

PARETO ANALYSIS OF PALEONTOLOGICAL DATA: A NEW METHOD OF WEIGHING VARIABLE IMPORTANCE

[Robert E.A. Boudreau](#), [George Carmody](#),

and

[James J. Cheetham](#)

ABSTRACT

Pareto analysis is a method to determine which few of many variables significantly affect a measured end result. A macro for Microsoft Excel™ was developed and successfully applied to a set of population abundances for three arcellacean and one foraminiferal species with respect to which environmental variable was most important in determining abundance; either pH, salinity or oxygen content. It was found that **Criboelphidium gunteri**, a foraminifera, was abundant under conditions of high oxygen content and high salinity. However, **C. gunteri** abundance decreased upon an increase in pH. **Centropyxis aculeata**, an arcellacean, decreased in abundance with an increase in salinity above 5‰. Another arcellacean, **Diffflugia corona**, experienced a decreased abundance when salinity increased. The arcellacean **Lagenodifflugia vas** was not sensitive to an increase in the three variables tested. This analysis allows rapid identification of the most important variables in determining abundances of protists in an environmental setting.

¹[Robert E.A. Boudreau](#), ²[George Carmody](#), and ^{2†}[James J. Cheetham](#), Departments of ¹Earth Sciences and ²Biology, College of Natural Sciences, Carleton University, 1125 Colonel By Drive, Ottawa, ON, K1S.

†To whom correspondence should be addressed.

KEYWORDS: pareto, foraminifera, arcellacean, variables, environment

Copyright: Paleontological Society, 22 October 1999

Submission: 1 October 1999, Acceptance: 18 October 1999

Next Section...

Close Window

Pareto Analysis of Paleontological Data: A New Method of Weighing Variable Importance

Robert E.A. Boudreau, George Carmody and James J. Cheetham

PLAIN-LANGUAGE SUMMARY:

Pareto analysis is a method that can determine which variable in a set of chemical or physical parameters is most important in determining measured output. A macro for Excel™ was developed and successfully applied to a set of population abundances for four protists in order to determine which variable was most important; pH, salinity or oxygen content. It was found that **Criboelphidium gunteri**, a foraminifera, thrived under conditions of high oxygen content and high salinity, but decreased in abundance with an increase in pH. **Centropyxis aculeata**, an arcellacean, decreased in abundance with an increase in salinity above 5 parts per thousand. **Diffflugia corona**, another arcellacean, experienced a decreased abundance when salt content increased. **Lagenodifflugia vas**, a third arcellacean, was not sensitive to an increase in the three variables tested. This macro allows rapid identification of the most important variables in determining abundances of protists (arcellaceans and foraminifera) in an environmental setting.

INTRODUCTION

Micropaleontological research can create great quantities of data. The use of computers has increased the ability to collect and manage this data. Since time and resources are limited it is important to extract and understand as much useful information as possible from these large data sets in a timely fashion.

Pareto analysis allows for determination of which few of the many variables significantly affect measured end results. This methodology has traditionally been used to design and optimize industrial processes ([Haaland 1989](#)). We have applied Pareto analysis to a set of data containing relative abundances of arcellaceans and formaminifera from sediment-water interface samples from different environments. Previous research on arcellaceans indicate that **Centropyxis aculeata** ([Ehrenberg 1832](#)) is tolerant of brackish conditions ([Scott 1977](#); [Scott and Medioli 1980](#); [Collins 1996](#)), while **Diffugia corona** ([Wallich 1864](#)) and **Lagenodiffugia vas** ([Medioli and Scott 1983](#)) are opportunistic species in stressed lacustrine environments ([Boudreau 1999](#)). **Criboelphidium gunteri** ([Cole 1931](#)) is a euryhaline marine foraminifera. However, it has also been found alive in non marine settings including the stressed environment caused by salt spring injection of waters, up to and including brines, in northern Lake Winnipegosis, Manitoba ([Boudreau 1999](#); [Patterson et al. 1997](#); [McKillop et al. 1992](#)). In this paper we will demonstrate how Pareto analysis allowed us to quickly and easily identify the main environmental factors contributing to the abundance of these different species and to determine whether the unusual conditions found in Lake Winnipegosis modified their preferred habitats.

Next Section...

