbf{x} - \mathbf{C}_m\|. \tag{1}$$

When facing incomplete vector, $\mathbf{x}$, the Euclidean distance $\|\mathbf{x} - \mathbf{C}_m(t)\|$ is computed with the valid components of $\mathbf{x}$ only ($x_k \notin M_x$). The weight updating of neurons (the best-matching neuron, $\mathbf{C}_{m_x}$, and its neighbors that belong to the set $N_{m_x}(t)$) affects the valid components only. In other words, by denoting $\mathbf{C}_m(t) = (c_{m;1},\ldots,c_{m;k},\ldots,c_{m;n})^t$ the components of weight vector associated to the neuron $\mathbf{C}_m$ at instant t and $x = (x_1,\ldots,x_n)^t$, then the weight updating process is accomplished by the equation:

$$c_{m;k}(t+1) = c_{m;k}(t) + h_{m,m_x}(t)\Big[x_k - c_{m;k}(t)\Big], \tag{2}$$

for $k \notin M_x$ (*i.e.* for valid components). Otherwise, no modification is performed:

$$c_{m;k}(t+1) = c_{m;k}(t). \tag{3}$$

1.1 Estimation of missing values

One of the interesting properties of the SOM algorithm for missing values 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:

$$\hat{x}_k = c_{m_x;k} \quad k \in \mathbf{M}_x. \tag{4}$$

When the Kohonen algorithm converges with a neighborhood of length 0, it is known that the code-vectors $\mathbf{C}_m$ converges asymptotically to the mean value of its class or subclass m. Therefore, this estimation method consists in estimating the missing values of a random variable by the mean value of its class, defined through a training set. It is obvious that the more the compactness and the separability of the classes, the more accurate the estimation. Eq. (4) may be turned to a fuzzyfication by using membership values of the observation x to the set of the code vectors Cottrell & Letrmy (2005). These membership values may also provide confidence intervals Cottrell & Letrmy (2005).

2. Case study

In this section we will study the application of this technique to a set of MODIS (Moderate Resolution Imaging Spectroradiometer) data dedicated to identifying bare soils in the Brittany region of France during the winter season. Most of observations are unusable due to the presence of clouds or shadows (knowing that this region is characterized by 200 rainy days per year). Therefore, in order to make bare soil monitoring in such conditions, it is necessary to process as most data as possible.
In this application, MODIS data are collected in time series of reflectances (near-infrared and red reflectances). In other words, each pixel in a multiband file is a temporal profile of a certain reflectance channel. The temporal profile of each pixel is considered to as an observation that contains as much components, x_k, as the number of dates, n, in the time series. Therefore, the SOM algorithm for the missing values will be applied twice on two different sets: one set for the near infrared channel of the MODIS image series, $\mathbf{X}_{NIR}$, that contains observations of the form $\mathbf{x} = NIR_1, \ldots, NIR_k, \ldots, NIR_n$, and another set for the red channel, $\mathbf{X}_R$, that contains observations of the form $\mathbf{x} = R_1, \ldots, R_k, \ldots, R_n$
If one temporal profile has cloud or shadow contamination in any date, then the index of this date is assigned to the set M_x and the whole observation is marked as having erroneous values. These erroneous values have to be detected and marked as missing values before proceeding with the SOM algorithm for missing values to recover them.

2.1 Data

We describe in this section the data used in this case study. These data contains two types of satellite data. The first type is the MODIS data that are contaminated by the presence of clouds and their associated shadows, while the second type is a set of high resolution data that has been used in the validation of the SOM algorithm for missing values. The following paragraphs provide more details about these data.

2.1.1 MODIS data

MODIS images of the season 2002-2003 were used in this case study. They were selected according to sensor zenithal viewing angle and cloud coverage criteria. Despite available cloud cover information in each MODIS tile, images have been selected visually. Selection based on MODIS tiles information appears to be too imprecise, since the information is not spatially distributed. Images with less than 50% of cloud coverage are selected. Concerning zenithal viewing angle selection, images acquired with an orbit track centered on Brittany and a radius of 200 km were selected. Low zenithal viewing angle values, inferior to 20, were selected only to avoid spatial resolution variations.
10 images are available from 11-25-2002 to 4-16-2003. Almost all of the images captured in these dates are contaminated by clouds, even the fact that these dates were selected, indeed, for their minimal amount of cloud coverage. Only band 1 and 2 of the MODIS data have been processed due to their 250m spatial resolution. Therefore, the SOM algorithm for missing values will be applied twice: one process dedicated to the data set constituted by the 10 dates of red reflectances and another independent process applied to the data set constituted by the 10 dates of near-infrared reflectances. In this study, processing steps dedicated to recover the erroneous data in the near-infrared channel is presented. The method can be applied in the same way to any other reflectance channel. Fig. 1 shows the near-infrared band at 11-25-2002 over the Brittany region.

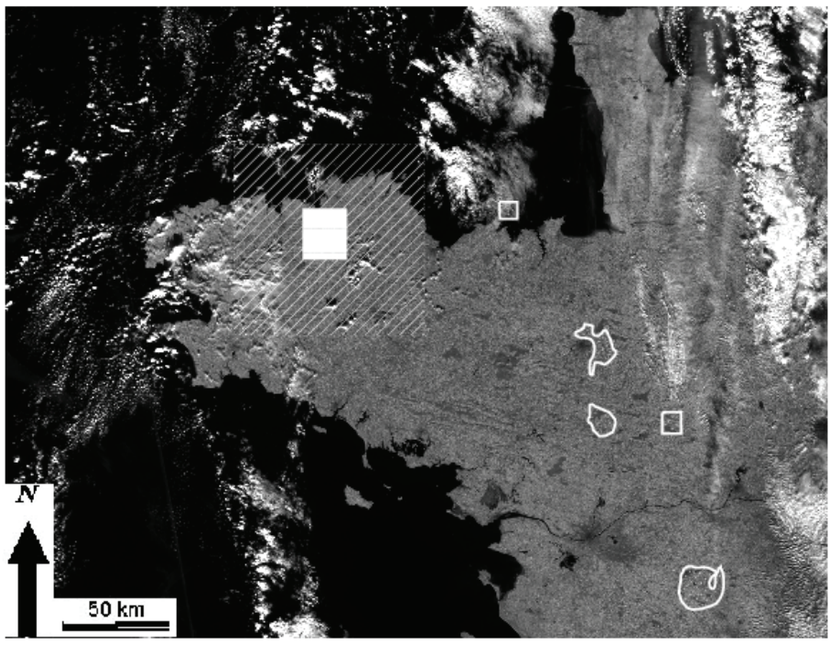


Fig. 1. Typical examples of data affected by clouds: the Brittany region as captured by MODIS near infrared band. Polygons represent areas that have been selected for training (these areas are almost clouds free for the temporal series used in the study). Rectangle hashed area shows the location of SPOT/HRVIR image acquired on 01-24-2003. White area corresponds to the deleted data of 01-24-2003 (see sec. 3.1).

In order to assess results of the recovering process, some original MODIS data has been replaced by missing values in the image of 01-24-2003. This date has been chosen because there is no clouds or shadows in the selected area and that high resolution data corresponding to the same area is available. The validation area is about 25×20km (see the white rectangle in Fig. 1). This validation area has been converted from reflectance values to zero values that are close to the deep shadow values.

2.1.2 High resolution data

Two high resolution images are used for validation. SPOT/HRVIR and Landsat/ETM+ images were corrected from atmospheric effects by using the 5S model Tanre et al. (1990) and then corrected from geometric distortion. High resolution images were aggregated to 250m resolution, by using the Point Spread Function of the MODIS sensor Huang et al. (2002), for the comparison to the original MODIS data. The accuracy of the recovering of missing data could have been evaluated by using original MODIS data. Nevertheless, data from alternate sensors is used indeed to evaluate the sensitivity of the recovering process with regards to the atmospheric and geometric difference introduced by those alternate sensors.

