



## Finding the Minimum Sample Richness (MSR) for multivariate analyses: implications for palaeoecology

K. J. TRAVOUILLO<sup>1,2</sup>, M. ARCHER<sup>1</sup>, S. LEGENDRE<sup>2</sup>, & S. J. HAND<sup>1</sup>

<sup>1</sup>*School of Biological, Earth and Environmental Sciences, University of New South Wales, New South Wales 2052, Australia, and* <sup>2</sup>*UMR 5125 PEPS, CNRS, France; Université Lyon 1, Campus de La Doua, Bt. Géode, 69622 Villeurbanne cedex, France*

### Abstract

Many techniques have been developed to estimate species richness and beta diversity. Those techniques, dependent on sampling, require abundance or presence/absence data. Palaeontological data is by nature incomplete, and presence/absence data is often the only type of data that can be used to provide an estimate of ancient biodiversity. We used a simulation approach to investigate the behaviour of commonly used similarity indices, and the reliability of classifications derived from these indices, when working with incomplete data. We drew samples, of varying number and richness, from artificial species lists, which represented original life assemblages, and calculated error rates for classifications of the parent lists and samples. Using these results, we estimated the Minimum Sample Richness (MSR) needed to achieve 95% classification accuracy. Results were compared for classifications derived from several commonly used similarity indexes (Dice, Jaccard, Simpson and Raup–Crick). MSR was similar for the Dice, Jaccard and Simpson indices. MSR for the Raup–Crick index was often much lower, suggesting that it is preferable for classifying patchy data, however the performance of this index was less stable than the other three in the simulations, which required an even lower MSR. MSR can be found for all presence/absence data from the contour graphs and equations as long as the absolute species richness and the beta diversity can be estimated.

**Keywords:** *Sample, richness, estimation, multivariate analysis, palaeoecology*

**Msc:** 62H30, 62D05

### Introduction

Dealing with incomplete data is a particular challenge in palaeontological studies. Travouillon et al. (2006) used several multivariate analyses to compare presence/absence data of the Riversleigh Local Faunas (Oligo-Miocene assemblages from north-west Queensland, Australia), to determine which fossil localities were members of the same assemblage. They noticed that many localities had very low species richness and concluded that they were probably not representative of the original life assemblage. This misrepresentation of the life assemblage resulted in the unexpected grouping of local faunas which are geologically different. Mares and Willig (1994) investigated this issue of representativeness using recent mammal faunas. They randomly

selected species from a known fauna and continued to increase sample size until the correct community was identified. Using this method, they answered two questions: (1) “How large a sample of species must be drawn from the fauna in order to arrive at a correct decision as to the fauna’s community association?” (2) “What percentage of the species from a fauna is required in a sample in order to yield a correct determination of the community from which those species were drawn?”

In this study, we use a simulation approach to investigate these questions. Multivariate analyses such as cluster analysis or ordination, which are often used to compare fossil localities, are not statistical tests. They are data reduction methods used to visualise relationships between objects and attributes in complex data. The results found by such analyses may be highly

Correspondence: K. J. Travouillon, School of Biological, Earth and Environmental Sciences, University of New South Wales, New South Wales 2052, Australia. E-mail: kennytravouillon@hotmail.com

skewed by missing data caused by undersampling or taphonomic processes. It is for this reason that we attempt to identify the relationship between classification accuracy and minimum sample richness. The Minimum Sample Richness (MSR) is defined here as the smallest number of taxa that must be present in a sample to achieve a given level of classification accuracy. We compare four similarity indices to see if different indices generated different MSR values. We used the following indices: Dice (also known as the Sørensen index), Jaccard, Simpson and Raup–Crick (Jaccard 1912; Dice 1945; Sørensen 1948; Simpson 1949; Raup and Crick 1979).

