Visual Analytics of Paleoceanographic Conditions

Roberto Theron*

Departamento de Informática y Automática
Universidad de Salamanca

Abstract

Decade scale oceanic phenomena like El Niño are correlated with weather anomalies all over the globe. Only by understanding the events that produced the climatic conditions in the past will it be possible to forecast abrupt climate changes and prevent disastrous consequences for human beings and their environment. Paleoceanography research is a collaborative effort that requires the analysis of paleo time-series, which are obtained from a number of independent techniques and instruments and produced by a variety of different researchers and/or laboratories. The complexity of these fenomena that consist of massive, dynamic and often conflicting data can only be faced by means of analytical reasoning supported by a highly interactive visual interface. This paper presents an interactive visual analysis environment for paleoceanography that permits to gain insight into the paleodata and allow the control and steering of the analitycal methods involved in the reconstruction of the climatic conditions of the past.

Keywords: Infovis, parallel coordinates, multiple linked views, exploratory analysis.

Index Terms: H.1.2 [User/Machine Systems]: Human information processing—Visual Analytics; J.2 [Physical Sciences and Engineering]: Earth and atmospheric sciences—Applications

1 Introduction

In the last few years, the world has suffered from some of the most catastrophic natural disasters in recent history. While some of them are of geologic origin, such as the Sumatra-Andaman earthquake (2004), which triggered the single worst tsunami in history, most of them are related to the weather. Examples of the latter case are Hurricane Katrina (2005), one of the costliest and deadliest hurricanes in American history; or the big El Niño (El Niño-Southern Oscillation, ENSO) in 1997-98, which cost hundreds of lives and caused 34 bn in damage worldwide, partly through flooding to Chile, Ecuador and Bolivia and partly through failed harvests in Australia, the Philippines and Indonesia. A more recent, milder one in 2002-03 caused the worst Australian drought in a century[1].

Other phenomena, less understood than ENSO, include the Arctic Oscillation (AO), the Pacific Decadal Oscillation (PDO), and the North Atlantic Oscillation (NAO). Oceanic features like sea surface temperature (SST) variations associated with these phenomena can significantly impact local, regional, and global climate conditions. Furthermore, fossil evidence has demonstrated that the Earth's climate can change within a decade, and those newly-established patterns can persist for decades or centuries.

While the need to foresee abrupt climatic changes is an urgent challenge for the society, paleoceanographic/paleoclimate research has shown that the causes and effects of these changes are very different, with extremely rapid variations even on a one-year basis. Computers have played a key role in our understanding of climatic dynamics. Nowadays, improved data acquisition methods offer us the opportunity to gain the needed depth of information to diagnose and prevent natural disasters. 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 [12]], a 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 ([23], [22]). But, in order to be useful, these mathematical or software tools must not remain only in the hands of specialists in statistics, but must also be made available and put to use 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 most recent statistical methods, and to conceive a user interface adapted to the specificities of their use in paleoclimatology.

The data registered over the course of thousands of years (mainly in ice and sediment cores) is an impressive source of information that, for instance, help us to model earth and ocean dynamics [19]. This raw data is the first step in making climatic predictions, and 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.), and are also produced by a number of different researchers and/or laboratories. This work is done with the aid of proper tools such as PaleoPlot [20] and AnalySeries [13].

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 ascertain environmental parameters, such as sea surface temperature (SST), at each given past moment. For the reconstruction of these features, two types of paleo climactic techniques have been used: on the one hand, isotope measurements (δ^{18}O) or biomarkers (U_{37}^{k} index), and on the other, quantitative reconstruction of environmental conditions of the past, the *Modern Analog Technique* (MAT, actually a nearest neighbor prediction) [5].

Although software tools for MAT have been developed [16], and some improvements have arisen such as SIMMAX [14], RAM [22] and artificial neural networks [10], all of which have one main drawback: once developed they are like black boxes which paleoclimatologists can use but from which no knowledge acquisition is involved. Paleoclimatologists trust in the reconstructions obtained, as they cannot know if the data used is valid from a geologic point of view. To complicate matters further, the classic MAT method inherently produces reconstructions whose precision is very difficult to estimate [11].

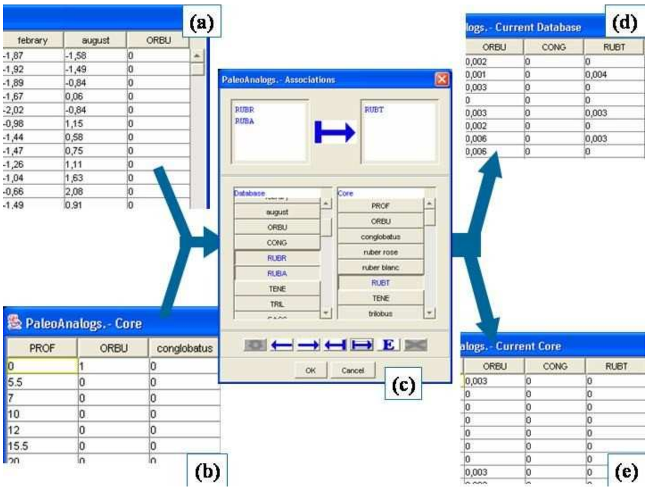
Visualization, which provides insight through images, can be

*e-mail: theron@usal.es

[1] *After Ivan: Prepare for Return of el Nio, by Geoffrey Lean, Environment Editor, The Independent - Sept 12, 2004*

considered a collection of specific mapping applications which take data from the problem domain and transform them into a visual range [7]. 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 a single glance [12]. Thus, this paper presents an interactive visual analysis environment that, through the combination of techniques from statistics, information theory, information visualization and visual data mining, enables paleoceanographers to discover unexpected relationships and that supports the reconstruction of climatic conditions of the past.

