Biodiversity and Conservation **9**: 1441–1466, 2000.

The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies

JORGE M. SOBERÓN[1,*], JORGE B. LLORENTE[2] and LEONOR OÑATE[3]
*[1]Instituto de Ecología, UNAM, Coyoacan, Mexico; [2]Museo de Zoologia, Facultad de Ciencias, UNAM, Coyoacan, Mexico; [3]Facultad de Ciencias, UNAM, Coyoacan, Mexico; *Author for correspondence: Instituto de Ecología, Ciudad Universitaria, UNAM, P.O. Box 70-275, Coyoacan, D.F., Mexico (fax: +52-5-4223531; e-mail: jsoberon@xolo.conabio.gob.mx)*

Received 23 March 1999; accepted in revised form 18 January 2000

Abstract. In recent years, use of databases of the labels of specimens deposited in museums and herbaria is becoming increasingly common as a tool for addressing biodiversity conservation and management problems. These databases are often large in size and complex in structure, and their application to conservation deserves a wider appreciation of some of the biases, gaps and potential pitfalls common to them. In this paper, we discuss some of the problems associated with using such databases for obtaining lists of species for arbitrary sites, as well as for the estimation of the distribution area of single species. The possibility of obtaining these closely related variables using specimen databases is shown to be scale-dependent. A tool based on mark-recapture techniques is applied to the problem of: (i) detecting sites with low number of species due to lack of adequate in-site sampling and, (ii) species with small estimated areas due to poor spatial coverage of samples.

Key words: butterflies, conservation, Mexico, museum databases

Introduction

In recent years, the increased use of database and geographical information system technology is changing the way in which taxonomic knowledge interacts with that of practitioners of ecology and biogeography. Large-scale databases of specimen labels, georeferenced to geographical coordinates, are being compiled by an ever-growing number of museums and herbaria, as well as by many governmental institutions all over the world (ICBP 1992; Scott et al. 1996; Miller 1994; Soberon et al. 1996; Umminger and Young 1997).

Typical databases use relational database software, although this does not mean that they comply with all the requirements of the relational model (Roman 1997). Taxonomic databases tend to be composed of three or four to 15 to 20 tables with several thousand georeferenced localities and from tens of thousands to hundreds of thousands of specimens (Pankhurst 1991). Currently, it is common to have isolated databases for single collections, or perhaps with one database centralizing the data coming

from many sources. However, prototypes already operational indicate that in the future many institutional databases, maintained and controlled by the curators of the scientific collections, will be linked through the Internet by powerful software agents.

Not only do such databases provide fast access to an unprecedented amount of information of interest for taxonomist and systematists, they can also be used in the work of ecologists and biogeographers as well as for applied purposes. For hundreds of species, new modeling techniques are allowing the quantification of those variables in the 'fundamental niche' (Hutchinson 1987; Holt and Gaines 1992) that have a geographical expression (Stockwell and Noble 1992, Peterson et al. 1999). Similar and other approaches allow calculation of the geographical ranges of species (Soto and Gomez-Pompa 1990; Carpenter et al. 1993; Butterfield et al. 1994; Jones et al. 1997) and the fitting of models that predict the number of species as a function of climatic parameters (Bojorquez-Tapia et al. 1995; Margules and Austin 1995; Llorente et al. 1994; Mourelle and Ezcurra 1996; Wohlgemuth 1998).

In principle, specimen databases should be able to answer two interrelated questions which are central to biogeography and macroecology: (1) What species are found in an arbitrary locality? and (2) What is the geographical distribution of each species? However, there are no universally accepted procedures to assess specimen databases as to the extent to which biases (spatial, temporal and taxonomic) in collecting effort hide the real patterns and might prevent answering those two questions, or even worse, provide false answers.

Therefore, the purpose of this paper is to analyze one medium-sized database from the perspective of its weakness in its use for two important conservation objectives: obtaining lists of species and the estimation of species' geographic range. The emphasis of the approach is methodological and related to the issues of 'data mining' and 'knowledge discovery' general to all large databases (Fayyad et al. 1996; Imielinsky and Mannila 1996).

Description of the database

Between 1978 and 1995 (Llorente et al. 1997) a compilation was made of the data in about 55,000 specimens in major American and Mexican butterfly collections. The institutions consulted appear in Llorente et al. (1997). The only two major collections left out were the de la Maza family's and the one at Instituto de Biologia, of the National University of Mexico, which is still being computerized. However, a significant part of the de la Maza collection information was included since the de la Mazas publish extensively, providing detailed information on localities (see Llorente and Luis 1993 and Llorente et al. 1997 for reviews).

