Data-Mining the Past Environment

R. Theron[1], D. Paillard[2], E. Cortijo[2], J. A. Flores[3], M. Vaquero[2], F. J. Sierro[3] and C. Waelbroeck[2]
[1]Departamento de Informática y Automática
Universidad de Salamanca, 37008 Salamanca, Spain
[2]Laboratoire des Sciences du Climat et de l'Environnement,
CEA-CNRS, Gif-sur-Yvette, France
[3]Departamento de Geología
Universidad de Salamanca, 37008 Salamanca, Spain

Abstract— **Currently, the Modern Analog Technique (MAT) is one of the most used techniques in paleoceanography and it is applied for the quantitative reconstruction of environmental conditions of the past. Through the calculation of distances between modern and paleo data, patterns (analogs) are found, and after an interactive analysis, paleoenvironmental features are reconstructed. PaleoAnalogs is a powerful and flexible tool, developed with Java technology, making it a multiplatform tool to be executed in any operating system; the tool can automatically take on the appearance and behavior of whatever operating system it happens to be running under; it is an interactive tool that permits 3D plots helping the analysis of three variables); it Includes 8 different types of distance measures; it is designed to carry out the reconstructions using different distance coefficients and for helping in the comparison between different results; and it provides a wizard for making associations (equivalent taxa, additions) of taxonomic categories between the modern data and the fossil data.**

I. Introduction

The data registered over thousands of years (mainly in ice and sediment cores) is an impressive source of information that, for instance, help us to model earth and oceans dynamics [1], first step to make climatic predictions. When looking for historic climatic data with durations exceeding decades, the largest and oldest record is found in the oceans. Palaeoceanographers need to manipulate, integrate and analyse time-series that are obtained from a number of independent techniques (such as ocean drilling, ocean tracers, AMC 14C datings, astronomic curves, etc.), which, moreover, are usually produced by different researchers and/or laboratories. This work is done with the aid of proper tools such as PaleoPlot [2] and AnalySeries [3].

Some of these data that are needed to understand paleoclimate are time-series of specific attributes related to the oceans. Thus, one problem scientists must face is how to know environmental parameters, such as Sea Surface Temperature (SST), that were present at each given past moment. For the reconstruction of this features, isotope measurements ($\delta^{18}O$) or biomarkers (U^{k}_{37} index) have been used. Some works [4] suggest that these techniques must be viewed with some degree of uncertainty. Gratefully, other independent techniques, based in microfossil assemblages, offer the possibility to obtain similar results. Since 1970s with CLIMAP [5], paleooceanographers have been trying to derive quantitative estimates using the distribution of fossil foraminifers present in the sedimentary record. After that, some better approaches have arisen, making the Modern Analogs Technique (MAT) [6] a standard in paleoceanography.

II. Finding Analogs

This section describes how PaleoAnalogs, a Java based program, carries out the method of modern analogs. It is assumed that the user has faunal census estimates of one or more fossil samples, the core file; and one or more sets of faunal data from modern samples with the related environmental features, the database file. Furthermore, the user must understand the taxonomic categories represented in the data sets, and be able to recognize taxa that are or may be considered equivalent in the analysis.