The remainder of this text is organised as follows: the second section explains how raw sediment core data is transformed into modern analogs (MAT), which are more appropriate for the analytical tasks which paleoclimatologists must conduct. Section 3 is devoted to explaining how visual representations and interaction technologies enable knowledge discovery and provide a means of analysis of spatial and temporal data, the way to facilitate expert-driven accurate reconstructions. The final section of this paper includes the main conclusions and proposals for future work.



febrary	august	ORBU
-1,87	-1,58	0
-1,92	-1,49	0
-1,89	-0,84	0
-1,67	0,06	0
-2,02	-0,84	0
-0,98	1,15	0
-1,44	0,58	0
-1,47	0,75	0
-1,26	1,11	0
-1,04	1,63	0
-0,66	2,08	0
-1.49	0,91	0

PROF	ORBU	conglobatus
0	1	0
5.5	0	0
7	0	0
10	0	0
12	0	0
15.5	0	0

RUBR
RUBA

RUBT

Database: august, ORBU, CONG, RUBR, RUBA, TENE, TRIL

Core: PROF, ORBU, conglobatus, ruber rose, ruber blanc, RUBT, TENE, trilobus

ORBU	CONG	RUBT
0,002	0	0
0,001	0	0,004
0,003	0	0
0	0	0
0,003	0	0,003
0,002	0	0
0,006	0	0,003
0,006	0	0

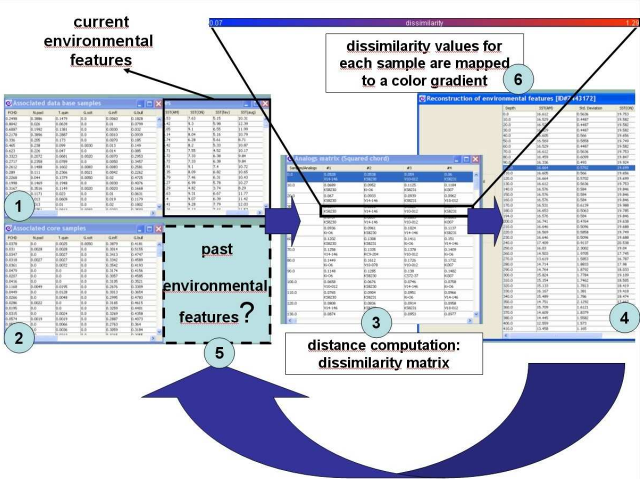
ORBU	CONG	RUBT
0,003	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0,003	0	0

Figure 1: Automated data transformation with the taxa association wizard: the different taxa used for current data and for past data (probably because they are from different sources) force a data recomputation in order to obtain comparable sets.

2 MODERN ANALOGS: DATA TRANSFORMATION

The National Visualization and Analytics Center, through its Research and Development Agenda for Visual Analytics [12], highlighted the key aspect of data representation and transformation as a way of supporting visualization and analysis.

This section describes how PaleoAnalogs, a Java-based program, makes use of the modern analog technique (MAT) [5] in order to provide faster and more accurate reconstructions of climatic conditions of the past. Initially, PaleoAnalogs used this k-nearest neighbor prediction as a stable machine learning classification method [2], which provided an automated tool for reconstruction of paleoenvironmental features such as sea surface temperatures (SST). Since MAT is a black-box method, where paleoclimatologists just collected the results of a mathematical algorithm that sometimes were not valid from a geologic point of view, a timid advance was introduced [21]: namely, a better means of understanding the reconstruction process via interactive visual analysis (colored scatter plots) of the analogs (neighbors) found by the algorithm.



current environmental features
dissimilarity values for each sample are mapped to a color gradient
0.07
dissimilarity
1.20
Associated data base samples
Associated core samples
Analog matrix (Squared chord)
Reconstruction of environmental features
past environmental features?
distance computation: dissimilarity matrix
1
2
3
4
5
6

Figure 2: Reconstruction of past conditions using the MAT technique (k-nearest neighbor prediction): steps 1 to 5; the computed values of dissimilarity are mapped to a color gradient that can be used in visual representations: step 6.

This was accomplished by means of a combination of new visual representations and interaction techniques, though spatial and temporal dimensions were not considered in the interactive analysis, both of which will be explained in section 3. Before that, however, the mechanism behind this transformation will be explained.

In k-nearest neighbor prediction, a database is used to predict the value of a variable of interest for each member of a target data set (SST, for instance) with respect to a number of additional predictor variables (e.g., microfossil species abundances). The k-nearest neighbor algorithm can be summarized as follows: a) for each case in the target data set, locate the k closest members (the k nearest neighbors) of the database and use a distance measurement to calculate how close each member of the training set is to the target row that is being examined; b) estimate the unknown variable of interest for that particular case as the average of the variables of interest for its k nearest neighbors.

At this point, in paleoclimatology a problem arises: the data from different sources prevent automated transformation and subsequent analysis. 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.

With PaleoAnalogs, 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. For example, in Figure 1, (a) has environmental information (february, august) and different species (ORBU, ...), while (b) has a sample label (PROF) and other different species names (ORBU, conglobatus, ...). 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 (in (c) it can be seen that the subspecies RUBR and RUBA of the database will be added to be compared with the species RUBT of the core and the latter will be used as the taxa), calculate proportions if needed, and identify the environmental features to be reconstructed. Thus, the database and the core data are transformed in such a way that they have the same number and equivalent taxa (thus, Figure 1.d and Figure 1.e contain exactly the same names for species; note that ORBU was present in both (a) and (b); instead of conglobatus (b), CONG is used; and the species RUBT is used instead of the subspecies).