METHODS

Data from previous research was used to test the Pareto macro using a variety of variables. Physical variables from northern Lake Winnipegosis, Manitoba, were compared to abundances of arcellacean and foraminiferal data that was compiled using multivariate analysis ([Boudreau 1999](#); [Fishbein and Patterson 1993](#)). A three factor analysis using as variables pH, oxygen content and salinity was conducted for each of the abundances of **Centropyxis aculeata**, **Diffflugia corona**, **Criboelphidium gunteri** and **Lagenodifflugia vas**.

Two-level factorial designs were used for the analyses. Each factor was analyzed at a high and low setting. For three factor analysis, the data set was parsed as shown in [Table 1](#) where A represents pH, B represents oxygen content and C represents salinity.

A macro for Excel™ was constructed to analyze the data sets. Two-level factorial designs (clear signal design) were used because they are easy to interpret and they are effective. This analysis includes $2 \times 2 \times 2 = 8$ combinations of high and low values for the parameters. The main effects of each parameter were estimated by evaluating the difference in foraminiferal or arcellacean abundance caused by changing from a low value of the given parameter to a high value. The main effects were taken as the differences in average response values between high and low values. The Excel™ macro constructs a Pareto Chart from the input data which identifies the parameter(s) and interactions between the parameters that contribute most to the abundance of the arcellaceans and the foraminifera. The data used in the analysis are shown in [Table 2](#) and an example of a test run is shown in [Table 3](#) using the data for **Criboelphidium gunteri**.

Pareto analysis can also be done graphically. First, the data are plotted as shown in [Figure 1A](#). Then, data in the cube plot are averaged across one dimension, for example [salt], to generate a square plot as shown in [Figure 1B](#).

The data is further reduced by averaging across another dimension, for example [O₂] to yield a line plot as shown in [Figure 1C](#).

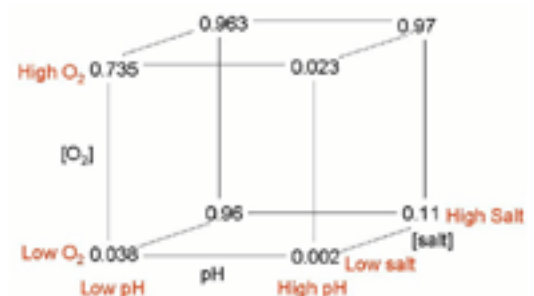


Figure 1A



The effect of going from low pH to high pH on the measured parameter, in this case relative abundance, is determined by subtracting the value at low pH from the value at high pH, in this case $0.277 - 0.674 = -0.397$. This result indicates that an increase in pH leads to a decrease in the abundance of this species. This technique is



Figure 1B

repeated in order to reduce the data to obtain the effects of the other variables; in this example, [O₂] and [salt].

Interaction between factors occurs when the effect of one factor depends on whether the other factor is at its high or low value.

Two-factor interactions can also be calculated using this analysis. Data is reduced to the square plot, and the average of runs in which the two factors are at extremes (ie. High-High and Low-Low) minus the average of the runs in which the factor levels are mixed (ie. High-Low and Low-High) will yield a number. If no two-factor interactions are occurring, this difference is zero.

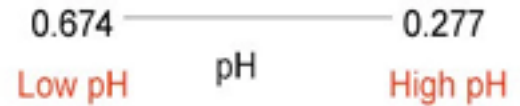


Figure 1C

The Excel spreadsheet allows one to conduct this analysis without the need to draw numerous graphs, and greatly speeds up the data analysis process.

Next Section...

RESULTS AND DISCUSSION

The results of the three-factor Pareto analysis for the three arcellaceans and the foraminifera are included as [Figure 2](#). This graph indicates that in the case of **Centropyxis aculeata**, oxygen is the most important variable, while increasing salinity would have decreased the population abundance. This is similar to results obtained in previous research, using multi-variate analysis, where **C.**

aculeata was found to be tolerant of salinities up to about 5‰. In addition, past research has shown that Centropyxids are excellent indicators of stressed environments with neutral pH conditions (5.5 - 7.5), and this is mirrored by Pareto analysis results, indicating that when pH is changed beyond this range, **Centropyxis aculeata** does less well, especially in combination with changes in salinity above 5‰.

