Palaeontologia Electronica
http://palaeo-electronica.org

fossil: Palaeoecological and palaeogeographical analysis tools

Matthew J. Vavrek

ABSTRACT

The *fossil* software package is a collection of analytical tools to synthetically analyse ecological and geographical data sets. The software is designed to be used with the R Statistical Language and is under an Open Source license, making it free to download, use or modify. The package includes functions for estimating species richness, shared species/beta diversity, species area curves and geographic distances and areas. The package also contains extensive documentation and examples of how to use all of the functions.

Matthew J. Vavrek. Redpath Museum, McGill University, 859 Sherbrooke St. W., Montreal, Quebec H4A 2K6, Canada. matthew@matthewvavrek.com

KEY WORDS: data analysis; software; palaeoecology; biogeography; palaeobiodiversity

INTRODUCTION

Multivariate analyses in palaeontology have become an increasing focus of many palaeontological research programs, especially with the development over the past decade of large datasets (e.g., Paleobiology Database; Carrano 2000; Alroy et al. 2001; Carrasco et al. 2005) and readily available computing power. A variety of statistical programs and software has been used and developed by and for palaeontologists, ecologists and evolutionary biologists as these massive data sets have become more commonplace (e.g., Hammer et al. 2001; Colwell 2009; Harrison and Larsson 2008; Maddison and Maddison 2009).

Large databases necessarily involve large numbers of collaborators, which may lead to an issue of heterogeneity and incompatibility of computing platforms and file formats. Despite the large number of freely available programs, there are few truly cross platform solutions available. One statistical environment gaining recognition over the last decade with its ability to perform intensive statistical analyses has been the R Statistical Language (R Development Core Team 2010; Ezard and Purvis 2009). This software is cross platform, freely available (Open Source) and has an extensive installed user and contributor base. While the base software when installed can perform many common statistical procedures, the software is easily extensible through packages, such as phylogenetic analysis (Paradis et al. 2004), time series analysis (Hunt 2008) and palaeobiological phylogenies (Ezard and Purvis 2009). These packages are available through a central repository called the Comprehensive R Archive Network, or CRAN. Additionally, data from virtually any source can be used, from plain text and Microsoft Excel tables to images and GIS shapefiles, and graphs and figures can be output in virtually any format. This flex-

PE Article Number: 14.1.1T
Copyright: Paleontological Society. March 2011
Copyright of *fossil* package and code: M.J. Vavrek
Submission: 9 April 2010. Acceptance: 30 December 2010

Vavrek, Matthew J., 2011. *fossil:* Palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica* Vol. 14, Issue1; 1T:16p;
http://palaeo-electronica.org/2011_1/238/index.html

ibility and availability is what has made it a growing success in the field of statistics and database analysis.

Here I present a new package that has been developed to enable a selection of ecological and geographic analysis tools to be added to the base R environment. The package was originally developed with palaeontologists in mind and is appropriately entitled *fossil*. As of this writing, it is in version 0.3.2, and although there are planned additions to the code, the functions already present allow for a large number of analyses to be performed.

Reasons for developing *fossil* are many fold. The underlying impetus was to create a single package to examine large datasets with up-to-date methods of biodiversity estimators and ecological pattern recognition that can be used in conjunction with geographic data over long time scales. Macroecological analysis is a growing area and palaeontologists now have a real opportunity to answer modern questions of biodiversity distributions, thanks in large part to the deep time of the fossil record. By providing powerful tools that integrate well, we can spend more time on the questions rather than the methods.

A number of the functions that have been implemented in *fossil* can also be found in the excellent package *vegan* (Oksanen et al. 2010). Many of the species diversity and species estimator functions are implemented in both packages. However, the *fossil* package was implemented to cover a number of use cases that *vegan* did not cover. Initially, the primary function that was needed was a way to estimate species diversity using a number of functions all at once. As well, the function to create distance matrices with user defined measures was at the time more difficult to use, and so I have tried to implement a more easily extensible method. The *fossil* package also implements a number of spatial analysis and export tools that are not found within *vegan*, such as methods to calculate geographic distances and areas from a set of points.

For example, the fossil record, while accurate, is by no means complete (Benton et al. 2000) yet can still provide important information on biogeographic patterns. Using *fossil,* we can compare sparse ecological data with a number of ecological similarity indices (e.g., Chao-Jaccard, Chao-Sorenson, Simpson) and then observe the patterns of connectivity using various types of neighbour joining techniques. These patterns can then be visualised in ecological space, using ordinations to group similar sites, and in geographic space, placing localities on a map and observing how this ecological connectivity relates to geography. Combining spatial, ecological and temporal data provides a more complete picture of the evolution of the biosphere than any one factor alone.

WHAT IS R AND WHY SHOULD WE USE IT?

The *fossil* package is constructed for use with the R Statistical Language. R owes its origins to the S Language, a program initiated at Bell Labs in the 1970s as a way to implement a computational statistical language (Becker et al. 1988). The S Language has been the basis for another well known statistical program, S-PLUS. In 1991 Ross Ihaka and Robert Gentleman at the University of Auckland began developing a statistical language for their teaching laboratory since no adequate commercial solution existed at the time. Their work mimicked many of the styles and methods of S, and eventually this package evolved into the R Language for Statistical Computing (Ihaka and Gentleman 1996). Since its origins, R has been open-sourced under the GNU Public License, meaning that anyone who chooses to use, redistribute or improve the software is free to do so provided they allow others the same rights (Stallman 1999). The program was originally written for a Macintosh system, but it has since been ported to virtually every computing architecture, both legacy and modern. This makes it an ideal candidate for a statistical system in many modern laboratories, where researchers possess their own (if not multiple) computers, often with different operating systems.