Figure 2 show the different steps followed in MAT. The goal of the technique is to reconstruct (predict) the environmental values that are missing in the core. Note that the database (Figure 2.1) contains data for the current environmental features of different locations in the ocean. The core (Figure 2.2) only has information about species abundances, i.e., we want to fill in the dotted rectangle. Once the database and the core have the same number and order of columns, each sample in the core is compared with each sample in the database using a dissimilarity coefficient. Thus, with this distance measure selected by the user, a dissimilarity matrix is built (Figure 2.3). For each core sample N dissimilarity values are given, with N being 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.

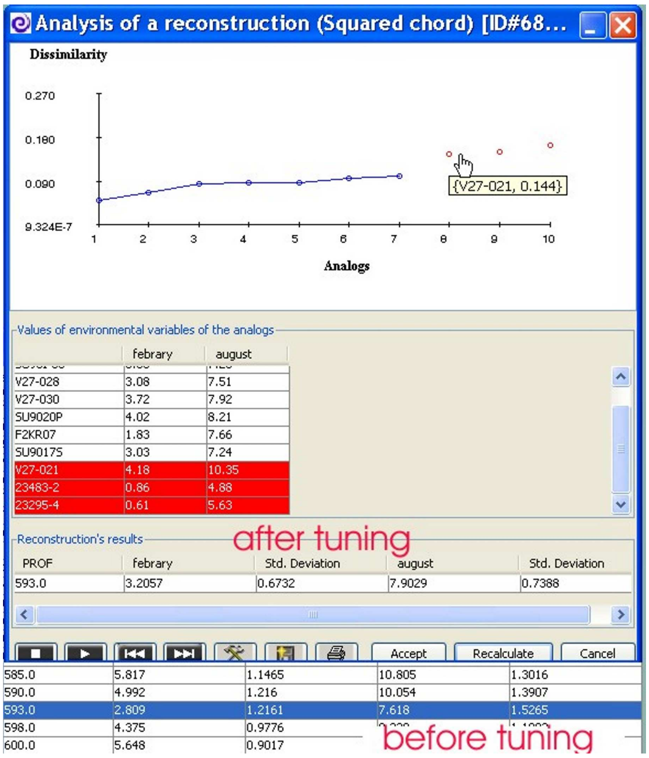
The final step in MAT is to reconstruct the environmental conditions of each core sample based on the environmental data of a number of best analogs (generally ten, although any number of analogs can be used). This can be done by calculating the average value of the paleovariable to be reconstructed or by weighting the analogs (the closer an analog is the more similar the reconstructed value should be to it, Figure 2.4). So after this computation is performed for each sample in the core, we have reconstructed the original missing values (Figure 2.5).

The problem at this point is that all we have is a black-box numerical method, so no matter how much information about time or location or other geological issues is available, the past conditions are reconstructed without the expert getting involved in the process. But the experts have much to say and that is the reason why an interactive analysis can be a powerful aid. The first approach is to map the information about dissimilarity to a color gradient (Figure 2.6). For each sample in the core, we now have an ordered list of samples in the database that are color coded (from blue to red, or from similar to dissimilar). This can be extremely useful for advanced visual representations.

3 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) [12]. Using the original and the 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 paleoclimatologists to decide upon using particular data or not, according to their field experience.

In the following subsection, a collection of visual representations and interaction techniques that enable analytical reasoning in paleoceanography are described.



	febrary	august
V27-028	3.08	7.51
V27-030	3.72	7.92
SU9020P	4.02	8.21
F2KR07	1.83	7.66
SU9017S	3.03	7.24
V27-021	4.18	10.35
23483-2	0.86	4.88
23295-4	0.61	5.63

PROF	febrary	Std. Deviation	august	Std. Deviation
593.0	3.2057	0.6732	7.9029	0.7388

585.0	5.817	1.1465	10.805	1.3016
590.0	4.992	1.216	10.054	1.3907
593.0	2.809	1.2161	7.618	1.5265
598.0	4.375	0.9776	[illegible]	[illegible]
600.0	5.648	0.9017		

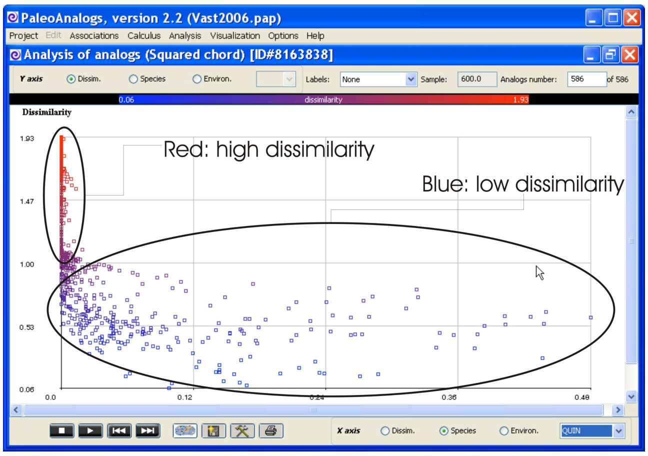
Figure 3: Reconstruction tuning: the most dissimilar analogs within the best ones might be inappropriate for the reconstruction of a given core sample. If the paleoclimatologist has gelogical evidence that suggests not to use these analogs, the reconstruction is computed again using only the selected analogs.