Diffflugia corona decreased in abundance when salinity increased and was tolerant of changes in pH. Previous research indicated that Diffflugids in general are intolerant of much salinity, but certain of these arcellaceans have the ability to indicate their environments with respect to pH. Research in Lake Winnipegosis has shown that **D. corona** tends to dominate assemblages in an environment with raised pH (8.0) and low salinity (0 - 3‰), which is again mirrored by the results of the Pareto analysis.

Populations of **Criboelphidium gunteri**, a euryhaline foraminifera, were abundant under conditions of high oxygen content and high salinity, and found to be able to adapt to changing conditions of salinity in northern Lake Winnipegosis. This is also indicated by the results of the three factor Pareto analysis, where **C. gunteri** is tolerant of changes in oxygen and salinity, but not pH.

The arcellacean, **Lagenodifflugia vas**, was not sensitive to an increase in the three variables tested by Pareto analysis. This is what was expected, as previous research in Lake Winnipegosis indicated that **L. vas** was able to discriminate its environment under high pH (8.3), brackish (0 - 3‰) and variable oxygen conditions.

The Excel™ macro has the capability of analyzing three, four and five variable systems. Updates will also be available at: <http://www.carleton.ca/~jcheetha/pareto>. [PE Note: To get the macro, click on the file name. The Internet browser (Netscape, Explorer, etc.) will save the file ([MACRO.XLS](#)) or open it in a specified application.]

In order to use the macro, set up a number of test runs with variable information set as high (H) or low (L), and enter the data as indicated in [Table 3](#). For instance, in the three variable Pareto Analysis, eight runs need to be set up with run #1 having high values for all variables.

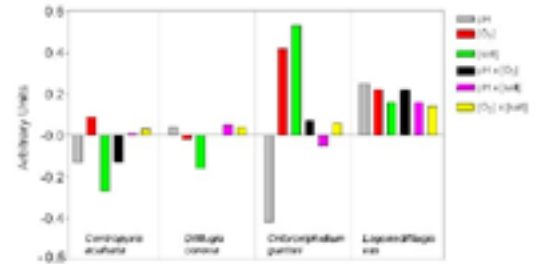


Figure 2

Close Window

REFERENCES

- Boudreau, R.E.A., 1999, Foraminifera and Arcellaceans from non-marine environments in northern Lake Winnipegosis, Manitoba, **M.Sc. Thesis, Carleton University**, 113p.
- Cole, W.S., 1931, The Pliocene and Pleistocene foraminifera of Florida: **Bulletin Florida State Geological Survey**, **6**, pp. 7 - 79.
- Collins, E.S., 1996, Marsh-estuarine benthic foraminiferal distributions and Holocene sea-level reconstructions along the South Carolina coastline, **Ph.D. Thesis, Dalhousie University**, Halifax, Canada, 240 p.
- Ehrenberg, C.G., 1832, Über die Entwicklung und Lebensdauer der Infusionsthier, nebst ferneren Beiträgen zu einer Vergleichung ihrer organischen Systeme, **Königliche Akademie der Wissenschaften zu Berlin Physikalische Abhandlungen**, 1831, pp. 1 - 154.
- Fishbein, E. and Patterson, R.T., 1993. Error-weighted maximum likelihood (EWML): a new statistically based method to cluster quantitative micropaleontological data. **Journal Of Paleontology**, **67(3)**, pp. 475 - 485.
- Haaland, P.D., 1989. Experimental Design in Biotechnology, Marcel Dekker Inc., New York and Basel.
- McKillop, W.B., Patterson, R.T., Delorme, L.D. and Nogrady, T., 1992, The Origin, Physico-chemistry and Biotics of Sodium Chloride Dominated Saline Waters on the Western Shore of Lake Winnipegosis, Manitoba, **The Canadian Field-Naturalist**, **106**, pp. 454 - 473.
- Medioli, F.S. and Scott, D.B., 1983, Holocene Arcellacea (Thecamoebians) from Eastern Canada, **Cushman Foundation for Foraminiferal Research**, Special Publication **21**, pp. 1 - 63.
- Patterson, R.T., McKillop, W.B., Kroker, S., Neilsen, E. and Reinhardt, E.G., 1997, evidence for rapid avian-mediated foraminiferal colonization of Lake Winnipegosis, Manitoba, during the Holocene Hypsithermal, **Journal of Paleolimnology**, **18**, pp. 131 - 143.
- Scott, D.B., 1977, Distributions and population dynamics of marsh-estuarine Foraminifera with applications to relocation Holocene sea-level: **Dalhousie University, Halifax, Nova Scotia, Ph.D. thesis**, 252 p.
- Scott, D.B. and Medioli, F.S., 1980, Quantitative studies of marsh foraminiferal distributions in Nova Scotia: their implications for the study of sea level changes, **Cushman Foundation for Foraminiferal Research**, Special Publication **17**, 58 p.
- Wallich, G.C., 1864, On the extent, and some of the principal causes, of structural variation among the difflugian rhizopods, **Annals and Magazine of Natural History**, series 3, **13**, pp. 215 - 245.

