Insights on the interpretation of SOM and U-Matrices with an example clustering based in oceanographic data

Fernando Jorge Pires[1], Victor Lobo[1,2], Fernando Bação[2]
[1]Portuguese Naval Academy, [2]ISEGI/UNL

SUMMARY

In this paper a process for the detection of clusters in oceanographic data is described. The application to oceanographic data is relevant as it allows the improvement of the understanding of the phenomena occurring in the Portuguese coast. Additionally, the application also illustrates how the self-organizing maps maybe used to explore and explain clusters, especially emphasizing the relevance of the visualization process in this context.

INTRODUCTION

Self-organizing maps (SOM) have been successful addressing many different types of problems. This is especially true in the context of clustering tasks (Kohonen 2001; Bacao, Lobo *et al.* 2005). Nevertheless, one issue has remained as a bottleneck in the dissemination of the SOM's as analysis tools: the subjectivity of the analysis leading to the definition of the number of clusters. The difficulty in interpreting the SOMs has lead to the use of more traditional clustering methods such as k-means and hierarchical clustering methods. These methods have some drawbacks, most notably they provide little indication about the data structure, making it difficult for the user to *understand* the data. In this paper we present some SOM based clustering techniques which enable a better understanding of the data structure.

The study of atmospheric and oceanographic phenomena is based on modelling techniques characterized by their strong nonlinearity and complexity. The models also need to incorporate large volumes of data, which cover large geographic areas, in order to capture the macro behaviour of the atmosphere, of the ocean and the interaction atmosphere-ocean. One way of reducing both the complexity and the amount of data needed to characterize the water and air masses, would be through the identification of homogeneous areas in terms of all the variables involved in the study.

Characterizing water and air masses, looking for regions of homogeneity, involves dealing with 3 dimensional data, which poses some problems as most of the geo-referenced applications deal with 2 dimensional data. Thus it is necessary to look for other approaches to accommodate the analysis and evaluation of clustering results in a 3D context. Good solutions for the analysis of 3D data can be relevant in a wide variety of problems such as atmospheric and ocean studies (pollution/meteorology), and also in emergent fields such as GIS in medicine in which the navigation through the human body is necessary.

In this paper we evaluate some SOM based clustering solutions using climatologic data from the Atlantic North from the National Oceanographic Data Centre (NODC) (NODC 2006), the geographic extent of the study region is presented in Figure 1. We choose 2 non-geographic features: salinity (S) and temperature (T). The reason for choosing these two variables is related with their importance in determining the oceanic circulation. Each data point is also characterized by it's x (longitude), y (latitude), and z (depth) coordinates. The selected data was separated into winter and summer data. The geographic region chosen has some interesting properties which will help evaluate the characteristics of the visualization tools that are proposed in this paper. The oceanographic characteristics of this area are particularly interesting as it is here that the water masses from the Mediterranean Sea enter the Atlantic Ocean. The collision of these two water masses creates a unique

setting characterized by a complex interaction in salinity, temperature, and depth, originating a rich "transition layer".

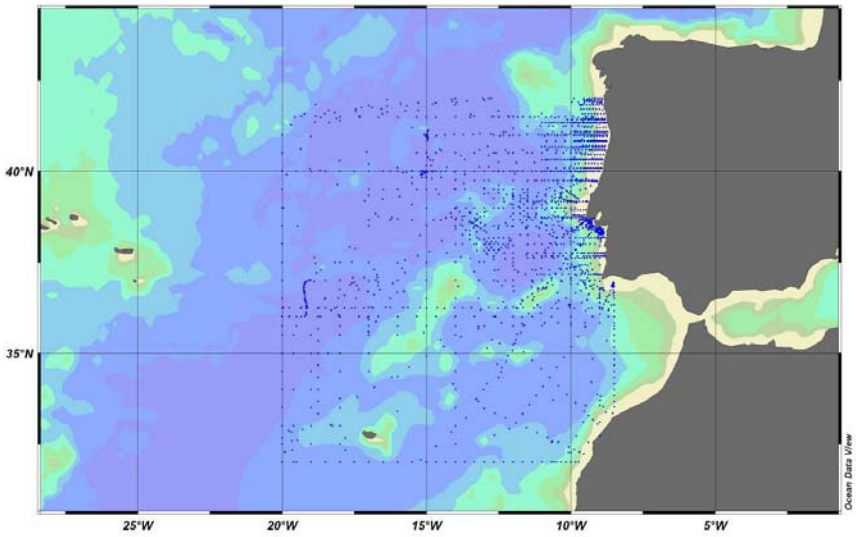

Figure 1: Geographic distribution of the data.

