The Palaeobotanist 57(2008) : 529-537
0031-0174/2008 $2.00

Five useful techniques for analysing palynological data

CARLOS JARAMILLO

Smithsonian Tropical Research Institute, P.O. Box 0843 - 03092, Balboa, Ancón, Republic of Panamá.
Email: jaramilloc@si.edu

(Received 18 May, 2007; revised version accepted 07 April, 2008)

ABSTRACT

Jaramillo C 2008. Five useful techniques for analysing palynological data. The Palaeobotanist 57(3) : 529-537.

Palynologists often produce large quantitative data sets that can seldom be matched by other types of palcontological data. Although palynological data are subject to study by analytical techniques to answer questions regarding evolution, paleoclimate, and biogeography, the use of palynological data has often been qualitative, thus limiting their interpretation. Here, five techniques that can be used with palynological data are presented. These deal with diversity (number of species, evenness, diversity indices, and abundance distribution models), comparing similarities among samples, building a composite section, constructing species ranges, and estimating edge effects. The code necessary to perform these techniques has been included using *R for Statistical Computing*. *R* is an open-source and powerful statistical software available freely to anyone worldwide.

Key-words—Palynology, Data, Analytical techniques.

परागाणविक आँकड़े विश्लेषित करने हेतु पाँच उपयोगी तकनीकें

कार्लोस जरामिल्लो

सारांश

परागाणुविज्ञानविद् प्रायः विशाल मात्रात्मक आँकड़े सेट पेश करते हैं जो कि यदा-कदा परागाणविक आँकड़े के अन्य प्ररुप से सुमेलित हो सकते हैं। यद्यपि परागाणविक आँकड़े, विकास, पुराजलवायु एवं जैवभूगोल संबंधी प्रश्नों के जवाब देना वैश्लेषिक तकनीकों द्वारा अध्ययन पर निर्भर हैं, परागाणविक आँकड़े का उपयोग प्रायः गुणात्मक रहा है इस प्रकार उनके भाषांतरण को सीमित कर रहा है। यहाँ, पाँच तकनीकें जो कि परागाणविक आकड़ों में प्रयोग की जा सकती हैं पेश हैं। ये विविधता (जाति की संख्या, समता, विविधता अक्षाँक, और वितरण प्रतिरुपण बहुलता), नमूनों के बीच सदृश्यता तुलना, संयुक्त खंड की रचना, जाति श्रेणियां निर्मित करना तथा उपांत प्रभावों के आकलन का कार्य करते हैं। मैं भी सांख्यिकीय अभिकलन हेतु आर का प्रयोग इन तकनीकों के उपयोग करने में आवश्यक कोड समाविष्ट करता हूँ। आर किसी को भी मुक्त रुप से विश्वव्यापी उपलब्ध एक मुक्त-स्त्रोत सांख्यिकीय सॉफ्टवेयर है।

मुख्य शब्द—परागाणुविज्ञान, आँकड़ा, वैश्लेषिक तकनीकें।

INTRODUCTION

PALYNOLOGICAL (pollen and spore) data can be useful for understanding plant distribution and evolution through time (Birks & Line, 1992; Harrington, 2004; Haskell, 2001; Morley, 2000; Odgaard, 1999). They can also be used to understand palaeoceanography, palaeoclimatology, and the evolution of unicellular organisms (using dinoflagellates). Palynological data are often quantitative (species counts), with large number of samples, ideal for many types of analyses. In the past decade, many analytical techniques have been developed to solve ecological and palaeoecological problems.

Changing Scenario in Palaeobotany and Allied Subjects
Editor : J.S. Guleria

However, although large numbers of palynological data are continuously produced, few analytical techniques are extensively used in the palynological literature.

In this paper, five simple techniques have been presented to address typical questions that a palynologist usually tries to answer. These techniques can be applied by using a free, open-source, statistical software called *R for Statistical Computing (R-Development-Core-Team, 2005)* and the R packages Vegan (Oksanen *et al.*, 2005) and Labdsv (Roberts, 2005). R is a powerful software that runs on Unix, Windows and Macintosh operating systems. All R Codes needed to apply these techniques to palynological data are also presented.

FIVE TECHNIQUES

The most typical questions asked by a palynologist are related to diversity (how many species), palaeoecology (how samples or species relate to each other), and biostratigraphy (age of the assemblage). Here, techniques to approach each of these questions are presented. But, first you must install R for statistical Computing in your machine (go to http://www.r-project.org/), then install the packages Vegan and Labdsv (using the *Package Installer* in R tools), and then load the packages Vegan and Labdsv (using the *Package Manager*, which is in the menu of R).

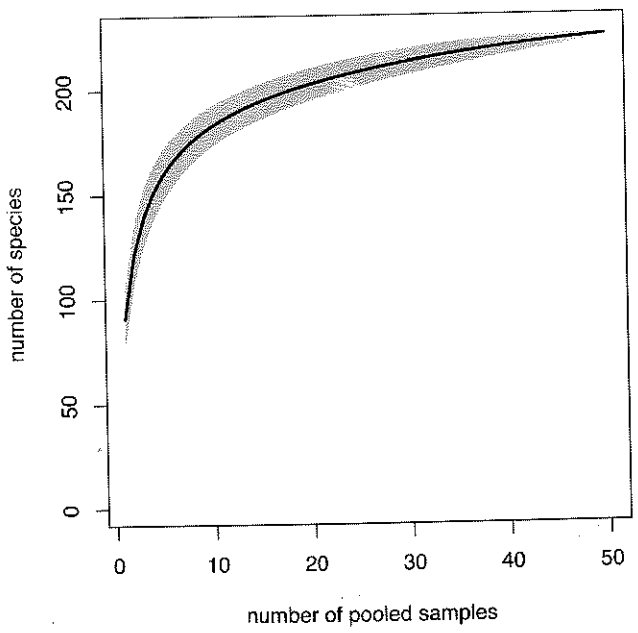

Fig. 1—Bootstrapped species accumulation curve, which shows how the number of species increases as the number of samples analysed increases. Shaded region shows the 95% confidence interval.