2.2 The use of the SOM

2.2.1 Finding outliers

We suppose here that within a temporal profile of normal cover, an outlier value represents an erroneous value (presence of clouds or shadows). This hypothesis may not be acceptable for other types of data, e.g. in high resolution data where a white car or a bright roof may induce

an outlier-like pixel. Therefore, a simple algorithm, to identify outliers in each temporal profile is used in this study. For operational use, however, one can use an appropriate cloud detection algorithm for the processed data, e.g. the standard MODIS 1km cloud mask.
One can use any classical test (for which an interest comparison may be found in Bakar et al. (2006)) to carry out this task. Nevertheless, most of these tests are based on the normal distribution assumption, which is no longer the case here. In fact, clouds and shadows may be thought of as *salt and pepper* noise. Then, the median filter should be more appropriate for this kind of noise. Therefore, the Box and Whisker method has been used, since it does not depend on a statistical model.
The technique states that x_k is mostly an outlier at date k if:

$$|x_k - 1.5\,(\mathbf{x}_{3/4} - \mathbf{x}_{1/4})| > |\mathbf{x}_{1/2}|. \tag{5}$$

It is based on rank statistics where: $\mathbf{x}_{1/2}$ is the median and $\mathbf{x}_{1/4}$ (resp. $\mathbf{x}_{3/4}$) is 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. Whenever an outlier is detected at one date or at several dates, the temporal signature is preserved and associated to the set of non-empty missing components $\mathbf{M}_x$.
This simple method could be improved, but satisfying results have been obtained when applied to the MODIS data set because it is capable of detecting both clouds and shadows and, hence, overcomes the standard MODIS 1Km cloud mask which does not detect shadows.

2.2.2 SOM implementation

A 50×20 neuron map has been implemented and trained on a specific area (shown in figure 1) where few clouds have been found in the time series.
The number of neurons in the SOM grid is a compromise between the processing time and the required quantization error (the average distance between each vector in the test data and its Best Matching Unit (BMU) in the SOM grid). One can begin by a reasonably small map, say 12×8, and increases the size of the map to reach a satisfactory quantization error. In this application, we reached an average quantization error of 0.0327 (in the sense of Mean Square error, MSE) with the proposed map size in a reasonable time: 45 sec. It is also advisable to have a rectangular shape map if the data has a correlation between its different components. So the choice of the rectangular shape 50×20 of the SOM grid.

2.2.3 Data projection on the SOM grid

Once the SOM grid has been trained until convergence, the complete time series has to be processed. Each pixel in the original data is projected on the SOM grid to find the weight vector that best matches it. The word projection is used here to indicate that the output image contains a subset of reflectance values that have been yielded by the SOM algorithm. It is worth mentioning here that the 50×20 neurons of the SOM grid are representative enough for each class, *i.e.* the variability of inter-class was taken into account.
Two different similarity measures have been used to find the winner neuron, that matches a given observation, in the projection phase. The Euclidean distance and the *spectral*[1] angle between each vector C_m in the SOM grid and the input vector x. The Euclidean distance was found to be more appropriated distance measure to recover the contaminated data by projection on the SOM grid. Further discussion about similarity measures will be follow in section 4

[1]It is a *temporal* application of the *Spectral* angle.

It is interesting to stress that the time series does not have to be uniformly sampled over the time. The SOM algorithm is dealing with vectors of $\mathbb{R}^{10}$ with no consideration to the gaps between those 10 dates.

2.3 Validation

Two methods have been used to validate recovered reflectance data. The first one is a simple difference between the so-called recovered image and the original one. The second method consists in mapping the residual errors of the linear regression between the recovered data and the original one. The linear relation is of the form:

$$y = ax + b \tag{6}$$

where x takes the value of the original MODIS pixels and y is associated to the pixels of the recovered image. In order to evaluate difference between the original image and the recovered one, residual errors (as denoted to as *residual* in the text) are calculated and mapped. Residuals express the distance between values of the recovered image and values estimated with the linear model:

$$\epsilon_i = y_i - ax_i - b \tag{7}$$

where i corresponds to a pixel index.
This method was originally used to detect changes with unscaled values Ingram et al. (1981); Jha & Unni (1994). In this study, this method is used to compare recovered MODIS image with high resolution images.

3. Results

Results yielded by the used method are exposed in two steps:

- validation of reflectance values after processing;
- validation of recovered spatial structures.

A comparison with the compositing method that deals with the problem of clouds and shadows in the low resolution images and which is the most used method in the relevant literature, is conducted in Abdel Latif et al. (2008).

3.1 Validation of reflectance values after processing

In this part, results are compared by considering:

- the projection of contaminated observations directly onto the SOM grid, referred to as SOM1 algorithm,
- the use the Box and Whisker technique to isolate outliers (erroneous values in each observation) and mark them as missing values before projection onto the SOM grid, referred to as SOM2 algorithm (cf. Fig. 2)

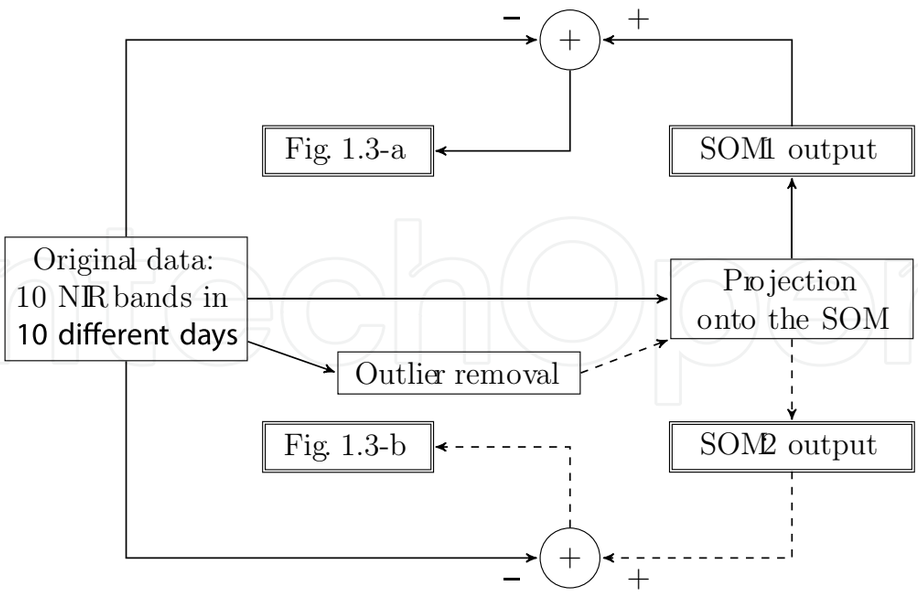
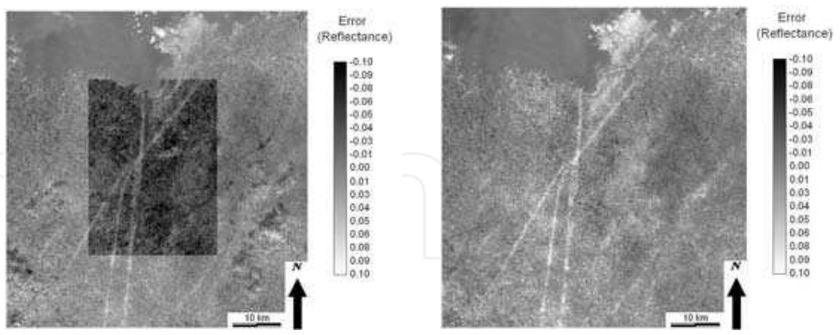


Fig. 2. A schematic representation for SOM1 and SOM2 algorithms and different processing steps that have been used to produce Fig. 3.

A simple difference is calculated to confirm the ability of SOM algorithms to recover reflectance values, (cf. Fig. 2 and 3). The visual analysis of Fig. 3 shows clearly that recovered images become more homogeneous when using SOM2 algorithm. Fig. 3-(a) shows that using SOM1 algorithm, recovered reflectance values of the artificial no-data rectangle (the white rectangle in Fig. 1) are lower than the rest of the image. Fig. 3-(b) shows the recovered reflectance values: no discontinuities may be distinguished between the no-data rectangle and its surroundings when SOM2 algorithm is used.