## Methods and materials

### Parent lists

Alternative sets of artificial parent lists of taxa, representing different original life assemblages, were generated for a range of values of richness ( $A$ ) and similarity between lists or Beta diversity ( $S$ ). Richness ( $A$ ) is the number of taxa in a parent list. The following  $A$  values were used; 25, 50, 100, 150, 200, 250, 300 and 350. The number of parent lists used in the analysis ( $N$ ) varies between analyses. All values of  $N$  between 2 and 16 (inclusive) were investigated and the following  $S$  values were used; 0, 10, 30, 50, 70, 90, 95 and 100. For example, with an  $S$  value of 50 and  $N$  value of 3, the first parent list shares 50% of its taxa with the second and the third parent lists. The second parent list shares only  $S/2$  (25%) with the third parent list. For  $N$  greater than 3, any subsequent parent lists would continue to be represented by decreasing values of  $S$ , i.e. they would share 12%, followed by 0% of the taxa with the first parent list. In each analysis, only one parameter was investigated at a time, so that the two not being investigated were represented by a standard value ( $A = 100$ ,  $S = 50$ ,  $N = 3$ ).

### Minimum Sample Richness (MSR)

The Minimum Sample Richness (MSR) is the smallest number of taxa in a sample needed for that sample to be a reliable indicator of its original life assemblage. We simulated presence/absence data for sites in a palaeontological field study by drawing replicate subsets from the parent lists. The subsets and parent lists were then compared using cluster analysis to see if they group together correctly (Figure 1). This comparison is replicated 1000 times. The error rate (proportion of replicates in which subsets were grouped with the wrong parent list) was then calculated for a range of subset sizes. If the error rate was greater than 0.05, then the procedure was repeated increasing the size of the subset by one species until an error rate of 0.05 or less was achieved. Data generation and clustering were performed with

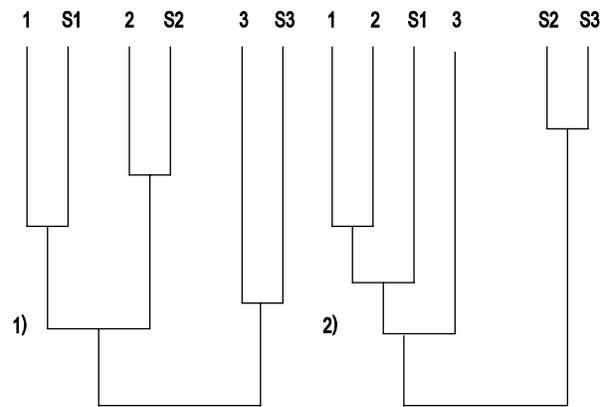


Figure 1. Example of cluster analysis with 1) correctly grouped subsets; and 2) incorrectly grouped subsets. Numbers 1, 2 and 3 represent the three parent faunas. S1, S2 and S3 are the corresponding subsets to the original faunas.

custom scripts in the *R* statistics environment (Gentleman and Ihaka 2005).

### Analyses

Four analyses were performed in order to investigate how changing the two parameters and the number of parent lists used ( $A$ ,  $S$  and  $N$ ) affected the MSR. In the first analysis, the value of  $A$  was investigated. The MSR for each value of  $A$  was recorded using four different similarity indices (Dice's, Jaccard's, Simpson's and Raup–Crick similarity indices), using the same dataset for each of the indices. The use of the four similarity indices follows Hammer and Harper (2006). We also compared the results using one subset per parent list at a time and using several subsets per parent list at a time (increasing from 6 to 12 subsets with increasing  $A$ ). In the second and third analyses, the values of  $S$  and  $N$  were investigated, respectively. For these analyses, we recorded the results using one subset per parent list and using each of the four similarity indices. In the fourth analysis, every possible combination of  $A$  and  $S$  values used in the first two analyses were investigated and the corresponding MSR values calculated.

## Results

The results of the first analysis are shown in Figure 2. For each value of  $A$ , there was no difference in MSR using the Dice, Jaccard or Simpson similarity indices. Using only one subset per parent list, there was a linear relationship between MSR and  $A$  ( $MSR = 0.34A$ ). Using several subsets drawn from each parent list, the relationship between MSR and  $A$  was more complex. For  $A$  values between 0 and 75, MSR was higher than for the analyses with a single subset per parent list, but the reverse was true for  $A$  values greater than 75. The Raup–Crick index generally achieved much lower MSR values than the other three indices.