1. Estimating Diversity

Three aspects are important to consider when dealing with diversity: the number of species, the evenness and the abundance distribution.

(a) Number of Species

In this paper, the word "diversity" is used in its original sense to denote the number of species (Rosenzweig, 1995), which is also called "richness". Pollen can be a useful tool for estimating plant diversity through time (e.g. Morley, 2000); pollen mostly reflects genera and families (Germeraad *et al.*, 1968; Jackson & Williams, 2004), indicating that it can be used to track the plant diversity at that taxonomic level through geologic time.

Within-sample diversity (the number of species in a given sample) can be assessed using a technique called rarefaction (Hurlbert, 1971; Sanders, 1968). In order to estimate the number of species in a sample, you may want to count just the number of species in a given sample. However, the number of species is controlled by the number of specimens counted; thus, as more grains are counted, more species are found. Therefore, in order to compare the diversity among several different samples, you must first standardize the counting of all samples. Rarefaction does this for you. This is a technique that calculates the number of species expected at a given sample size smaller than the actual sample (Sanders, 1968). This technique is used to account for differences in diversity resulting from different sample sizes.

How to apply it? For every sample, a rarefaction must be performed to calculate the number of morphospecies found at a given count (e.g. 200, 300 grains). All samples that have smaller counts than the established cutoff count must be excluded.

R-code

First, let us assume a matrix *x*, which has two samples; the first one, *X1*, has a total count of 205 grains belonging to 50 species. The second sample, *X2*, has a total count of 250 grains belonging to 60 species. Species are in columns, samples in rows.

X1=c(21,3,14,2,1,1,1,1,2,3,4,7,2,7,1,1,1,1,1,1,20,
15,12,15,11,21,1,1,2,2,2,2,3,4,4,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,0,0,0,0,0,0,0,0,0,0)

X2=c(21,3,14,2,1,1,1,1,2,3,14,7,2,7,1,1,12,1,15,1,
20,15,12,15,11,21,1,1,2,2,2,2,3,4,4,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1)

x = rbind(X1, X2)

The rarefaction is performed as:

rarefy(x, sample, se = TRUE)

Sample is the subcounting size you want to rarefy to, and *se* is the standard error of the rarefaction.

e.g.

rarefy(x, 205, se=TRUE)

However, although large numbers of palynological data are continuously produced, few analytical techniques are extensively used in the palynological literature.

In this paper, five simple techniques have been presented to address typical questions that a palynologist usually tries to answer. These techniques can be applied by using a free, open-source, statistical software called *R for Statistical Computing (R-Development-Core-Team, 2005)* and the R packages Vegan (Oksanen *et al.*, 2005) and Labdsv (Roberts, 2005). R is a powerful software that runs on Unix, Windows and Macintosh operating systems. All R Codes needed to apply these techniques to palynological data are also presented.

FIVE TECHNIQUES

The most typical questions asked by a palynologist are related to diversity (how many species), palaeoecology (how samples or species relate to each other), and biostratigraphy (age of the assemblage). Here, techniques to approach each of these questions are presented. But, first you must install R for statistical Computing in your machine (go to http://www.r-project.org/), then install the packages Vegan and Labdsv (using the *Package Installer* in R tools), and then load the packages Vegan and Labdsv (using the *Package Manager*, which is in the menu of R).

1. Estimating Diversity

Three aspects are important to consider when dealing with diversity: the number of species, the evenness and the abundance distribution.

(a) Number of Species

In this paper, the word "diversity" is used in its original sense to denote the number of species (Rosenzweig, 1995), which is also called "richness". Pollen can be a useful tool for estimating plant diversity through time (e.g. Morley, 2000); pollen mostly reflects genera and families (Germeraad *et al.*, 1968; Jackson & Williams, 2004), indicating that it can be used to track the plant diversity at that taxonomic level through geologic time.

Within-sample diversity (the number of species in a given sample) can be assessed using a technique called rarefaction (Hurlbert, 1971; Sanders, 1968). In order to estimate the number of species in a sample, you may want to count just the number of species in a given sample. However, the number of species is controlled by the number of specimens counted; thus, as more grains are counted, more species are found. Therefore, in order to compare the diversity among several different samples, you must first standardize the counting of all samples. Rarefaction does this for you. This is a technique that calculates the number of species expected at a given sample size smaller than the actual sample (Sanders, 1968). This technique is used to account for differences in diversity resulting from different sample sizes.

How to apply it? For every sample, a rarefaction must be performed to calculate the number of morphospecies found at a given count (e.g. 200, 300 grains). All samples that have smaller counts than the established cutoff count must be excluded.

R-code

First, let us assume a matrix *x*, which has two samples; the first one, *X1*, has a total count of 205 grains belonging to 50 species. The second sample, *X2*, has a total count of 250 grains belonging to 60 species. Species are in columns, samples in rows.

X1=c(21, 3, 14, 2, 1, 1, 1, 1, 2, 3, 4, 7, 2, 7, 1, 1, 1, 1, 1, 1, 20, 15, 12, 15, 11, 21, 1, 1, 2, 2, 2, 2, 3, 4, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

X2=c(21, 3, 14, 2, 1, 1, 1, 1, 2, 3, 14, 7, 2, 7, 1, 1, 12, 1, 15, 1, 20, 15, 12, 15, 11, 21, 1, 1, 2, 2, 2, 2, 3, 4, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)

x = rbind (X1, X2)

The rarefaction is performed as:

rarefy (x, sample, se = TRUE)

Sample is the subcounting size you want to rarefy to, and *se* is the standard error of the rarefaction.

e.g.

rarefy (x, 205, se=TRUE)

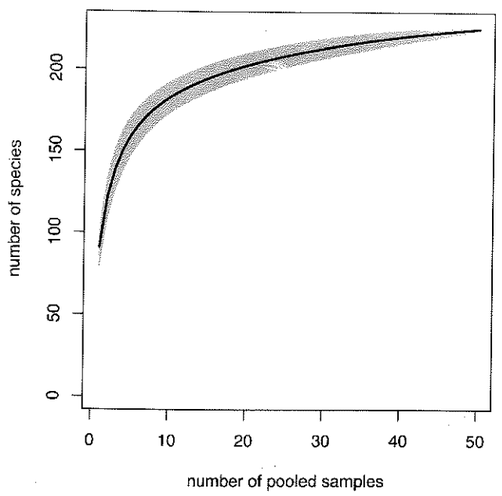

Fig. 1—Bootstrapped species accumulation curve, which shows how the number of species increases as the number of samples analysed increases. Shaded region shows the 95% confidence interval.

the result is

	X1	X2
S	50	53.458921
se	0	2.138707

The result is the number of species at a counting size of 205 grains. Sample X1 has 50 species, and sample X2 has 53.4 species with a standard error of 2.1. You have to remember that samples that do not reach the sample counting level must be discarded from the analysis, because rarefaction is not useful for estimating diversity beyond actual counts.