(a) (b)

Fig. 3. Difference between recovered MODIS image and original MODIS image, 01-24-2003. (a) with SOM1 algorithm; (b) with SOM2 algorithm.

Considering that the difference between original reflectance values and recovered ones are associated to errors, Fig. 4 shows the distribution of these errors obtained when processing the area defined on Fig. 1. It is worth noting that, with SOM1 algorithm, reflectance values are underestimated with a mean around 0.05 (the mean of error distribution is at about -0.05).

When using SOM2 algorithm, the mean of errors is around 0.02. Moreover, 80% of errors are located between −0.04 and 0.07.

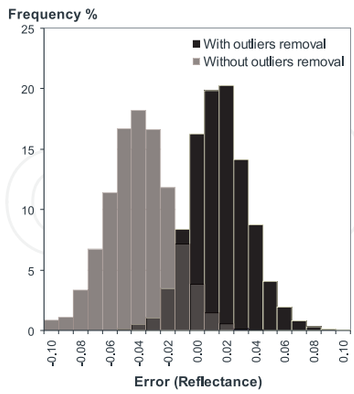
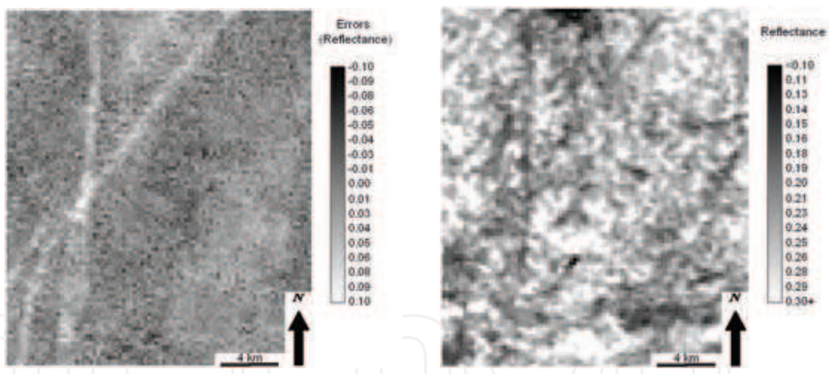


Fig. 4. Distribution of errors between recovered MODIS image and original MODIS image, evaluated on the no-data rectangle, on 01-24-2003.

Fig. 5-(a) focuses on some atmospheric artifacts seen in Fig. 3: white vertical and diagonal lines. These lines are whiter in the difference image, Fig. 5-(a), which means that the SOM algorithm chose whiter samples (weight vectors) to replace darker ones in the original MODIS image Fig. 5-(b). These lines are due to shadows of airplane trails which were not visually detectable from the NIR band but from the visible bands only.



(a) (b)

Fig. 5. Atmospheric artifacts (shadows of airplane trails) corrected by the SOM algorithm, 01-24-2003. (a) Difference between recovered MODIS image using SOM2 and original MODIS image; (b) Original MODIS image.

These results show that, by using the Box and Whisker technique to isolate contaminated values, the retrieved reflectance values are scaled and located in the range of valid reflectance values. In the next part, the study focuses on the spatial structure of recovered data.

3.2 Validation of recovered spatial structures

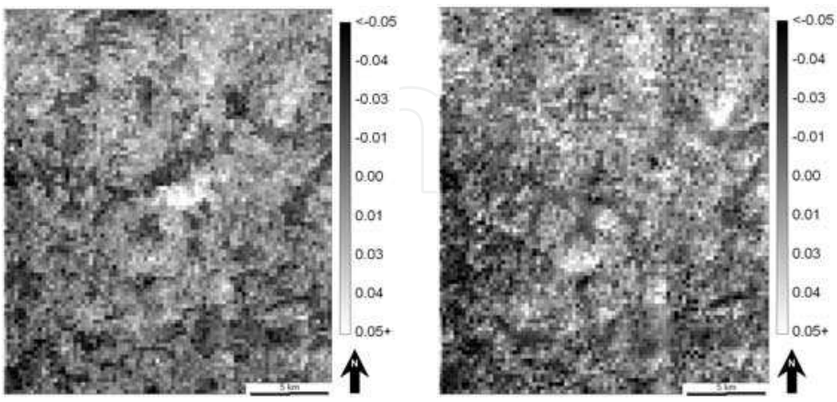
The next stage of the validation points up the correctness of recovered spatial structures by comparing the recovered images with 1) original MODIS images and 2) high resolution images aggregated at 250m spatial resolution.

The correlation coefficients between the original MODIS image, SPOT/HRVIR image and the recovered MODIS image, all on 01-24-2003, are shown in table 1. It appears that isolating contaminated values, by the Box and Whisker method, before the projection on the SOM grid increases the correlation between recovered MODIS and original MODIS (from 0.83 to 0.87) and increases the correlation between recovered MODIS and SPOT/HRVIR (from 0.81 to 0.86). Residuals of the two linear relations highlight the atmospheric artifacts that have been substituted by the SOM algorithm. Fig. 6 shows the residuals of the relation between the recovered MODIS image and the SPOT/HRVIR image. Whiter colors mean that these values were recovered to higher values than the simple linear relation, of eq. (6), did (cf. section 2.3). Darker colors mean that SOM algorithm estimated values lower than the simple linear relation did. Hence, these differences in the estimated values are mainly due to difference of atmospheric contamination such as haze and shadows, and show the relevance of the SOM estimation.

Fig. 7 shows the residuals of the relation between original and recovered MODIS image. As before, whiter colors means that these pixels were estimated by the SOM algorithm to higher values than in the original image.

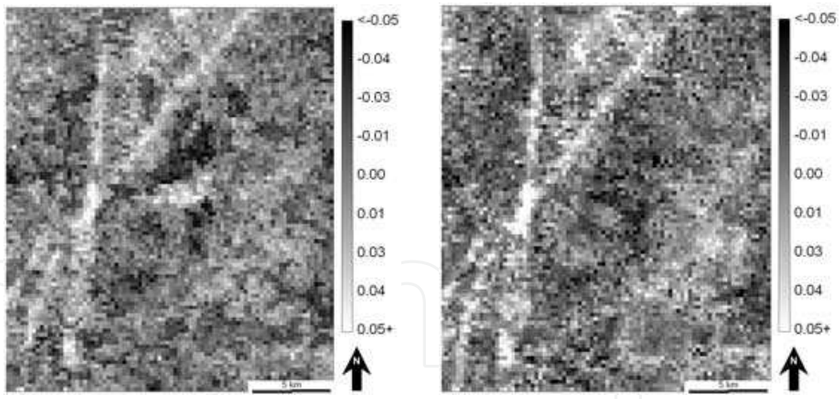
r	original MODIS	recovered MODIS using algorithm:	
		SOM1	SOM2
SPOT/HRVIR	0.84	0.81	0.86
original MODIS	1	0.83	0.87

Table 1. Correlation coef. between recovered MODIS data and SPOT/HRVIR data, 01-24-2003.



(a) (b)

Fig. 6. Maps of residuals between recovered MODIS image and SPOT/HRVIR image, 01-24-2003. (a) SOM1 algorithm; (b) SOM2 algorithm.



(a) (b)

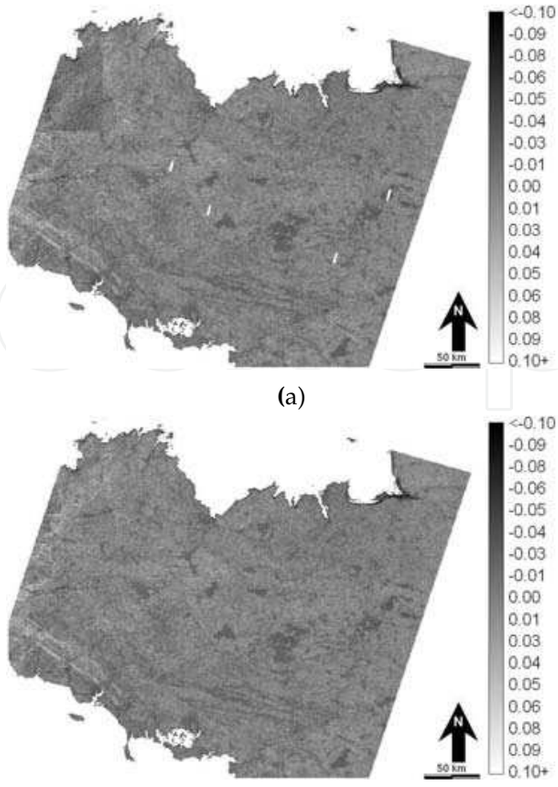
Fig. 7. Maps of residuals between recovered MODIS image and original MODIS image, 01-24-2003. (a) SOM1 algorithm; (b) SOM2 algorithm.