3.1 Improving Reconstructions with Visual Insight

As explained above, the classical MAT reconstructed the environmental conditions of each core sample based on the environmental data of a number of best analogs (generally ten). This could 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.

It must be noted that the classical MAT technique only permitted users to select the number K of analogs (neighbors), normally 10, that are 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; therefore, from a geological point of view, it would be a mistake to use them for reconstruction.

Figure 3 shows the PaleoAnalogs interactive reconstruction tool. Here one can see which sites were the closest analogs (neighbors) for each sample in the core. For the example in the figure, centimeter 593.0 of the core (which is a particular year in the past, depending on its sedimentation rate) is being studied. In this case the user has considered that the difference between the dissimilarity values of the seventh and eighth analogs is not acceptable; interactively, by clicking on each point of the plot, the analog is deselected (red circles) and the average recalculated using only the remaining selected analogs. On the other hand, the information of each analog is exposed, e.g. the seventh analog is site V27-021, so the paleoclimatologist may decide to reject a particular analog using geological criteria that recommend against using the data from that site. Note that after the recalculation, the temperatures of February and August have changed from 2.809 and 7.618, respectively (blue highlighted row in the *Reconstruction of environmental features window*), when the 10 best analogs were used, to 3.2057 (February) and 7.0029 (August), using only 7 analogs. Also, standard deviations are reduced 50%, approximately.


PaleoAnalogs, version 2.2 (Vast2006.pap)

Analysis of analogs (Squared chord) [ID#8163838]

Y axis ● Dissim. ○ Species ○ Environ.

Labels: None

Sample: 600.0

Analogs number: 586 of 586

X axis ○ Dissim. ● Species ○ Environ. QUIN

Figure 4: QUIN Species as an indicator of analog dissimilarity: absence (or very low abundance) of QUIN in a sample of the database will produce a high value of dissimilarity with sample 600.

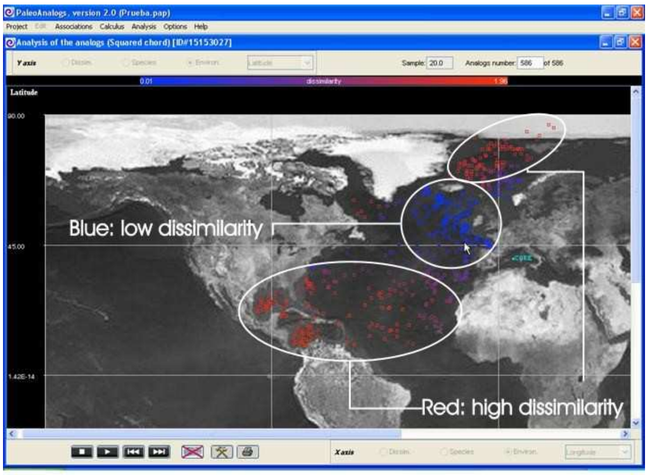
Another very valuable example of knowledge discovery for paleoclimatology (Figure 4) would be that a particular species (QUIN, for instance) is valid as an analog indicator, since the dissimilarity for the sample is very high when the proportion of this species is close to zero (all analogs along the y-axis are red and have a high dissimilarity value, while the rest of analogs along the x-axis turn bluer and bluer, which means they are increasingly similar as the proportion of the species grows). This kind of analysis can be done for any pair of variables, such as fossil species or environmental variables (latitude, longitude, and so forth). The dissimilarity is always present in these scatter plots as a third, color coded dimension.

3.2 Space and Time Reasoning

Since going down in the core means going backwards in time, another important step that has been added is to provide a way to leverage the ability of paleoclimatogists for reasoning about time. On the other hand, all the analysis is done for geospatial information. Thus, both space and time are considered in the visual representations, in order to gain insight into the core at hand.

Actually, time is a variable that can be considered in the previous examples. Both the interactive reconstruction tool and the scatter plots can be animated. This way, the experts can analyse the whole history of the core. In the first case, it is possible to go further down into the core, and visualize the dissimilarity plots of the analogs used for the reconstruction of each sample (depth/age) of the core. This way, visually, a different dissimilarity pattern can be easily discovered. In the second case, the relationship between QUIN species, for instance, and dissimilarity through time can be inspected, and paleoclimatogists can learn if this species behaves in the same way throughout the geological history of that particular site.

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 proceeding with the algorithmic reconstruction further knowledge can be easily discovered from the calculated set of analogs.


PaleoAnalogs, version 2.0 (Prueba.pap)

Analysis of the analogs (Squared chord) [ID#15153027]

Y axis ○ Dissim. ○ Species ● Environ. Latitude

Sample: 20.0

Analogs number: 586 of 586

X axis ○ Dissim. ○ Species ● Environ. Longitude

Figure 5: Analogs geographic distribution: it can be seen that for sample 20 of this core (see cyan label) all the similar analogs are located within a range of latitudes, while the dissimilar ones are grouped in warmer or colder zones. This means that sample 20 of the analysed site, *t* kiloyears ago, had colder temperatures than the current temperatures at the Mediterranean sea.

As stated in the previous sections, the problem is that paleontologists can obtain reconstructions as outputs using techniques such as nearest neighbor prediction, but these techniques provide no way for acquiring new knowledge. However, in the particular case of paleoclimatology, ad hoc visualization tools that do provide insight into these forests of numeric data may be developed.