Many other statistical programs encourage their users to manually select their data and choose the analyses to be run with a mouse cursor. At first glance, this is a much simpler way of interacting with the data, but it suffers from a major drawback; analyses of this type are not truly reproducible (Leisch and Rossini 2003; Green 2003). Although descriptions of statistical procedures used in refereed papers are a must, trying to record exact mouse clicks and button selections is virtually impossible. R on the other hand encourages users to record each and every step of the process used. Most users of R will write their methods of analysis out in a text editor of some kind and then proceed to run this code in the R environment, with every step, from analysis to figure creation, fully documented.

The deeper benefits of this method may not be obvious. I have personally experienced situations where mistakes were made early on in the process of data analysis and not found until much later. While in a graphical, mouse driven environment trying to repeat all the steps necessary is often time consuming, well written R code can be easily modified and re-run with minimal fuss. Further, as the program is consistent across platforms, collaborators can run the code on their platform of choice, without having to worry if their version of a program has the same available functions. This benefit also extends to other scientists, who by taking other researchers' code can re-run published findings exactly, without having to purchase software of any kind.

What follows is not an in-depth introduction to R; there have already been many books written on the subject. For a good start, the original text by Becker et al. (1988) and a more recent text by Braun and Murdoch (2008) are highly recommended. Rather the focus of this paper is the use of the functions found within the *fossil* package.

SETTING UP THE ENVIRONMENT

R is available for virtually any platform and can be installed from the R Project website, www.r-project.org/. Please note that throughout this paper all R commands are distinguished from ordinary text using a bold-face font, and blocks of R commands are set apart in monospace font. All commands are preceded by a chevron (>) that does not need to be entered, but simply represents the beginning of a new command.

Throughout much of this paper, I use a theoretical data set called **fdata,** consisting of three parts. **fdata.list** is a table with each row representing an individual species occurrence and columns for locality name, species name, species abundance, latitude and longitude. **fdata.mat** is a matrix (12 by 12) with each unique species as a row and each locality as a column. The last part is **fdata.lats,** a SpatialPoints object containing the longitude and latitude for each locality. All of this data is found as part of the *fossil* package. As well, the entirety of the code used to analyse the data and create figures for this paper is available as a supplementary file, along with full instructions on how to use it.

To begin using the *fossil* package in an interactive session, you must first ensure the package has been installed on your computer. It is available online from **CRAN** and can be downloaded from within an R session by typing **install.packages ('fossil')** at the command prompt. You will be prompted to choose a download location; simply try to choose one closest to your location. Once the *fossil* package is available on your computer, you can load it in to R using the command **library (fossil).** Every time you start a new session, you will have to load the package again using the **library()** command as extra libraries are not loaded by default to keep the memory use as low as possible.

```
>library(fossil)
```

LOADING YOUR DATA IN R

Large databases used in palaeoecology studies are often simply tables, whether in plain text files or Excel tables, where every row consists of a unique observation, usually of a species at some location in space and time. However, the species, locations and times in these lists are rarely unique, and often consolidation of the data into usable matrices of species versus location is needed. There are two functions that aid in the conversion of lists of points into two types of matrices that will be referred to throughout the remainder of the paper. The first function is the **create.matrix()** function, which takes a list of species and their occurrences and converts it to a matrix of species (rows) by localities (columns). With the commands

```
>data(fdata.list)
>create.matrix(fdata.list,tax.name="s
pecies",locality="locality")
```

we can create an occurrence matrix from the fdata.list example data set; alternatively, if we wish to create an abundance matrix, we use virtually the same command, but include the option **abund = TRUE** and give the name of the abundance column (in this case, **'abundance'**) for the **abund.col** option. This method will give us an abundance matrix identical to **fdata.mat.**

```
>data(fdata.list)
>create.matrix(fdata.list,tax.name="s
pecies",locality="locality",
+abund=TRUE,abund.col="abundance")
```

For the *fossil* package, data follows the convention of species as rows and localities as columns. Data that is in matrix format already but with species as columns and localities as rows can be transposed with the **t()** command.

Similarly, much palaeontological data comes with some sort of spatial data about its provenance integrated with the occurrence data. As such, the locality data is often duplicated for each unique species at a certain site. In order to simplify plotting

TABLE 1. Names, formulas and alternate names for included similarity coefficients. Variables in the formulae are: a = number of shared species, b = number of species found only in the first sample, and c = the number of species found only in the second sample.

Coefficient Name	Formulae	Alternate Name	Function Call
Jaccard	a/(a+b+c)	Coefficient of Community	jaccard()
Sorenson	2a/(2a+b+c)	Dice, Czekanowski, Coincidence Index	sorenson()
Simpson	a/(a+min(b,c))	-	simpson()
Braun-Blanquet	a/(a+max(b,c))	-	braun.blanquet()
Ochiai	a/sqrt{(a+b)(a+c)}	Coefficient of Closeness	ochiai()
Kulczynski	[a/(a+b)+a/(a+c)]/2	-	kulczynski()

georeferenced data, a function called **create.lats()** can be used to extract the site coordinates from a list, eliminating duplicate entries.

```
>data(fdata.list)
>create.lats(fdata.list,loc="locality
",long="longitude",
+lat="latitude")
```

DISTANCE/SIMILARITY/BETA DIVERSITY INDICES

Measuring the ecologic distance between sets of samples is often a necessary first step in many multivariate analyses (Green 1980; Shi 1993). As such, it also is often a contentious one, with different researchers advocating different measures, with at times multiple correct arguments. Although I do not wish to provide a full explanation here of every single measure, I will provide a brief overview of those included in the *fossil* package. Some of these measures are best described as indices of beta diversity, although they are grouped here with other similarity measures for convenience since they are typically used in a similar fashion.

