Visual Discovery and Reconstruction of the Climatic Conditions of the Past*

Roberto Therón

Departamento de Informática y Automática,
Universidad de Salamanca, Salamanca, 37008, Spain
theron@usal.es

Abstract. The development of new tools and methodologies is necessary in order to better understand current and past climatic changes,. To be useful, these mathematical or software tools must not remain only in the hands of specialists in statistics, but must also be usable by the larger community of paleoclimatologists. It is therefore necessary to conceive a user interface adapted to the specificities of their use in paleoclimatology. Here, we propose the development of new tools of interactive analysis. Through the combination of techniques coming from knowledge discovery and information visualization (visual data mining), rapid and accurate paleoclimatic reconstructions will be easier to produce.

1 Introduction

While the need to foresee abrupt climatic changes is an urgent challenge for the society, paleoclimate research has shown that the causes and effects of these changes are very different, with extremely rapid variations even on one-year basis. Computers have played a key role in our understanding of the climatic dynamics. Nowadays, the improvement of data acquisition methods offer us the opportunity to gain the needed depth of information to diagnose and prevent any natural disaster. However, although data are available, the development of new tools and new methodologies is necessary. If very high precision physical or chemical measurements are necessary to reconstruct paleoenvironments, they often need to be accompanied by sophisticated statistical analysis methods ([1],[2]). But, to be useful, these mathematical or software tools must not remain only in the hands of specialists in statistics, but must also be usable by the larger community of paleoclimatologists. It is therefore necessary to foster an optimal use of these mathematical tools, by establishing methodological choices among the most relevant and the most recent statistical methods, and to conceive a user interface adapted to the specificities of their use in paleoclimatology.

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 [3], first step to make climatic predictions. When looking for historic climatic data with durations exceeding decades, the largest and

* This work was supported by the MCyT of Spain under Integrated Action (Spain-France) HF2004-0277 and by the Junta de Castilla y León under project SA042/02.

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part III, LNCS 3993, pp. 32–39, 2006.

oldest record is found in the oceans. Palaeoceanographers need to manipulate, integrate and analyze 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 [4] and AnalySeries [5].

Some of these data 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), at each given past moment. For the reconstruction of this features, isotope measurements (δ^{18}O) or biomarkers (U^k_{37} index) have been used. On the other hand, for the quantitative reconstruction of environmental conditions of the past, currently the *Modern Analog Technique* (MAT, actually a nearest neighbor prediction)[6], is one of the most commonly used techniques in paleoclimatology.

Although software tools for MAT have been developed [7], and some improvements have arisen such as SIMMAX [8], RAM [2] and artificial neural networks [9], they all have a main drawback: once developed they are black boxes. Paleoclimatologists can use them but no knowledge acquisition is involved; they just trust in the reconstructions obtained, they cannot know if the data used is valid from a geologic point of view. Furthermore, the classic MAT method inherently produces reconstructions whose precision is very difficult to estimate [10].

Visualization provides insight through images and can be considered as a collection of application of specific mappings from the problem domain to a visual range [11]. Thus, it is our aim to design new tailor-made methods of analysing and viewing the paleoclimatic data that would have different advantages and one goal: avoiding blind reconstructions by means of well designed user driven procedures. Through the combination of techniques coming from statistics, information theory, information visualization and visual data mining, rapid and accurate paleoclimatic reconstructions will be easier to produce. In this paper we show how new methods of interactive analysis can be extremely useful for the paleoclimatic field.

The rest of this paper is organised as follows: in section two how modern analogs (neighbors) are found through the calculation of dissimilarity coefficients between modern and paleo data (MAT) is shown. Third section is devoted to explain how a proper interactive analysis enables knowledge discovery and permits more accurate reconstructions. To finalize, the main conclusions and future work are described.

2 Calculating the Distances

This section describes how PaleoAnalogs, a Java based program, improves the modern analog technique in order to provide faster and more accurate reconstructions [12]. 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.

The process begins after the selection of the core and database files; in general, these files will contain different taxa (figure 1.a and figure 1.b), 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. With the help of the taxa association wizard (figure 1.c) this problem is easily worked out, allowing the user to determine which taxa from both the modern and fossil data files are compared, calculate proportions if needed, and identify the environmental features to be reconstructed.