Among-sample diversity (how the diversity increases as you pool together many samples) can be calculated using bootstrapped species accumulation curves (Gilinsky, 1991). There may be two different scenarios: In one case, a region where the number of species per sample is very high, but when more samples are analysed from the same region or stratigraphic section, the number of new species (not found in the first sample), will not increase very much. In a second case, a sample might have few species, but every time a new sample is analysed, the species found are different from those in the previously analysed sample.

To compare among-sample diversities standardization of the sample size is essential, because as more samples are analysed, the probability of finding new species increases. Bootstrapping is a technique that facilitates comparison of intervals with different sample densities (Gilinsky, 1991). A single sample is selected at random and the number of species is counted based on that sample; a second sample is selected and the number of species is recalculated using the pooled data from both samples; a third is selected and the process continues until all samples are included (Colwell & Coddington, 1994). The whole process is repeated hundreds of times, and mean and standard deviation are calculated for each sampling level.

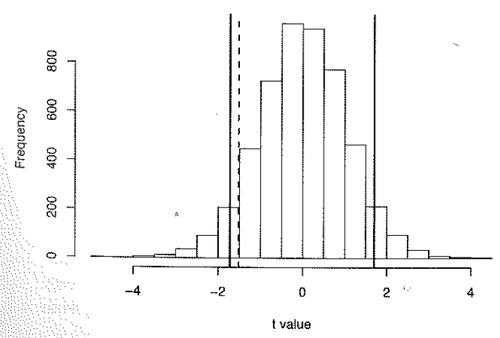

Fig. 2—Histogram of the *t* value derived from the null hypothesis (no difference in diversity between two group of samples), with the 95% confidence limits of the resulting distribution (continuous vertical lines). The dashed vertical line shows the *t* value of the original data.

R-Code

Let us assume a matrix *BCI* of abundance data from the Barro Colorado Island forest; species are in columns (225 species), samples in rows (50 samples).

data (BCI)

We calculate the species accumulation curve:

specaccum(BCI)

We can then plot the accumulation curve (Fig. 1):

plot(specaccum(BCI), ci.type="poly", col="black", lwd=2, ci.lty=0, ci.col="gray", ylab="number of species", xlab="number of pooled samples")

Plot is a function that graphs the species accumulation curve and its 95% confidence interval.

Comparing two results—A common situation is to have two different groups of samples, *x* and *y*, that you want to compare to see if they differ in diversity or any other metric, e.g. diversity changes across a major geological boundary like the Cretaceous-Palaeogene, or Permo-Triassic. Usually you would calculate the average rarefied diversity for each group of samples and then compare their results using a statistical test (e.g. student *t-test*). But, how do we know if the difference found is really significant? In other words, what is the probability of obtaining by pure chance the difference between *x* and *y* that you found? A useful technique to assess the degree of significance of a given difference is using a randomization analysis. The randomization procedure simulates the null hypothesis that both data sets, *x* and *y*, are from the same population. First you pool both the *x* and *y* data sets into a single set. Then you create two different sets by randomly sampling with replacement diversity estimates from the pooled samples. These two sets have the same number of samples as the original *x* and *y* groups. Then, a student *t-test* is calculated for the difference of the two sets. This procedure is repeated 5000 times. The resulting histogram of difference in *t*-tests, which simulates a *t*-test between two samples from the same population (the null hypothesis), is compared with the original *t*-test to evaluate its degree of significance.

R-Code

Let us assume two sets of samples, *x* and *y*. Set *x* has 15 samples, and set *y* has 20 samples. Each sample is a rarefied diversity at a counting of 200 grains, Set *x* has an average of 22 species per sample, and set *y* has an average of 25 species per sample.

x=rnorm(15, mean=22, sd=5)

y=rnorm(20, mean=25, sd=6)

We are going to compare the difference in average diversity among the two sets using a student *t*-test. First, we

the result is

	X1	X2
S	50	53.458921
se	0	2.138707

The result is the number of species at a counting size of 205 grains. Sample X1 has 50 species, and sample X2 has 53.4 species with a standard error of 2.1. You have to remember that samples that do not reach the sample counting level must be discarded from the analysis, because rarefaction is not useful for estimating diversity beyond actual counts.

Among-sample diversity (how the diversity increases as you pool together many samples) can be calculated using bootstrapped species accumulation curves (Gilinsky, 1991). There may be two different scenarios: In one case, a region where the number of species per sample is very high, but when more samples are analysed from the same region or stratigraphic section, the number of new species (not found in the first sample), will not increase very much. In a second case, a sample might have few species, but every time a new sample is analysed, the species found are different from those in the previously analysed sample.

To compare among-sample diversities standardization of the sample size is essential, because as more samples are analysed, the probability of finding new species increases. Bootstrapping is a technique that facilitates comparison of intervals with different sample densities (Gilinsky, 1991). A single sample is selected at random and the number of species is counted based on that sample; a second sample is selected and the number of species is recalculated using the pooled data from both samples; a third is selected and the process continues until all samples are included (Colwell & Coddington, 1994). The whole process is repeated hundreds of times, and mean and standard deviation are calculated for each sampling level.