All of the similarity functions can be used in the same way. The functions need two arguments representing the two samples. It is important that the species occurrences are arranged in the same way for each site, and that any absent species are represented by a zero.

```
>sampleA<-c(1,1,0,1,1,1,1)
>sampleB<-c(0,1,1,0,0,1,1)
>sorenson(sampleA,sampleB)

[1]0.6
```

The species estimator functions included can be broadly grouped into two categories, those that use occurrence data and those that use abundance data. As abundance data is not always available, especially in palaeontology, more measures that use occurrence data are included in the package. Occurrence based measures can also be used with abundance data, but the abundance matrix is converted to an occurrence matrix by the function.

One of the oldest and best known occurrence measures is the Jaccard measure, also known as the Coefficient of Community (Table 1; Jaccard 1901; Shi 1993). The measure has seen extensive use, largely due to its simplicity and intuitiveness (Shi 1993; Magurran 2004). A similar measure also in common use is the Sorenson measure (also known as Dice, Czekanowski or Coincidence Index), which places more emphasis on the shared species present rather than the unshared, as can be seen in the difference in values for the example data set. Again, the calculation is relatively simple and intuitive, and both indices have been shown to provide useful results (Wolda 1981; Hubálek 1982). Two other similar indices that are occasionally used are the Ochiai and Kulczynski measures. While Hubálek (1982) lists the Ochiai and Kulczynski indices as providing good results, the Jaccard or Sorenson are typically more recommended if only because they are more commonly used.

One of the most common problems in palaeontology, and indeed in many ecological studies, is that of differing sample sizes. Comparing two sites of very unequal sampling intensities can give a biased view of the actual species overlap. For example, a subsample of a site could be considered identical to the original site, as all the species in the subsample will be within the original. However, all the previous measures would show less than complete similarity due to their mathematical properties. With this in mind, Simpson (1960) developed a measure, which can account for variability of sample sizes. His formula scales the value by the number of species from the least sampled site, so that the subsample in this case would have full similarity with the original. The Simpson measure is often used with data that is highly variable in sampling intensity, such as fossil datasets, for this very reason.

While the *fossil* package contains a number of occurrence based similarity indices, by no means are they all included. For example, Shi (1993) lists 39 and Hubálek (1982) lists 43 different variations of the similarity index, many of which are rarely used outside their original papers.

While not as common in palaeontological data sets, abundance values can provide valuable information about a community that is not possible using occurrence data. Analyses of community structure are very limited without abundance data, and abundance data can provide more subtle distinctions between communities. As well, species abundances can provide some measure of sampling intensity.

Possibly the most widely used abundance based measure is the Bray-Curtis measure, due to its strong relationship to ecological distance under varying conditions (Bray and Curtis 1957; Faith et al. 1987; Minchin 1987; Clarke 1993). The measure is equivalent to the Sorenson coefficient when used as a similarity measure with occurrence data. The Morisita-Horn index, while not as common as the Bray-Curtis, is also a highly recommended measure due to its relative independence from sample size and diversity (Wolda 1981; Magurran 2004). While there are several variations of the measure, I have used the version found within Magurran (2004).

Luckily, though the diversity of indices may seem somewhat overwhelming, the package provides an easy way to use them with large data sets. An included function called **dino.dist()** will take a matrix of species occurrences versus locality (or any analogous groupings) and return a full pairwise distance matrix as output. This function is written such that any other similarity index, including those defined by other packages or by the user, can be specified and used to calculate the matrix.

NON-PARAMETRIC SPECIES ESTIMATORS AND RAREFACTION

An obvious problem in palaeontology is the incompleteness of the record, and therefore our incomplete knowledge of the number of species present, whether locally or globally. Modern ecologists suffer from the same problem, whereby it is impractical to sample every single member of even relatively small communities of organisms (Chazdon et al. 1998). However, smaller samples still contain important information about the community and can be extrapolated from to provide estimates of the true richness of the total community. Of course, such extrapolations must account for sampling intensity and area (Gleason 1922; Preston 1948).

One of the most commonly used methods for dealing with unequal sampling intensity is rarefaction, or interpolation of the data (Sanders 1968). Rarefaction provides a method of comparison between different communities, whereby each community is ”rarefied” back to an equal number of sampled specimens (Heck et al. 1975; Foote 1992; Colwell and Coddington 1994). Within the *fossil* package is a method for rarefaction known as a Coleman Curve (Coleman 1981; Coleman et al. 1982). This type of rarefaction is carried out through a resampling method rather than a rarefaction formula; resampling is computationally much simpler and faster, and provides indistinguishable results from the formula based method (Coleman 1981; Coleman et al. 1982; Colwell and Coddington 1994; Magurran 2004). The Coleman Curve is an empirical measure of the rarefied number of individuals, while the rarefaction function is a theoretical model of what the empirical curve would look like. Although rarefaction can be useful, it is very sensitive to the underlying pattern of species abundance, such that collections with much lower species evenness will often give lower estimates of species diversity than those with very even abundances, regardless if species diversities in reality are equal (See Gotelli and Colwell [2001] for an in-depth treatment of the issue.).