Close Window

Figure 1A. Cube plot of data for *Criboelphidium gunteri*.

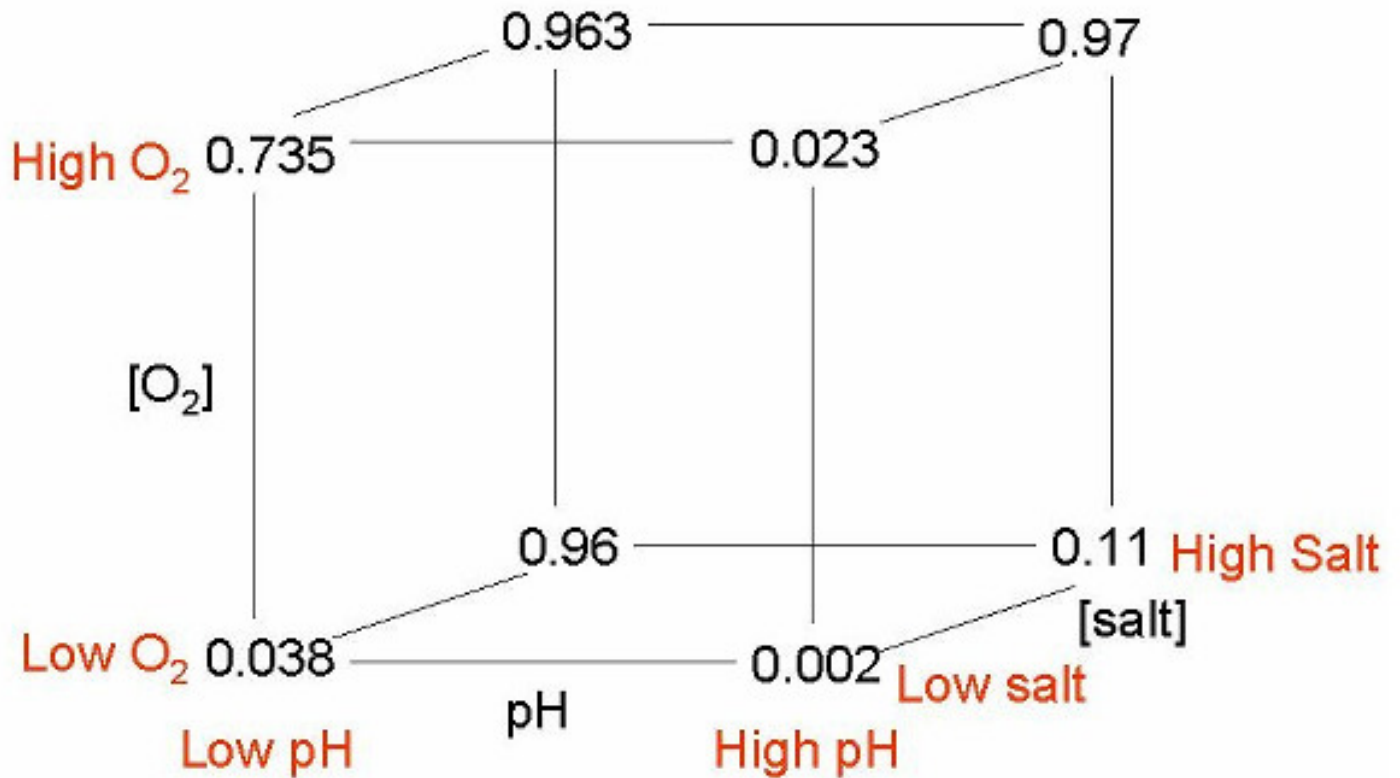


Figure 1B. Square plot for *Criboelphidium gunteri* averaged across salinity. i.e., The upper right edge of Figure 1A was averaged to produce the upper right corner of Figure 1B $[(0.023 + 0.97)/2 = 0.993/2 = 0.497]$.