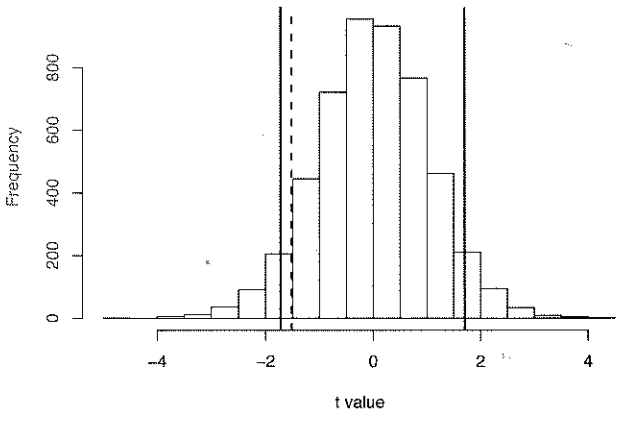

Fig. 2—Histogram of the t value derived from the null hypothesis (no difference in diversity between two group of samples), with the 95% confidence limits of the resulting distribution (continuous vertical lines). The dashed vertical line shows the t value of the original data.

R-Code

Let us assume a matrix *BCI* of abundance data from the Barro Colorado Island forest; species are in columns (225 species), samples in rows (50 samples).

```
data(BCI)
```

We calculate the species accumulation curve:

```
specaccum(BCI)
```

We can then plot the accumulation curve (Fig. 1):

```
plot(specaccum(BCI), ci.type="poly", col="black", lwd=2, ci.lty=0, ci.col="gray", ylab="number of species", xlab="number of pooled samples")
```

Plot is a function that graphs the species accumulation curve and its 95% confidence interval.

Comparing two results—A common situation is to have two different groups of samples, x and y, that you want to compare to see if they differ in diversity or any other metric, e.g. diversity changes across a major geological boundary like the Cretaceous-Palaeogene, or Permo-Triassic. Usually you would calculate the average rarefied diversity for each group of samples and then compare their results using a statistical test (e.g. student *t-test*). But, how do we know if the difference found is really significant? In other words, what is the probability of obtaining by pure chance the difference between x and y that you found? A useful technique to assess the degree of significance of a given difference is using a randomization analysis. The randomization procedure simulates the null hypothesis that both data sets, x and y, are from the same population. First you pool both the x and y data sets into a single set. Then you create two different sets by randomly sampling with replacement diversity estimates from the pooled samples. These two sets have the same number of samples as the original x and y groups. Then, a student *t-test* is calculated for the difference of the two sets. This procedure is repeated 5000 times. The resulting histogram of difference in t-tests, which simulates a t-test between two samples from the same population (the null hypothesis), is compared with the original t-test to evaluate its degree of significance.

R-Code

Let us assume two sets of samples, x and y. Set x has 15 samples, and set y has 20 samples. Each sample is a rarefied diversity at a counting of 200 grains, Set x has an average of 22 species per sample, and set y has an average of 25 species per sample.

```
x=rnorm(15, mean=22, sd=5)
y=rnorm(20, mean=25, sd=6)
```

We are going to compare the difference in average diversity among the two sets using a student t-test. First, we

need to produce a new vector, *bootsp200*, pooling the *x* and *y* sets together. This vector simulates a single population that contains samples from both *x* and *y*.

```
bootsp200=c(x,y)
```

Second, we replicate the original sets, *x* and *y*, by randomly choosing samples from the pooled *x* and *y* data sets (the vector that we just created, *bootsp200*). Thus, we are creating a null hypothesis (both *x* and *y* sets are coming from the same population). Then, we compare the average of the two simulated data sets using a student *t*-test, and use the parameter *t* produced by the test as the metric to evaluate the differences. We repeat this process 5000 times. The resulting value is stored in the vector *xy.boot*.

```
nrand<-5000
xy.boot =numeric(nrand)
for(i in 1:nrand){
x.boot=sample(bootsp200,15, replace=TRUE)
y.boot=sample(bootsp200,20, replace=TRUE)
xy.boot[i]= t.test(x.boot,y.boot)$statistic
}
```

Then we plot a histogram of the *t* value derived from the null hypothesis, and draw the 95% confidence limits of the resulting distribution (Fig. 2). Finally we plot the parameter *t* produced by the *t*-test when the original data sets, *x* and *y*, are compared. Thus, we can evaluate the probability of finding the *t* value of the original data set by chance (the null hypothesis).

```
hist(xy.boot, xlab="t value")
clim<-quantile(xy.boot, c(0.05, 0.95))
abline(v=clim, lwd=2)
abline(v= t.test(x,y)$statistic, lwd=2, col="red", lty=2)
```

(b) Evenness

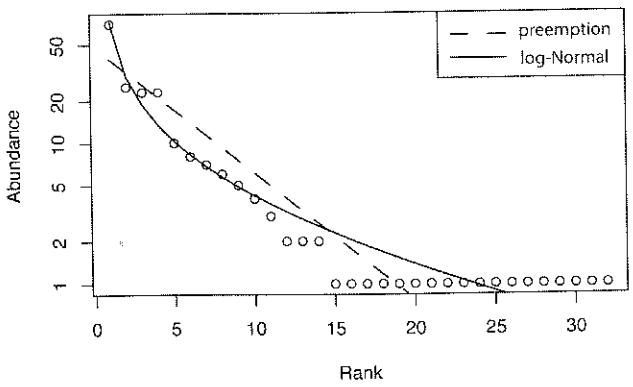

Fig. 3—Species abundance distribution compared to two abundance models (preemption and veiled log-Normal). Species are ordered along x axis, from the most abundant to the right, to the less abundant to the left. The deviance is used to evaluate what model fits the data better (in this case the log-Normal fits the data better).

It is also useful to evaluate the variation in the species abundances among a community. Pielou's evenness (J) summarizes in a single metric the abundance distribution of a population (Hayek & Buzas, 1997). J=(H/log S) where H= Shannon Index ="S $p_i \log_{10} p_i$, where p_i is the proportional abundance of species i, and S= number of species. The highest possible evenness ($J = 1$) occurs when all species have the same number of individuals. A low value indicates that most individuals belong to very few species.