Although rarefaction interpolates data back, non-parametric species estimators extrapolate from the data to find what the 'true' number of species may have been (Colwell and Coddington 1994). The typical way these estimators operate is by using the number of rare species that are found in a sample as a way of calculating how likely it is there are more undiscovered species. As an example, the Chao 1 estimator (Chao 1984; Colwell and

Coddington 1994) calculates the estimated true species diversity of a sample by the equation:

$$S_1 = S_{obs} + \frac{F_1^2}{2F_2}$$

where *Sobs* is the number of species in the sample, *F*1 is the number of singletons (i.e., the number of species with only a single occurrence in the sample) and *F*2 is the number of doubletons (the number of species with exactly two occurrences in the sample). The idea behind the estimator is that if a community is being sampled, and rare species (singletons) are still being discovered, there is likely still more rare species not found; as soon as all species have been recovered at least twice (doubletons), there is likely no more species to be found. Tests of the estimator have shown that it does provide reasonable estimates, at least for modern data sets (Chao 1984; Colwell and Coddington 1994; Chazdon et al. 1998). Of course, as the value is an estimate there is a degree of uncertainty, and a method to calculate the variance for the estimators has been provided by Chao (1987) in the form of

$$var(S_1) = F_2\left[\left(\frac{(F_1/F_2)}{4}\right)^4 + (F_1/F_2)^3 + \left(\frac{F_1/F_2}{2}\right)^2\right]$$

Although the Chao 1 estimator works for abundance data, often only occurrence data are available. There is another estimator, named conveniently Chao 2 (Chao 1987; Colwell and Coddington 1994), which uses occurrence data from multiple samples in aggregate to estimate the species diversity of the whole. This estimator is defined as:

$$S_2 = S_{obs} + \frac{Q_1^2}{2Q_2}$$

which is virtually identical to the Chao 1 estimator, with singletons (*Q*1) being species occurring in only one sample and doubletons (*Q*2) occurring in two samples. This estimator can also make use of the Chao 1 variance formula provided above, with the substitution of *F*1 and *F*2 for *Q*1 and *Q*2, respectively.

Chao and colleagues (Chao and Lee 1992; Chao et al. 1993; Lee and Chao 1994) have also published another pair of estimators, called the Abundance Coverage Estimator and the Incidence Coverage Estimator, which use abundance and occurrence based data sets, respectively. These estimators are much more complex; the Abundance-based Coverage Estimator takes the form

$$S_{ace} = S_{common} + \frac{S_{rare}}{C_{ace}} + \frac{F_1}{C_{ace}}\gamma_{ace}^2$$

where *Scommon* are the species that occur more than 10 times in the sampling, *Srare* are those species which occur 10 times or less, *Cace* is the sample abundance coverage estimator, and finally *γace* is the estimated coefficient of variation for *F*1 for rare species (See Chazdon et al. 1998, for a full explanation and definition of the estimator). In simpler terms, the formula uses the number of rare species (>= 10) and the number of singletons (*F*1) to estimate how many more undiscovered species there might be. Although this formula is for the abundance estimator, virtually the same holds true for the incidence based estimator, except that instead of the species abundance, it uses the number of samples each species occurs in. Both of the coverage estimators have been found to give good results and are highly recommended (Chazdon et al. 1998; Hortal et al. 2006)

Another estimator provided is the Jackknife estimator, developed by Burnham and Overton (1978, 1979) originally for use with capture/recapture studies. The formula

$$S_{jack1} = S_{obs} + Q_1\left(\frac{m-1}{m}\right)$$

represents the first order version of the estimator; the variable *m* represents the total number of samples. Smith and van Belle (1984) also provided a second order variation, with the formula

$$S_{jack2} = S_{obs} + \left[\frac{Q_1(2m-3)}{m} - \frac{Q_2(m-2)^2}{m(m-1)}\right]$$

The second order Jackknife has shown to be one of the most effective estimators and may be the best estimator at the moment for highly sparse palaeontological collections since it is the least susceptible to sampling bias (Chazdon et al. 1998; Hortal et al. 2006).

Finally, for completeness I also provide the bootstrap estimator

$$S_{boot} = S_{obs} + \sum_{k=1}^{S_{obs}} \left(1 - p_k\right)^2$$

developed by Smith and van Belle (1984). The bootstrap richness estimator has been generally regarded as one of the poorer species estimators, and Chazdon et al. (1998) in fact recommend against using it.

Though the various estimators vary greatly in their formulae, the functions within *fossil* take care of most of the nuances and generally require only one argument, that being a species occurrence matrix or species abundance vector or matrix.

```
>data(fdata.mat)
>chao1(fdata.mat)

[1]12.25

>jack1(fdata.mat) [1]12.98980
```

It is often best to use a number of these estimators in concert, as concurrence between their individual values can lend support to their results. Colwell (2009) has released a program for Windows called *EstimateS* which does exactly this; it can calculate multiple species estimators for a data set, along with their variances and a species accumulation curve. Since Colwell's program is so useful, it was used as a template to create the function **spp.est().** The function has several important options, namely the number of randomizations and whether or not to use abundance data. The **spp.est()** function calculates a rarefaction curve, the Chao, Coverage Estimators and Jacknife, as well as standard deviations for all the estimates. As a default the function will run 10 randomizations of the data, however for more accurate estimates a much larger number of randomizations should be run. It should be noted that with a large data set and a large number of randomizations that the function may take a long time to complete. At this time, work has been undertaken to parallelize this function, enabling a large increase in processing time when using a multicore or multiprocessor system.

MINIMUM SPANNING TREES

Minimum Spanning Trees (MST) and the associated Minimum Spanning Networks/Forests (MSN) are a useful method of visually displaying relationships between samples, whether those samples are biogeographic or taxonomic in nature (Figure 1; Gower and Ross 1969). The MST is closely related to the final product of a Single Linkage Cluster Analysis (SLCA; Sneath 1957; Gower and Ross 1969) and connects all the points in a sample with the minimum number of connections (*n* - 1). The method used to find the tree—also the most common method—is to begin with a single point at random, and begin connecting to the closest point not already in the tree. When there is more than one equally close point, one will be chosen at random. The randomness aspect of the connections can be disabled in the options for the function, if so desired, such that the first listed point will be used as the start for the tree and if more than one point is equally close, the first listed will be chosen. Although there are other MST functions available for R (Oksanen et al. 2010), those other methods did not allow for a random start or random selection of equally minimal branches. The MSN is closely related to the MST; the MSN is a combination of all the possible MSTs. This could mean that if there was only one shortest MST that the MSN would be identical.

