Using Data Mining and Visualization Techniques for the Reconstruction of Ocean Paleodynamics

R. Theron[1], J. A. Flores[2], F. J. Sierro[2], C. Pelejero[3], J. Grimalt[3] and M. Vaquero[1]
[1]Departamento de Informática y Automática
[2]Departamento de Geología
Universidad de Salamanca, 37008 Salamanca, Spain
[3]CSIC, Barcelona, Spain

Abstract—**Ocean dynamics modeling is essential for predicting the impact of climatic change on human activities. The need of large time series (e. g., paleoclimatic data) has been exposed as one of the challenges to embark on decadal climate predictability. The goal of this study was to reconstruct the surface water dynamics in the China sea during the last 130,000 years. In order to do that it was necessary the assembly of integrated comprehensive datasets obtained from a gravity core recovered in the Sunda Slope (South China Sea) using different techniques such as quantitative analyses from coccolithophores, stable isotopes and biomarkers. Thanks to data mining, through a variety of data analysis tools, from Silicon Graphics' Mineset, such as Decision trees and Clustering we have established the variations in the water column stratification (relative position of nutricline/thermocline). This study is a good example of combining data mining and paleoceanography to explain some general paleodynamics, including short-time events, showing the potential to monitor and predict in the context of decadal time-series.**

I. Introduction

Over the last years we have witnessed how the claim of society for accurate climate prediction has increased; therefore, the climate predictability has emerged as one of the most powerful areas of research. Two approaches may be used: the development of new methods that will provide “best-guess” predictions; and a better understanding of the climate changes in the past that will lead to a more accurate ability of prediction. Along with atmospheric and land processes, ocean dynamics modeling is essential for predicting the impact of climatic change on human activities. A very interesting point is to understand how some mechanisms has contributed in the past to sudden climate changes and the need of large time series has been exposed as one of the challenges to embark on decadal climate predictability. There is only one record of climatic data with durations exceeding decades: the paleoceanographic record.

The goal of this study was to reconstruct the surface water dynamics in the South China sea during the last 130,000 years. In order to do that it was necessary the assembly [1] of integrated comprehensive datasets obtained from gravity core SU17961-2 (8º30.4'N, 112º 19.9'E; 1968 m water depth; 992 cm of core length), recovered in the Sunda Slope (South China Sea) using different techniques such as quantitative analyses from coccolithophores [2][3], Uk37 technique for the Sea Surface Temperature (SST) and $\delta^{18}O/\delta^{16}O$ ratio for the sea-surface salinity (S) [2]. This core shows a continuous record for the last 6 Marine Isotope Stages (MIS).

Data mining, through a variety of data analysis tools, can help to discover pattern relationships, exceptional and missing values in data that may be used to make predictions. These discoveries are easy to perceive when shown graphically, although we are restricted to showing many variables on a two-dimensional computer screen or paper. The new and powerful visualization tools require people to train their eyes in order to understand the information being conveyed [4].

In the next section it is outlined a data mining approach that helps paleoceanographers to model past events.

II. Mining Paleorecords

In the past decades a lot of research has been done on time-series and data mining has found a new opportunity to apply its methods and algorithms. Climatic change is not an uncommon field for data mining, and some [5] have treated the case of ocean time-series, but we have no news of these techniques applied to paleorecords study, that have been traditionally done through factor and spectral analysis [6].

Previous works in the area [3][4] guided the data analysis. The goal was to compare three variables trough time: temperature and salinity were already available, but there was the need to find a means of establishing the variations in the water column stratification. In order to do that, we analyzed the cocolitophore assemblage and estimated the ratio between upper photic layer inhabitants vs. the lower photic layer inhabitants. Equation (1) shows the water column stratification index, N.

$$N = \frac{\sum (placoliths < 3\mu m)}{F.profunda + \sum (placoliths < 3\mu m)} \qquad (1)$$

This study was partially supported by CICYT 1FD97-1148 and CLI-1002-CO2-02 grants.