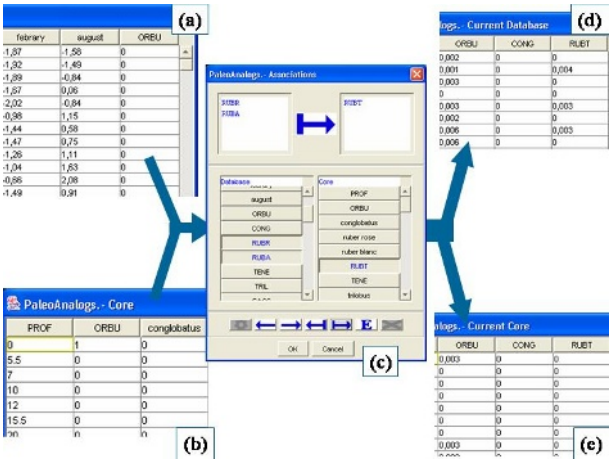

Fig. 1. Taxa association

Once the database and the core data are transformed to have the same number and equivalent taxa (figure 1.d and figure 1.f), each sample in the core is compared with each sample in the database using a dissimilarity coefficient.

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.

The next step in MAT is to reconstruct the environmental conditions of each core sample based on the environmental data of a number of best analogs (generally ten). This can be done by calculating the average value or by weighting the analogs. However, this is somehow very strict, because some of the used analogs could not be valid from a geologic point of view and should be eliminated.

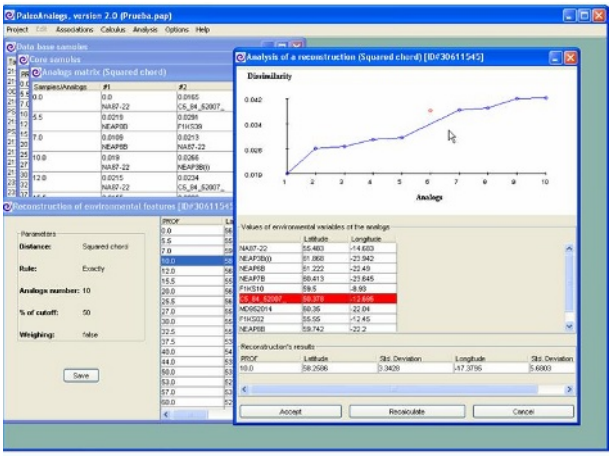

Fig. 2. Reconstruction tuning

It must be noted that the classical MAT technique only permitted to select the number K of analogs (neighbors), normally 10, used to calculate the averaged variables. But consider the case where only 3 of these 10 modern analogs were actually similar to the sample being reconstructed, while the 7 remainders were only the following most similar, and, therefore, it would be a mistake to use them for the reconstruction from a geological point of view.

Figure 2 shows the PaleoAnalogs interactive reconstruction tool. It enables for each sample to observe which sites have been closest. For the example in the figure, centimeter 10 of the core (which is a particular year in the past, depending on the sedimentation rate) is being studied. In this case the user has considered that the difference between the dissimilarity values of the fifth and sixth analogs is not acceptable; interactively, by clicking on each point of the graphic the analog is deselected, and the average recalculated using only the remaining selected analogs. On the other hand, the information of each analog is exposed, e.g. sixth analog is site C5-84-S2007, so the paleoclimatologist may decide upon rejecting any particular analog using geological criteria that may suggest not to use the data from that site.

3 Discovering the Past

Although the MAT method is very useful for paleoclimatic reconstruction there is much more information that can be provided than a mere neighbor distance calculation. Thus, before proceed with the algorithmic reconstruction further knowledge can be easily discovered from the calculated set of analogs.

As it has been stated in the previous sections, the problem is that paleontologists can obtain reconstructions as outputs from techniques such as nearest neighbor prediction, but no ways of knowledge acquisition are possible. However,

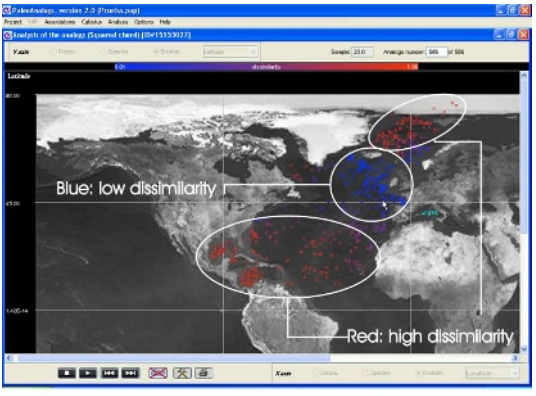

Fig. 3. Analogs geographic distribution

if the particular case of paleoclimatology is considered, *ad hoc* visualization tools may be developed, that will indeed provide insight in that forest of numeric data.