BIOGEOGRAPHY AND GIS

Biogeography is concerned with locations of organisms in space. The *fossil* package implements a number of functions to assist in converting georeferenced datasets into formats useful for both graphing within R and exporting to GIS programs. R was originally created as a statistical language, but its ability to use and display geographic data is quite advanced for a non-GIS system. The *sp* package (Pebesma and Bivand 2005) along with a number of geographic libraries allows a user to put in data in a number of projections and change projection and datum. For a thorough treatment of spatial data analysis with R, I highly recommend Bivand et al. (2008); here I provide only a cursory description of the topic.

The simplest geographic function to use is likely **create.lats(),** which as mentioned previously can extract the locality data from a list of taxa occurrences. With the output from this function, a number of further analyses can be done. For example, it is often useful to have the distances between two points in space; this can be easily

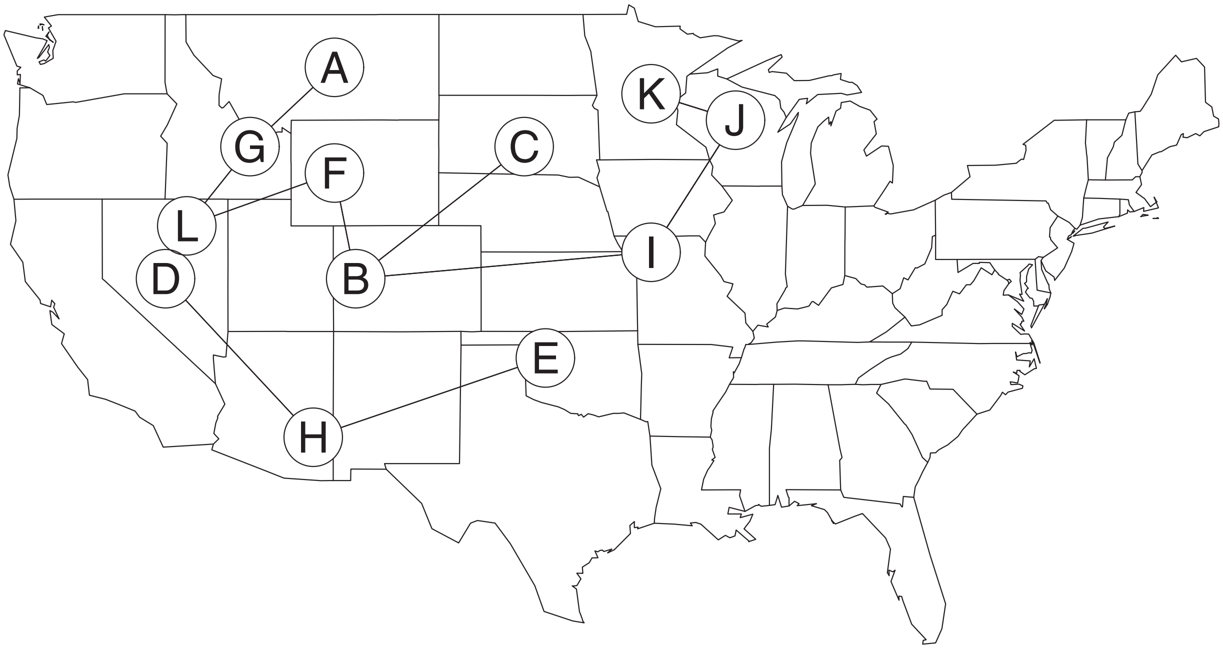


FIGURE 1. Minimum Spanning Tree for the **fdata** example data set from the *fossil* package, overlain over a map of the USA. Letters correspond to locality name.

```
>data(fdata.mat)
>fdata.dist<-dino.dist(fdata.mat)
>fdata.mst<-dino.mst(fdata.dist)
>data(fdata.lats)
>library(maps)
>map("state")
>mstlines(fdata.mst,coordinates(fdata.lats))
>points(coordinates(fdata.lats),pch=16,col="white",cex=3)
>points(coordinates(fdata.lats),pch=1,cex=3)
>text(coordinates(fdata.lats),labels=LETTERS[1:12])
```

accomplished with the **earth.dist()** function, which returns a matrix of pairwise distances in kilometres (Figure 2). One note, however, is that the original matrix of locations must be in decimal degrees. Of course, the sp package provides functions to convert between coordinate systems if necessary.

Biogeography is concerned with species locations in space, and the sampling distributions of those species can cause some interesting effects in diversity calculations, namely the well researched species/area effect (Arrhenius 1921; Gleason 1922; Preston 1960; Connor and McCoy 1979; Rosenzweig 1995). Although palaeontology often pays little attention to this effect, Carrasco et al. (2005) have shown that it does hold true in fossil data sets. As a way to observe these effects efficiently, I have created the **function sac()** that can create a summary species area curve for a data set (Figure 3). As its arguments, it takes a table of longitude/latitude and a species occurrence matrix. It makes use of another function called **earth.poly(),** which can take a table of locations and calculate which points create the vertices for a minimum spanning polygon/convex hull, as well as calculate the true geographic area of the polygon.