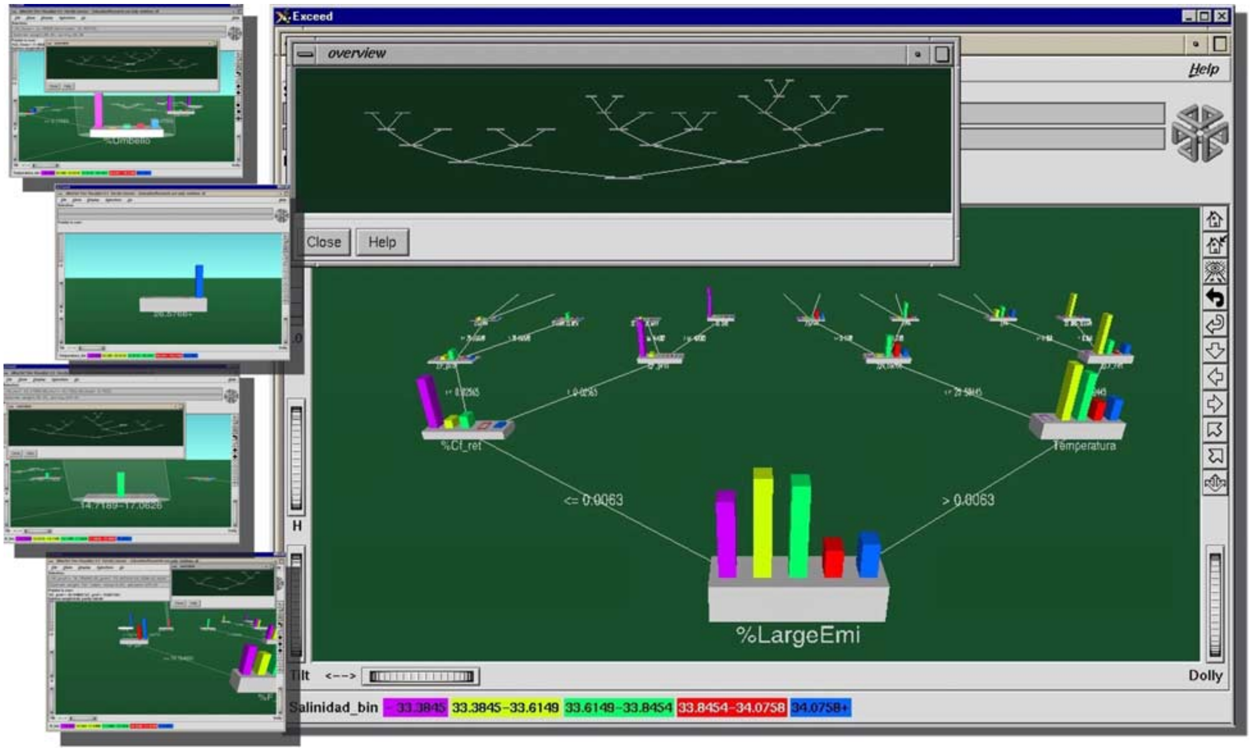

Fig. 1. Use of decision trees to find a water column stratification index

It was necessary to determine which particular coccolitophores species contribute more to the biological signal of water column stratification. We managed to do that using two techniques: a clustering algorithm [7], with which we obtained clusters of species behaving similarly trough time; and decision trees (Fig. 1), which led to rules establishing how different species where more or less affected by different conditions of temperature and salinity during different time spans. In the main window of Fig. 1 an example of the later is shown: as a first step, five levels of salinity were defined, then a decision tree (classifier) algorithm was used over the data of every isotope stage, eliminating the temperature data. In this example we can see that a low proportion of "Large" *Emiliania* (first branch on the left) during MIS 2 is a signal of low salinity (first column on the left of the histogram is only present on this branch). The same process was applied to obtain a decision tree for the temperature over each MIS studied (eliminating the salinity data). And, finally, the same was done for the index, N, eliminating both the temperature and the salinity data.

III. Results

As a final point, the three variables: salinity, temperature and water column stratification (N) were represented, using a splat (spatial aggregation that shows the distribution of the samples) visualization tool, through time. This way, we could study the correlation of the three variables through time, the three of them together, or by couples, using the shades of the samples for each plane. The results show a clear relationship between the different analyzed variables and samples. As a summary we consider here only from MIS 1 to 6 (including MIS 5 substages), as reference intervals (Fig. 2).

During the cold MIS 6 we observed a moderate to high surface water stratification and high salinity, although it shows fluctuations. Unfortunately, sampling resolutions in this interval prevent an accurate interpretation. During MIS 5e high temperature and salinity in a relatively stratified (deep thermocline position) environment, are the most characteristic features. MIS 5d show a similar pattern, but with high variability in the water-column stratification. During MIS 5c, a good correlation between low values of temperature and salinity, but relatively swallow thermocline. The same pattern observed for MIS 5d is shown during MIS 5b. MIS 5a is characterized by high fluctuations in temperature and salinity, but the thermocline was placed relatively deep. During MIS 4 and 3 a very weak surface water stratification and small changes in salinity were observed. This situation was amplified during MIS 2. During MIS 1 similar conditions to those observed in MIS 5 were found: high values of temperature and salinity, and a relatively deep thermocline.

