ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Images Reconstruction using an iterative SOM based algorithm.

M.Jouini[1], S.Thiria[2] and M.Crépon[3]*

1- LOCEAN, MMSA team, CNAM University, Paris, France
2- LOCEAN, MMSA team, UVSQ University Paris, France
3- LOCEAN, MMSA team, CNRS, Paris, France

Abstract. The frequent presence of clouds in optical remotely sensed imagery prevents space and time continuity and limits its exploitation. The aim of this study is to propose a new statistical processing approach for the reconstruction of areas covered by clouds in a time sequence of optical satellite images. The approach is an iterative SOM based algorithm and is applied here to reconstruct ocean color images. It used the information contained in ocean color images and a set of satellite-derived dynamic ocean products (sea surface temperature: SST, altimetry: SSH) to reproduce the local spatio-temporal relationships of the cloudy images. The reconstruction method is general and can be extended to similar problems.

1 Introduction

Optical satellite observations are limited by clouds preventing space and time continuity of such data. Estimating missing values is therefore a crucial problem when working with optical satellite images. In this context, many techniques for recovering missing data have been proposed. The standard one is the construction of spatial or temporal composite images by combining images from multiple sensors or aggregating them over time. Some other methods based on optimal interpolation (Smith et al. [1], (1996), Beckers and Rixen [2]), or on multi-resolution analysis dealing with turbulence properties (Pottier et al [3]) have been developed to reconstruct under clouds optical satellite images. The performances of all these reconstruction methods are usually limited by the percentage and diversity of cloud coverage. These methods need 60% of non cloud contaminated pixels to be efficient at least. An alternative approach to reconstruct satellite images contaminated by clouds whose cloud coverage is greater than 60%, is to take into account their statistical properties in space and time and to interpolate them by using appropriate methods. In the present study, we explored the potentiality of Self Organizing maps (SOM, Kohonen [4]), which are efficient unsupervised neuronal classifiers commonly used to solve environmental problems, for filling the gaps due to clouds. Our approach is general and can be applied to a large variety of images in geophysics and more generally when multi-dimentional and correlated information can be used. In this paper, we focus on a particular problem that is the reconstruction of ocean color images in order to retrieve the missing Chlorophyll-a concentration (CHL). The paper contains 3 sections. The first one presents the SOM based algorithm for the reconstruction of missing data in ocean color images, the second deals with the

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

performances obtained on a set of daily ocean color images (CHL) through the Ouest of Africa during the winter season of 2003. Section 3 is devoted to a conclusion.

2 The reconstruction methodology

2.1 Approach

Chlorophyll (CHL) variability at the ocean surface is measured by ocean color satellites and is strongly coupled to ocean dynamics (Abraham et al [5]). As physical (dynamical) parameters and biological (here Chlorophyll) information at comparable temporal and spatial scales are highly related and can be observed from space, we used this knowledge to learn a set of typical situations (combining dynamical and biological information). In this context, we associated ocean dynamics parameters as sea surface temperature (SST) and sea surface height (SSH) with the CHL values. The principle is to find the most probable ocean situation matching the noisy chlorophyll present situation (the situation containing missing CHL values) from a learning dataset that includes a set of large variety of possible situations of (CHL, SST, SSH). The dynamics and biological information are clustered into a large number of significant classes reproducing at most the learnt situations. Each class represents a specific spatio-temporal environment whose knowledge allows the CHL missing data estimation. Since spatial structures of the fluid play an important role in the equations governing the evolution of ocean variables, we associated the (CHL, SST, SSH) variables characterizing an ocean situation with their spatial context. We considered, therefore, that the three variables CHL, SST and SSH are distributed on a 3x3 spatial window that implicitly models theses parameters. Besides, since CHL, SST and SSH are time dependent variables, we decided to characterize a [CHL, SST, SSH] situation by 2 successive images observed at time t0-4, t0. Thus a situation is a vector of dimension 54. The classification of the (CHL, SST, SSH) situations is done using the SOM (Self organizing map) algorithm which is based on a neural network structured into two layers. The first layer represents the input layer that receives the data (here the [CHL, SST, SSH] situations). The second one is a neurons grid, usually 2 dimensional, with a topological ordering of [CHL, SST, SSH] typical situations. It summarizes the information contained in the multivariate learning set $L \subset D \subset \Re^N$ by producing a small number m of reference vectors W_j ($0<j\leq m$) that belongs to D and are statistically representative of the learning set. Each neuron represents a subset (or a class) of L that gathers data having common statistical characteristics, which are synthesized by its reference vector W_j($0<j\leq m$). The reference vectors W_j ($0<j\leq m$) are determined from L, through a learning process (Kohonen [4]) by minimizing a non-linear cost function. For a given training pattern $p \in L$ presented to the network, the Euclidian distance with all the reference vectors are computed and the closest reference vector W_j is selected. This reference vector is called the best matching unit (BMU) and its associated neuron is denoted the winning neuron. After finding one BMU, all the reference vectors W_j of the SOM are updated: the BMU and its topological neighbors are moved in order to better match the input vector. At the end of the training process, the SOM map provides topological relationships between all the different neurons (or classes). A classification can be thus applied to analyze new situations or incomplete (noisy situations) situations of [CHL, SST, SSH]. The