Though the R environment is powerful when analysing GIS data, it lacks a large amount of visual interactivity with the data. Often, it is simply easier to use a GIS program to view geographic data, and as such I have tried to make it as simple as possible to move geographic data out of R. Currently the package provides helper functions for exporting both geographic points (**lats2Shape()**) and MSTs/MSNs (**msn2Shape**) to shapefile format using the package **shapefiles** (Stabler 2006). To use the functions, you need the shapefile package available on your system; the package can be downloaded using the install.packages(shapefiles)

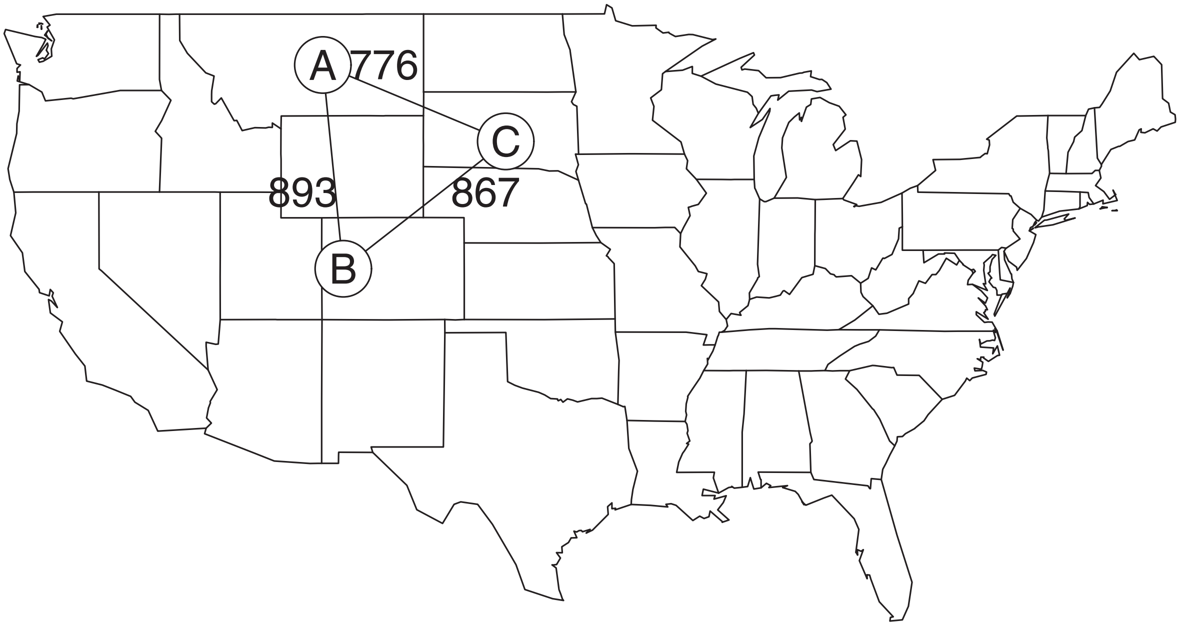


FIGURE 2. Distances between three selected locations from the **fdata** sample data. Distances given between points are in km.

```
>data(fdata.lats)
>fd.subset<-coordinates(fdata.lats)[1:3,]
>earth.dist(fdata.lats[1:3,])

      locA        locB
locB 893.4992
locC 776.3101867.2648

>map("state")
>polygon(fd.subset)
>text(c(-110,-101,-106),c(42,42,47),labels=round(earth.dist(fd.subset)[c(1,
+3,2)]))
>points(fd.subset,pch=16,col="white",cex=3)
>points(fd.subset,pch=1,cex=3)
>text(fd.subset,label=LETTERS[1:3])
```

command. Once the shapefiles have been created, they can be saved using the **write.shapefile()** command. The **shapefiles** can then be loaded in any GIS program (Figure 4).

```
>data(fdata.lats)
>shape.lats<-lats2Shape(fdata.lats)
>fdata.dist<-dino.dist(fdata.mat)
>fdata.mst<-dino.mst(fdata.dist)
>shape.mst<-
msn2Shape(fdata.mst,fdata.lats)
```

CONCLUSION

I optimistically envisage the *fossil* package growing larger and larger in both function and use. As the project is Open Source, I encourage others to help aid in its development both by simply using it in various and novel situations, as well as suggesting new possible methods, indices and functions that may be useful. As well, I readily encourage others to use the original source code for their own purposes, with the only caveat that attribution is given where appropriate. I hope that encouraging the recopying and reuse of this code will save others time while developing their meth-

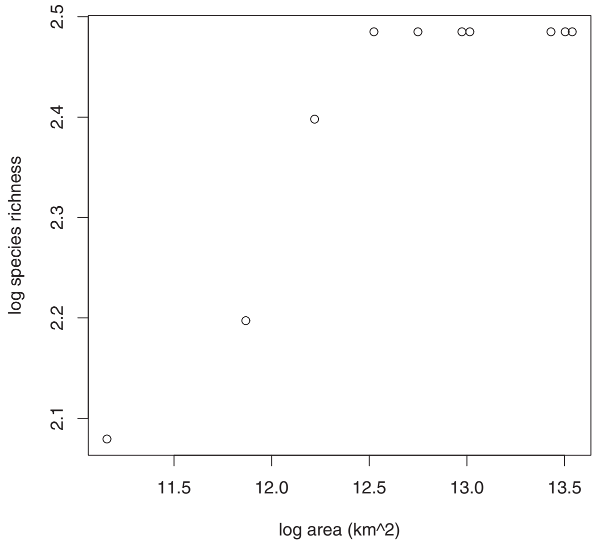


FIGURE 3. Species area curve for the **fdata** sample data.

```
>plot(log(sac(fdata.lats,fdata.mat)[
[1]]),ylab="logspeciesrichness",
+xlab="logarea(km^2)")
```

ods and allow more time for the actual data analysis.

ACKNOWLEDGMENTS

I would like to thank H. Larsson for his helpful reviews and comments on this manuscript and for introducing me to the R Language. Also, this manuscript was greatly improved by comments from P.D. Polly and from two anonymous reviewers. Work on this project was supported by a National Sciences and Engineering Research Council Canada Graduate Scholarship (Masters and Doctoral) and a Fonds Québécois de la Recherche sur la Nature et les Technologies Doctoral scholarship.