R-Code

Let us assume an abundance data matrix x *that has samples in rows and species in columns. The Pielou's index J is calculated as:*

```
J<- diversity(x)/log(specnumber(x))
```

(c) Diversity Indices

Ecologists have developed several metrics that attempt to summarize in a single number both the number of species and the evenness. These metrics try to answer, using a single number, how well the abundances among the species in a sample are distributed and how many species the sample has. They have called these metrics "diversity indices". There are many indices to choose from, and each one has both strengths and weaknesses, making the choice of a particular index very subjective. A useful index is the Shannon index (H= -"$p_i \log_{10} p_i$, p_i=proportion of individuals that belong to species i). It is a measure of uncertainty of a selection process (Hayek & Buzas, 1997; Zar, 1999). A maximum value of H occurs when species are equally abundant in a sample, and the uncertainty of knowing which species will be observed next would therefore be highest (Hayek & Buzas, 1997). H is a function of the number of species and the abundance distribution of individuals within those species (Hayek & Buzas, 1997).

R-Code

Let's assume an abundance matrix *x*, where species are in columns and samples in rows. The Shannon index H is calculated as:

```
H<- diversity(x)
```

(d) Abundance Distribution Models

It is useful to evaluate the shape of the abundance distribution of a sample, and compare it with several statistical models of species abundance distribution (Magurran, 2004). There is usually a strong bias in pollen abundance toward wind-pollinated taxa, but still, the shape of the abundance distribution could be quite informative. It also can be very useful when comparing dinoflagellate cyst communities from stressed versus non-stressed environments.

There are many models of species abundance distribution. The most common types are the niche preemption model (also called geometric series or Motomura model), and

the veiled log-Normal model (Wilson, 1991). The empirical abundance distribution is compared with those models, and the best-fitting model is often calculated using deviance as a parameter for fitness between the empirical data and a given abundance model.

The niche preemption model calculates the expected abundance, a_r of species at rank r as $a_r = J\,a(1 - a)^{r-1}$. Only one parameter is estimated, the preemption coefficient a, which gives the decay rate of abundance per rank (Oksanen *et al.*, 2005; Wilson, 1991). There is also a fixed scaling parameter J, which is the total abundance. In the preemption model, the most abundant species takes a proportion k of some limiting resource, the second most dominant takes the same fraction k of the remainder, and so on until all species are accommodated in the community (Magurran, 1988). This type of abundance distribution is found primarily in harsh environments dominated by few species such as alpine forests (Wilson, 1991). The veiled log-Normal model assumes that the logarithmic abundances are distributed normally, or $a_r = \exp(\log \hat{\imath} + \log s\,N)$, where a_r is the expected abundance a of species at rank r; $\log \hat{\imath}$ is the fitted mean of log abundance, s is the fitted standard deviation of Ln abundance, and N is a Normal deviate that includes a parameter, a veil, that assumes that only a proportion of the most common species were observed in the community (Oksanen *et al.*, 2005; Wilson, 1991). This type of abundance distribution is very common among most plant communities from non-extreme environments (Magurran, 2004).

R-Code

Let us assume a sample x that contains the species abundances of 32 species.

x = c(70, 25, 23, 23, 10, 8, 7, 6, 5, 4, 3, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)

The deviance of the preemption model versus the empirical data is calculated as (Fig. 3):

```
rad.preempt(x)
plot(rad.preempt(x))
```

and the deviance of the veiled log-Normal model versus the empirical data is calculated as (Fig. 3):

```
rad.veil(x)
plot(rad.veil(x))
```

2. Similarity among samples

Often, it is required to find out differences or similarities among groups of samples along a stratigraphic profile, or across a region. A useful technique to solve this problem is using similarity indices . Many of them have been proposed in the ecological literature over the years. One of the most effective similarity measures is the Sorensen index (Magurran, 2004; Sorensen, 1948). It is a presence/absence index that is simple to calculate and interpret. It ranges from one, when two samples have the same species, to zero, when no species are in common among two samples.

Sorensen= 2a / (2a + b +c), a= total number of species present in both samples, b= number of species present only in sample 1, c= number of species present only in sample 2.

However, this index does not take species abundance into account. Two samples could be very different when the most abundant species are compared. A similarity index that takes abundances into account is the Morisita-Horn index, a widely used index that is not strongly influenced by the number of species and sample size, but is sensitive to the most abundant species (Wolda, 1981, 1983). It also ranges from zero (no shared species) to one (identical species and abundances). It is often recommended to transform the