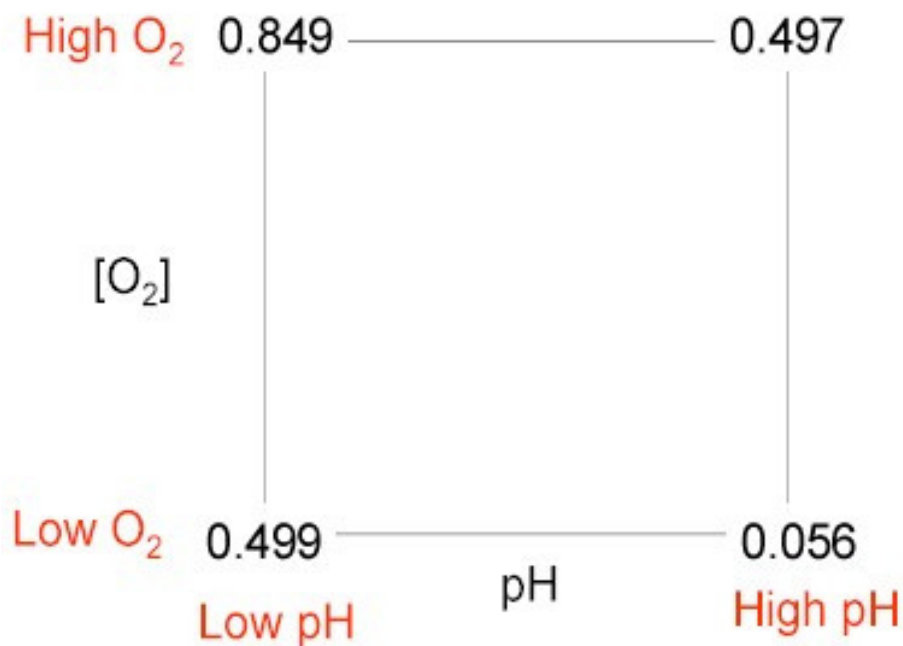
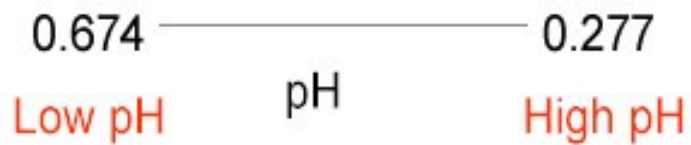
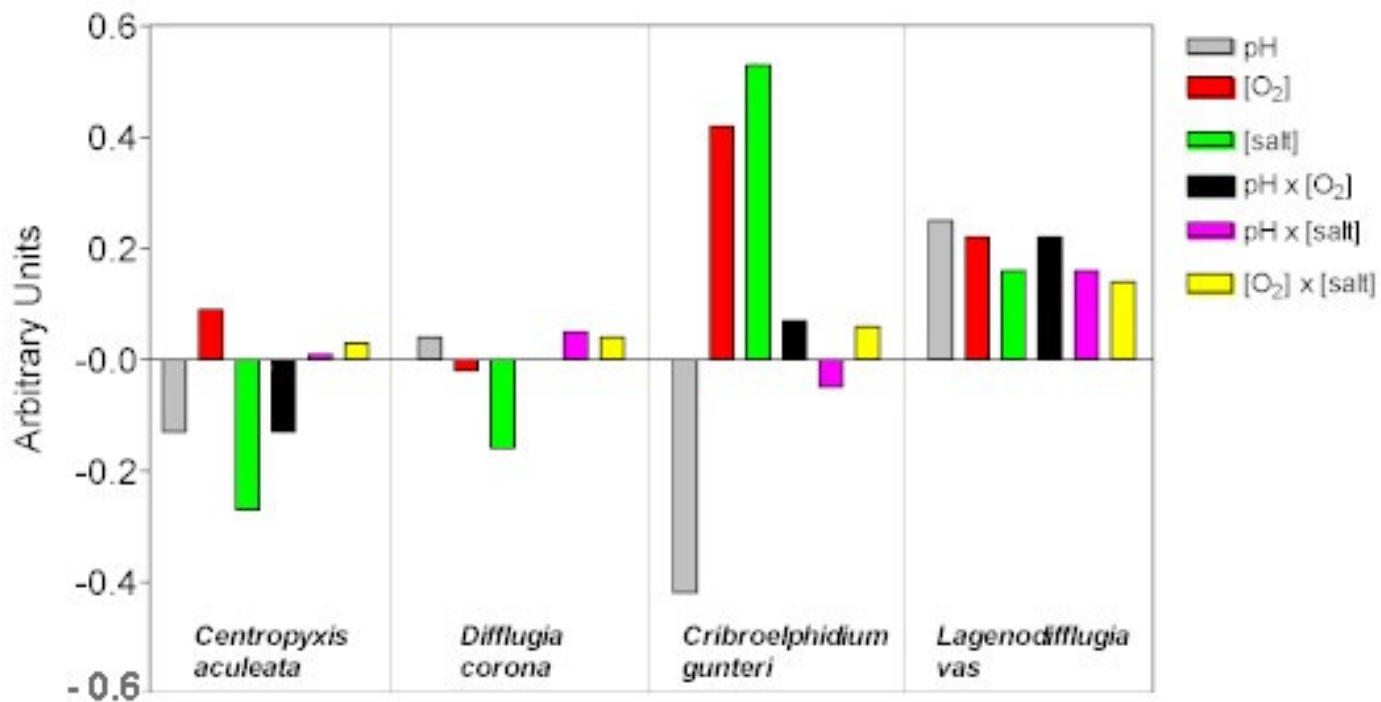