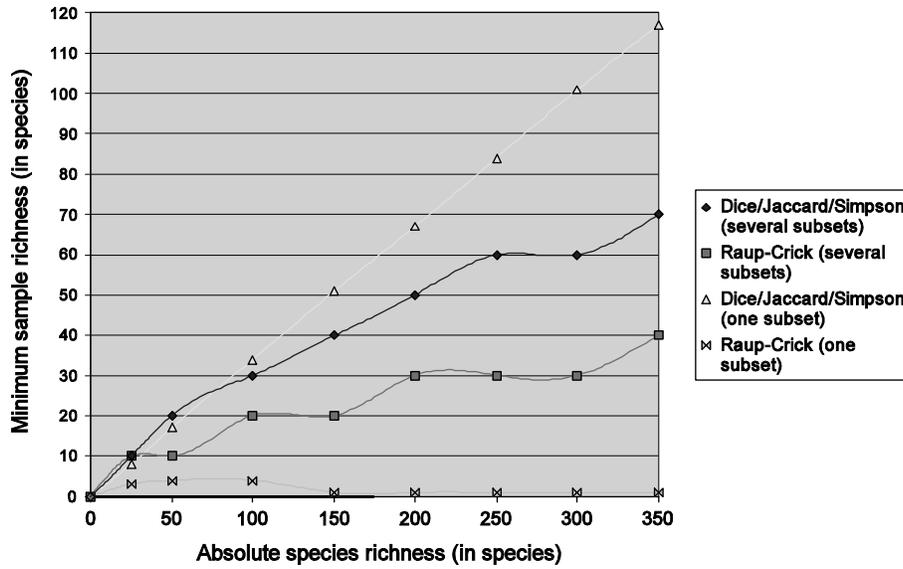


Figure 2. Plot of minimum sample richness (MSR) against absolute species richness ( $A$ ) where  $N = 3$  and  $S = 50$ , using Dice's, Jaccard's, Simpson's and Raup-Crick's similarity indices.

For analyses with several subsets drawn from each parent list, the curve resembles that for the other indices but with a substantially lower MSR.

In the second analysis, we varied the similarity between parent lists and only one subset was used per list. The results of this analysis are shown in Figure 3. Classifications derived from the Raup-Crick index correctly identified the parent lists with a much smaller MSR than required with the other indices. For, Dice, Jaccard and Simpson indices, MSR increased steadily with increasing  $S$ . When  $S = 50$ ,  $MSR = 34$ . In contrast, the MSR using the Raup-Crick index remained quite low when  $S$  was less than 50. When  $S$  was greater than 50, MSR rapidly increased but remained lower than for the other indices.

The third analysis aimed at testing whether the number of parent lists ( $N$ ) used in the analysis influenced

the MSR. The results are shown in Figure 4. Increasing  $N$  from 2 to 15 did not affect the results for classifications derived from the Dice, Jaccard and Simpson indices and MSR remained stable at 34 taxa. In contrast, the Raup-Crick measure displayed a sudden jump in MSR at  $N = 11$ .

Figure 5 shows the joint effects of richness ( $A$ ) and similarity between parent lists ( $S$ ) on MSR for the Dice, Jaccard and Simpson indices, while Figure 6 shows the corresponding results for the Raup-Crick index. MSR increases with increasing  $A$  and  $S$  regardless of the similarity index used but Raup-Crick achieved a lower MSR, especially when  $S$  was less than 50.

Contour graphs (Figures 7 and 8) are obtained by Delaunay triangulation of the MSR Surface and projection onto the  $A$ - $S$  plane. Least-squares 3D-surface fitting is calculated for the Dice/Jaccard/Simpson indices as the combination of an  $A$ -MSR linear relation and a  $S$ -MSR second order polynomial

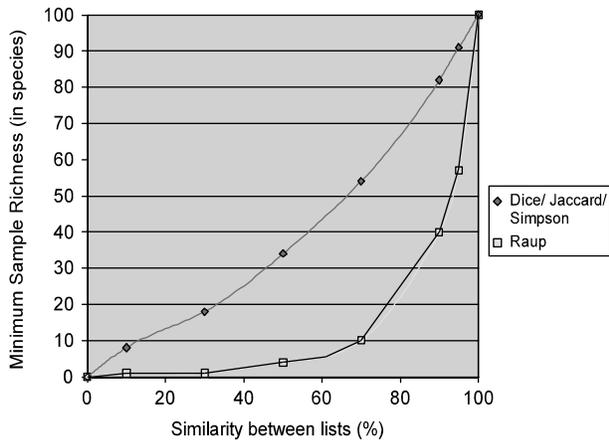


Figure 3. Plot of minimum sample richness (MSR) against similarity between parent lists ( $S$ ) where  $N = 3$  and  $A = 100$  using Dice's, Jaccard's, Simpson's and Raup-Crick's similarity indices.

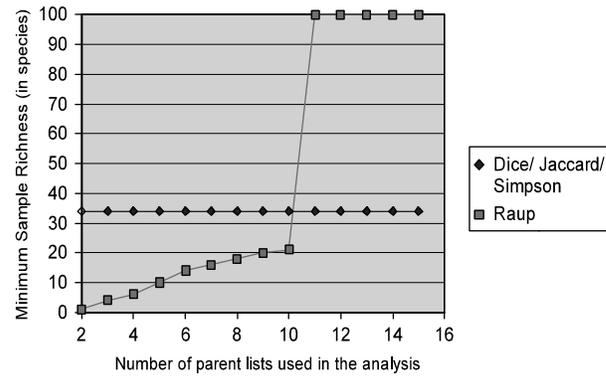


Figure 4. Plot of minimum sample richness (MSR) against number of lists ( $N$ ) where  $S = 50$  and  $A = 100$  using Dice's, Jaccard's, Simpson's and Raup-Crick's similarity indices.

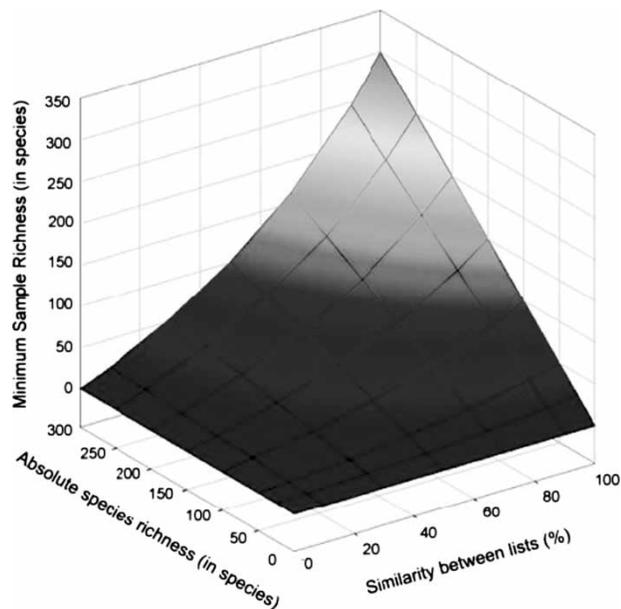


Figure 5. 3D plot of Minimum Sample Richness (MSR), absolute species richness ( $A$ ) and similarity between parent lists ( $S$ ) using Dice's, Jaccard's and Simpson's similarity indices.

relation. In order to optimize the correlation between simulated and fitted MSR estimates, the fitted estimates are rounded to the upper integer (ceil-function) when  $S \leq 50\%$  and to the lower integer (floor-function) when  $S > 50\%$ . The resulting prediction functions are:

$$\text{MSR} = \text{ceil}\{[(7 \cdot 10^{-5} \times A) \times S^2] + [(3 \cdot 10^{-3} \times A) \times S]\} \text{ if } S \leq 50\%$$

$$\text{MSR} = \text{floor}\{[(7 \cdot 10^{-5} \times A) \times S^2] + [(3 \cdot 10^{-3} \times A) \times S]\} \text{ if } S > 50\%$$

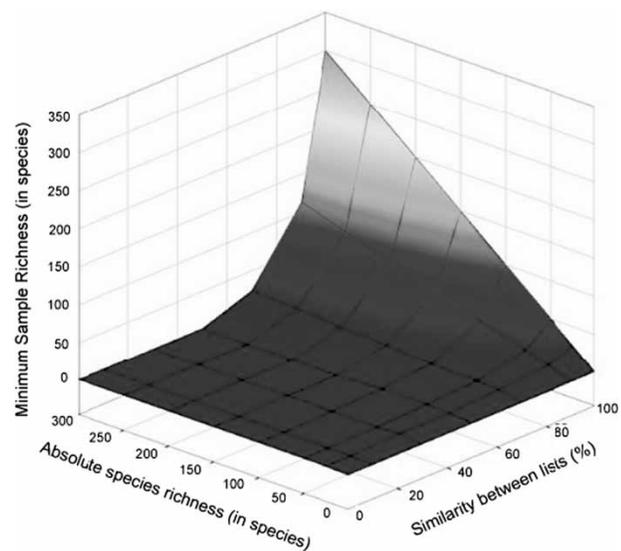


Figure 6. 3D plot of minimum sample richness (MSR), absolute species richness ( $A$ ) and similarity between parent lists ( $S$ ) where  $N = 3$  using Raup-Crick's similarity indices.

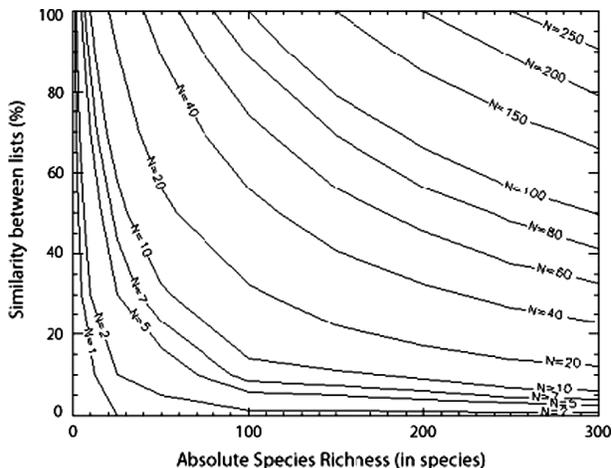


Figure 7. Contour graph of the MSR surface using Dice/Jaccard/Simpson's indices.

The resulting determination coefficient ( $R^2 = 0.999$ ) indicates that the fitted surface perfectly matches the observed one, thus enabling the use of these two complementary functions to estimate MSR-value for any  $[S, A]$ -values. Due to the very nature of the Raup-Crick's index (a type-I error rate of a significance test; see below), its MSR-surface was not modelled in the same way; its very high concavity not corresponding to any simple combination of linear, polynomial, exponential or power function. Nevertheless, a rough prediction of a MSR-value for any  $[S, A]$ -values can be graphically obtained from the contour graph (Figure 8).

### Discussion

Palaeontological data is by nature incomplete (Hammer and Harper 2006). However, an appreciation of how this "incompleteness" can affect analyses and interpretations can be achieved through simulation studies such as that presented here. Mares and Willig

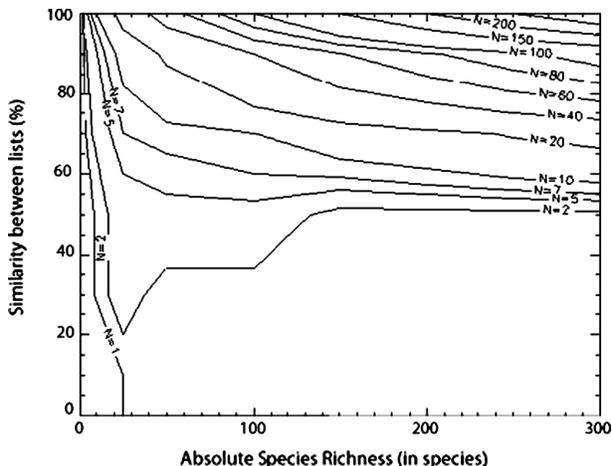


Figure 8. Contour graph of the MSR surface using Raup-Crick's index.

(1994) investigated the minimum sample size which would correctly group a sample to its original faunal list. However, application of their method to other types of data is difficult. We developed a method to provide palaeoecologists with a way to check whether their fossil samples are statistically representative of the original life assemblage.

Very early in the analysis, we came across a challenge: should we be using one subset or several subsets per parent list to determine the MSR? The first analysis we performed was aimed at answering this question. We noticed that when we used Dice's, Jaccard's or Simpson's similarity indices, using one subset required a higher MSR than using several subsets. In addition, we found that using one subset gave a linear relationship whilst several subsets did not. We found that using several subsets gave unstable results (varying with the number of subsets included in the clustering) because the subsets were creating links between each other and the parent lists, lowering the chances of missgrouping the subsets and the parent lists. The values for MSR can be predicted using one subset and the results show that increasing the number of subsets can only increase the chance of a correct grouping. For this reason, we chose to use only one subset per parent list for the remaining analyses.

The results of the analyses showed that the Raup–Crick similarity index behaved very differently to the Dice, Jaccard and Simpson indices in that it performed worse with increasing numbers of subsets or parent lists. The Raup–Crick index is a probabilistic measure which assesses the pattern of species between two samples in terms of a “random sprinkling of species” hypothesis (Legendre and Legendre 1998). The original formulation of the index (Raup and Crick 1979) used a randomisation procedure to estimate the probability of the observed data, although by recognising that the calculation is equivalent to a “sampling without replacement” problem an exact value can easily be derived from the hypergeometric distribution (M. Bedward personal communication 2006). The probabilistic nature of the Raup–Crick index is probably the reason why it performed less well than the other indices, as the chance of correctly grouping similar lists decreases with increasing numbers of similar lists. Travouillon et al. (2006) noticed in their cluster analysis that Raup–Crick's index clustered together local faunas that had no species in common. The relationship between the same local faunas was unresolved using the Dice, Jaccard and Simpson indices. Raup–Crick's index can, therefore, be potentially misleading. The index can also be unstable depending on the data and may lead to error.

Nonetheless, we recommend use of this index, as it does perform better than the other indices tested. Raup–Crick's index performs much better than the other indices test here when data are sparse. Its probabilistic nature is most likely the reason why

it performs better than distance measures in this case. The fewer major groups (i.e. environments or time periods, which are the equivalent to the parent lists in the analysis) and the fewer faunal/floral samples being compared (equivalent to the subsets in the analysis), the better.

The downside to the method described here to calculate the MSR, is that it requires the absolute species richness and the percentage similarity (Beta diversity) of the faunas being compared to be known. However, there are a number of ways to acquire estimates of absolute species richness. Estimating absolute species richness can be done using accumulation curves or extrapolating from species abundance distribution or from non-parametric estimators (Colwell and Coddington 1994; Chazdon et al. 1998; Magurran 2004). Walther and Martin (2001) reviewed many of these estimators and considered that the two Chao estimators (Chao 1984, 1987; Shen et al. 2003) were the least biased, followed by the two jackknife estimators (Burnham and Overton 1979). Chao et al. (2006) applied Laplace's boundary-mode approximations (See Erkanli 1994, 1997) to the Chao estimators to improve the accuracy of their estimations of species richness and allow their use with replicated incidence data (presence/absence data). For palaeontological data, for which abundance data is often unavailable, non-parametric estimators (Chao2 estimator) are the only tools that can be used, as they are the only estimators that are able to use presence/absence data. Colwell (2000) released a computer program called EstimateS which not only provides the tools to estimate the absolute species richness but also provides tools to estimate the similarity between two or more sites (Beta diversity). Again, most of the techniques used to estimate the beta diversity use abundance data (ICE and ACE, Chao et al. 2000; Chao's Abundance-based Jaccard and Sørensen indexes, Chao et al. 2005) but beta diversity can be estimated for presence/absence data using classic diversity indices such as Jaccard, Sørensen (same as Dice) or Bray–Curtis (Magurran 2004) or using Chao et al. (2006)'s estimators of beta diversity. Using estimates for absolute species richness and similarity will necessarily only provide an estimate of MSR. MSR is therefore as accurate as the estimates of absolute species richness and similarity. However, even an estimate of MSR will provide palaeoecologists with a useful tool to check whether their data is representative of the original life assemblage. Future work will aim at applying the method to palaeontological data and testing the strengths and weaknesses of the method.

### Acknowledgements

We thank M. Bedward of the New South Wales National Parks and Wildlife Service for writing the *R* routines essential to this paper. We also want to thank him for his advices and for reading the manuscript. We also would like to thank Gilles Escarguel for providing his help with

the fitted functions of MSR. We also want to thank P. Brewer, M. Bassarova, J. Louys (University of New South Wales) and K. Giles Sproule (University of Sydney) for reading and commenting the manuscript. Finally, we thank Bill Sherwin (University of New South Wales) for his advice. Contribution UMR5125-07.020.

## References

- Burnham KP, Overton WS. 1979. Robust estimation of population size when capture probabilities vary among animals. *Ecology* 60: 927–936.
- Chao A. 1984. Nonparametric estimation of the number of classes in a population. *Scand J Stat Theory Appl* 11:265–270.
- Chao A. 1987. Estimating the population size for capture–recapture data with unequal catchability. *Biometrics* 43:783–791.
- Chao A, Chazdon RL, Colwell RK, Shen T-J. 2005. A new statistical approach for assessing compositional similarity based on incidence and abundance data. *Ecol Lett* 8:148–159.
- Chao A, Hwang W-H, Chen Y-C, Kuo C-Y. 2000. Estimating the number of shared species in two communities. *Statist Sin* 10: 227–246.
- Chao A, Shen T-J, Hwang W-H. 2006. Application of Laplace's boundary-mode approximations to estimate species and shared species richness. *Aust NZ J Stat* 48:117–128.
- Chazdon RL, Colwell RK, Denslow JS, Guariguata MR. 1998. In: Dallmeier F, Comiskey JA, editors. *Forest biodiversity research, monitoring and modelling: Conceptual background and old world case studies*. Paris: Parthenon Publishing.
- Colwell RK, Coddington JA. 1994. Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond B* 345: 101–118.
- Colwell RK. 2000. EstimateS—statistical estimation of species richness and shared species from samples Version 7.5, Available online at: <http://viceroy.eeb.uconn.edu/EstimateS> (accessed 15 July 2006)
- Dice LR. 1945. Measures of the amount of ecological association between species. *Ecology* 26:297–302.
- Erkanli A. 1994. Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *J Am Stat Assoc* 89:250–258.
- Erkanli A. 1997. Boundary-mode approximations for posterior expectations. *J Stat Plan Infer* 58:217–239.
- Gentleman R, Ihaka R. 2005. The R Project for statistical computing, Available online at: <http://www.r-project.org/> (accessed 10 November 2005)
- Hammer O, Harper D. 2006. *Paleontological data analysis*. UK: Blackwell Publishing.
- Jaccard P. 1912. The distribution of the flora of the alpine zone. *New Phytol* 11:37–50.
- Legendre P, Legendre L. 1998. *Numerical ecology*. Amsterdam: Elsevier.
- Magurran AE. 2004. *Measuring biological diversity*. UK: Blackwell Publishing.
- Mares MA, Willig MR. 1994. Inferring biome associations of recent mammals from samples of temperate and tropical faunas: Paleocological considerations. *Hist Biol* 8:31–48.
- Raup DM, Crick RE. 1979. Measurements of faunal similarities in Paleontology. *J. Paleontol.* 53:1213–1227.
- Shen T-J, Chao A, Lin J-F. 2003. Predicting the number of new species in further taxonomic sampling. *Ecology* 84:798–804.
- Simpson EH. 1949. Measurement of diversity. *Nature* 163:688.
- Sørensen T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Det Kgl Danske Vidensk Selsk Biol Skr* 5:1–34.
- Travouillon KJ, Archer M, Hand SJ, Godthelp H. 2006. Multivariate analyses of Cenozoic mammalian faunas from Riversleigh, north-western Queensland. *Alcheringa Special Issue* 1: 323–349.
- Walther BA, Martin J-L. 2001. Species richness estimation of bird communities: How to control for sampling effort? *Ibis* 143: 413–419.