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

analysis of a new situation is done by introducing it to the input layer and computing its BMU. If we consider a situation p for which the CHL is unknown at a pixel i, its CHL value is determined by:
1. Computing the distance to the referent vectors using a "restricted euclidian distance" which only considers the existing components of p.
2. Selecting the BMU for this restricted distance
3. Assigning the CHL content of the BMU affected to the pixel i to the situation p.

2.2 **The learning process**

We selected a set of daily satellite CHL, SST and SSH sub-images at a resolution of 9km of 75x104 pixels overlapping the intergyre zone for the winter season of 2003. CHL and SST images were observed by the MODIS satellite (http://oceancolor.gsfc.nasa.gov/reference) while SSH ones were provided by the Mercator model (www.mercator-ocean.fr/) assimilating Topex and Jason altimeter observations. We dedicated, for each variable (CHL, SST and SSH), 68 images to the learning process, for which the cloud coverage in SST and CHL images is less than 80%. We kept 8 images for which the cloud coverage is less than 30% to quantify the method reconstruction performances. The learning data set (L) is thus constituted by randomly selecting 1/3 of the situations (197,600 situations) constructed from the 68 dedicated images, the remaining ones (2/3 of the initial situations) being used for sensitivity tests. As, heavy compact clouds can cover CHL and SST images during a long period, the [CHL, SST, SSH] situations hence generate incomplete vectors whose size is too small for an accurate retrieval process. The BMU selected by applying the "restricted euclidian distance" may not find the best matching situation (the best referent vector). At the end of the learning phase, each neuron was associated with a referent vector whose component computed from remote sense observations might be noisy. Indeed, the referent vectors may have clustered incomplete situations for which some (even all) components are missing. Figure 1 (on the left) shows the percentage of captured vectors for which CHL values are missing, for each neuron of the SOM map obtained from the first Learning phase. This percentage averages the 60 %. To improve this learning process, we have developed an iterative algorithm, which aims to estimate, in an iterative way, missing data in the learning data set (CHL and SST data). Let us denote SOM-S0 the first map learnt above. We assume that the [CHL, SST, SSH] images related to a given day have an "internal coherence". We projected all the situations of that day onto SOM-S0; we were therefore able to replace missing CHL or SST values of one situation for the considered day and affected to one BMUd by the mean of known CHL or SST values of the same day and projected on the same BMU (BMUd). The learning data set with these reconstructed CHL or SST values is considered as a new learning data set. We then learned a second SOM map (SOM-SIter1) using this completed learning data set. The process was repeated until the percentage of CHL missing values, which is initially of about 60%, reached 15% for the CHL (Figure1). It basically took three iterations to reach this minimum percentage. Then, the last SOM trained map SOM-SIter3 was used to reconstruct test images.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