Trained in geographic visualization, geologists 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 5): a paleoclimatologist is studying the data obtained from a particular point in the Mediterranean Sea (the cyan label CORE in the Figure shows that point) and the modern data comes from the North Atlantic Ocean. Instead of just calculating 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 5, which is the three dimensional (longitude(x-axis), latitude(y-axis) and dissimilarity(color)) representation for the sample at 20 cm[2]. Thus, the expert would easily 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).

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

A combination of visualization approaches may unearth a wealth 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 at some interesting conclusions. Thus, in Figure 6 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

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

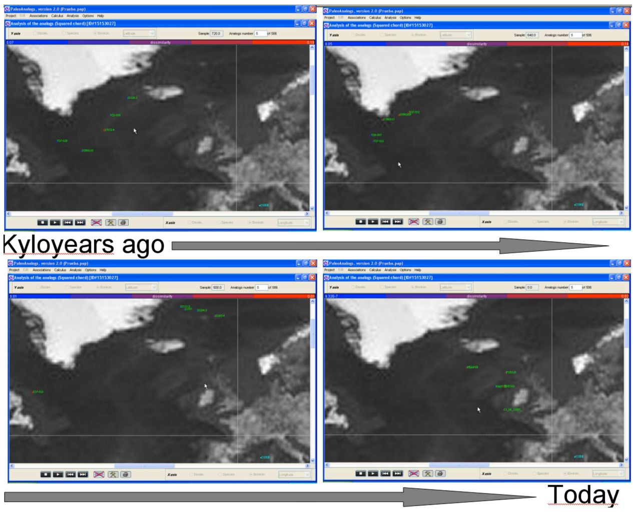


Figure 6: Best analogs over time: the animation will show that at one moment, 1 of the most 5 similar analogs for a given sample fall very far from the rest of the group. Using this analog for reconstructions would lead to inaccurate results.

are distributed across a wide range of latitudes. As the animation shows the evolution, we can see that the best analogs are grouped, reflecting the typical distribution we should expect. The snapshot on the bottom-left shows a particularly 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 that in all probability the outsider is only the fifth closest neighbor, but not a *real neighbor*, so that particular site should not be taken into consideration in the reconstruction for that sample.

3.3 A Novel Method for Paleoenvironmental Reconstructions: Interactive Parallel Coordinates Plots

Scatter plots, maps and animations are common methods for geovisualization that have a long history in cartography and information visualization (Figure 7 shows an environment for visual analytics in paleooeanography that makes use of these techniques). Parallel Coordinates Plots (PCP) [6] which is used for the representation of multidimensional data, is also a common method of information visualization and is an emerging practice in geovisualization [9][15]. In the PaleoAnalogs framework, the use of PCP (see Figure 8) as part of the interactive visual analysis of multidimensional data can lead to paleoclimatic knowledge discovery.

PaleoAnalogs presents a unique use of interactive PCP; instead of just using it as another way of visualizing the data, it is used as a highly interactive tool that allows scientists both to gain insight into the paleodata and also visually reconstruct the paleoenvironmental features.

The data transformations (MAT) described above provide a mechanism for extracting patterns from raw data. The output of this process is depicted using interactive PCP to facilitate the exploration of relationships among attributes (see Figure 8): i.e., each site of the database is drawn as a polyline passing through parallel axes, which represent the species, and the environmental variables that we want to reconstruct (last four axes on the right, in Figure 8). 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 axes. Each polyline of the database is color-coded and the MAT technique is used for that purpose (mapping the dissimilarity values to a color gradient, as explained above), i.e., the redder the polyline is, the more dissimilar it is with respect 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 (note that in this example, all environmental features are seasonal sea surface temperatures (SST)), which means that sample 20 corresponds to a cold period or a cold site.

However, several interaction techniques [18][8] have been integrated with this PCP to allow brushing [1], linking, animation, focus + context, etc., for exploratory analysis and knowledge discovery purposes.

3.3.1 Focus+context

In the Figure 8, a focus + context [18] technique can be seen for the labels of each site in the database. On the left side, each polyline is connected outside the axes with its label, allowing easy identification of each site. The labels are ordered top down, depending on different criteria chosen by the user, whether alphabetically, by dissimilarity index, according to latitude or no-crossing (the labels are arrahnged in order to prevent lines connecting with the values on the first axis from crossing over them). This way the expert can easily select the polyline of a particular site, which is highlighted 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 ordering criterium chosen and the position of the current focus. Since there is a shortage of space for so many labels, a fisheye approach [4], a powerful technique for organizing the suppresion of irrelevant data, was developed.

3.3.2 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 [3][17]. The range of an attribute can be specified by moving the handles at the top and bottom of a range slider (see Figure 8). 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 so that users can see the position of each one, and labels are also maintained on the background. Figure 9 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 axes. 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 highlighted. In the snapshot, the expert is comparing a site (orange polyline) with the reconstructed core sample.

Also note that in Figure 9, all axes 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 interactive visual analysis is that any axis can be dragged and dropped, so the order of the axis is changed. This way the shape of the polyline also changes, helping to reveal hidden patterns and making analysis easier. In Figure 10, the axis order has been altered slightly so the PCP is uncluttered. The elimination of selected variables (axes) can also be helpful.