REFERENCES

Alroy, J., Marshall, C.R., Bambach, R.K., Bezusko, K., Foote, M., Fursich, F.T., Hansen, T.A., Holland, S.M., Ivany, L.C., Jablonski, D., Jacobs, D.K., Jones, D.C., Kosnik, M.A., Lidgard, S., Low, S., Miller, A.I., Novack-Gottshall, P.M., Olszewski, T.D., Patzkowsky, M.E., Raup, D.M., Roy, K., Sepkoski, J.J., Jr., Sommers, M.G., Wagner, P.J., and Webber, A. 2001. Effects of sampling standardization on estimates of phanerozoic marine diversification. *Proceedings of the National Academy of Sciences*, 98:6261-6266.

Arrhenius, O. 1921. Species and area. *The Journal of Ecology*, 9:95-99.

Becker, R.A., Chambers, J.M., and Wilks, A.R. 1988. *The New S Language*. Wadsworth, Pacific Grove, California.

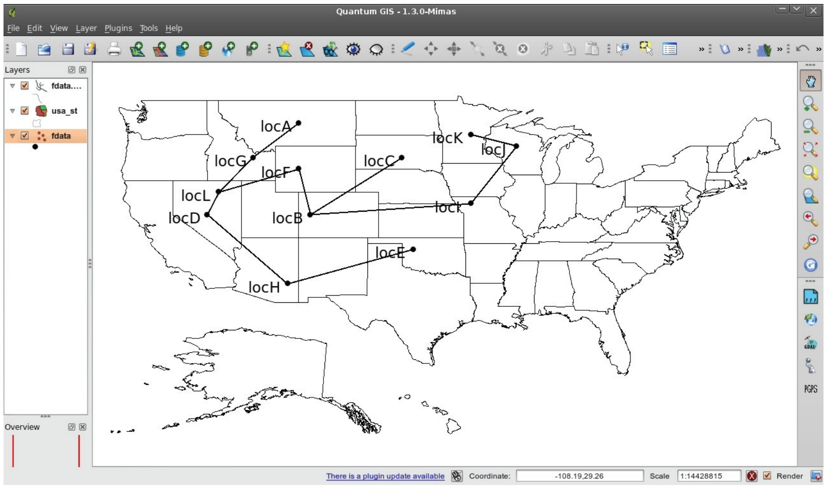
Benton, M.J., Wills, M.A., and Hitchin, R. 2000. Quality of the fossil record through time. *Nature*, 403:534-537.

Bivand, R.S., Pebesma, E.J., and Gómez-Rubio, V. 2008. *Applied Spatial Data Analysis with R*. Springer, New York.

Braun, W.J. and Murdoch, D.J. 2008. *A First Course in Statistical Programming with R*. Cambridge University Press, New York, NY, USA.

Bray, J.R. and Curtis, J.T. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27:326-349.

Burnham, K.P. and Overton, W.S. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65:625-633.



Coordinate: -108.19,29.26

Scale 1:14428815

FIGURE 4. A screen shot from Quantum GIS, showing the exported latitude and MST shapefiles on a map of the USA.

Burnham, K.P. and Overton, W.S. 1979. Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60:927-936.

Carrano, M. 2000. Taxonomy and classification of non-avian Dinosauria: Online Systematics Archive 4: *The Paleobiology Database*. http://paleodb.org.

Carrasco, M., Kraatz, B., Davis, E., and Barnosky, A. 2005. *Miocene Mammal Mapping Project (MIOMAP)*. Technical report, University of California Museum of Paleontology. www.ucmp.berkeley.edu/miomap/

Chao, A. 1984. Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11:265-270.

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43:783-791.

Chao, A. and Lee, S.M. 1992. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87:210-217.

Chao, A., Ma, M.C., and Yang, M.C.K. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, 80:193-201.

Chazdon, R., Colwell, R., Denslow, J., and Guariguata, M. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica, p. 285-309. In Dallmeier, F. and Comiskey, J. (eds.), *Forest Biodiversity Research, Monitoring and Modeling: Conceptual Background and Old World Case Studies*. Parthenon Publishing, Paris.

Clarke, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18:117-143.

Coleman, B.D. 1981. On random placement and species-area relations. *Mathematical Biosciences*, 54:191-215.

Coleman, B.D., Mares, M.A., Willig, M.R., and Hsieh, Y.H. 1982. Randomness, area, and species richness. *Ecology*, 63:1121-1133.

Colwell, R. 2009. *EstimateS: Statistical estimation of species richness and shared species from samples. Version 8.2.* http://viceroy.eeb.uconn.edu/EstimateS

Colwell, R.K. and Coddington, J.A. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 345:101-118.

Connor, E.F. and McCoy, E.D. 1979. The statistics and biology of the species-area relationship. *The American Naturalist*, 113:791-833.

Ezard, T.H. and Purvis, A. 2009. paleoPhylo: free software to draw paleobiological phylogenies. *Paleobiology*, 35:460-464.

Faith, D., Minchin, P., and Belbin, L. 1987. Compositional dissimilarity as a robust measure of ecological distance. *Plant Ecology*, 69:57-68.

Foote, M. 1992. Rarefaction analysis of morphological and taxonomic diversity. *Paleobiology*, 18:1-16.

Gleason, H.A. 1922. On the relation between species and area. *Ecology*, 3:158-162.

Gotelli, N.J. and Colwell, R.K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4:379-391.

Gower, J.C. and Ross, G.J.S. 1969. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18:54-64.

Green, P.J. 2003. Diversities of gifts, but the same spirit. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 52:423-438.

Green, R.H. 1980. Multivariate approaches in ecology: The assessment of ecologic similarity. *Annual Review of Ecology and Systematics*, 11:1-14.