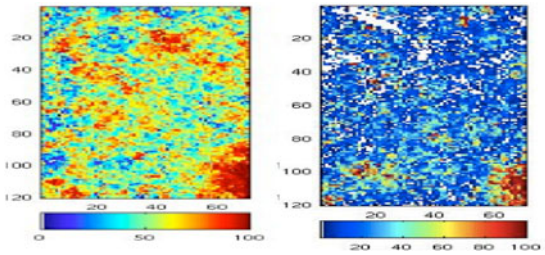

Fig. 1: Percentage of CHL missing values captured by each neuron on SOM-S0 (left) and SOm-Siter3 (right). The white colors correspond to neurons with 0% of CHL missing data.

2.3 **The reconstruction phase**

We used SOM-SIter3 to complete the cloudy situations of the test images. SOM-SIter3 was able to give CHL values for almost all situations (15% of missing values at the end of the reconstruction). Computation of performances has shown that the CHL value at time t0 at the central pixel of the BMU is not necessarily the best estimate of the CHL situation. This best value might be in the SOM-SIter3 neighborhood of the BMU in most of the cases. In order to increase the retrieval performances, we decided to use the topological order induced on SOM-S-Iter3 in combination with the spatio-temporal neighborhood relationships of the satellite images.

• We considered that the BMU for pixel i given by the "restricted Euclidian distance" is not the most reliable one and that its neighbors on the SOM map are also possible candidates. More precisely, we considered that the six first BMUs given by the "restricted Euclidian distance" and for each one, the eight nearest neighbor neurons (3x3 window) on the SOM map are possible candidates. Doing so, we have at least 6x(8+1) possible neurons as BMU. We decided to take the most frequently solicited neuron among that 54 neurons during the procedure as the candidate BMU (denoted BMU* in the following) for one pixel I to restitute.

• Let us consider a pixel i at t0 for which we want to estimate the CHL value. According to the coding used for an input vector, the pixel i can be associated with two different input vectors representing the ocean situations respectively at (t0-4, t0) and at (t0, t0+4). For both vectors, we now consider the situations associated with the 8 pixels of the spatial neighborhood of i (3x3 window) on the satellite image. Each pixel of this neighborhood can be considered as closely related to one situation representing the pixel i (Figure 2). Doing so, each situation at (t0-4, t0) and (t0, t0+4) generates respectively 8 new situation vectors. Finally, a pixel i is associated with (1+8)x2 distinct situations. By projecting these 18 ocean situation vectors on SOM-SIter3, we obtain 18 BMU* at the most, which are related to the same pixel i. Each BMU* corresponds to a referent vector associated with a subset of situations captured during the learning phase. The most probable CHL value of one pixel i is then the CHL median value of the CHL situations captured by the 18 BMU*.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

3 Application: reconstruction performances in the case of heavy clouds coverage

To evaluate the reconstruction performances of the proposed approach, several sensitivity tests were made on respectively the 2/3 situations kept in the learning process (section 2.2) and the 8 test images. The example of one image of the test data set shown below demonstrate the feasibility of the method on real data and provide a qualitative (visual) and quantitative evaluation of the reconstructed image quality in the case of heavy compact coverage (100% clouds). We focus thus on an image of the test data set taken the 21st December (Figure 3).

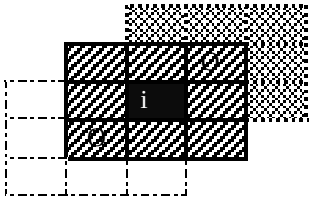

Fig. 2: Representation of the possible situations associated with one CHL pixel i to restitute. The eight pixels surrounding i (deep grey) are candidates. The eight pixels surrounding each candidate are potential candidates also. We have drawn the potential candidates (light grey and white) associated with two candidates (denoted O) surrounding i.