To confirm these results, recovered image was compared to Landsat/ETM+ and MODIS images with a larger area to evaluate how will be the correlation coefficient for spatially large areas. According to table 2 the relation between the original MODIS and the Landsat/ETM+ image shows a correlation coefficient of 0.91. The recovering process decreases slightly the relation with the aggregated Landsat/ETM+ image with the SOM2 algorithm (from 0.91 to 0.89) and with the SOM1 algorithm (from 0.91 to 0.88). This decrease in the correlation with the Landsat/ETM+ image is interpreted by the fact that this image is not as clear as the SPOT/HRVIR image. Therefore, the recovery process changes some values in the MODIS image which were well correlated, as clouds or shadows, in the Landsat/ETM+ image. Residuals between recovered MODIS image and Landsat/ETM+ image are shown in Fig. 8.

r	original MODIS	recovered MODIS using algorithm:	
		SOM1	SOM2
Landsat/ETM+	0.91	0.88	0.89
original MODIS	1	0.96	0.97

Table 2. Correlation coef. between recovered MODIS and Landsat/ETM+ data, 03-15-2003.

The comparison of results, obtained from 01-24-2003 and 03-15-2003 cases, proves that the use of an outlier detector and removal, SOM2 algorithm, improves the performances of the non-parametric estimator.
Using residuals of a linear relation to validate recovered reflectance values, the spatial structure of recovered images is considered to be corresponding to the landscape structure. These analysis allow to conclude that the spatial structure is well recovered. Moreover, the SOM algorithm removes small atmospheric artifacts that are not visually detectable and replaces them by valid reflectance values (cf. Fig. 5).



(a)

(b)

Fig. 8. Maps of residuals between recovered MODIS image and Landsat/ETM+ image, 03-15-2003. (a) SOM1 algorithm; (b) SOM2 algorithm.

4. An appropriated similarity measure

In section 3, we saw that results obtained using the SOM2 algorithm is better than those obtained by the SOM1 algorithm. This is because in the projection phase, SOM2 algorithm uses only the valid components in searching the best matching temporal profile, from the SOM grid, for a given contaminated temporal profile. Unfortunately, the performance of the SOM2 algorithm will be identical to that of SOM1 algorithm if the outlier detector fails to detect the outlier values in the temporal profile. The question now is: are there any better similarity measures that better match the contaminated profiles with the code vectors? In other words, is there any similarity measure that is robust against the presence of outliers in the temporal profile, before the projection onto the SOM, even with the use of the simple outlier detector of Box and Whisker?

We present hereafter a similarity measure that will make the SOM2 algorithm robust against the presence of any erroneous values that remain without detection prior to the projection phase. This robustness makes it possible to apply the SOM algorithm for missing values without the need to the intermediate stage of detecting and marking erroneous values to recover these erroneous values.

4.1 Similarity measures

We can find 4 different similarity measures that are heavily used in the literature of remote sensing applications (the thematic field of the case study). They are: (a) the Euclidean distance, (b) the spectral angle mapper, (c) the spectral correlation measure and (d) the spectral information divergence measure and (e) a proposed similarity measure.

4.1.1 Euclidean distance

The Euclidean distance (ED) between two temporal profiles portrayed as vectors $(\mathbf{x})$, $(\mathbf{y})$ lie in $(\mathbb{R}^n)$ is given by:

$$\mathrm{ED}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{8}$$

As a similarity measure, we can drop out the square root function because it is a monotonic increasing function. Its presence or absence will not affect the results.

4.1.2 Spectral angle measure

The Spectral Angle Measure (SAM) Kruse et al. (1993), will be applied to a temporal profiles and not to a spectral signatures, is the measure of the angle between two vectors $(\boldsymbol{x})$ and $(\boldsymbol{y})$ which lie in $(\mathbb{R}^n)$ and is as follows:

$$\mathrm{SAM}(\boldsymbol{x},\boldsymbol{y}) = \cos^{-1}\left(\frac{\sum_{i=1}^{n} x_i y_i}{[\sum_{i=1}^{n} x_i^2]^{1/2}\sum_{i=1}^{n} y_i^2]^{1/2}}\right). \tag{9}$$

The difference between the Euclidean distance and spectral angle is that the later is not affected by the magnitude of the involved vectors (if two vectors have the same direction, then it will not matter if their magnitudes are different).

4.1.3 Spectral correlation measure

The Spectral Correlation Measure (SCM) van der Meer (2006) is calculated as the correlation coefficient between two temporal profiles $(\mathbf{x}\text{and}\mathbf{y} \in \mathbb{R}^n)$ as:

$$\mathrm{SCM}(\mathbf{x},\mathbf{y}) = \frac{n\sum_1^n x_i y_i - \sum_1^n x_i \sum_1^n y_i}{\sqrt{[n\sum_1^n x_i^2 - (\sum_1^n x_i)^2][n\sum_1^n y_i^2 - (\sum_1^n y_i)^2]}}. \tag{10}$$

This similarity measure has the advantage that it takes into account the relative shape of the two vectors as well as the component matching. The correlation can be both positive and negative. Such a measure takes brightness difference and shape difference between vectors into account.

4.1.4 Spectral information divergence

The Spectral Information Divergence (SID) measure Chang (2000) calculates the distance between the probability distribution produced by two vectors $(\mathbf{x})$ and $(\mathbf{y})$ (here, temporal profiles) by firstly found the probability distribution $(\mathbf{p} = \{p_i\}_1^n,\ p_i = x_i/\sum_1^n x_i)$ and $(\mathbf{q} = \{q_i\}_1^n,\ q_i = y_i/\sum_i^n y_i)$. Then the spectral information divergence is given by:

$$\mathrm{SID}(\mathbf{x},\mathbf{y}) = D(\mathbf{x}\|\mathbf{y}) + D(\mathbf{y}\|\mathbf{x}). \tag{11}$$

Where $(D(\mathbf{x}\|\mathbf{y}) = \sum_1^n p_i \log(p_i/q_i))$ and $(D(\mathbf{y}\|\mathbf{x}) = \sum_i^n q_i \log(q_i/p_i))$. It should be noted that $(D(\mathbf{x}\|\mathbf{y}))$ is called the relative entropy of $(\mathbf{y})$ with respect to $(\mathbf{x})$ which is also known as Kullback-Leibler information function.

4.1.5 A proposed similarity measure

This subsection presents a similarity measure for the first time in the literature to be used with the SOM algorithm, that is why we use the word "proposed" in this book chapter. The proposed similarity measure is suitable when searching a candidate codebook for a given vector that has some erroneous components. This measure tries to decrease the effects of erroneous components to the resulted measure with respect to the non-erroneous components. The proposed similarity measure was firstly used in Chapelle et al. (1999) as a heavy-tailed radial base function (RBF) that has been used in a classification of data base images by comparing their histograms using the Support Vector Machine (SVM) technique. This non-Gaussian RBF kernel is given by:

$$K(\mathbf{x},\mathbf{y}) = e^{-\rho \sum_i |x_i^a - y_i^a|^b}, \tag{12}$$

where $(a \leq 1)$ and $(b \leq 2)$. Since the exponential function is a monotonic function, hence its presence or absence will not affect the comparison between a vector $(\mathbf{x})$ and a set of vectors $(\mathbf{y}_i)$ to choose the best match vector $(\mathbf{y}^*)$ to the given input vector $(\mathbf{x})$. Therefore, the proposed similarity measure is given by:

$$\mathrm{D}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i^a - y_i^a|^b. \tag{13}$$

It remains the determination of the two parameters a and b. For the special case of $a = 1$ and $b = 2$ this similarity measure performs exactly as the Euclidean distance similarity measure.