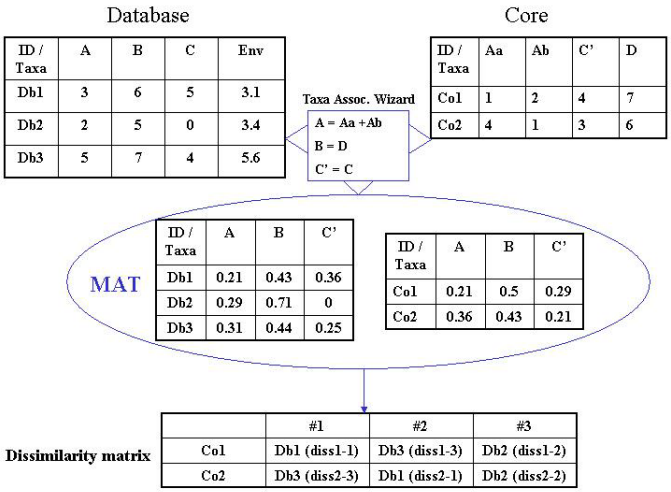

Figure 1. Finding modern alanalogs

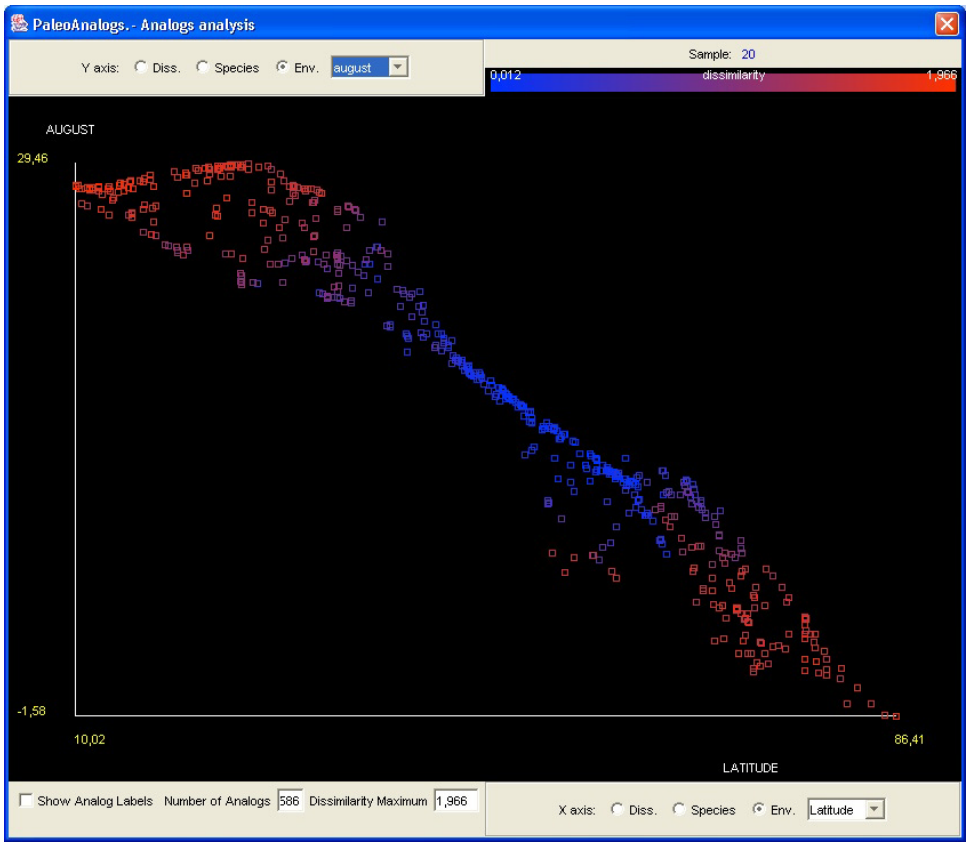

Figure 2. PaleoAnalogs interactive analogs analysis

The process begins after the selection of the core and database files (Figure 1); in general, these files will contain different taxa, both because different taxa are prevalent in different regions and because data providers use varying taxonomic categories (species and subspecies), names, and abbreviations. MAT requires that corresponding variables in different data sets be recognizable as such, otherwise it would be impossible to calculate the distance measures. Other tools, like ANALOG [7], solve this problem trough the use of complex files of transformation rules; with the help of the taxa association wizard (Figure 3) this problem is easily worked out, allowing the user to determine which taxa from each modern and fossil data files are compared, calculate proportions if needed, and identify the environmental features.

Once the database and the core data are transformed to have the same number and equivalent taxa, each sample in the core is compared with each sample in the database using a dissimilarity coefficient. PaleoAnalogs provides the user with the choice of distance measure among eight coefficients; this is particularly important, because with the bibliography standard in the field, most of the paleoceanographers do use only the squared chord distance, because either it is the unique distance offered in the program they have, or they don't know exactly what they are doing and simply collect the results.

Finally, using the distance measure selected by the user, a dissimilarity matrix is built. For each core sample *N* dissimilarity values are given, being *N* the number of samples in the modern database; these values are ordered increasingly so that each row of the matrix contains, left-to-right, the list of the *N* best analogs, that is, the database samples ordered by their alikeness to that particular core pattern.

III. Interactive Analysis and Recontructing Parameters

By clicking on a row of the dissimilarity matrix a 2D-plot for that core sample is shown. Figure 2 represents the analogs distribution for core sample '20'. This representation is enriched by adding color (blue-to-red) covering the range of dissimilarity values for the core sample's analogs.

