Visual Knowledge Discovery in Paleoclimatology with Parallel Coordinates

Roberto Therón

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

Abstract. Paleoclimatology requires the analysis of paleo time-series, obtained from a number of independent techniques and instruments, produced by several researchers and/or laboratories. Analytical reasoning techniques that combine the judgment of paleoceanographers with automated reasoning techniques are needed to gain deep insights about commplex climatic phenomena. This paper presents an interactive visual analysis method based in Parallel Coordinates that enables the discovery of unexpected relationships and supports the reconstruction of climatic conditions of the past.

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. By means of an analysis of such data, paleoceanographers are expected to assess (understand the past) and forecast (estimate the future). Although massive amounts of data are available, the development of new tools and new methodologies is necessary to help the expert extract the relevant information. This is the approach of Visual Analytics [11] as the science of analytical reasoning supported by highly interactive visual interfaces.

If very high precision physical or chemical measurements are necessary to reconstruct paleoenvironments, they often need to be accompanied by sophisticated statistical analysis methods ([21], [20]). 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 [18], 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 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 [19] and AnalySeries [12].

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) [4], is one of the most commonly used techniques in paleoclimatology.

Although software tools for MAT have been developed [15], and some improvements have arisen such as SIMMAX [13], RAM [20] 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 [6]. These visual representations combined with interaction techniques that take advantage of the human eye's broad bandwidth pathway to the mind allow experts to see, explore, and understand large amounts of information at once [11]. Thus, this paper presents an interactive visual analysis method that, through the combination of techniques coming from statistics, information theory, information visualization and visual data mining, enables the discovery of unexpected relationships and supports the reconstruction of climatic conditions of the past.

2 Visual Analytics: visual representations and interaction

This section explains how the use of proper interactive visual representations foster an analytical discourse (a dialogue between the analysts and the information) [11]. Using the original and transformed data, it is possible to automatically find patterns in information, and represent such information in ways that are meant to be revealing to the analyst. On the other hand, by interacting with these representations, using their expert knowledge, it is possible to refine and organize the information more appropriately. This way, it is possible, not only to reconstruct paleoenvironmental features, such as SST or salinity, but to visualize what information is being used to estimate these variables, and help the paleoclimatologists to decide upon using particular data or not, according to their field experience.

2.1 A novel method for paleoenvironmental reconstructions: interactive parallel coordinates plots

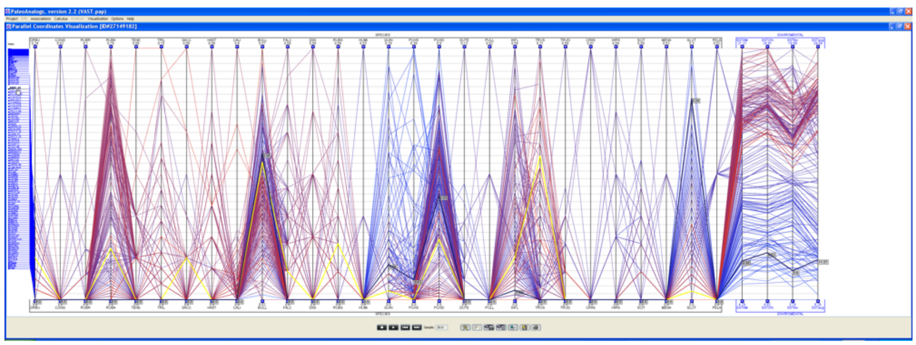

Fig. 1. Reconstruction visually driven by Parallel Coordinates

Scatter plots, maps and animations are common methods for geovisualization that have a long history in cartography and information visualization. Parallel Coordinates Plots (PCP) [5] are also a common method of information visualization, used for the representation of multidimensional data, and is an emerging practice in geovisualization [8][14]. In the PaleoAnalogs framework, the use of PCP (see figure 1) as part of the interactive visual analysis of multidimensional data can provide paleoclimatic knowledge discovery.