We assume that this image is totally covered by clouds at *t0* (*t0* corresponds to the 21st of December 2002). For reconstructing the CHL values at *t0*, we used the CHL images at respectively *t0-4* (17th of December) and *t0+4* (26th of December) and the SST, SSH images at *t0-4*, *t0*, *t0+4*. We first visually compared the CHL reconstructed image (CHL-SRe, on the right of Figure 3) to the CHL Modis image (CHL-S on the left of Figure 3), on the left of Figure 13). Original and reconstructed images are very similar showing that the SOM estimation works correctly and is able to reproduce most of the patterns of the original image in the case of CHL images heavily affected by clouds. Table 1 compares the initial CHL image (CHL-S) of the 21th December to the reconstructed one (CHL-SRE) using statistical estimators (mean CHL concentrations, relative error err-rel and the Kullback distance D (Kullback et al. [6])) which measures the similarity between two probability distributions p and q (here p for the distribution of the Modis CHL image and q the distribution related to the CHL reconstructed image, equation 1). Similar reconstruction performances were obtained for the 8 test images.

$$D(p,q) = \sum_i p_i \log \frac{p_i}{q_i} \quad (1)$$

4 Conclusion

We have applied, in the present study, an iterative SOM based algorithm to reconstruct CHL satellite observations masked by clouds, which is a serious problem for observing the CHL patterns and their associated quantities with satellite ocean

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

color sensors. The SOM iterative process performed well and was able to reconstruct the missing CHL concentration with a good accuracy comparing to the "standard SOM algorithm". In this context, different time samplings of the observations with respect to the missing data were investigated on both learning and test data set.

	Log_{10}CHL mean concentration ($Log_{10}mg/m^3$)	CHL mean concentration (mg/m^3)	Relative error err_rel (mg/m^3)	Kull. Dist. D (mg/m^3)
CHL-S	-0.84	0.17	REF	REF
CHL-SRe	-0.81	0.19	0.25	0.07

Table 1: Similarity criteria (Mean CHL concentration expressed in respectively Log_{10} (mg/m3) and (mg/m3), relative error and Kullback distance calculated both on CHL (mg/m3)), comparing the initial CHL image of the 21th December (1st row) and the reconstructed one (2nd row). The cells denoted REF indicate that the CHL Modis image is taken as reference of comparison for both relative error and Kullback distance.

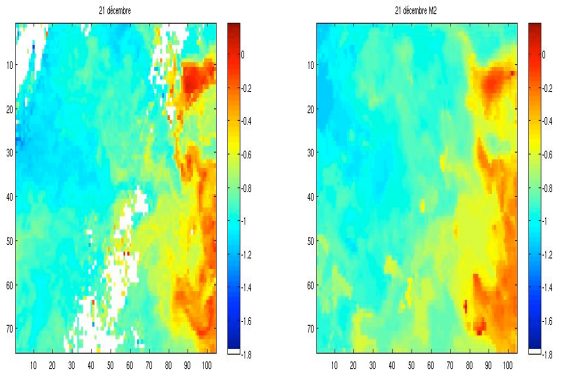

Fig. 3: CHL-S image of the 21st December (left), CHL-SRe image (right). The CHL values are Log_{10} transformed.

6 References

[1] Smith, T. M., Reynolds, R. W., Livezey, R. E., & Stokes, D. C. (1996). Reconstruction of historical sea-surface temperatures using empirical orthogonal functions. Journal of Climate, 9, 1403−1420.

[2] Beckers, J., & Rixen, M. (2003). EOF calculations and data filling from incomplete oceanographic data sets. Journal of Atmospheric Oceanography Technologies, 20, 1839−1856.

[3] Pottier C, A. Turiel, V. Garçon (2008). Inferring missing data in satellite chlorophyll maps using turbulent cascading. Remote sensing of Environment, 112, 4242-4260

[4] Kohonen T. (2001). Self Organizing Maps (3rd ed.). Berlin Heidelberg: Springer Verlag ; (501 pp)

[5] Abraham, E. R. 1998 The generation of plankton patchiness by turbulent stirring. *Nature* **391**, 577–580.

[6] Kullback,S. Haykin, editor. Unsupervised *Adaptive Filtering vol.1: Blind Source Separation*, John Willey ans Sons, New York, 2000.