4.2 The proposed similarity measure

We will proceed to determine best values for the parameters a and b in an empirical way. This empirical way uses a set S_{MODIS} of 5000 MODIS temporal profiles. Each temporal profile consists of 10 dates of near-infrared reflectance values. These temporal profiles have been selected randomly from the reconstructed MODIS time series that has been yielded in section 2. To come over the quantization effects of the SOM algorithm for missing values, a random reflectance value between -0.02 and 0.02 has been added to all reflectance values in the S_{MODIS}. Reflectance values in this set are supposed to match real objects on the land surface.

A perturbed set , S_p, has been simulated using the S_{MODIS} set. From the 5000 temporal profiles in each band in S_{MODIS}, 1500 locations (at each individual date) are selected randomly to simulate perturbed reflectance values and have been assigned to a random value between 0 and 0.9. The 1500 locations in each date are independent of other 1500 locations in other dates. In this way, the set S_p has 4733 temporal profiles that have perturbed values in at least one date. There are only 20 temporal profiles that has a perturbation in more than 6 dates with 17 one perturbed at 7 positions, 3 at 8 positions and 1 temporal profile that has a perturbation at 9 dates.

Therefore, the best values for a and b parameters will be those which maximize the correct matching from the set S_p to the set S_{MODIS} (if a perturbed temporal profile is mapped to its original, non perturbed temporal profile, this mapping or matching is considered as to be correct). Fig. 9 shows the percentage of correct matching in function of one parameter, here variable, of a or b while fixing the other parameter to 1. The figure shows also that at a and b equal 1, the percentage of correct matching is 23.86%. For $a = 1$ and $b = 2$ (Euclidean distance similarity measure) the percentage of correct matching is only 13.04%.

If the value of a or b get larger than 1, the performance is dramatically decreased. For values less than 1, the percentage of correct matching increases. At $a = 0.1, b = 1$ the correct matching

is 48.98%, while at $a = 1$ and $b = 0.1$ the correct matching is 99.92%. In other words, the similarity measure with $a = 1$ and $b = 0.1$ has correctly matched 4996 temporal profiles from the 5000 one in the test. That is to say this arrangement has succeeded to correctly match temporal profiles perturbed at up to 70% of its components. The rest 4 temporal profiles that have not been matched correctly are the 3 temporal profiles that have been perturbed in 8 out of 10 dates (80% of the data are wrong) and the temporal profile that have been perturbed in 9 out of 10 dates.

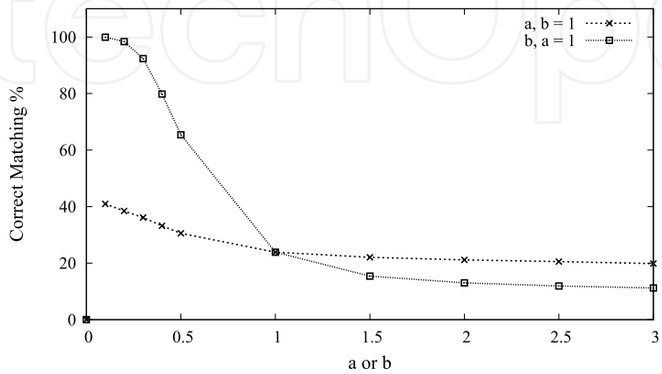


Fig. 9. The percentage of correct matching as a function of one of the parameters a or b while fixing the other at 1.

We notice also from Fig. 9 that decreasing b is much profitable than decreasing a. When a or b approaches 0, the correct matching becomes 0. Fig. 10 shows that highest correct matching percentage is attained at $b = 0.1$ regardless of the value of a. Therefore, a value of $a = 1$ and $b = 0.1$ will be used as best values of a and b in the proposed similarity measure. The choice of $a = 1$ is to speed up the calculations. Therefore, the proposed similarity measure, for our application, is given by:

$$\mathrm{D}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|^{0.1}. \tag{14}$$

It is worth mentioning here that we stopped at $(b = 0.1)$ because it gives us the highest possible correct matching in our simulation. For other applications, a smaller values of a and b may be used to increase the correct matching percentage.

4.3 Comparison of different similarity measures using MODIS data

In this section, we will compare the proposed similarity measure with the most used similarity measures found in the literature and that have been reported in section 4. The comparison will be done by applying the SOM algorithm for missing values to MODIS data.

Therefore, we will compare different reconstructions of a simulated contaminated version of a MODIS time series of the near infrared channel to its clear version. Due to the difficulty of obtaining a clear MODIS time series on the winter season of the Brittany region in France, we will use a reconstruction version of the time series of 10 dates as yielded in sec. 2.

The procedure of comparison may be described by the schematic representation of Fig. 11. The TS0 in this figure refers to a subset of 401×401 pixels of the original 10 dates time series of near infrared channel that has been used in sec. 2. TS1 in the same figure refers to a reconstruction of TS0 using SOM2 algorithm as in sec. 2. Due to the quantization effect of the Kohonen SOM

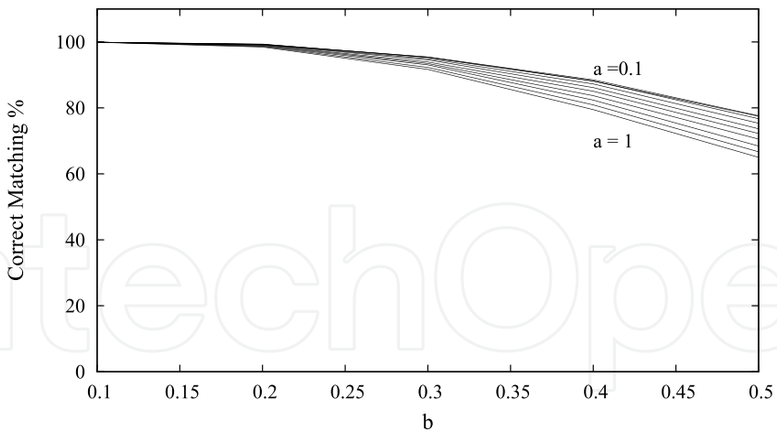
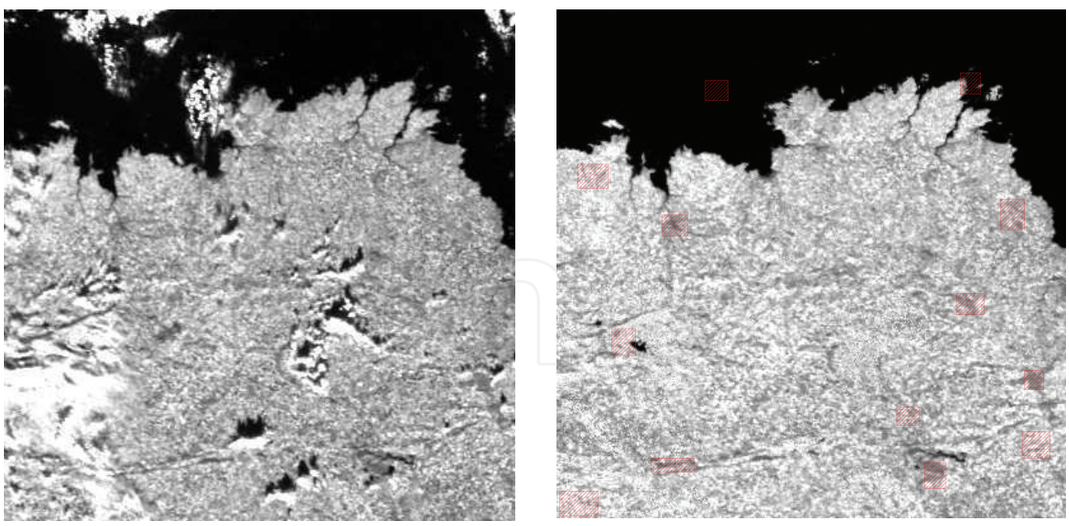


Fig. 10. The percentage of correct matching in function of *b* while varying *a* from 0.1 to 1.

algorithm, we will add a random noise in the range between -0.02 and 0.02 reflectance values to the TS1 to produced the time series TS2. The added noise is a random noise drawn from an independent and identical distribution (iid). Images on 01-24-2003 are shown in Fig. 12 for both TS0 and TS1. To simulate clouds and associated shadows, 20% of all pixels in all dates have been selected randomly and have been replaced by a random value between 0 and 0.9 to produce TS3 time series.
In this way a temporal profile of 10 dates, in TS3, has in average two erroneous reflectance values. There are 14375 temporal profiles (from 160801 temporal profiles) that are contaminated in more than 4 dates.