We shall cluster this data using two different approaches. First we will plot the data points in the S/T plain (salinity versus temperature), to see if any well defined clusters are visible. This is what we call "visual clustering". It must be noted that this "visual clustering" is only possible because the data is 2-dimensional. This allows us to understand the clusters that will later be obtained by the SOM, which normally would be used to process higher dimensional data.

Next we use a Self-Organizing Map (SOM) to process the data. The SOM used for this purpose has a large number of units (far more than the expected number of clusters), constituting what is sometimes called an "Emergent SOM" (Ultsch 2005). The SOM will project the original data points onto an "output space", which is 2-dimensional even when the original data is not, and where we hope to identify the data more clearly. To visualize the "output space", and identify the clusters, we use a U-Matrix (Ultsch and Simeon 1989). We shall see that using this approach it is easier to detect the clusters, and that by visualizing the clusters detected by the U-Matrix in the original S/T plain we can understand what they correspond to.

VISUAL CLUSTERING OF THE DATA USING ONLY TEMPERATURE (T) AND SALINITY (S)

Visual clustering (when possible), despite being very simple and "low tech", is the most intuitive and actually the best way of identifying clusters. We may simply plot the data, and see if a clear pattern of "groups" emerges. Our data is characterized by only two non-geographical variables (S and T), so we may plot each datum as a point in the plain defined by S and T. In Figure 2 we can see two graphs depicting the distribution of the data in that plane in Winter (left) and in Summer (right).

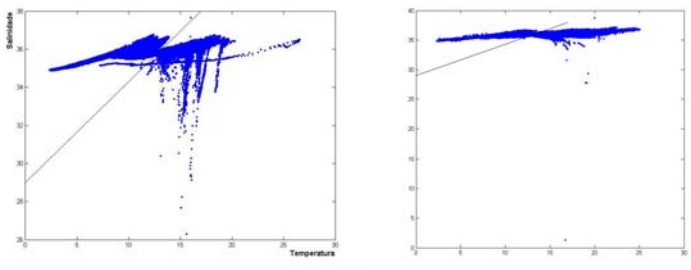

Figure 2: Distribution of data in the S/T plane for Winter (left) and Summer (right) the line indicates de partition found between two main clusters, in both cases.

In the plot of the Winter data, we can visually identify two "blobs" of data, that correspond to two distinct clusters, and we may draw a line (represented in black) separating those two clusters. Although we can visually separate these two clusters, the distinction between them is quite fuzzy. In the case of the Summer data the distinction is almost impossible without prior information. If we use *k*-means clustering (with *k*=2), the clusters found are basically the same, although the angle of the dividing line will be slightly different.

These clusters, obtained by visual inspection of the Winter data, make sense in the context of traditional oceanographic analysis: the two large clusters correspond to the well known surface and deep water layers. To confirm that the clusters do indeed correspond to these layers, we may plot their data points onto a geographical map (with x,y,z coordinates), using red to represent the points of one cluster, and blue for the other. This map is shown in Figure 3. We can clearly see that one cluster corresponds to the surface layer (in red) and the other to the deep water layer (in blue).

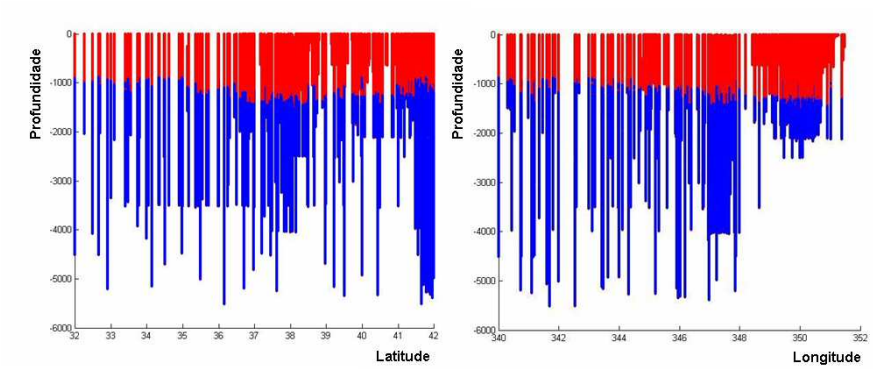

Figure 3: Projection of the winter data clusters onto different views of the 3D geographical space. The data points of one cluster are represented in red, and the other in blue. On the left we may see depth versus latitude, and on the right depth versus longitude. It is evident that depth is the main factor in determining the clusters, although the border between classes also depends on other factors.