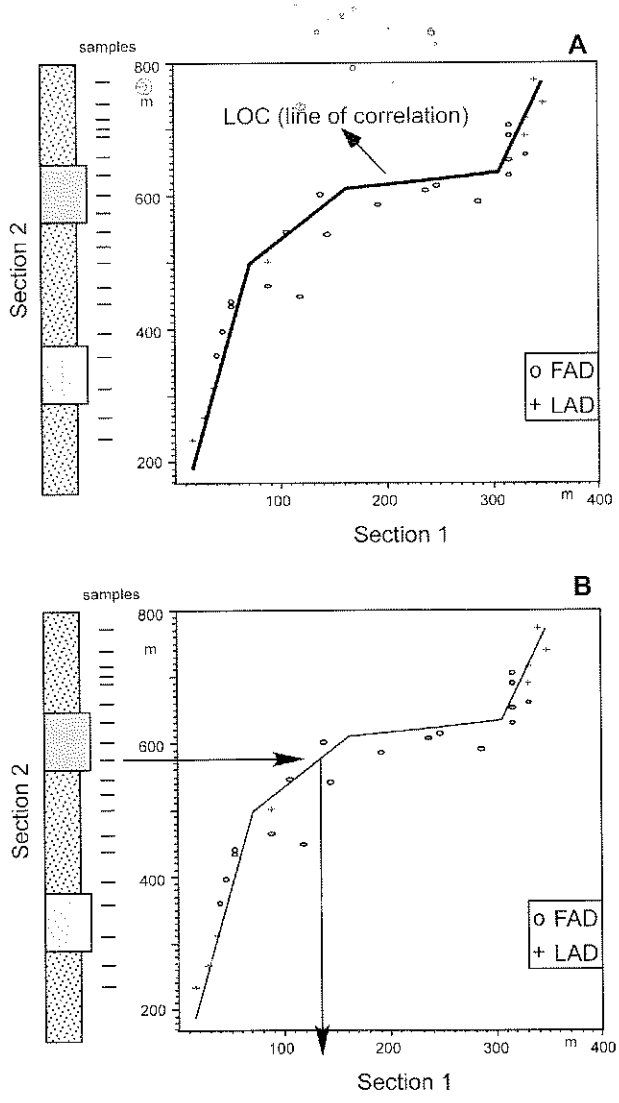

Fig. 4—Graphic correlation and building a composite section. (4A) The line of correlation between section 1 and 2 is traced based upon the distribution of FAD (first appearance datum) and LAD (last appearance datum). (4B) Samples from section 2 are extrapolated into section 1, using the line of correlation. This procedure allows the location of all samples, from all sections, into a single section, the Composite Section.

abundance data before the analysis by taking the square root of each value, to minimize the effect of very abundant species on the Morisita-Horn index.

Morisita-Horn = 2*"(x[ij]*x[ik])/((d[j]+d[k]) * "(x[ij])*"(x[ik])), where

d[j] ="(x[ij]^2)/Nj^2)

d[k] ="(x[ik]^2)/Nk^2)

Nj= total number of individuals at sample J

Nk= total number of individuals at sample K

x[ij]= number of individuals in the ith species in sample J

x[ik]= number of individuals in the ith species in sample K

R-Code

Let's assume an abundance matrix *x* that contains species in columns and samples in rows. Sorensen and Morisita-Horn are calculated as dissimilarities (1 minus the index).

```
dsvdis(x, index="sorensen")
vegdist(x, method = "horn")
```

3. Building a Composite Section

It is often the case that there may have several stratigraphic sections across a region. How could one build up an overall summary of the pattern seen in the fossil record? One tool is producing a composite section. A good method to construct a composite section is graphic correlation (Edwards, 1984, 1989; Shaw, 1964). A detailed explanation of this technique was published by Edwards (1989). Bivariate plots are made where the points of origination and extinction of the taxa present in one section are compared against the same taxa in another section. Based upon the distribution of the origin and extinction points of the plot, a line of correlation is traced (Fig. 4a). This line represents time equivalence among the two sections that are being compared. Using this line, all samples from one section can be extrapolated to the other section. This procedure is applied to every available section, until all samples from all sections are extrapolated to a single section, thus constructing a composite section (Fig. 4b).

No code has been implemented in R to perform graphic correlation. Such a code is expected in the near future. However, one can use a simple bivariate plot (x versus y), and perform the graphic correlation by hand.

4. Constructing species ranges, the Range-through Method

It is often necessary to calculate species ranges, either for biostratigraphic purposes or to estimate standing diversity (number of species at a given time). The range-through method (Boltovskoy, 1988) is very useful for such purposes. This method assumes that a taxon is present in all samples that lie between its first and last appearance datums. It minimizes the effect of facies-related fossils and differences in capture probability on biostratigraphic ranges and standing diversity. Calculation of standing diversity often excludes unique taxa (those that are present in only one sample), because they can introduce noise into the diversity pattern (Wing, 1998).

R-Code

First, we define a function (*fill.occur*) that will perform the range-through method

```
fill.occur=function(sp)
{
occur=which(sp>0) ##array of row numbers where sp>0
fad=occur[1] #row of first number in array
numboccur=length(occur)# how many numbers in the array
lad=occur[numboccur]#row number of the last position in the array
alloccur=rep(0,length(sp))## produces a matrix of zeroes similar in size to original
i=1:length(sp)## all positions in the matrix are now labeled i
alloccur[i>=fad & i<=lad]=1## replaces all i in between fad and lad by 1
return(alloccur)#gives the matrix out
}
```

Then, we apply the function *fill.occur* to an abundance matrix *x*, which has species in rows, and samples in columns. It produces a presence/absence matrix, *y*, with the range-through already applied

```
y<- apply(x,2,fill.occur)
```

Standing diversity can then be calculated by summing the diversity of each sample of matrix *y*

```
apply(y,2,sum)
```

The positions of the first appearance datum (FAD) and last appearance datum (LAD) of each species in the composite section can be calculated using the function *fadlad* and a vector containing the stratigraphic position of each sample (*depthtotal*)

```
fadlad=function(sp,depth)#funtion to calculate FAD and LAD
{
occur=which(sp>0) ##array of row numbers where sp>0
fad=occur[1] #row of first number in array
numboccur=length(occur)# how many numbers in the array
lad=occur[numboccur]#row number of the last position in the array
```