PaleoAnalogs presents a unique use of interactive PCP; instead of just use it as another way of visualizing the data, it is used as a highly interactive tool that permits both gaining insight about the paleodata and visually reconstruct the paleoenvironmental features.

The data transformations (MAT) described above provides a mechanism for extracting patterns from the data. The output of this process is depicted using interactive PCP to facilitate the exploration of relationships among attributes. That is (see figure 1), each site of the database is drawn as a polyline passing through parallel axis, which represent the species, and the environmental variables that we want to reconstruct (last four axis on the right, in the example of the figure). The polyline corresponding to a particular sample (20 cm of depth in this example) in the core is represented as a yellow polyline. Note that, since the core only have the species data and we want to reconstruct the environmental variable for each sample, there are no yellow segments in the environmental axis. Each polyline of the database is color coded and the MAT technique is used for that purpose, i.e., the more red the polyline is, the more dissimilar is to that particular sample of the core.

This static picture is already showing many things that were hidden in the previous approach. For instance, it can be observed that the most similar sites for sample 20, are clustered in the low temperatures. This means that sample 20 corresponds to a cold period or a cold site.

However, several interaction techniques [17][7] have been integrated with this PCP to allow brushing [1], linking, animation, focus + context, etc., for exploratory analysis and knowledge discovery purposes enhance exploration.

focus+context In the figure 1, a focus + context [17] technique can be seen for the labels of each site in the database. On the left side, each polyline is connected outside the axis with its label, allowing easy identification of each site. The labels are ordered top down, depending on different criteria chosen by the user: alphabetically, dissimilarity index, latitude or no-crossing (the labels are ordered to avoid that the lines connecting with the values on the first axis produce any crossing). This way the expert can easily select the polyline of a particular site, that is highlited in black and the values for each axis are shown. Since the context is maintained, the expert can access the label faster, depending on the chosen ordering criterium and the position of the current focus. Since there is a space shortage of space for such a number of labels, a fisheye approach [3], a powerful technique for organizing the suppersion of irrelevant data, was developed.

Filtering and axis interaction Another powerful feature in PaleoAnalogs is dynamic filtering. PaleoAnalogs provides a dynamic query on the PCP in the form of axis filtering [2][16]. The range of an attribute can be specified by moving the handles at the top and bottom of a range slider. The range sliders are embedded within the PCP.

To prevent users from losing global context during dynamic filtering, all the polylines are maintained on the background. Users can see the position of a polyline. Also labels are maintained on the background. Figure 2 shows a reconstruction already computed. After filtering the sites (the current ranges are shown in the handles) that were too dissimilar (maintained in dark gray on the background), the expert decided to reconstruct the SSTs for sample 20. Note that now the yellow segments of the polyline for that sample also occupy the environmental axis. As expected, the values are an average of the values of the blue polylines. On the left hand side, only the interesting site labels are highlited. In the snapshot, the expert is comparing a site (orange polyline) with the reconstructed core sample.

Also note that in figure 2 , all axis have the same scale (a percentage) in order to compare the relative abundances of the species, and help discover relationships between species and climatic features.

Another feature that helps in the interactive visual analysis is that any axis can be dragged and dropped, so the order of the axis is changed. This way