Figure 1C. Line plot of pH for **Criboelphidium gunteri**. The right hand side of Figure 1B was averaged to produce the right side of Figure 1C $[(0.497 + 0.056)/2 = 0.5525/2 = 0.277]$.



Close Window

Figure 2. Histograms of effects of pH, [O₂], [salt] and variable interactions on the abundances of four protists.



Close Window

Table 1 (Figure 1). Set up of Excel macro for three factor screening analysis. A represents pH, B represents oxygen content and C represents salinity.

	A	B	C	Observation
Run #1	Hi	Hi	Hi	
Run #2	Hi	Hi	Lo	
Run #3	Hi	Lo	Hi	
Run #4	Hi	Lo	Lo	
Run #5	Lo	Hi	Hi	
Run #6	Lo	Hi	Lo	
Run #7	Lo	Lo	Hi	
Run #8	Lo	Lo	Lo	

Close Window**Table 2.** Abundances of three arcellaceans and one foraminifera used with the variables pH, oxygen and salinity in the three factor Pareto analysis.

3 FACTOR ANALYSIS				pH = A	oxygen content = B	salinity = C	
run A B C				Centropyxis aculeata	Diffflugia corona	Criboelphidium gunteri	Lagenodiffflugia vas
1	H	H	H	0.04	0.12	0.97	0.76
2	H	H	L	0.25	0.20	0.023	0.17
3	H	L	H	0.036	0.11	0.11	0.05
4	H	L	L	0.333	0.25	0.002	0.007
5	L	H	H	0.265	0.038	0.963	0.00
6	L	H	L	0.54	0.20	0.735	0.005
7	L	L	H	0.038	0.00	0.96	0.00
8	L	L	L	0.328	0.27	0.038	0.002

Close Window

Table 3. Entry of data into the EXCEL macro for three factor screening analysis using the data for **Criboelphidium gunteri** from [Table 2](#). A represents pH, B is oxygen content and C is the salinity.

	A	B	C	Observation
Run #1	Hi	Hi	Hi	0.97
Run #2	Hi	Hi	Lo	0.023
Run #3	Hi	Lo	Hi	0.11
Run #4	Hi	Lo	Lo	0.002
Run #5	Lo	Hi	Hi	0.963
Run #6	Lo	Hi	Lo	0.735
Run #7	Lo	Lo	Hi	0.96
Run #8	Lo	Lo	Lo	0.038

C l o s e W i n d o w



Discussion: [PARETO ANALYSIS OF PALEONTOLOGICAL DATA: A NEW METHOD OF WEIGHING VARIABLE IMPORTANCE](#)

Robert E.A. Boudreau, George Carmody, and James J. Cheetham

To send comments, contact:

Norman MacLeod and Tim Patterson
Executive Editors



Palaeontologia_Electronica@nhm.ac.uk

(Norman)



Palaeontologia_Electronica@carleton.ca

(Tim)

Postings from recent e-mail messages regarding this article (none available at this time).