Hammer, Ø., Harper, D.A., and Ryan, P.D. 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica*, 4:9.

Harrison, L.B. and Larsson, H.C.E. 2008. Estimating evolution of temporal sequence changes: A practical approach to inferring ancestral developmental sequences and sequence heterochrony. *Systematic Biology*, 57:378-387.

Heck, K.L., van Belle, G., and Simberloff, D. 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, 56:1459-1461.

Hortal, J., Borges, P., and Gaspar, C. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology*, 75:274-287.

Hubálek, Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, 57:669-689.

Hunt, G. 2008. *paleoTS: Modeling evolution in paleontological time-series*. R package version 0.3.1. http://cran.r-project.org/web/packages/paleoTS/

Ihaka, R. and Gentleman, R. 1996. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299-314.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547-579.

Lee, S.M. and Chao, A. 1994. Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, 50:88-97.

Leisch, F. and Rossini, A.J. 2003. Reproducible statistical research. *Chance*, 16:46-50.

Maddison, W.P. and Maddison, D. 2009. *Mesquite: a modular system for evolutionary analysis. Version 2.72*. http://mesquiteproject.org

Magurran, A.E. 2004. *Measuring Biological Diversity*. Blackwell, Oxford.

Minchin, P.R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. *Plant Ecology*, 69:89-107.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.G., Simpson, G.L., Solymos, P., Stevens, M.H.H., and Wagner, H. 2010. *vegan: Community Ecology Package*. R package version 1.17-0. http://cran.r-project.org/web/packages/vegan/

Paradis, E., Claude, J., and Strimmer, K. 2004. Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289-290.

Pebesma, E.J. and Bivand, R.S. 2005. Classes and methods for spatial data in R. *R News*, 5:9-13.

Preston, F.W. 1948. The commonness, and rarity, of species. *Ecology*, 29:254-283.

Preston, F.W. 1960. Time and space and the variation of species. *Ecology*, 41:611-627.

R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenzweig, M.L. 1995. *Species Diversity in Space and Time.* Cambridge University Press, Cambridge, UK.

Sanders, H.L. 1968. Marine benthic diversity: A comparative study. *The American Naturalist*, 102:243-282.

Shi, G.R. 1993. Multivariate data analysis in palaeoecology and palaeobiogeography - a review. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 105:199-234.

Simpson, G. 1960. Notes on the measurement of faunal resemblance. *American Journal of Science*, 258:300-311.

Smith, E.P. and van Belle, G. 1984. Nonparametric estimation of species richness. *Biometrics*, 40:119-129.

Sneath, P.H.A. 1957. The application of computers to taxonomy. *Journal of General Microbiology*, 17:201-226.

Stabler, B. 2006. *shapefiles: Read and Write ESRI Shapefiles*. R package version 0.6. http://cran.r-project.org/web/packages/shapefiles/

Stallman, R. 1999. The GNU Operating System and the Free Software Movement, p. 53-70. In DiBona, C., Ockman, S. and Stone, M. (eds.), *Open Sources: Voices from the Open Source Revolution*. O'Reilly, Sebastopol, California.

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

APPENDIX

Appendix: R source code

```
###################################################
###chunknumber1:data-in
###################################################
#linesstartingwitha#arecomments,andareignoredbytheRinterpreter
#install.packages('fossil')
library(fossil)



###################################################
###chunknumber2:list-to-occ-mat
###################################################
data(fdata.list)
create.matrix(fdata.list,tax.name='species',locality='locality')



###################################################
###chunknumber3:list-to-abund-mat
###################################################
data(fdata.list)
create.matrix(fdata.list,tax.name='species',locality='locality',
abund=TRUE,abund.col='abundance')



###################################################
```

```
###chunknumber4:list-to-lats
###################################################
data(fdata.list)
create.lats(fdata.list,loc='locality',long='longitude',
lat='latitude')


###################################################
###chunknumber5:sim-measure
###################################################
sampleA<-c(1,1,0,1,1,1,1)
sampleB<-c(0,1,1,0,0,1,1)
sorenson(sampleA,sampleB)


###################################################
###chunknumber6:spp-ests
###################################################
data(fdata.mat)
chao1(fdata.mat)
jack1(fdata.mat)


###################################################
###chunknumber7:shapefiles
###################################################
data(fdata.lats)
```

```
shape.lats<-lats2Shape(fdata.lats)
fdata.dist<-dino.dist(fdata.mat)
fdata.mst<-dino.mst(fdata.dist)
shape.mst<-msn2Shape(fdata.mst,fdata.lats)


###################################################
###chunknumber8:mst-map
###################################################
data(fdata.mat)
fdata.dist<-dino.dist(fdata.mat)
fdata.mst<-dino.mst(fdata.dist)
data(fdata.lats)
library(maps)
map('state')
mstlines(fdata.mst,coordinates(fdata.lats))
points(coordinates(fdata.lats),pch=16,col='white',cex=3)
points(coordinates(fdata.lats),pch=1,cex=3)
text(coordinates(fdata.lats),labels=LETTERS[1:12])


###################################################
###chunknumber9:geo-dist-map
###################################################
data(fdata.lats)
fd.subset<-coordinates(fdata.lats)[1:3,]
earth.dist(fdata.lats[1:3,])
```

```
map('state')

polygon(fd.subset)

text(c(-110,-101,-106),c(42,42,47),

labels=round(earth.dist(fd.subset)[c(1,3,2)]))

points(fd.subset,pch=16,col='white',cex=3)

points(fd.subset,pch=1,cex=3)

text(fd.subset,label=LETTERS[1:3])



###################################################

###chunknumber10:spp-area-fig

###################################################

plot(log(sac(fdata.lats,fdata.mat)[[1]]),

ylab='logspeciesrichness',xlab='logarea(km^2)')
```