3.3.3 Animation

Contrary to other approaches [15], time is not represented on an axis. As in the previous cases, time is one of the most relevant variables in the analysis. PaleoAnalogs, which can show an animation

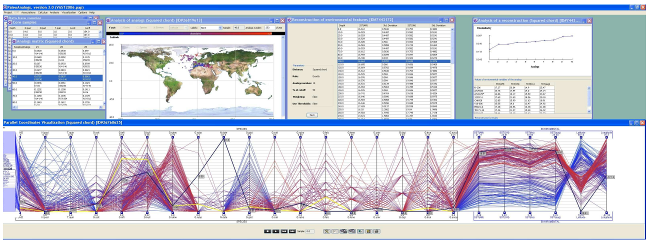
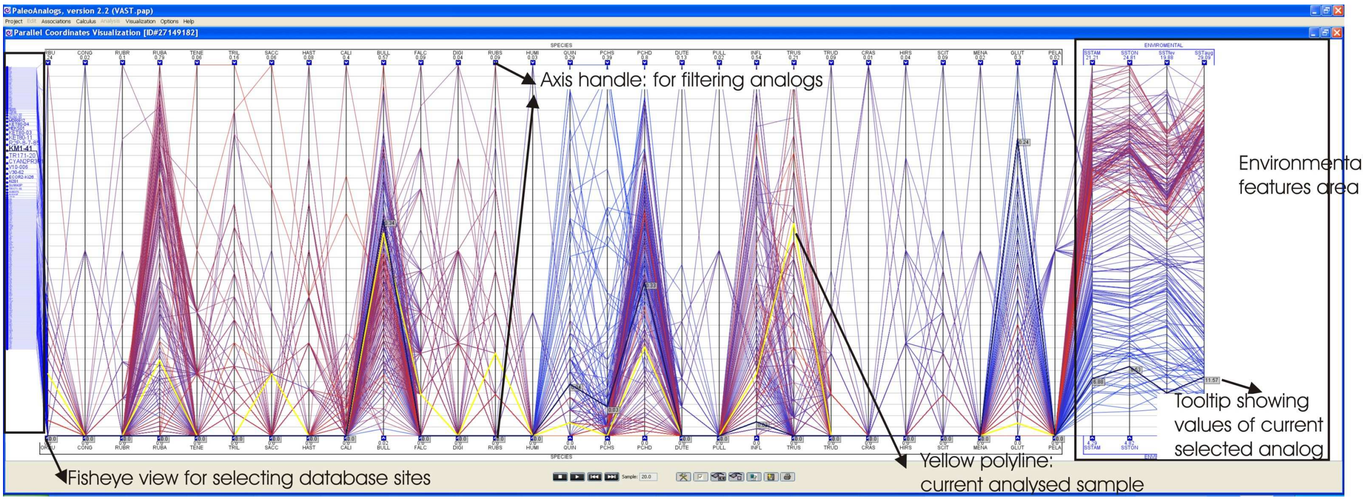


Figure 7: The PaleoAnalogs Visual Analytics Environment



PaleoAnalogs, version 2.2 (VAST.pap)
Project Edit Associations Calculus Analysis Visualization Options Help
Parallel Coordinates Visualization [ID#27149182]
SPECIES
ENVIROMENTAL
BU
CONG
RUBR
RUBA
TENE
TRIL
SACC
HAST
CALI
BULL
FALC
DIGI
RUBS
HUMI
QUIN
PCHS
PCHD
DUTE
PULL
INFL
TRUS
TRUD
CRAS
HIRS
SCIT
MENA
GLUT
PELA
SSTAM
SSTON
SSTfeb
SSTaug
KM1-41
Sample: 20.0
Axis handle: for filtering analogs
Environmental features area
Tooltip showing values of current selected analog
Yellow polyline: current analysed sample
Fisheye view for selecting database sites

Figure 8: Reconstruction visually driven by Parallel Coordinates: the yellow polyline show a pattern for the anlysed sample, the color of the rest polylines show how different species or environmental variables are related to that particular sample. By different means of interactions, those analogs that best fit the expert reasoning will be used to reconstruct the variables for that sample (the yellow segments will continue the polyline in the environmental features area).

of the PCP, enable paleoclimatologists to visualize the evolution of different species over the course of geological time, and study their relationships, both among themselves and between some species and SST, for instance.

3.3.4 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 useful in order to highlight how temperature varies over 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 most logical approach for the problem at hand, as all the benefits explained above can be put together in order to provide paleoceanographers with the best interactive visual tool for discovering knowledge and supporting decisions about climatic reconstructions.

In Figure 11 the brushing and multiple linked views approach of PaleoAnalogs can be seen. As in [3], there are three modes of brushing and linking that can be coordinated in all the multiple views described in the previous sections:

- probing: this mode is used to view more details about an object (e.g. site labels and dissimilarity values) and to gain an understanding of the relationships between the different views. Probing is a transient operation conducted by moving the mouse pointer over an object, highlighting 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 for two sites in the database). Objects remain painted until they are reset explicitly.

This way, a complete analysis that takes advantage of all the discused features can be performed. Paleontologists can have an animation going down the core, observing how the analogs distribution changes with time, latitude and longitude (map and/or axes). The evolution of the species patterns can also be seen. Thus, all the information with geological meaning can be used by the expert, who decide, by means of the different interactions, how each sample is reconstructed.