(1) (2)

Fig. 12. Near InfraRed channel of MODIS images on 01-24-2003: (1) the original one and (2) the reconstructed one by the SOM algorithm for missing values.

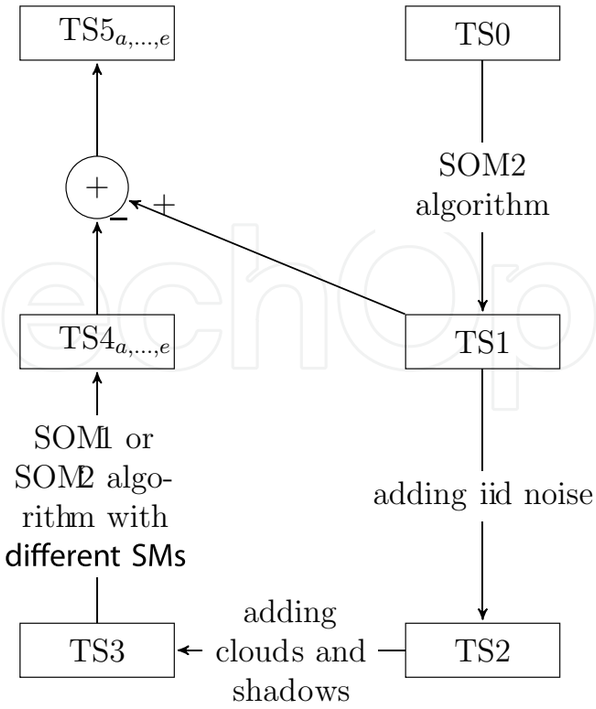


Fig. 11. A schematic representation of steps used to compare the different similarity measures

A comparison of different similarity measures is now possible considering the TS1 as the original time series and a set of the different reconstructions of the MODIS data using different similarity measures($TS4_{a,...,e}$ in Fig. 11).
It remains now to decide what criteria will be used to compare the different reconstructions. We found that statistics of first and second order of a difference time series (difference between original and different reconstructions: $TS5_{a,...,e}$ in Fig. 11) were sufficient to monitor the best similarity measure to be used in such applications. Therefore the average mean and the average standard deviation of each difference time series will be used to find the best similarity measure. An ideal reconstruction will produce a difference image with 0 mean and 0 standard deviation. But due to the fact that the SOM algorithm for missing value has a quantization effect and that 14375 temporal profiles contaminated in 4 dates or more, the 0 mean and standard deviation is not realistic in this comparison. Therefore the nearer the average mean of TS5 to 0 and the smaller the standard deviation are, the better the reconstruction is.

4.3.1 Difference Time Series

In this section we will present results yielded by applying the SOM algorithm for missing values to MODIS data using different similarity measures in the projection phase of the algorithm. Fig. 13 shows the first date of the different reconstructed time series (TS4 in Fig. 11) using the SOM1 algorithm.

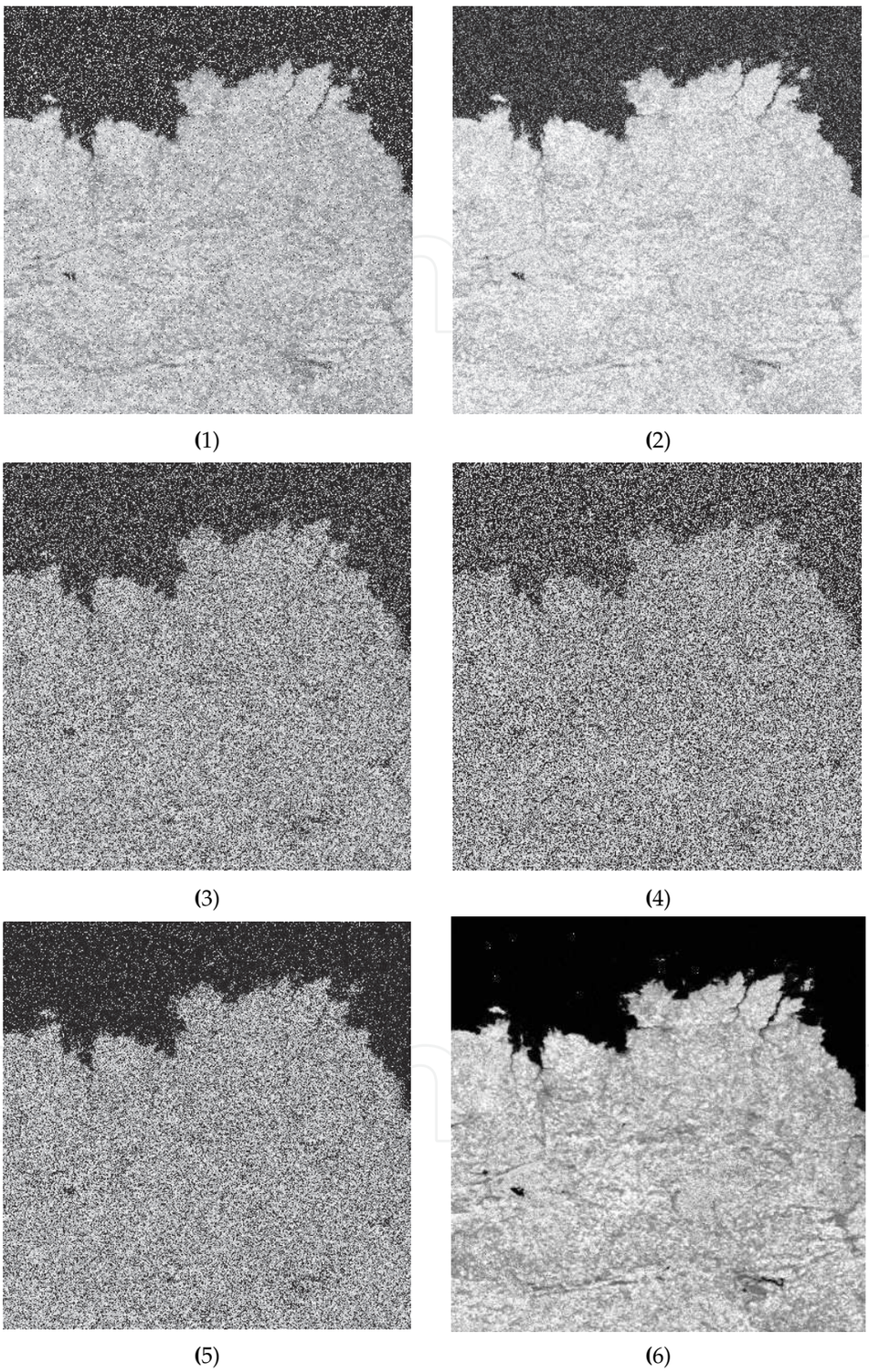
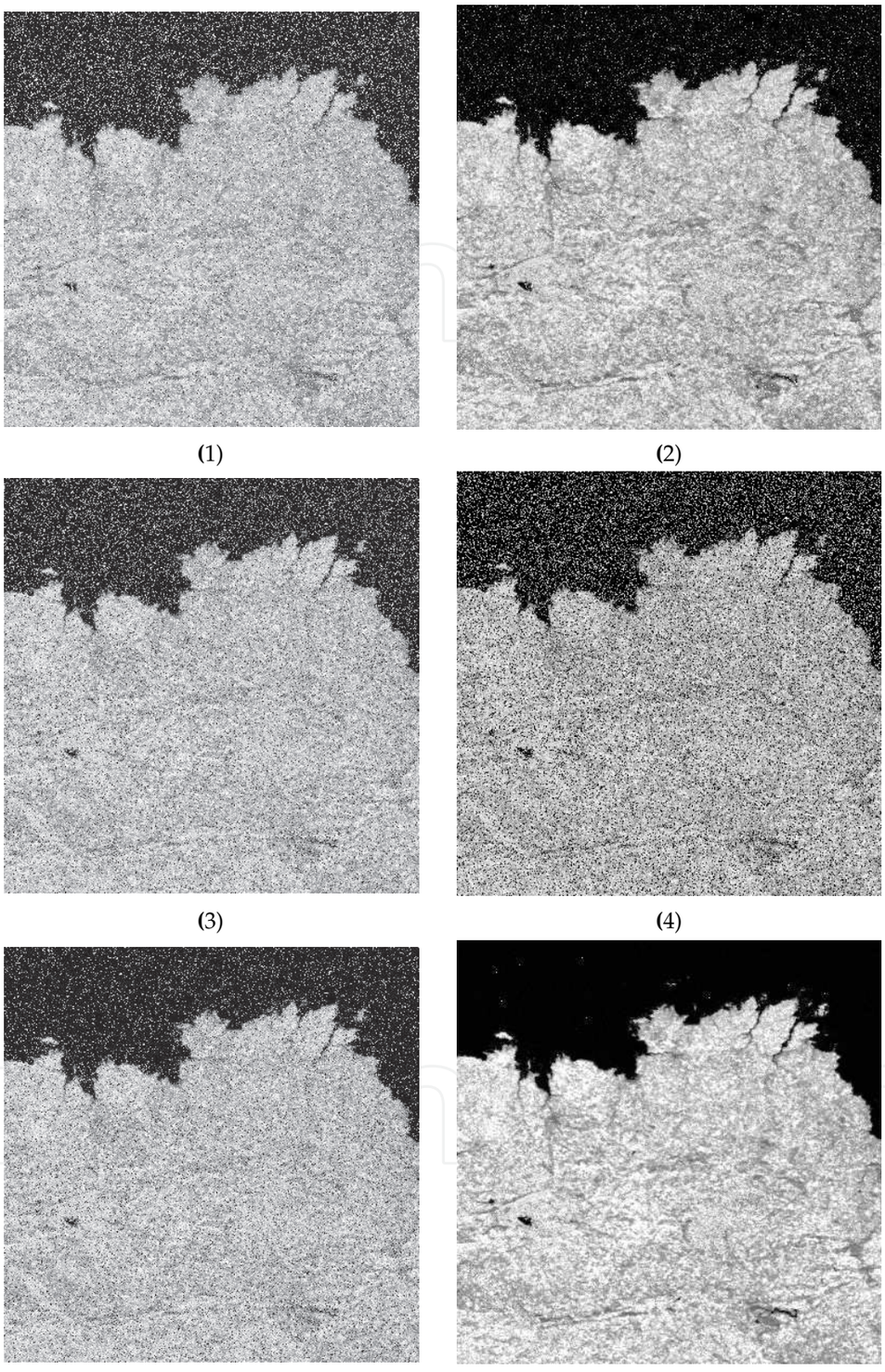