In Summer the distinction between these two clusters is not so abrupt, since there is a smooth continuum between them. Although in this case it is difficult to identify the clusters, we may still confirm that the cluster boundaries (in terms of S and T) determined for Winter do make sense in this case.

In both, Winter and Summer data, "visual clustering" does not identify another well known, but rather small, layer: the transition layer (which is particularly important for sound transmission problems). Thus we can conclude that although 2D visual clustering can be quite useful it also may be insufficient in cluster detection, even when dealing with low dimensional data, such as the case here. In the next section we use a SOM to cluster the data and consider the gains that it can bring to the cluster detection process.

CLUSTERING WITH A SOM, BY VISUALIZING THE U-MATRIX

Next we processed the data using a SOM in order to assess if it would improve our understanding of this dataset. More specifically, we would like to identify the clusters in Summer, and detect the transition layer that we know exists. To this end we trained a SOM with the available data, and then obtained and visualized it's U-Matrix (Ultsch, Guimarães *et al.* 1993), or UMAT for short. The SOM used had 20*10 units, and the UMAT and Component Planes obtained are presented in Figure 4.

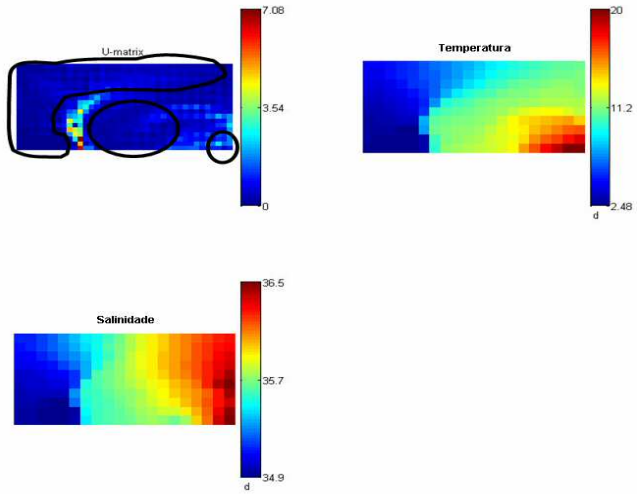

Figure 4: U-Matrix (UMAT) and Component Planes of the data concerning Summer.

In that UMAT we may see three regions of dark blue, that correspond to low values in the UMAT, and hence to clusters in the data. These regions are separated by lighter colours all the way up to bright red, which correspond to separations between the clusters. Additionally, the analysis of the Component Planes helps not only confirm the existence of the clusters but also characterize its nature. The Component Plane maps the value of each unit in each of the input variables. In Figure 4 it is quite obvious the distribution of Salinity and Temperature, for instance we can see that the lower right cluster is characterized by high values of temperature and salinity. Using the Component Planes it becomes apparent that in between the two main layers, with contrasting values in terms of temperature and salinity, a third one emerges and is characterized by a transition in the variables values.

The three regions visible in the UMAT correspond to (from the left to the right):

1- The Depth layer (low T and S values)

2- The Transition layer (intermediate and less correlated T and S values)

3- The Surface layer (high T and S values)

Thus through the SOM it is possible to visually identify an additional cluster, which was not apparent in the visual inspection of the input space. Since the data is 2-dimensional and we can easily see the clusters by plotting them, this identification is particularly relevant *per se*. Nevertheless the method which allowed us to identify this third cluster can be important in the exploration of complex multidimensional datasets, where plotting the data in any 2-D sub-space does not reveal the clusters. For an experienced user of clustering with emergent SOMs, the interpretation is almost intuitive, and the actual data values that characterize the clusters can be easily obtained. However, it would be interesting to plot this UMAT, used for visual inspection of data and detection of clusters, into the original input space. This is what we do in the next section.

PROJECTING SOM AND UMAT IN THE INPUT SPACE

To understand the clusters obtained in the previous section, we may use dynamically linked windows, shown in Figure 5, that represent:

1- The UMAT, as it is normally shown. The axis of this UMAT have no direct connection with the axis of the input space. In this graph we visually identify clusters and separation zones.