3.4 Presentation and Dissemination

As part of a future effort on providing a way for paleoceanographers to capture their analytic assessments, it is possible to work

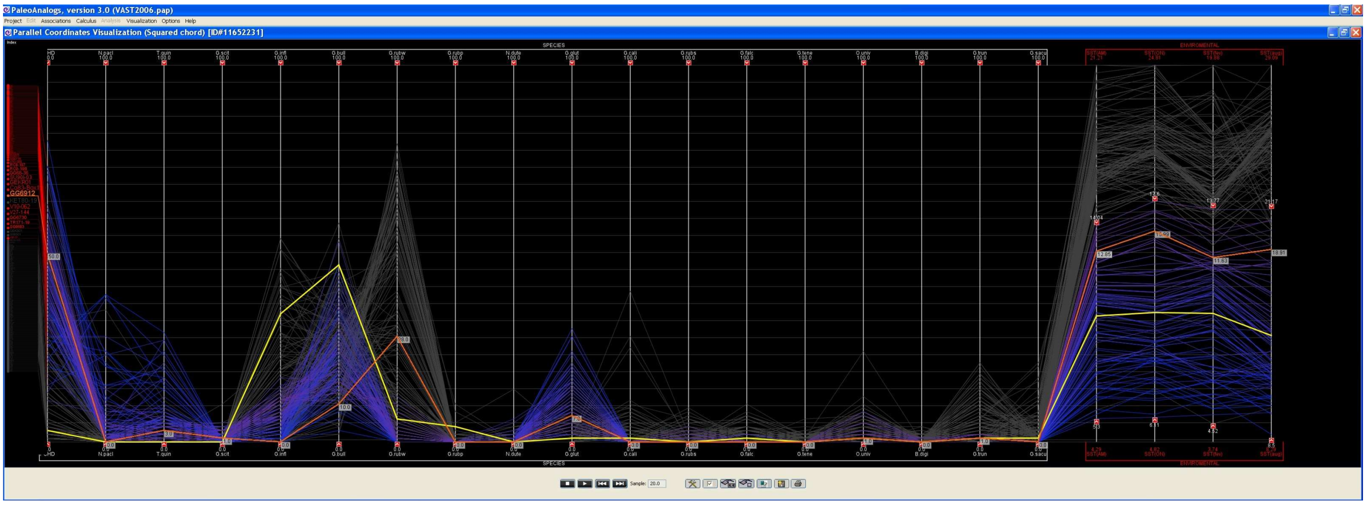


Figure 9: Analytical reasoning and reconstruction by means of interaction: the expert has decided to filter the analogs based on the similarity of SSTs, so the handles of environmental axes select only the blue polylines; the shape of the area covered by them is very similar to the sample pattern (yellow line). All the species axes now have the same scale, showing the relative importance among species.

with two different sets of colors, one for visual interaction and one for printing and image exportation (see Figure 8). Also, a paleoceanographer's analysis can be saved in a project, and the different stages can be independently saved and recovered. Also summaries, such as the associations used for one particular study, are automatically produced.

4 Conclusions and Future Work

In conclusion, the interactive analysis features introduced to PaleoAnalogs for understanding and visualizing complex dynamics in the fields of paleoclimatology and paleoceanography have produced promising and encouraging results. Together with the expert judgment of paleoceanographers, the provided visual analytics techniques foster a deeper understanding of the process of reconstruction of paleoenvironmental conditions. By introducing user-driven reconstructions the curse of black-box analytical methods can be broken and the domain implications can be assesed by the experts; as a result, more accurate models of the past oceanographic and climatic conditions can be produced. We have demonstrated the potential of an integrative approach to the visualization and analysis of the evolution of a given geographic location. In particular, we have focused on various practical issues concerning detecting relationships between species and environmental features in one oceaninc site expanding thousands of years.

As for future work, one major challenge has been identified: other classical analytical methods, which are currently used in the paleocenography field (such as principal components analysis or spectral analysis), are highly related to the MAT technique. The design and development of appropiate visualization tools, linked with the current ones, would provide an excellent environment for the analysts in which they could perform quantitative and qualitative studies, driving all the reconstruction process with their expertise.

Finally, we can add that the encouraging results indicate that this is a promising line of research with potential benefits to users from others disciplines that traditionally face similar prediction problems by means of machine learning methods.

Acknowledgements

This work was supported by the MCyT of Spain under project TIN2006-06313 and 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] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[3] 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.

[4] George W. Furnas. Generalized fisheye views. In *Human Factors in Computing Systems CHI '86 Conference Proceedings*, pages 16–23, 1986.

[5] W. H. Hutson. The agulhas current during the late pleistocene: Analysis of modern faunal analogs. *Science*, 207:64–66, 1980.

[6] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985.

[7] Alfred Inselberg. Conflict detection and planar resolution for air traffic control. In *Intelligent Transportation Systems*, 2001.

[8] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[9] 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.

[10] 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.

[11] 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.

[12] National Visualization and Analytics Center. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.

[13] D. Paillard, L. Labeyrie, and P. Yiou. Macintosh program performs time-series analysis. *Eos, Transactions, American Geophysical*