The 55,000 specimens were clumped in 36,685 records, that is, groups of specimens with the same name (i.e., of the same species), date, collector and associated georeferenced locality.

The taxonomy follows Tyler et al. (1994) and Llorente et al. (1997). Different subspecies were regarded as different entities for a total of 176 different subspecies, 70 of the Papilionidae and 106 of the Pieridae.

The locality table has 2261 different names. Some of them are easily identified and well-defined sites (field stations, for example) but others are more subject to interpretation. Specimens with broadly defined 'localities', like 'Mexico' or 'State of Chiapas' were not used. All localities in the table can be traced to the name of a city, village, river, lake or road, with some extra data describing distance and direction from that reference. The process of adding geographic coordinates to the localities was time consuming and difficult. Essentially it was done using 1:50,000 and 1:250,000 cartography and the official census nomenclator for Mexico, and sometimes field books.

After georeferencing and checking by hand, an automatic procedure was run to find inconsistencies between the coordinates fields and the 'political entity' field. This is, a register with a given State should have coordinates included in that State polygon, and all of them should have their coordinates in the terrestrial part of the country. This simple check revealed more than one hundred inconsistencies that had to be corrected again. In the end, all the 2261 localities were assigned geographical coordinates with an estimated resolution of 1 min of arc, or pixels of about 1.1 km of side, at the latitudes of Mexico.

A report on a previous version of the database, together with a detailed printout of all the geographical information as well as illustrations of each species appear in Llorente et al. (1997).

Analysis of the database

The first question we will try to answer is whether the 2261 different sampling localities and the 36,685 records are unbiased spatially and temporally.

In Figure 1 we display the pattern of collecting over time. We grouped records in decades, starting in 1900 (there are too few records previous to 1900 to appear in the graph). The graph displays both the absolute numbers of records for that decade as well as the accumulated fraction of records relative to the total of 24,509 post 1900 records. There is a peak of collecting effort located in the decades of 1970 and 1980. During 1990, collecting effort as registered in our database decreased to levels similar to the average between 1910 and 1950, so collecting effort is not regularly distributed in the last few decades.

From a spatial perspective the data are not well distributed either, as Figure 2 shows. It is apparent that the 2261 localities are not randomly or regularly distributed over the country. One way to show this is to follow Bojorquez-Tapia et al. (1994) to detect so-called road effects. Using a Geographical Information System we put a 10 km buffer around all the federal highways of Mexico and then counted how many

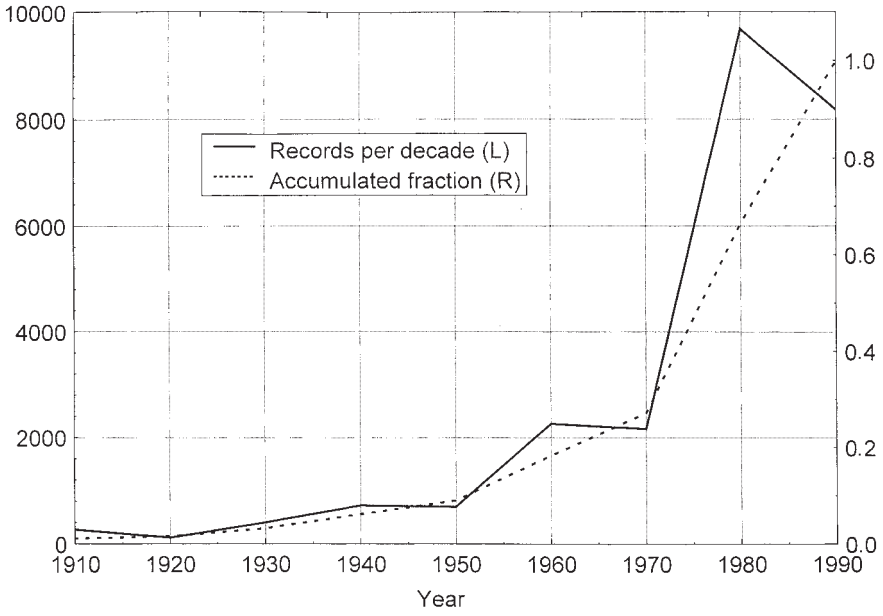

Figure 1. Distribution of collecting efforts over time. Data accumulated by decades starting in 1900.

localities were located over the buffer. The proportion of country area inside and outside the buffer was used to calculate the expected number of localities inside and outside the highway buffer assuming the localities were randomly distributed. The

Figure 2. Localities in the database and the 'highway effect'. The bands represent a 10 km buffer on each side of the federal highways of Mexico. The open circles are localities inside and the dark circles are localities outside the buffer.

resulting value of χ^2 (656.75, d.o.f. = 1, $P < 0.000001$) is consistent with the widely held belief that most biological collecting in Mexico has been done along major roads and around cities and field stations.