2- The UMAT units and colors, superimposed on the original S/T plain. To perform this visualization, we simply identify the S/T coordinates of each unit, plot it in the S/T plain, and interpolate the color-code to obtain a smooth map. It must be noted that this step still needs improvement. Because the UMAT only has true data for the intervals between units, the UMAT presented is the result of interpolations.

3- The positions of the SOM units in the S/T plain.

This dynamical linking is shown (statically) with the arrows of the figure.

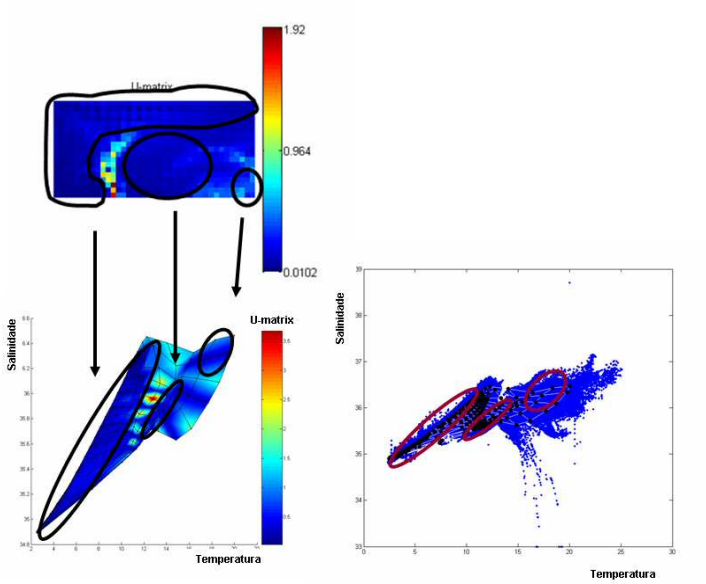

Figure 5: Dynamically linked graphs of the UMAT in som-space (top), in the input space of S/T (imediatly below), and the positions of the som units themselves in the S/T plane.

We can select a unit on the UMAT and it is automatically identified in the UMAT projection and the points classified in it are also identified in the 2D graph. We can see the largest cluster identified in the UMAT represented on the top and left of the SOM, corresponds to the low values of T and S characteristic of the deep water layer. But we can now distinguish two sub-regions in what visually we identified as a single cluster: the sub-group that is visible in the central zone, and the one visible on the bottom right corner of the UMAT. By seeing where the units of these two clusters lie, we can see that they do indeed correspond to the "right-hand-side" mega-cluster, but that there is a clear separation (represented by a clear blue), between them.

CONCLUSION

In this paper we showed the importance and usefulness but also the fragility of preliminary data visual inspection, which was possible, in this case, due to the low dimensionality of the dataset. We also showed that the SOM in conjunction with the UMAT allows an improved analysis of the data structure and characteristics. Without waving the usefulness of the visual analysis the SOM permits the identification of clusters that otherwise would pass unnoticed. This is especially important in high dimensional spaces. Greater dimensionality would only emphasise the relevance of the SOM in the cluster identification tasks.

Finally, we also showed that the Portuguese coast is quite homogeneous, both in latitude and longitude, and that the separation between characteristic types is done in depth. We observed, as expected, that there are two main water layers and that sometimes a third one can also be identified. The proposed method allowed a clear identification of the boundaries between these layers.

BIBLIOGRAPHY

Bacao, F., V. Lobo, et al. (2005). Self-organizing Maps as Substitutes for K-Means Clustering. Lecture Notes in Computer Science,. V. S. Sunderam, G. v. Albada, P. Sloot and J. J. Dongarra. Berlin Heidelberg, Springer-Verlag. 3516: 476-483.

Kohonen, T. (2001). Self-Organizing Maps. Berlin-Heidelberg, Springer.

NODC (2006). National Oceanografic Data Centre, http://www.nodc.noaa.gov.

Ultsch, A. (2005). Clustering with SOM: U*C. WSOM 2005, Paris.

Ultsch, A., G. Guimarães, et al. (1993). Knowledge Extraction from Artificial Neural Networks and Applications. Transputer-Anwender-Treffen, Aachen, Springer Verlag.

Ultsch, A. and H. P. Simeon (1989). Exploratory Data Analysis Using Kohonen Networks on Transputers, Department of Computer Science, University of Dormund, FRG.