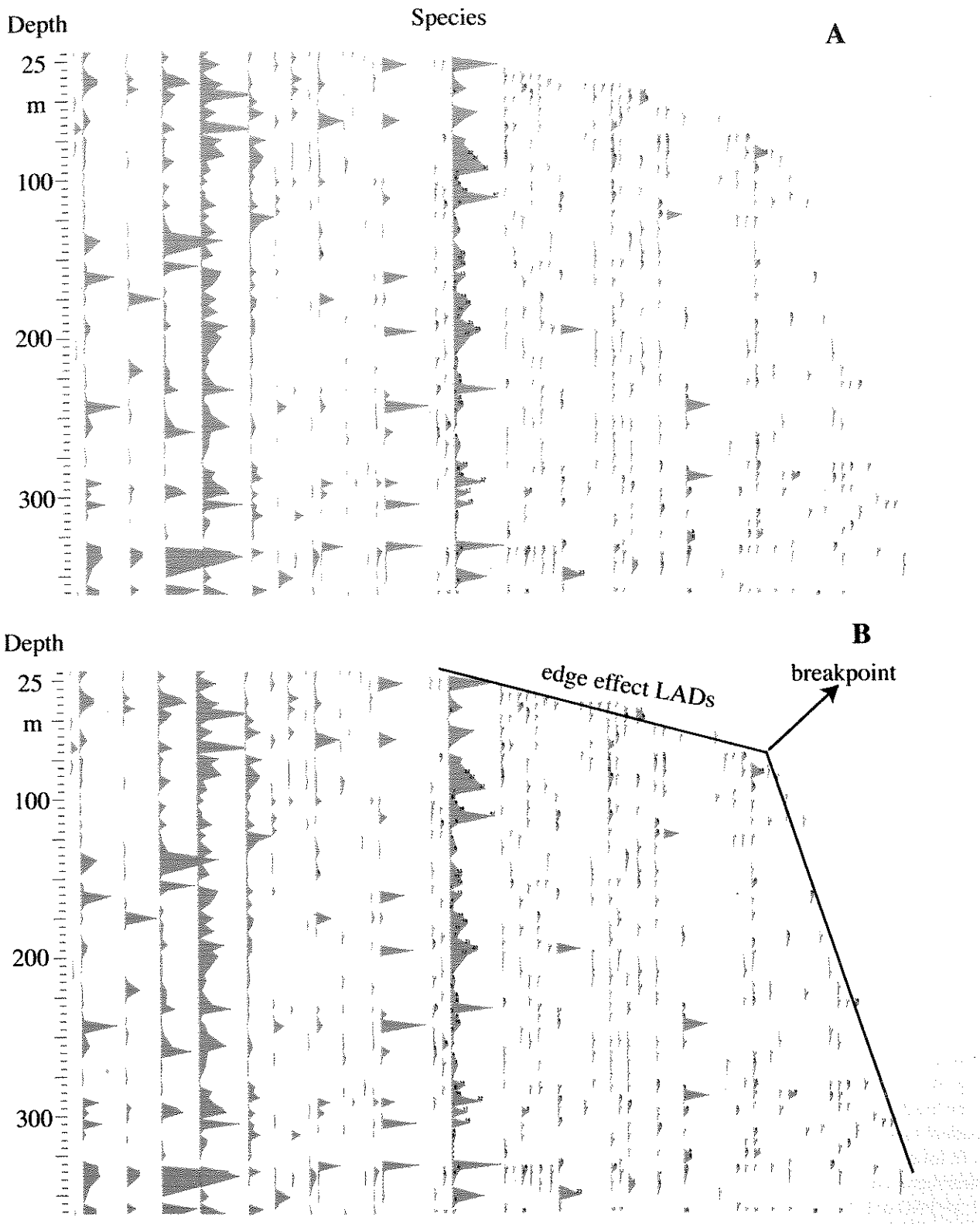

Fig. 5—Biostratigraphic range chart. (5A) Species are ordered based upon the stratigraphic position of their LAD, from the youngest to the left, to the oldest to the right. Notice that there is a linear arrangement of FADs at the top of a section. This is produced by the edge effect. (5B) Piecewise regression to find the breakpoint, the point that indicates where the edge effect weakens.

```
return(c(depth[fad],depth[lad]))
}
```

The FAD and LAD of each species in matrix *y* (species must be in rows, samples in columns) are then calculated as:

```
apply(y,1,fadlad,depthtotal)
```

5. Estimating Edge Effects

Edge effects are very problematic in biostratigraphy, palaeoecology and palaeodiversity. They are produced because not all species have the same probability of being found (Foote, 2000). A rare species appears in the fossil record after its time of origination, and it also disappears from the fossil record before its time of extinction. On the other hand, abundant species usually begin and end close to their real origination and extinction events. In a typical palynological sample most of the species have rare to moderate abundances. The FAD and LAD calculated for those species in a section are going to be offset from their real point of origination and extinction. The edge effect is more significant when the section analysed has few samples or is stratigraphically very short. The edge effect is a major problem in palynostratigraphy and few analytical solutions have been proposed to deal with it. A capture-recapture model, often used in zoological studies, would be ideal for this purpose, but it still has not been developed.

One tool, albeit not perfect, to estimate the edge effect is using a piecewise regression. If we arrange the species according to their LADs, we are going to notice that there is often a linear arrangement at the top of a section (Fig. 5a). Something similar happens at the bottom of the section if we arrange the species according to their FADs. It seems reasonable to assume that the linear pattern is produced by the edge effect. Often the most abundant species had a LAD very close to the end of the section, and species less abundant tend to have their LAD farther apart from the top of the section; the rarer the species, the farther apart (Fig. 5a). The position of the linear pattern could be found by performing a piecewise regression. This regression assumes that there are two different regression functions to the same data (SPSS, 1999) and attempts a two-segment fit of the data. The breakpoint is the intersection of the two fitted regression lines and would represent the point where the edge effect becomes minimal (Fig. 5b). The regression iteratively tries all possible positions of the breakpoint and chooses the one that produces the lowest residual sum of squares (Yeager & Ultsch, 1989). The model to fit follows Duggleby and Ward (1991) for a two-segment linear regression. $y = y_T + [(m_L + m_R)(x - x_T) - (m_L - m_R)\,½x\text{-}x_T½] / 2$ y= FAD or LAD, x = species, x_T = breakpoint species, y_T = breakpoint FAD or LAD, m_L = slope left of breakpoint, m_R = slope right of breakpoint.