Geologists are trained for geographic visualization, and they basically face a problem of evolution through time, so we can design an interactive visual interface that takes advantage of both location and time.

Let's consider the following situation (figure 3): a paleoclimatologist is studying the data obtained from a particular point in the Mediterranean Sea (the label CORE in the figure shows that point) and the modern data comes from the North Atlantic ocean. Instead of just calculate the temperature reconstruction, he/she can analyse first how the analogs for a given sample (depth/age) are distributed geographically. This can be seen in figure 3, which is the three dimensional (longitude(x axis), latitude(y axis) and dissimilarity(color)) representation for the sample at 20 cm[1]. Thus, the expert would easily **see**(discover) that the studied site, t kiloyears ago had temperatures much more similar to those of cold sites of today (blue zone of analogs in the picture) than those in warm or polar latitudes (red zone of analogs).

Also, this representation can be done for a selected number of analogs (just the number of analogs that will be used for the reconstruction, for instance, or for those with dissimilarity values smaller than a cutoff. Analogs can be labeled with the associated database sample name so the expert might decide that a particular analog is not valid for the reconstruction due to a geological reason.

A combination of visualization approaches may discover a lot of information: choosing to show only 5 analogs, labeling each analog, zooming in and animating the evolution of the whole core, i.e., visualizing the analog evolution through time, we may arrive to some interesting conclusions. Thus, in figure 4 we could start with the deepest sample in the core, i.e., t kiloyears ago: since at that age the planet was covered with ice, we can see that the best 5 analogs are distributed within a wide range of latitudes. As the animation is showing the evolution, we

[1] Depending on the particular age model this depth will be a number of kiloyears in the past.

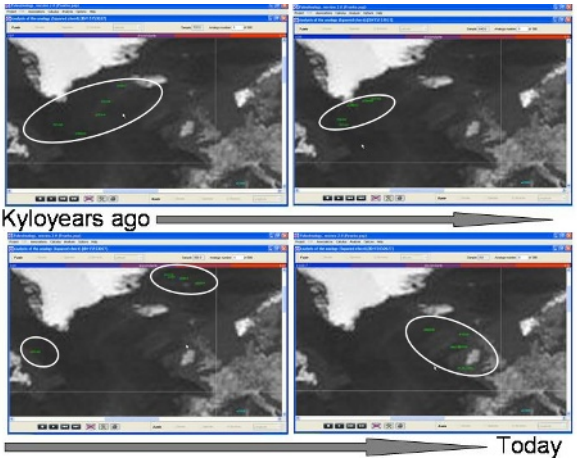

Fig. 4. Best analogs trough time

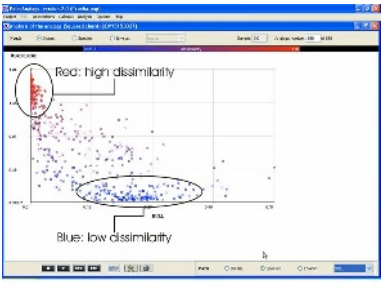

Fig. 5. Bull Species as an indicator of analog dissimilarity

can see that best analogs are grouped, which is the typical distribution we should expect. The snapshot on the bottom-left shows a particular interesting situation: four of the five analogs are grouped up north (note the blue color), while the fifth one is located at a much warmer latitude (note the red color). This analog distribution should warn the paleoclimatologist, the most probable reason is that the outsider is only the fifth closest neighbor, but not a *real neighbor*, so that particular site should not be considered in the reconstruction for that sample.

Another example (figure 5) of knowledge discovery would be that a particular species (BULL, for instance) is valid as an analog indicator, since the dissimilarity for the sample is very high when the proportion of this species is close to zero (all analogs in the x axis are red and have a high dissimiliraty value, while the rest of analogs are getting more and more blue, i.e., more similar, as the proportion of the species grows).