Fig. 13. The simulated cloudy near infrared channel of a MODIS image on 01-24-2003 along with its reconstruction by the SOM1 using different similarity measures: (1) simulated cloudy image, (2) using the Euclidean distance, (3) using the SAM measure, (4) using the SCM, (5) using the SID measure, (6) using the proposed similarity measure.



(1) (2)

(3) (4)

(5) (6)

Fig. 14. The simulated cloudy near infrared channel of a MODIS image on 01-24-2003 along with its reconstruction by the SOM2 algorithm using different similarity measures: (1) simulated cloudy image, (2) using the Euclidean distance, (3) using the SAM measure, (4) using the SCM, (5) using the SID measure, (6) using the proposed similarity measure.

We can notice visually that the near infrared channel of Fig. 13-6 is the only output that may be accepted as a reconstructed version from the one of 12-(2). This image is the one reconstructed by the proposed similarity measure without the need to apply an outlier detector to the contaminated temporal profiles before the projection onto the code vectors.

TS5	(a) ED	(b) SAM	(c) SCM	(d) SID	(e) Proposed SM
mean	-.03575	0.03665	0.03606	0.03878	-0.0002
std	.05594	0.13173	0.14953	0.1218	0.01607

Table 3. Mean and standard deviation of difference images ($TS5_{a,\ldots,e}$) using different similarity measures in the projection phase of the SOM1 algorithm for missing values

Table 3 shows the superiority of the proposed distance measure to all other distance measures used in this experiment. The mean of the time series $TS5_e$, expressed in reflectance value, is two order of magnitude less than all other TS5s. Also the standard deviation is one order of magnitude better than all other TS5s produced by all other similarity measure except the one produced by the Euclidean distance similarity measure.
Fig. 14 shows results using the SOM2 algorithm. We can notice better performance of SOM2 algorithm with respect to the SOM1 algorithm.
Again, the most similar time series to the one in Fig. 12-(2) is 14-(6). It is almost free from the salt and pepper noise contrarily to all other images. The remaining salt and pepper noise, seen clearly in the Pacific ocean, are mainly due to the high contamination of their temporal profiles. Table 4 shows the average of the mean value and the average standard deviation, expressed in reflectance values, of each difference time series ($TS5_{a,\ldots,e}$) using the SOM2 algorithm.
We can also note from the two tables 3 and 4 that the proposed similarity measure gives the best results when using SOM1 or SOM2 algorithm. Moreover, all other similarity measures failed to reconstruct a reliable time series using the SOM1 algorithm. In other words, we have to use SOM2 algorithm when using any similarity measure except with the proposed similarity measure which may be used directly with SOM1 algorithm and still give better results than other similarity measures even with the SOM2 algorithm.

TS5	ED	SAM	SCM	SID	Proposed SM
mean	-0.00736	-0.00233	0.00228	-0.00246	-0.00018
std	0.03516	0.08345	0.09807	0.07718	0.01609

Table 4. Mean and standard deviation of difference images ($TS5_{a,\ldots,e}$) using different similarity measures in the projection phase of SOM algorithm for missing values and without taking into account erroneous components (SOM2 algorithm).

4.3.2 Temporal Profiles

In this section, results yielded from applying the proposed similarity measure and the Euclidean distance to selected temporal profiles are presented in Fig. 15. Each graphic in this figure represents 4 temporal profiles: a real temporal profile (that belongs to TS1), a contaminated temporal profile (that belongs to TS2), a reconstructed temporal profile using SOM1

algorithm along with Euclidean distance similarity measure in the projection phase and a temporal profile yielded from applying the SOM1 algorithm along with the proposed similarity measure in the projection phase.