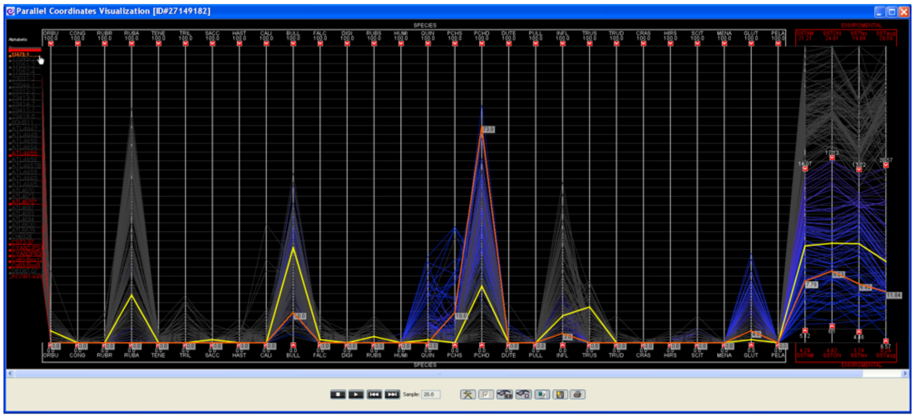

Fig. 2. Analytical reasoning and reconstruction by means of interaction

the shape of the polyline also changes, helping to reveal hidden patterns and making analysis easier. In 3, axis order have been altered slightly so the PCP is uncluttered. The elimination of selected variables (axis) can also be helpful.

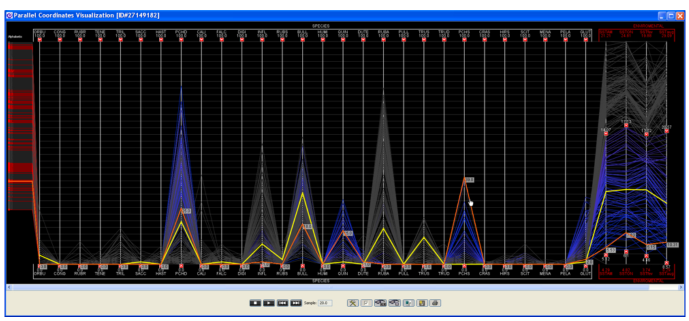

Fig. 3. Uncluttering Parallel Coordinates by changing the axis order

Animation On the contrary to other approaches such as [14], time is not represented on an axis. As in the previous cases, time is one of the most relevant variables in the analysis. PaleoAnalogs can show an animation of the PCP. By doing this, it is possible to visualize the evolution of different species trough geological time, and its relationships, both among themselves and between some species and SST, for instance.

Brushing and multiple linked views Geolocation plays a key role in paleoceanography analysis. One possibility is to represent the latitude and/or longitude values for each database site. This can be interesting in order to highlight how temperature varies with time for the same latitude, for example.

In previous sections the benefits of using interactive maps in PaleoAnalogs have been described. A common coordination technique is brushing and linking [1], where users can select objects in one view and the corresponding objects in all the other views are also automatically selected. This technique is the natural approach for the problem at hand. This way, all the benefits explained above can be put together in order to provide the paleoceanographers with the best interactive visual tool to discover knowledge and support decisions about climatic reconstructions.

In figure 4 the brushing and multiple linked views approach of PaleoAnalogs can be seen. As in [2], three modes of brushing and linking interaction that are coordinated among all the views described in the previous sections are possible:

- probing: this mode is used to view more details about an object (e.g. site labels and dissimilarity values) and to get an understanding of the relationships between the different views. Probing is a transient operation. Moving the mouse pointer over an object, highlight the object (e.g, a polyline) and as the mouse pointer is moved away, the highlighting disappears.
- selecting: this mode is used to mark objects that are of short-term interest, in order to further examine or perform operations on them (e.g see the values on every axis of a selected polyline). Clicking on an object selects it and marks it. If a selected object is filtered, then it becomes deselected.
- painting: this mode is used to mark objects that are of long-term interest, in order to use them as references for comparisons (e.g compare two polylines of two sites in the database). Objects remain painted until they are reset explicitly.