Finally, using specific visual techniques for the interactive analysis of multidimensional data such as Parallel Coordinates [11] (see figure 6) will improve the

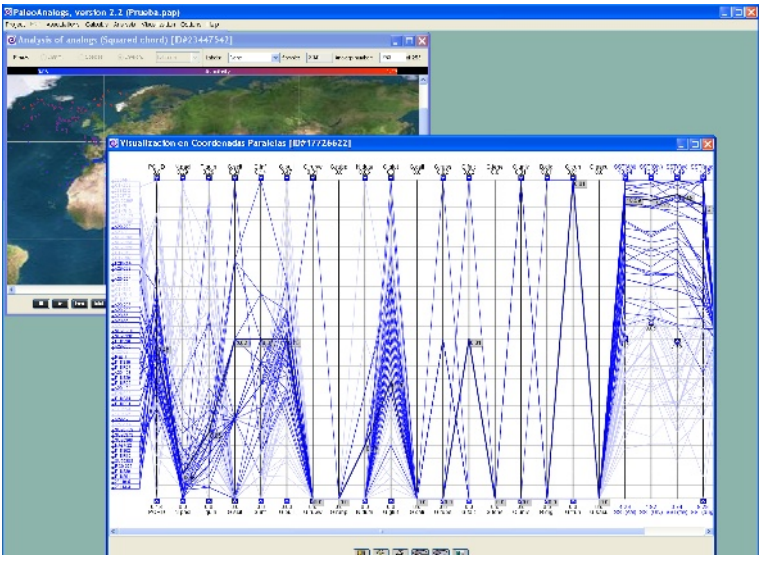

Fig. 6. Reconstruction visually driven by Parallel Coordinates

paleoclimatic knowledge discovery; this practice is an emerging usability issue in geovisualization[13].

4 Conclusions and Future Work

This work is an example of how interactive analysis can help knowledge discovery in the paleoclimatology field. We have shown how a well-known standard technique of the field (MAT) can be greatly improved so the reconstructions of paleoenvironmental conditions can be more accurate. This is accomplished by fostering a user driven reconstruction procedure where the expert get more insight from the data and can decide on the validity of potentials reconstructions. Finally, we can add that more complex interactive analysis can be designed that will help to gain a deeper knowledge about the climatic evolution of a given area.

References

1. Yiou, P., Baert, E., Loutre, M.F.: Spectral analysis of climate data. Surveys of Geophysics **17**(6) (1996) 619–663
2. Waelbroeck, C., Labeyrie, L., Deplessy, J., Guoit, J., Labracherie, M., Leclaire, H., Duprat, J.: Improving past sea surface temperature estimates based on planktonic faunas. Paleoceanography **13** (1998) 272–283
3. Therón, R., Flores, J.A., Sierro, F.J., Pelejero, C., Grimalt, J., Vaquero, M.: Using data mining and visualization techniques for the reconstruction of ocean paleodynamics. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Volume IV. (2002) 2382–2384

4. Therón, R., Flores, J.A., Sierro, F.J., Vaquero, M., Barbero, F.: Paleoplot: A tool for the analysis, integration and manipulation of time-series paleorecords. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Volume VI. (2002) 3528–3530
5. Paillard, D., Labeyrie, L., Yiou, P.: Macintosh program performs time-series analysis. Eos, Transactions, American Geophysical Union **77** (1996) 379
6. Hutson, W.H.: The agulhas current during the late pleistocene: Analysis of modern faunal analogs. Science **207** (1980) 64–66
7. Schweitzer, P.N.: Analog: A program for estimating paleoclimate parameters using the method of modern analogs. Technical Report 94-645, U. S. Geological Survey Open-File (1994)
8. Pflaumann, U., Duprat, J., Pujol, C., Labeyrie, L.: Simmax: A modern analog technique to deduce atlantic sea surface temperatures from planktonic foraminifera in deep-sea sediments. Paleoceanography **11** (1996) 15–35
9. Malmgren, B.A., Kucera, M., Nyber, J., Waelbroeck, C.: Comparison of statistical and artificial neural network techniques for estimating past sea surface temperatures from planktonic foraminifer census data. Paleoceanography **16**(5) (2001) 520–530
10. Mannila, H., Toivonen, H., Korhola, A., Olander, H.: Learning, mining or modeling? a case study from paleoecology. In: Discovery Science. (1998) 12–24
11. Inselberg, A.: Visualization and knowledge discovery for high dimensional data. In: Proceedings of the Second International Workshop on User Interfaces to Data Intensive Systems. (2001) 5–24
12. Theron, R., Paillard, D., Cortijo, E., Flores, J.A., Vaquero, M., Sierro, F.J., Waelbroeck, C.: Rapid reconstruction of paleoenvironmental features using a new multiplatform program. Micropaleontology **50** (2004) 391–395
13. Robinson, A.C., Chen, J., Meyer, E.J., MacEachren, A.M.: Combining usability techniques to design geovisualization tools for epidemiology. Cartography and Geographic Information Science, **32** (2005) 243–255