R-Code

Piecewise regression for FAD. It calculates a piecewise regression of the vector *x*, which is a set of FAD values given for a section. It returns the position of the breakpoint, *breakFAD*, which fits the data better. It is important to exclude from the analysis all species that are present in the oldest sample analysed (for FAD analysis), and all species that are present in the youngest sample (for LAD analysis).

```
fad.edge<-sort(x,decreasing = TRUE)
cumuedge.fad<-c(1:length(fad.edge))
step1<-numeric(length(fad.edge))
for (i in (1:length(fad.edge))){
step1[i]<-sum(resid(lm(fad.edge[1:i]~cumuedge.fad[1:i]))
^2)
}
step2<-numeric(length(fad.edge))
for (i in (1:length(fad.edge))){
step2[i]<-sum(resid(lm(fad.edge[i:length(fad.edge)]
~cumuedge.fad[i:length(fad.edge)]))^2)
}
piecewise<-step1+step2## sum of first and second
segment
breakFAD= fad.edge[which(piecewise==min(piecewise
))]
```

Piecewise regression for LAD. It calculates a piecewise regression of the vector *y*, which is a set of LAD values given for a section. It returns the position of the breakpoint, *breakLAD*, which fits the data better.

```
lad.edge<-sorty,decreasing = TRUE)
cumuedge.lad<-c(1:length(lad.edge))
step1<-numeric(length(lad.edge))##first segment of
piecewise
for (i in (1:length(lad.edge))){
step1[i]<-sum(resid(lm(fad.edge[1:i]~cumuedge.fad[1:i]))
^2)
}
step2<-numeric(length(lad.edge))##last segment of
piecewise
for (i in (1: length(lad.edge))){
step2[i]<-sum(resid(lm(fad.edge[i:218]~cumuedge.
fad[i:218]))^2)
}
piecewiselad<-step1+step2## sum of first and second
segment
breakLAD= lad.edge[which(piecewiselad==min(
piecewiselad))]
```

CONCLUSIONS

Five simple analytical techniques to be used with palynological data were presented. These techniques deal with

diversity issues, similarity comparisons, building a composite section, constructing species ranges, and estimating edge effects. The source code to implement these techniques in the free, open-source *R for Statistical Computing* software is given. There is still much room for improving the handling of data among the palynological community.

Acknowledgements—*The presented work is supported by the Smithsonian Paleobiology Endowment Fund, the Fondo para la Investigación de Ciencia y Tecnología Banco de la Republica of Colombia, and the Unrestricted Endowments SI Grants. Richard Condit helped with the R code used in the analysis. Shivani Moodley, Iann Sanchez, and Natasha Atkins gave useful comments to the manuscript. Surangi Punyasena and Walton Green also provided input. Special thanks are due to M.I. Barreto for her continuous support and source of ideas.*

REFERENCES

Birks HJB & Line JM 1992. The use of rarefaction analysis for estimating palynological richness from Quaternary pollen-analytical data. The Holocene 2: 1-10.

Boltovskoy D 1988. The range-through method and first-last appearance data in paleontological surveys. Journal of Paleontology 62: 157-159.

Colwell RK & Coddington JA 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London Series B 345: 101-118.

Duggleby RG & Ward LC 1991. Analysis of physiological data characterized by two regimes separated by an abrupt transition. Physiological Zoology 64: 885-889.

Edwards LE 1984. Insights on why graphic correlation (Shaw's method) works. Journal of Geology 92: 583-597.

Edwards LE 1989. Supplemented graphic correlation: A powerful tool for paleontologists and nonpaleontologists. Palaios 4: 127-143.

Foote M 2000. Origination and extinction components of taxonomic diversity: general problems. *In:* Erwin DH & Wing SL (Editors)—Deep time: Paleobiology's perspective: 74-102. The Paleontological Society, Lawrence.

Germeraad JH, Hopping CA & Muller J 1968. Palynology of Tertiary sediments from tropical areas. Review of Palaeobotany and Palynology 6: 189-348.

Gilinsky NL 1991. Bootstrapping and the fossil record. *In:* Gilinsky NL & Signor PW (Editors)—Analytical paleobiology: 185-206. Paleontological Society, Pittsburgh.

Harrington G 2004. Structure of the North American vegetation gradient during the late Paleocene/early Eocene warm climate. Evolutionary Ecology Research 6: 33-48.

Haskell J 2001. The latitudinal gradient of diversity through the Holocene as recorded by fossil pollen in Europe. Evolutionary Ecology Research 3: 345-360.

Hayek LC & Buzas MA 1997. Surveying natural populations: New York, Columbia University Press, 563 p.

Hurlbert SH 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecological Monographs 54: 187-211.

Jackson ST & Williams JW 2004. Modern analogs in Quaternary paleoecology: Here today, gone yesterday, gone tomorrow? Annual Review of Earth and Planetary Sciences 32: 495-537.

Magurran AE 1988. Ecological diversity and its measurement: New Jersey, Princeton University Press, 179 p.

Magurran AE 2004. Measuring Biological Diversity: Malden, MA, USA, Blackwell Publishing, 256 p.

Morley RJ 2000. Origin and evolution of tropical rain forests: New York, John Wiley & Sons, 362 p.

Odgaard BV 1999. Fossil pollen as a record of past biodiversity. Journal of Biogeography 26: 7–17.

Oksanen J, Kindt R & O'Hara B 2005. Community Ecology Package, Package VEGAN, R for Statistical Computing.

R-Development-Core-Team 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Roberts D 2005. Laboratory for Dynamic Synthetic Vegephenomenology, R for Statistical Computing.

Rosenzweig ML 1995. Species diversity in space and time: Cambridge, Cambridge University Press, 433 p.

Sanders HL 1968. Marine benthic diversity: a comparative study. American Naturalist 102: : 243-282.

Shaw AB 1964. Time in stratigraphy: New York, McGraw-Hill, 365 p.

Sorensen T 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biologsike Skrifter 5: 1-34.

SPSS 1999. Systat 9, statistics II: Chicago, SPSS Inc., 552 p.

Wilson JB 1991. Methods for fitting dominance/diversity curves. Journal of Vegetation Science 2: 35-46.

Wing SL 1998. Late Paleocene-early Eocene floral and climatic change in the Bighorn Basin, Wyoming. *In:* Berggren W, Aubry MP & Lucas S (Editors)—Late Paleocene-early Eocene biotic and climatic events: 371-391. Columbia University Press, New York.

Wolda H 1981. Similarity indices, sample size and diversity. Oecologia 50: 296-302.

Wolda H 1983. Diversity, diversity indices and tropical cockroaches. Oecologia 58: 290-298.

Yeager DP & Ultsch GR 1989. Physiological regulation and conformation: A BASIC program for the determination of critical points. Physiological Zoology 62: 888-907.

Zar JH 1999. Biostatistical Analysis: Englewood Cliffs, N.J., Prentice-Hall, 663 p.