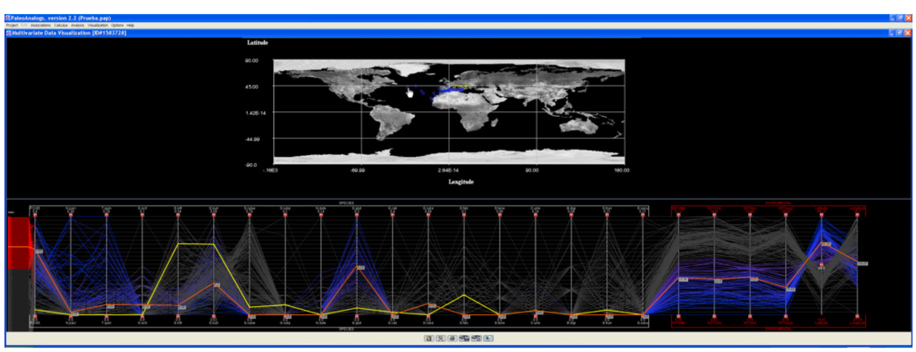

Fig. 4. Multiple linked views in PaleoAnalogs

3 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.

Acknowledgements

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. The author would like to acknowledge Ana Mayordomo García for her assistance in the implementation of this work.

References

1. Richard A. Becker and William S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.
2. D. Brodbeck and L. Girardin. Design study: using multiple coordinated views to analyze geo-referenced high-dimensional datasets. In *Proceedings. of the International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 104–111, 2003.
3. George W. Furnas. Generalized fisheye views. In *Human Factors in Computing Systems CHI '86 Conference Proceedings*, pages 16–23, 1986.
4. W. H. Hutson. The agulhas current during the late pleistocene: Analysis of modern faunal analogs. *Science*, 207:64–66, 1980.
5. Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985.
6. Alfred Inselberg. Conflict detection and planar resolution for air traffic control. In *Intelligent Transportation Systems*, 2001.
7. Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.
8. E.L. Koua and M.-J. Kraak. A usability framework for the design and evaluation of an exploratory geovisualization environment. In *Proceedings. Eighth International Conference on Information Visualisation*, pages 153–158, 2004.
9. B. A. Malmgren, M. Kucera, J. Nyber, and C. Waelbroeck. Comparison of statistical and artificial neural network techniques for estimating past sea surface temperatures from planktonic foraminifer census data. *Paleoceanography*, 16(5):520–530, 2001.
10. H. Mannila, H. Toivonen, A. Korhola, and H. Olander. Learning, mining or modeling? a case study from paleoecology. In *Discovery Science*, pages 12–24, 1998.

11. National Visualization and Analytics Center. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.
12. D. Paillard, L. Labeyrie, and P. Yiou. Macintosh program performs time-series analysis. *Eos, Transactions, American Geophysical Union*, 77:379, 1996.
13. U. Pflaumann, J. Duprat, C. Pujol, and L. Labeyrie. Simmax: A modern analog technique to deduce atlantic sea surface temperatures from planktonic foraminifera in deep-sea sediments. *Paleoceanography*, 11:15–35, 1996.
14. A. C. Robinson, J. Chen, E. J. Meyer, and A. M. MacEachren. Combining usability techniques to design geovisualization tools for epidemiology. *Cartography and Geographic Information Science,*, 32:243–255, 2005.
15. P. N. Schweitzer. Analog: A program for estimating paleoclimate parameters using the method of modern analogs. Technical Report 94-645, U. S. Geological Survey Open-File, 1994.
16. Jinwook Seo and Ben Shneiderman. Interactively exploring hierarchical clustering results. *IEEE Computer*, 35(7):80–86, 2002.
17. Robert Spence. *Information Visualization*. Addison-Wesley, 1 edition, 2001.
18. 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*, volume IV, pages 2382–2384, 2002.
19. 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*, volume VI, pages 3528–3530, 2002.
20. C. Waelbroeck, L. Labeyrie, J.C. Deplessy, J. Guoit, M. Labracherie, H. Leclaire, and J. Duprat. Improving past sea surface temperature estimates based on planktonic faunas. *Paleoceanography*, 13:272–283, 1998.
21. P. Yiou, E. Baert, and M. F. Loutre. Spectral analysis of climate data. *Surveys of Geophysics*, 17(6):619–663, 1996.