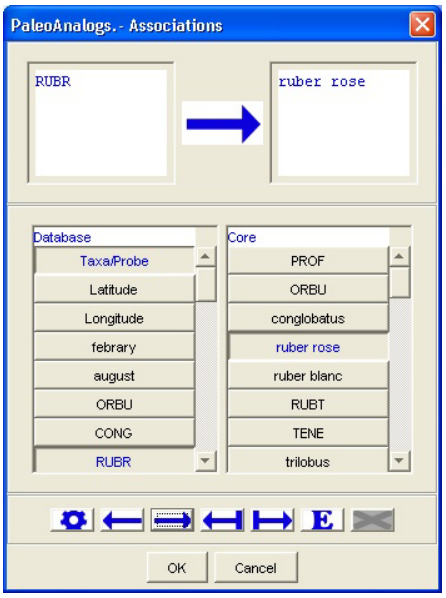

Figure 3. Taxa associations wizard

On each axis values are represented, from the modern database, of one of three categories: dissimilarity values, species or environmental features.

This 3-dimensional plots will help paleceanographers to have a better understanding of the reconstructions they are dealing with., In Figure 2 it can be seen a Temperature-in-August/Latitude/Dissimilarity plot; it shows how the best analogs for core sample '20' are concentrated in latitudes between 30º and 50º and temperatures between 9ºC and 19ºC, approximately.

Also, this representation can be done for a selected number of analogs or for those with dissimilarity values smaller than a cut-off. Analogs can be labelled with the associated database sample name, which might be of help when showing small number of analogs.

Finally, as it can be seen in Figure 4, a reconstruction using different parameters is produced.

IV. Conclusion and Future Work

This work is an example of how pattern recognition can be un invaluable tool in the paleoclimatology field. Thanks to it, reconstructions of paeloenvironmental conditions can gain in feasibility, simplicity and expeditiousness. However, a means of better understanding of the process must be established; this has been accomplished by an interactive analysis of the patterns (analogs) found. As future improvements of PaleoAnalogs, currently research interests are: including new micropaleontologic groups and assessing the results in comparison with biogeochemical techniques.

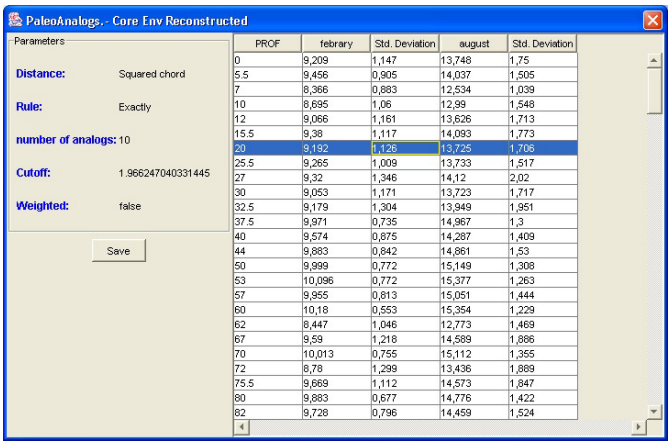

Figure 4. Reconstruction of paleo sea surface temperatures

Also, adding more standard methods of the field (RAM, ANNs, etc.), considering time as a fourth dimension in the interactive analysis, and developing and validating new methods that include Data Mining techniques are part of current research.

Acknowledgment

This study was partially supported by MCyT REN2002-11126-E / ANT grant.

References

[1] R. Theron, J. A. Flores, F. J. Sierro, C. Pelejero, J. Grimalt and M. Vaquero, “Using data mining and visualization techniques for the reconstruction of ocean paleodynamics,” in proceedings of the *IEEE International Geoscience and Remote Sensing Symposium*, vol. IV, pp. 2382-2384, 2002.

[2] 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 proceedings of the *IEEE International Geoscience and Remote Sensing Symposium*, vol. VI, pp. 3528-3530, 2002.

[3] D. Paillard, L. Labeyrie and P. Yiou, “Macintosh program performs time-series analysis,” in *Eo, Transactions*, American Geophysical Union, vol. 77, 379, 1996.

[4] H.J. Dowsett and M.M. Robinson, “Application of the Modern Analog Technique (MAT) of Sea Sea Surface Temperature estimation to Middle Pliocene North Pacific Planktonic Foraminifer Assemblages” in *Palaeontologia Electronica*, vol. 1, issue 1, art.3 22 pp., 1998.

[5] CLIMAP Project Members, “The Surface of the ice-age Earth” in *Science*, vol. 191, pp. 1131-1144, 1976.

[6] W.H. Huston, “The Agulhas current during the Late Pleistocene: Analysis of modern faunal analogs” in *Science*, vol. 207, pp. 64-66, 1980.

[7] P. Schwitzer, “ANALOG: A program for estimating paleoclimate parameters using method of modern anlaogs,”” in *U.S. Geological Survey Open file Report* 94-645, 1994.