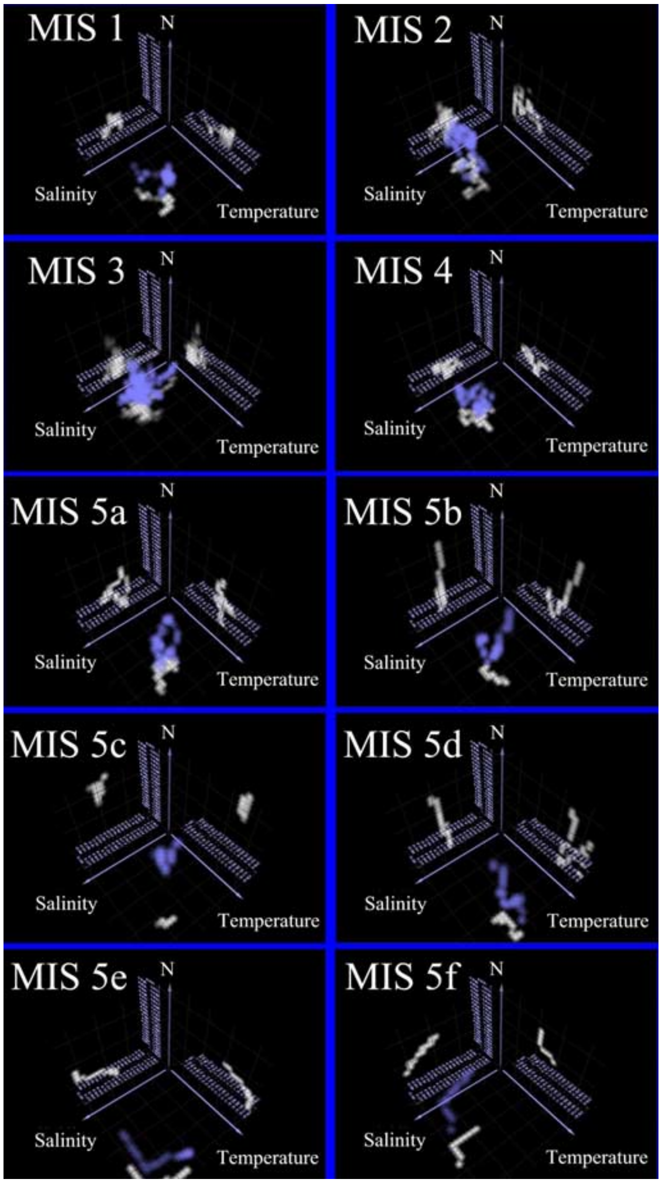

Fig. 2. Temperature-Salinity-N variation through MIS

The variability of the above-mentioned parameters, as well as the micropaleontological content is analyzed in the monsoon-dynamics context, combined with other factors such as the closure/opening of different straits in the region.

According with [2], glacial stages led to a stable estuarine circulation and a strong O_2-minimum layer via a closure of the Borneo sea strait. At the same time a strong seasonality and a large river input from Sunda shelf occurs during cold periods. Interglacials were marked by a strong inflow of warm water via Borneo sea strait, with a marked low seasonality. Our data suggest that for MIS 2, 3, 4 and 6 (cold periods), salinity show the lower values, in agreement with an increase in the fresh-water input from the Sunda river and closure of the Borneo strait. The position of the thermocline (nutricline) for these intervals was relatively deep (well stratified system). For the warmer periods MIS 1, 5a and 5e, salinity shows the highest values, and the thermocline position was also deep, maintaining a similar pattern to that observed for cold periods. The input of warm water through the Borneo strait, after its opening, can produce a similar effect to the stated before. During MIS 5b, 5d, and specially during 5c, we interpret a shallow position of the nutricline with the subsequent increase of mixing in the surface-waters.

IV. Conclusion and Future Work

In the paper we described how paleoceanographers can benefit from the use of data mining an visualization techniques. We have shown an example of reconstruction of ocean paleodynamics in the Soth China Sea. Due to the coarse resolution of the paleoceanographic record, sometimes we are unable to give further explanation of detected short-time events. As the spatial and temporal resolution of paleoceanographic data improves, tailor-made algorithms can provide better understanding.

References

[1] R. Theron, J. A. Flores, F. J. Sierro, M. Vaquero, and F. Barbero, “PaleoPlot: A tool for the analysis, integration and manipulation of time-series paleorecords,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2002, in press.

[2] Wang et al., “East Asianmonsoon climate during the late pleistocene: high-resolution sediment records from the South China Sea,” in *Marine Geology*, vol. 156, pp. 245-284, 1999.

[3] C. Pelejero, J. Grimalt, M. Sarnthein, L, Wang and J. A. Flores, “Molecular biomarker record of sea surface temperature and climatic change in the South China Sea during the last 140,000 years,” in *Marine Geology*, vol. 156, pp. 109-121, 1999.

[4] —, *Introduction to data mining and knowledge discovery*, 3rd edition, Two Crows Corporation, 1999.

[5] M. Steinbach, P. N. Tan, V. Kumar, S. Klooster, C. Potter, “Data mining for the discovery of ocean climate indices,” In *ACM international conference on knowledge discovery and data mining*, Workshop on Temporal Data Mining, 2001.

[6] R. Vautard and M. Ghil, “Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series,” in *Physica*, vol. 35D, pp. 395-424, 1989.

[7] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hal, 1998.