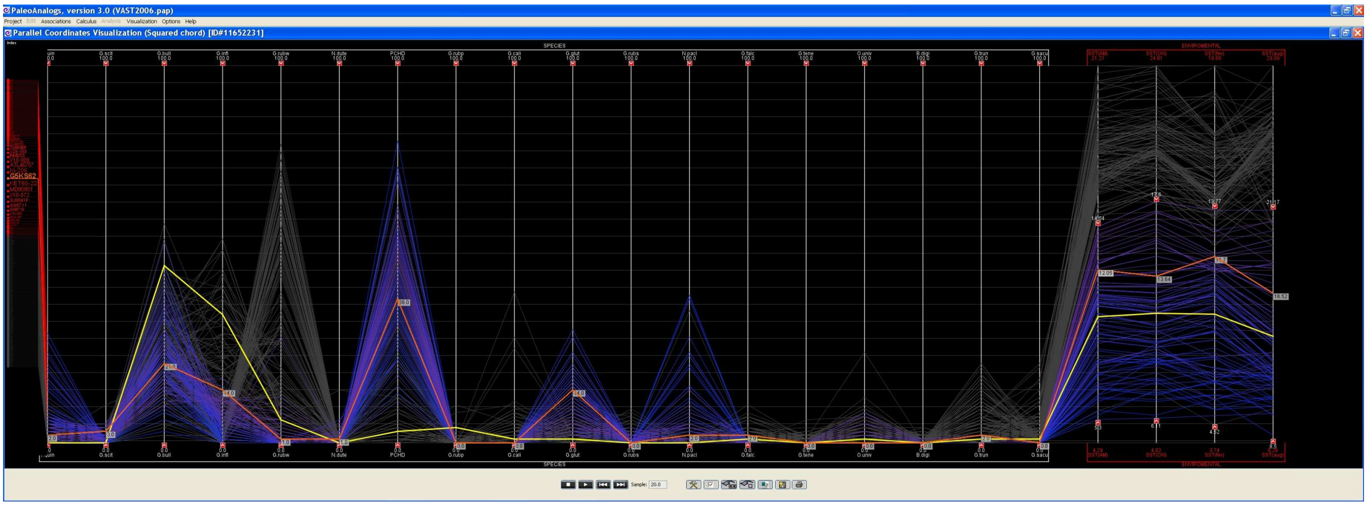

PaleoAnalogs, version 3.0 (VAST2006.pap)
Project Edit Associations Calculus Analysis Visualization Options Help
Parallel Coordinates Visualization (Squared chord) [ID#11652231]
SPECIES
ENVIROMENTAL
SST(AM) SST(ON) SST(fev) SST(aug)
PCHD N.dute G.ruber G.bull G.infl G.scit G.glut G.cali G.rubs N.pach G.falc G.tene G.trun G.sacu
Sample: 20.0

Figure 10: Uncluttering Parallel Coordinates by changing the axis order

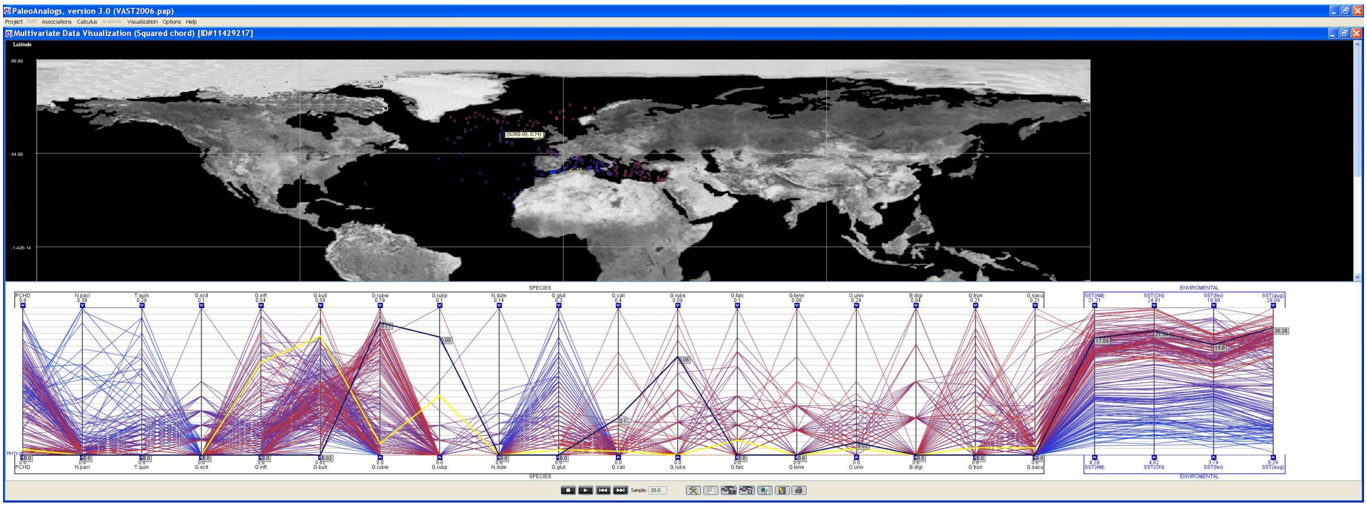

PaleoAnalogs, version 3.0 (VAST2006.pap)
Project Edit Associations Calculus Analysis Visualization Options Help
Multivariate Data Visualization (Squared chord) [ID#11429217]
Latitude
SPECIES
ENVIROMENTAL
SST(AM) SST(ON) SST(fev) SST(aug)
Sample: 20.0

Figure 11: Multiple linked views in PaleoAnalogs

Union, 77:379, 1996.

[14] 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.

[15] 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.

[16] 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.

[17] Jinwook Seo and Ben Shneiderman. Interactively exploring hierarchical clustering results. *IEEE Computer*, 35(7):80–86, 2002.

[18] Robert Spence. *Information Visualization*. Addison-Wesley, 1 edition, 2001.

[19] 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.

[20] 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.

[21] R. Theron, D. Paillard, E. Cortijo, J. A. Flores, M. Vaquero, F. J. Sierro, and C. Waelbroeck. Rapid reconstruction of paleoenvironmental features using a new multiplatform program. *Micropaleontology*, 50:391–395, 2004.

[22] 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.

[23] P. Yiou, E. Baert, and M. F. Loutre. Spectral analysis of climate data. *Surveys of Geophysics*, 17(6):619–663, 1996.