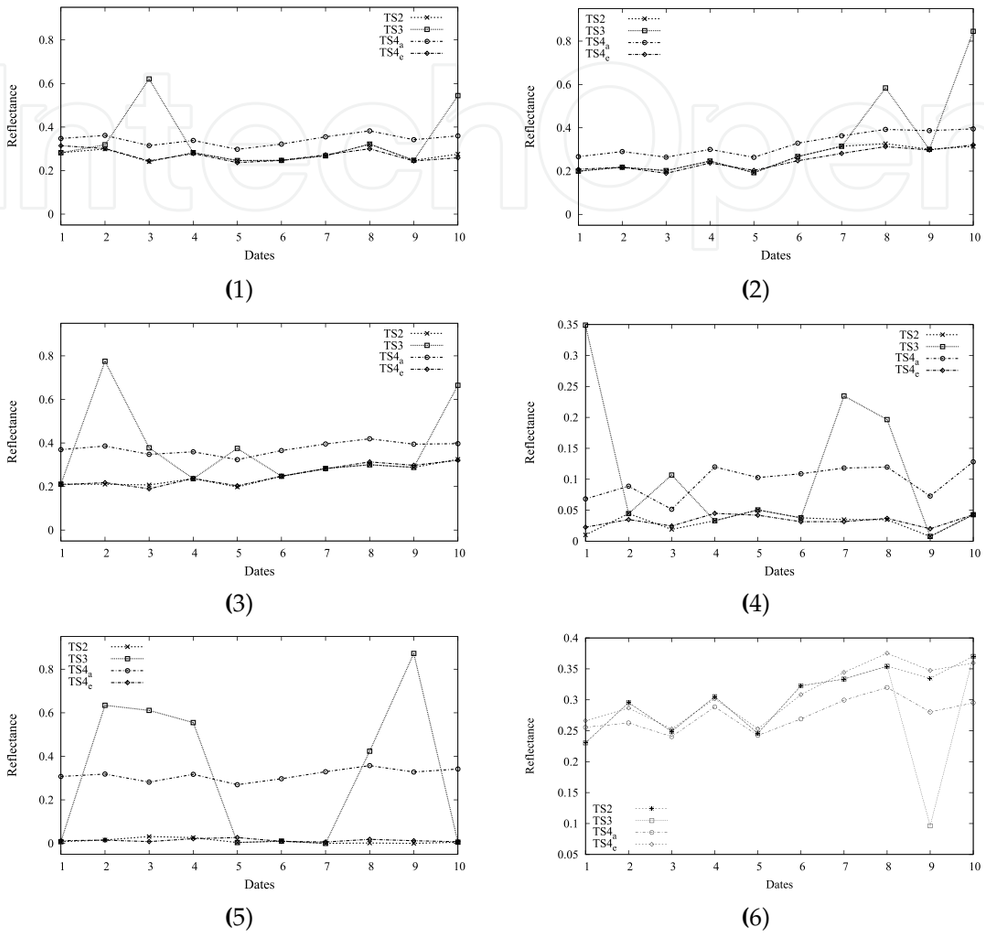


Fig. 15. Examples of matching the temporal profiles using the Euclidean distance and the proposed similarity measure. TS2 represents the original TP, TS3 represents the TP with the added clouds and shadows, $TS4_a$ represents the best matching TP from the SOM using the Euclidean distance and TS_e represents the candidate TP using the proposed similarity measure.

Temporal profiles have been selected to show different possible situations. Fig. 15-(1) and 15-(2) show the reconstruction of a temporal profile which was contaminated by a simulated cloud on two dates at different positions. Fig. 15-(3), 15-(4) and 15-(5) are the same as 15-1 but their original temporal profile was contaminated at four or five dates. The temporal profile in 15-(6) shows another situation where the simulated contamination was a shadow.
We can notice two things: firstly, the proposed similarity measure outperforms the Euclidean distance similarity measure in all situations. Secondly, the Euclidean distance similarity measure is affected by the large differences with respect to small differences and that temporal

profiles reconstructed using the Euclidean distance similarity measure is biased to the direction of contamination (clouds or shadows).

5. Conclusion

A technique to recover erroneous data was presented in this chapter. The application of the technique to temporal series of low resolution images affected by clouds and associated shadows has been studied. This non-parametric algorithm, which is based on the Kohonen's SOM, does not require any statistical models to fit the data. The training phase depends on data issued from the scene to be processed for a certain thematic analysis. The input data is to be considered as reflectance observations with as many spectral bands as allowed by the sensor. Significant results have been found by using a set of 10 images only. Some of these images were very close to each other in time, three of them has two days in-between only. In other words, no uniform sampling is necessary to perform such a temporal profile reconstruction.
In comparison to other methods used to compose time series in order to observe land use and land cover changes, the SOM algorithm along with an outlier detector, SOM2 algorithm, shows several advantages. On the contrary, the MVC has several drawbacks that can be found in the literature Cihlar et al. (1994); Eklundh (1995); Chen et al. (2003) even if it is the most popular technique. In fact, the association of pixels acquired with different zenithal angles and observation conditions creates patchwork artifacts. The MVC is commonly used on NDVI time series and not on reflectance time series whereas the SOM algorithm can be used on reflectance images. Some observations that have no need to be replaced are often replaced with the NDVI maximum value, which is not necessarily the best NDVI value acquired within a period.
A new Similarity measure has been presented to used in the projection phase in the SOM algorithm for missing values. The proposed similarity measure is derived from a non-Gaussian RBF kernel. This proposed similarity measure is performing better than four similarity measures that are widely used in the literature. It performs also with normal distributions as well as with the heavy-tailed distributions.

6. References

Abdel Latif, B., Lecerf, R., Mercier, G. & Hubert-Moy, L. (2008). Preprocessing of Low Resolution Time Series Contaminted by Clouds and Shadows, (7).

Bakar, Z., Mohemad, R., Ahmad, A. & Deris, M. (2006). A comparative study for outlier detection techniques in data mining, *Conference on the Cybernetics and Intelligent Systems*, pp. 1–6.

Chang, C.-I. (2000). An Information-Theoretic Approach to Spectral Variability, Similarity, and Discrimination for Hyperspectral Image Analysis, **46**(5): 1927–1932.

Chapelle, O., Haffner, P. & Vapnik, V. N. (1999). Support Vector Machines for Histogram-Based Image Classification, **10**(5): 1055–1064.

Chen, P., Srinivasan, R., Fedosejevs, G. & Kiniry, J. (2003). Evaluating different NDVI composite techniques using NOAA-14 AVHRR data, *International Journal of Remote Sensing* **24**(17): 3403–3412.

Cihlar, J., Manak, D. & Voisin, N. (1994). AVHRR bidirectional reflectance effects and compositing, *Remote Sensing of Environment* **48**(1): 77–88.

Cottrell, M. & Letrmy, P. (2005). Missing values: processing whith Kohonen algorithm, *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, Brest, France, pp. 17–20.

Eklundh, L. R. (1995). Noise estimation in NOAA AVHRR maximum-value composite NDVI images, *International journal of remote sensing(Print)* **16**(15): 2955–2962.

Fessant, F. & Midenet, S. (2002). Self-Organising Map for Data Imputation and Correction in Surveys, *Neural Computing & Applications* **10**(4): 300–310.

Huang, C., Townshend, J., Liang, S., Kalluri, S. & Defries, R. (2002). Impact of sensor's point spread function on land cover characterization, assessment and deconvolution, *Remote Sensing of Environment* **80**: 203–212.

Ingram, K., Knapp, E. & Robinson, J. (1981). Change-detection technique development for improved urbanised area delineation, *CSC/TM-81/6087*, Report prepared for NASA, Computer Sciences Corporation, Springfield, Maryland.

Jha, C. S. & Unni, N. V. M. (1994). Digital Change Detection of Forest Conversion of a Dry Tropical Forest Region, *International Journal of Remote Sensing* **15**(13): 2543–2552.

Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heidebrecht, K. B., Shapiro, A. T. & Barloon, P. J.and Goetz, A. F. H. (1993). The spectral image processing system (SIPS)–interactive visualization and analysis of imaging spectrometer data, **44**: 145–163.

Tanre, D., Deroo, C., Duhaut, P., Herman, M. & Morcrette, J. J. (1990). Description of a computer code to simulate the satellite signal in the solar spectrum - The 5S code, *International Journal of Remote Sensing* **11**: 659–668.

van der Meer, F. (2006). The Effectiveness of Spectral Similarity Measures for the Analysis of Hyperspectral Imagery, *Internation Journal of Applied Earth Observation and Geoinformation* **8**: 3–17.

Self-Organizing Maps

Edited by George K Matsopoulos

ISBN 978-953-307-074-2
Hard cover, 430 pages
Publisher InTech
Published online 01, April, 2010
Published in print edition April, 2010

The Self-Organizing Map (SOM) is a neural network algorithm, which uses a competitive learning technique to train itself in an unsupervised manner. SOMs are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space and they have been used to create an ordered representation of multi-dimensional data which simplifies complexity and reveals meaningful relationships. Prof. T. Kohonen in the early 1980s first established the relevant theory and explored possible applications of SOMs. Since then, a number of theoretical and practical applications of SOMs have been reported including clustering, prediction, data representation, classification, visualization, etc. This book was prompted by the desire to bring together some of the more recent theoretical and practical developments on SOMs and to provide the background for future developments in promising directions. The book comprises of 25 Chapters which can be categorized into three broad areas: methodology, visualization and practical applications.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen