Ann. Bot. Fennici 37: 289–296
Helsinki 15 December 2000
ISSN 0003-3847
© Finnish Zoological and Botanical Publishing Board 2000

Estimating the quality of the records used in quantitative biogeography with presence–absence matrices

Miguel Murguía & José Luis Villaseñor

Murguía, M. & Villaseñor, J. L., Instituto de Biología, Universidad Nacional Autónoma de México, Departamento de Botánica, Apartado postal 70-367, 04510 México, D.F. México

Received 13 March 2000, accepted 6 November 2000

Murguía, M. & Villaseñor, J. L. 2000: Estimating the quality of the records used in quantitative biogeography with presence–absence matrices. — *Ann. Bot. Fennici* 37: 289–296.

We propose two new methods to estimate the error associated with data in presence–absence matrices commonly used in quantitative biogeographic analyses. The first is based on the estimated richness as compared with the known richness; the second is based on presence–absence frequencies addition in the data matrix. We propose an approach to integrate both criteria by using an index that determines the quality of the records as specified in the data matrix for each geographic region. It is assumed that the errors more commonly associated with the scoring of the biodiversity are those related to the absence of records in the subregions. On the other hand, it is also assumed that those mistakes caused by the scoring of elements in areas where they are not found are minimal. We propose to relate the quality of the records to the level of resolution or precision of the biogeographical analysis, assuming that the lower resolution of the analysis, the better the quality of the records. When a presence–absence data matrix is reordered, considering larger subregions and respectively a lower number of them, the possibility of errors due to lack of records of elements diminishes. This situation can be viewed as a particular case of the so-called modifiable areal unit problem (MAUP) of the geographic information systems.

Key words: biodiversity, biogeography, MAUP, presence–absence records, richness estimation

INTRODUCTION

Quantitative biogeography (Crovello 1981, Bricks 1987) is an important approach in biogeographical analyses. It is based on the use of similarity or dissimilarity indices to compare the records of the biological elements in different regions or geographic units (Operational Geographic Units, or OGUs) with the ultimate goal of grouping the regions according to similarities or dissimilarities (e.g., Hagmeier & Stults 1966, McLaughlin 1989, Nimis *et al*. 1995.) However, there are several decision problems that have not been successfully resolved when carrying out a biogeographical analysis of this kind (Bricks 1987). The choice of a similarity index (Hubálek 1982), the definition of the OGUs (Anderson & Marcus 1993, Wong 1996), and the choice of a clustering method (Bricks 1987, Crovello 1981) are among the decisions that should be made at the onset. Still another question to be answered is: is it really possible to apply a similarity index into the analysis? In other words, how reliable the results of the analysis based on the selected parameters will be (for example the number and shape of the OGUs and the selected similarity index).

There are several analyses concerning similarity indices behavior (*see* for example Hubálek 1982, Sánchez & López 1988). However, to our knowledge, there are no studies exploring the consequences of the application of such indices to an incomplete data set. Likewise, the definition of the form and size of the OGUs is another problem that is related to the sample representativity. When a study area is subdivided, the sampling units are re-defined, either as regular figures such as squares (e.g., McAllister *et al*. 1994) or pentagons (e.g., Griffith 1993), or as irregular areas such as political boundaries or biogeographical or biotic regions (e.g., McLaughlin 1989). Most often the definition of OGUs is carried out after sampling. Phipps (1975) proposed a method of defining OGU in a way that concentrates on maximizing the number of matches; however, he did not provide a justification for the proposed criteria.

When a similarity analysis is carried out based on a presence–absence data matrix that scores the species or taxa occurring in a set of OGUs, it is necessary to estimate the errors produced by different collecting efforts in each area. Generally, knowledge of species richness comes from efforts carried out in different periods of time, and by different teams of researchers with different purposes. There is no single "sampling strategy" and, accordingly, it is not known how complete a given sample is.

Present estimates and analyses of biodiversity should consider the possible omissions of species and other taxonomic groups. Robust estimates should include quantitative parameters that show data on the confidence of discussed results. Taylor (1977) provided a pioneer exercise encompassing such ideas in quantitative biogeography; the author used statistical tests to diminish the error in the similitude-dissimilitude calculations among OGUs due to innacuracies in the scoring of species.

These tests are important, especially because areas that are supposedly well-known floristically, such as North America, continue to reveal new species and genera (Ertter 2000).

This paper aims to define an approach to estimate the error associated with the quality of records (sampling) that document the biodiversity of a region as is scored in a presence–absence data matrix. We also examine the effects of repeated subdivision of regions. An estimate of this kind may help to determine a "minimum limit" of error associated with the results of a biogeographical analysis.

This paper also proposes a strategy for evaluation how representative the records of a sample are in a biogeographical analysis. This is done by estimating the total richness based on the known richness, and considering the number of species presences recorded in the data matrix.

CRITERIA TO ESTIMATE QUALITY OF THE RECORDS

To produce an index that estimates the quality of the sample records for a biogeographical analysis, two factors must be considered: the floristic or faunistic richness, and the number of incidences in the presence–absence data matrix. Non-parametric methods can be used to estimate the richness based on collecting data, assuming they represent samples of populations (Colwell & Coddington 1994). These estimates provide an ap-

proximate value of the "real richness" from the partially known richness. Thus a parameter of the quality of the records (or the sampling in the area) can be obtained from the difference between the known (sampled) and the estimated richness.

Another way to estimate the quality of the records is by presence per species in the data matrix (positive records or hints). A data matrix in which all species are recorded from all the regions can be considered complete, because there are no possible additions to the record. On the other hand, a data matrix with few species records for each region may be faulty when scoring the data for several OGUs. Thus, it is possible to define a maximum and a minimun of the presences in the data matrix and to prefer those data matrices with the high scoring.

Richness estimators

Colwell and Coddington (1994) analyzed several proposals to estimate the total richness of a region from the known richness. They discussed the behavior of several of these richness estimators and list the formulae to calculate them.

Richness estimators require a data matrix of the species by OGUs. There are estimators that only require a presence–absence data matrix and do not require abundance data, while others require both. Among the first kind (as named by Colwell & Coddington 1994) can be cited those of Chao 2, Jackknife 1, Jackknife 2, Bootstrap or ICE (incidence-based coverage estimator), while those of Chao 1 and ACE (abundance-based coverage estimator) are of the second kind. There is evidence that estimators based only on incidences predict richness better than those based on abundance (Colwell 1996 and R. K.Colwell pers. comm).

We propose to evaluate the confidence of the data to be used in a quantitative biogeographical analysis by estimating the richness of the region under study and using the parameter E_s. This parameter measures the proportion of the known richness related to the estimated richness; it is calculated with some of the formulae that estimate the richness (*see* Colwell & Coddington 1994).

Let S_{est} be defined as the estimated (floristic or faunistic) richness for the whole region under study (obtained, for example, by using Chao 2 as defined in Colwell & Coddington 1994), and S_{obs} as the known richness for the same region. Neither S_{est} nor S_{obs} represent the "real" richness of the region; strictly speaking both are estimates. The error E_s of the known richness S_{obs} with respect to the estimated one can be calculated according to the following formula:

$$E_s = [(S_{est} - S_{obs})/S_{est})] - 1 = 1 - S_{obs}/S_{est} \quad (1)$$

E_s takes values from 0 to 1, because the known richness S_{obs} (we assume) should never surpass the estimated richness S_{est}; accordingly, the quotient S_{obs}/S_{est} will never be larger than 1.

The E_s parameter encompasses all the regions (OGUs) and, because a biogeographical analysis may include multiple comparisons within the data matrix (among all the OGUs), the error in the results will be higher, perhaps a potency of E_s:

$$\text{Error} = C(E_s)^k \quad (2)$$

with C and k constants.

To obtain the E_s value the Chao 2 estimator formula can be used:

$$S_{est} = S_{obs} + Q_1^2/(2Q_2) \quad (3)$$

where S_{est} is the estimated richness according to the Chao criterion, Q_1 is the number of species (or taxa) that occur in only one sample (i.e., in only one region) and Q_2 is the number of species (or taxa) that occur in two samples or regions. Chao 2 is one among several proposed estimators; however, we have used it because is easy to interpret, it provides a better estimate of richness than others and it uses only presences and absences, not requiring abundance data, as Chao 1 (Colwell & Coddington 1994).

E_s is a value that determines the robustness of the results from different studies; the lower the E_s, the higher the quality of the results. In general, it is expected that the confidence of the results is greater when spatial resolution diminishes, that is, when the OGUs include larger areas.

Presence–absence data matrix frequencies addition

A second criterion to evaluate the confidence of the results of a similarity analysis is to consider the way species distribute among the sampled ar-

eas or OGUs, and the kind of mistakes most commonly made when a presence–absence data matrix is assembled. In terms of information content, two kinds of mistakes can be produced: (1) a particular species is scored when it really does not occur (errors of commission), and (2) a particular species is not scored when it does occur (errors of omission.)

These two cases are viewed in the data matrix as "presences" (or 1's) where they should be "absences" (or 0's) and vice versa. When all species are present in all OGUs, it can be considered that a good sampling representation has been made, because there is no evidence of missed scores for any species. On the other hand, if there are restricted distributions, for example, if certain species are only present in one OGU, the probability that presences have not been scored for several OGUs increases. In the first case, results that can be derived from the similarity analysis are extremely poor, whereas in the second case results should seem very informative. Thus, the more robust the results are, the less informative, and vice versa. In this way, we consider that the errors most common in a presence–absence data matrix are of the latter kind.

A way to evaluate the confidence of the distributions of species in a presence–absence data matrix is by analyzing their frequency number. Consider an histogram of species frequencies in j areas (i.e., in 1, 2, 3, ..., m areas). The sum (F) of frequencies of all classes multiplied by all classes is defined as follows:

$$F = \sum(jQ_j); j = 1,..., m \qquad (4)$$

where m is the total number of areas (OGUs), Q_j is the number of species found in j areas.

F-values increase when OGUs become smaller; in other words, when spatial resolution is greater. If it is assumed that estimating error of richness is 0, then F has its upper limit at $S_{obs}m$ and its lower limit is either S_{obs} or m; that is, when not one species is shared between any pair of sites (restricted distribution).

If the data matrix has only 1's, then $F = S_{obs}m$. Assuming that neither a column nor a row have only 0's, then the minimum possible of 1's is S_{obs} or m, and strictly the maximum of S_{obs} and m corresponds to $\max(S_{obs}, m)$. If there are no mistakes in the calculation of F, that is, if it is assumed that S_{obs} is correct, then the maximum F-value corresponds to $S_{obs}m$. This means that there are no missing species in the studied area because all of them are present in every OGU.

In this way, a measure of the quality of the records (Q_S) can be defined as:

$$Q_S = F/[S_{obs}m - \max(S_{obs}, m)] \qquad (5)$$

This quotient can take values between 0 and 1, because F is divided by the interval of values it can take.

If there are no mistakes in the calculation of S_{obs}, F may be the source of error; that is, may be all species of the region are scored but not in each OGU where they occur. The former assuming no mistakes are produced when scoring presences (the 1's are well placed in the data matrix) and the source of error is only the lack of records (could be 1's where there are 0's). On the other hand, an error in S_{obs} very probably causes, as a consequence, an error in F; if a row is absent in the data matrix, there will be no 1's to add.

Figure 1 shows different forms an histogram of frequencies of species in j sites can take, assuming a constant S_{obs} and a variable F. If the data matrix is re-built in order to have a smaller number of OGUs, the trend is to get histograms like those on the right-hand side in Fig. 1. Although the absolute value of F diminishes, its relative value with respect to m will probably increase.

COMBINING RICHNESS ESTIMATORS AND SUM OF FREQUENCIES

An additional alternative to evaluate the quality of the records in presence–absence data matrices is combining the E_s and F parameters. The limits of F can be re-written both in terms of the estimated richness (S_{est}) and in terms of the error E_s of the richness estimation.

From Eq. 1 above, we have that

$$S_{est} = S_{obs}/(1 - E_s) \qquad (6)$$

As discussed above, F has a minimum value of $\max(S_{obs}, m)$ and a maximum value of $S_{obs}m$. However, if it is considered that the richness can take values between S_{obs} and S_{est}, and the former is substituted by $S_{obs}/(1 - E_s)$ as the upper limit of F, then the new values the F range can take are

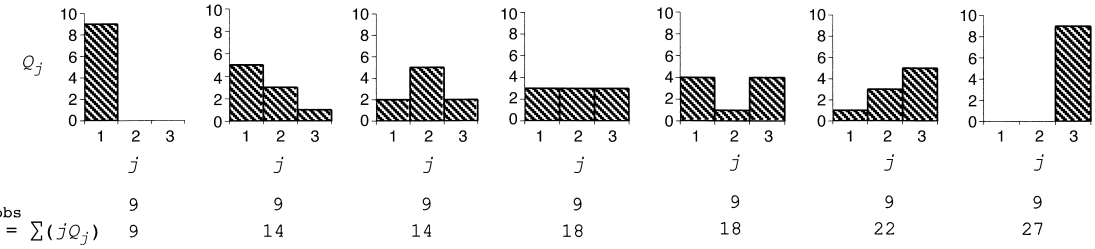

Fig. 1. Seven forms an histogram of number of species by regions with constant known richness (S_{obs}) and variable F can take ($F = \Sigma(jQ_j)$; $j = 1,\ldots,$ m; m = total number of OGUs and Q_j = frequencies of species in j sites). To the right the histogram forms with higher F-values are illustrated. The first histogram has nine taxa, all of them occurring in only one site, while the last histogram also has nine taxa, but all of them occur in three sites. The former has a low F-value (9 = 9(1)) while the latter has a high F-value (27 = 9(3)). F-values increase when OGUs become smaller, that is when resolution is greater. Histograms on the right are from the data matrices with most 1's and a low error in recording data is suspected. The histograms on the left are from the data matrices with less 1's and a larger error in recording data is suspected.

defined as the minimum value $\max(S_{obs}, m)$ and as the maximum value $S_{obs}m/(1 - E_s)$.

If F is divided by the interval $S_{obs}m/(1 - E_s) - \max(S_{obs}, m)$ a number between 0 and 1 is obtained. A number close to 1 indicates a lesser probability of error in the species sampling, and a number close to 0 indicates a greater probability of error. Based on this, an estimate of the quality of the sampling records (Q_S) is:

$$Q_S = F/(S_{obs}m/(1 - E_s) - \max(S_{obs}, m)) \quad (7)$$

A PARTICULAR CASE OF MODIFIABLE AREAL UNIT PROBLEM (MAUP)

There is a problem when one attempts to describe the behavior of F when a particular area is subdivided into smaller areas of different sizes. In general, statistical variance of values associated with geographic units rises when an area is fragmented into smaller ones. This fact has been defined as the modifiable areal unit problem (MAUP, *see* Wong 1996). MAUP implies that while a correspondence (if it does occur) between variance and size of geographic units is found the results of studies in the same region using different sized units are not comparable.

Biogeographical data are also subject to MAUP because statistical variance is larger when smaller OGUs are considered. Q_S defines the effect of changing size and/or shape of OGUs, helping to decide what configuration to use in order to carry out a quantitative analysis.

A CASE STUDY USING DATA OF SELECTED GROUPS OF THE FLORA OF MEXICO

To evaluate the formula for estimating the quality of the sampling records (Q_S) in a presence–absence data matrix, we analyzed the information on the distribution of different floristic groups of the flora of Mexico. We used the information from several databases: (1) the families of flowering plants (Magnoliophyta), (2) the genera of Asteraceae, Fabaceae, Malvaceae, Poaceae, and Rhamnaceae, and (3) the species of the genera *Ageratina* (Asteraceae) and *Desmanthus* (Fabaceae). The first two data sets come from unpublished information compiled by the junior author (except for Malvaceae, obtained from Fryxell (1988) and Rhamnaceae, obtained from Fernández (1993)). The data for *Ageratina* were taken from Turner (1997) and for *Desmanthus* from Luckow (1993). Accordingly, we analysed data for three different taxonomic levels (families, genera, and species). Tables 1 and 2 summarize the information of the number of taxa considered (*see* S_{obs}).

To define the OGUs, Mexico was first subdivided into its 32 political states (Fig. 2 and Table 2) and then those 32 states were re-ordered into 8 regions (Table 2). Accordingly, two sets of OGUs were used, each one having its own data matrix.

Different parameters used to estimate the Q_S values for each taxonomic group in the two sets of OGUs are presented in Table 1. As there is a hierarchical arrangement among the groups analyzed, it is natural to see higher Q_S values at

Table 1. Parameters calculated to determine the quality of records (Q_S) for different angiosperm taxa of the flora of Mexico (number of OGUs: 32/number of OGUs: 8). Q_1= number of species (or taxa) that occur in only one sample (i.e., in only one region); Q_2= number of species (or taxa) that occur in two samples or regions; S_{obs}= the known richness; F = sum of frequencies of all classes multiplied by all classes of the histogram of species frequencies by sample; S_{est} = the estimated richness according to the Chao 2 criterion; $E_s(\%) = 100[(S_{est} - S_{obs})/S_{est}]$; $Q_S(\%)$ = the proposed quality of the sampling records index.

Plant group	Q_1	Q_2	S_{obs}	F	S_{est}	$E_s(\%)$	$Q_S(\%)$
Magnoliophyta	7/7	9/16	257/257	4864/1633	259.7/258.3	1.0/1.0	60.4/90.2
Asteraceae	68/84	32/45	369/369	3977/1537	441.2/447.4	16.4/17.5	28.9/47.9
Fabaceae	11/14	8/8	139/139	1980/761	146.6/151.2	5.2/8.1	43.5/71.1
Malvaceae	4/7	2/6	53/53	660/252	57.0/57.1	7.0/7.1	37.3/62.4
Rhamnaceae	0/1	1/0	11/11	230/76	11.0/11.0	0.0/0.0	67.4/98.7
Poaceae	19/12	6/19	159/159	1856/807	189.1/162.8	15.9/2.3	31.5/70.6
Ageratina	44/3	28/51	138/138	720/498	172.6/138.9	20.0/0.1	13.4/51.5
Desmanthus	6/1	2/5	18/18	100/74	27.0/18.1	33.3/0.5	12.0/58.4

Table 2. Political division of Mexico and the region into which each state was assigned. *See* Fig. 2.

	State	Region
AGS	Aguascalientes	4
BCN	Baja California	1
BCS	Baja California Sur	1
CAM	Campeche	8
CHIS	Chiapas	7
CHIH	Chihuahua	2
COAH	Coahuila	3
COL	Colima	4
DF	Distrito Federal	6
DGO	Durango	2
GTO	Guanajuato	6
GRO	Guerrero	7
HGO	Hidalgo	5
JAL	Jalisco	4
MEX	México	6
MIC	Michoacán	6
MOR	Morelos	6
NAY	Nayarit	4
NLE	Nuevo León	3
OAX	Oaxaca	7
PUE	Puebla	7
QRO	Querétaro	6
QROO	Quintana Roo	8
SLP	San Luis Potosí	5
SIN	Sinaloa	2
SON	Sonora	2
TAB	Tabasco	8
TAM	Tamaulipas	3
TLAX	Tlaxcala	6
VER	Veracruz	5
YUC	Yucatán	8
ZAC	Zacatecas	4

higher taxonomic levels, for example families, and lower Q_S values at lower taxonomic levels.

At the family level, 257 families of Magnoliophyta occurring in Mexico were analyzed. It is considered that taxonomic and floristic knowledge at this level is adequate; therefore, results may serve to calibrate the upper limits of expected Q_s values at lower taxonomic levels. When a set of 32 OGUs (Fig. 1 and Table 2) is used, a Q_S of 60.4% is obtained; with 8 OGUs (Table 2), the Q_S value rises to 90.2%.

At the genus level, the three families with the largest number of species in Mexico (Asteraceae, Fabaceae, and Poaceae) plus two additional taxonomically well-known families (Malvaceae and Rhamnaceae) were evaluated. Q_S at the genus level was, as expected, intermediate between those at family and species level (Table 1). The Q_S values ranged from 29% to 67% when 32 OGUs were used, and from 47.9% to 98.7% when a set of 8 OGUs was employed. At the species level the two genera analyzed, with contrasting numbers of species, had Q_S values ranging from 12% to 13.4% with 32 OGUs, and from 51.5% to 58.4%, when the set of OGUs was 8.

The results point out that Q_S values will be influenced by the size of the taxonomic group as well as by the quality of taxonomic and distributional knowledge. Those groups with a large number of species are taxonomically less well known and thus show lower Q_S values (for example Asteraceae) than those with fewer species and

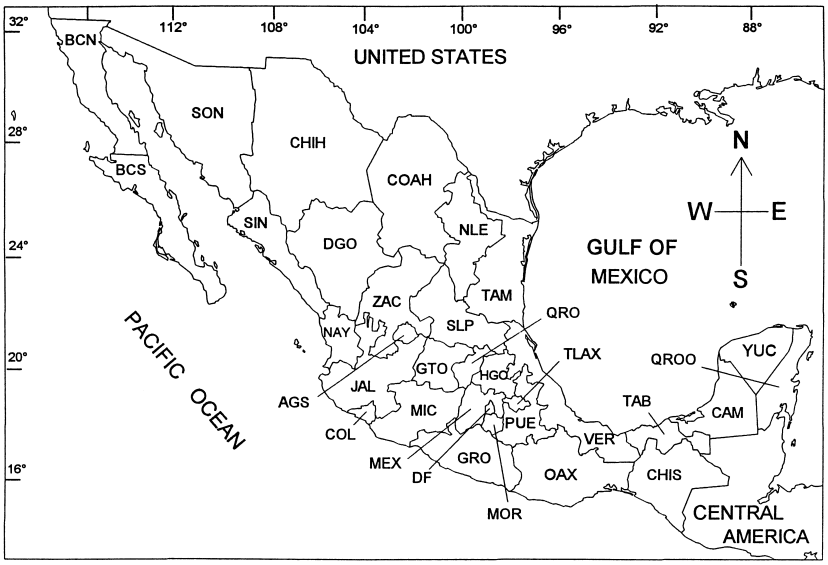

Fig. 2. Political states of Mexico (abbreviations as in Table 2.)

taxonomically better known (for example Rhamnaceae).

At the genus level, where the expected Q_S values are intermediate between family and species levels, estimated values can be proposed either for taxonomically well-known groups or to poorly-known ones. The Asteraceae have Q_S values lower than those obtained at species level (Table 1); on the other hand the Rhamnaceae have values higher than those obtained for all the families. It can then be proposed that when Mexico is divided into 32 OGUs (state level), values higher than 50% indicate a good "representativity" of the data set at the genus level, whereas values lower than 30% indicate a poor data set to be analyzed. On the other hand, when the country is subdivided into 8 OGUs, intervals of Q_S values could be established as follows: 50% or less are "poor" (poor data quality in the data matrix), 50% to 70% are "good", and more than 80% are "very good".

DISCUSSION AND CONCLUSION

Biogeographical analyses based on similarity coefficients must be performed with the realization that they are based on incomplete data. Therefore, it is important to estimate the quality of the data matrix in order to assign an error to the results obtained. In this paper two different criteria are proposed. Also, a combination of both is proposed, resulting in a formula applicable to any data matrix based on presence–absence data. It has the advantage of comparing the quality of the records, either from different regions or taxonomic groups or from data matrices from the same region and the same taxonomic groups but with different OGUs configuration.

Several elements affect the quality of biogeographical data represented in a presence–absence data matrix. Among the most important are the size of the taxonomic group, the taxonomic level, and the kind and size of the OGUs. To define the quality of the data, it is important to consider all of these factors. Q_S coefficient can be useful to estimate how complete the data are, in order to make comparable the values between areas.

The Q_S coefficient provides a way to compare the status of biodiversity knowledge from different regions. The higher the Q_S values, the better known the region. Likewise, it provides a way to estimate the biodiversity knowledge of a region at different times. An increase in this knowledge is not only the scoring of new taxa in a region; it includes their scoring in different places.

An index to estimate the records' quality is useful because it provides an indirect way to decide the resolution level of biogeographical analyses. Our Q_S parameter may be useful to measure how this quality diminishes when resolution increases.

ACKNOWLEDGEMENTS: We thank David Charlet, Claudio Delgadillo, Jorge Llorente, Luis E. Eguiarte, Juan José Morrone, Lawrence Kelly, and one anonymous reviewer for careful reviews of the manuscript and for their valuable comments. The Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) and the Asociación de Biólogos Amigos de la Computación, A. C. (ABACo, A.C.) provided the hardware and software facilities. The senior author thanks the Dirección General de Asuntos del Personal Académico (DGAPA-UNAM) for support through a Doctoral Fellowship.

REFERENCES

Anderson, S. & Marcus, L. F. 1993: Effect of quadrat size on measurements of species density. — *J. Biogeogr.* 20: 421–428.

Bricks, H. J. B. 1987: Recent methodological development in descriptive biogeography. — *Ann. Zool. Fennici* 24: 165–178.

Colwell, R. K. 1996: User's guide to the richness estimator program, Estimates. — Univ. Connecticut. MS available via the internet at http://viceroy.eeb.uconn.edu/EstimateS.

Colwell, R. K. & Coddington, J. A. 1994: Estimating terrestrial biodiversity through extrapolation. — *Phil. Trans. Royal Soc. London B* 345: 101–118.

Crovello, T. J. 1981: Quantitative biogeography: an overview. — *Taxon* 30: 563–575.

Ertter, B. 2000: Floristic surprises in North America north of Mexico. — *Ann. Missouri Bot. Garden* 87: 81–109.

Fernández, R. 1993: *La familia Rhamnaceae en México*. — Ph.D. Thesis. Escuela Nacional de Ciencias Biológicas, I.P.N. México, D.F. 345 pp.

Fryxell, P. A. 1988: Malvaceae of Mexico. — *Syst. Bot. Monogr.* 25: 1–522.

Griffith, D. A. 1993: Advanced spatial statistics for analysing and visualizing geo-referenced data. — *Int. J. Geogr. Inf. Syst.* 7: 107–123.

Hagmeier, E. M. & Stults, C. D. 1966: A numerical analysis of the distribution patterns of North American mammals. — *Syst. Zool.* 15: 125–155.

Hubálek, Z. 1982: Coefficients of association and similarity, based on binary (presence–absence) data: an evaluation. — *Biol. Rev.* 57: 669–689.

Luckow, M. 1993: Monograph of *Desmanthus* (Leguminosae-Mimosoideae). — *Syst. Bot. Monogr.* 38: 1–166.

McAllister, D. E., Schueler, F. W., Roberts, F. W., Roberts, C. M. & Hawkinsiller, J. P. 1994: Mapping and GIS analysis of the global distribution of coral reef fishes on an equal-area grid. — In: Miller, R. I. (ed.), *Mapping the diversity of nature*: 155–175. Chapman & Hall, Oxford.

McLaughlin, S. P. 1989: Natural floristic areas of the western United States. — *J. Biogeogr.* 16: 239–248.

Nimis, L., Malyshev, L., Bolognini, G. & Friesen, N. 1995: Phytogeographic diversity of the Putorana flora (N. Siberia). — *Ann. Bot. Fennici* 32: 1–17.

Phipps, J. B. 1975: BestBlock: optimizing grid size in biogeographic studies. — *Can. J. Bot.* 53: 1447–1452.

Sánchez, O. & López, G. 1988: A theoretical analysis of some indices of similarity as applied to biogeography. — *Folia Entomológica Mexicana* 75: 119–145.

Taylor, D. W. 1977: Floristic relationships along the Cascade-Sierran Axis. — *Am. Midl. Natur.* 92: 333–349.

Turner, B. L. 1997: The Comps of Mexico. Vol. 1. Eupatorieae. — *Phytologia Mem.* 11: 1–272.

Wong, D. 1996: Aggregation effects in geo-referenced data. — In: Arlinghaus, S. L. (ed.), *Practical handbook of spatial statistics*: 83–106. CRC Press, New York.