Finally, to obtain a picture of how the overall effort has been distributed among species, in Figure 3 the distribution of records per species is displayed. The histogram displays a highly right-skewed distribution. Most species (53%) are known from less than one hundred records.

The above points begin to depict a biased and non-uniform distribution of the collect efforts. With the above in mind, we will now analyze the database from the perspective of its capacity to give information on two complementary themes: (1) providing complete listings for localities (the 'alpha diversity' view), and (2) providing data for the estimation of geographical ranges (the 'beta diversity' view).

The names for the two views came from Whittaker's (1972) subdivision of total species richness in a large region in an 'alpha' component, given by the local diversity, and a 'beta' component, given by the turnover of species among habitats. It is an easy task to show that Whittaker's original measure for the beta component is just the reciprocal of the average number of localities in which they are present (Schluter and Ricklefs 1993; J. Soberon and P. Rodriguez, unpublished manuscript).

Although of course this type of database can be used for many tasks not related to the above views, many pressing conservation questions are crucially dependent on being able to answer 'alpha' (location of hotspots, design of reserves, restoration assessment) or 'beta' (specific species protection, reintroduction programs) types of

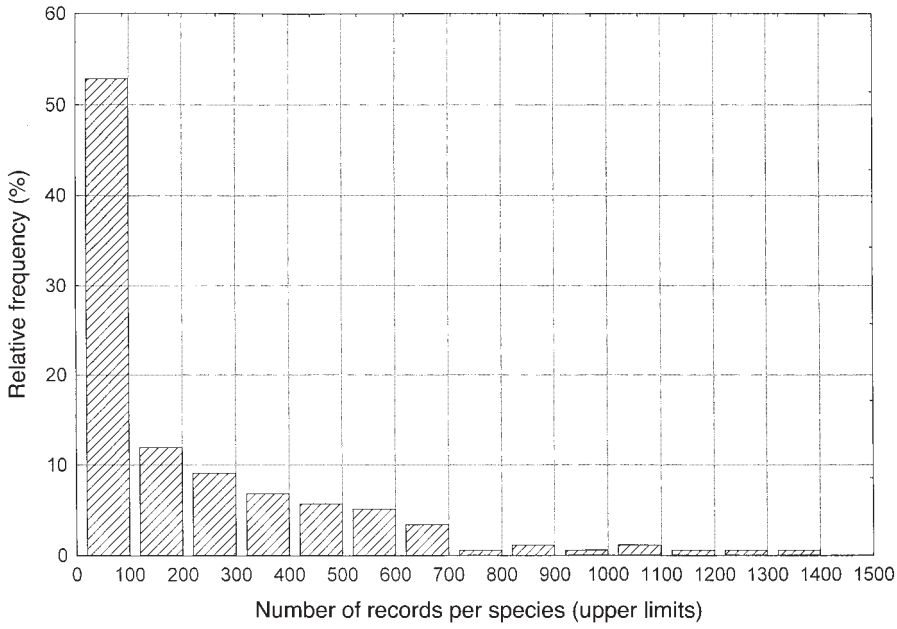

Figure 3. Distribution of number of records among species.

questions. We put the emphasis on those perspectives because they are important for conservation purposes and because some governmental agencies in many parts of the world are using this type of database to address such questions (Soberon et al. 1996).

We shall study the database at three different scales. (1) The localities, which have a resolution of a few square kilometers; (2) a grid of 1/2 degree cells superimposed on the map of Mexico, with a resolution of about 2800 km^2; and (3) a subdivision of Mexico on eight vegetation types with an average surface of 241,300 km^2.

The 'alpha diversity' point of view

The scale of the localities

To assess how complete the database is for getting lists of names of species in the localities we obtained distributions of records and species per locality. The distribution of records among localities displays a pattern of extreme skewness, since even the distribution of the logarithms is skewed (Figure 4). Almost 50% of the localities have 1 to 3 records, and only about 3% of the localities have more than ninety records.

This results clearly point out to a very biased and unsatisfactory distribution of the sampling effort in space and it is not surprising then that the number of species per locality also show a markedly skewed distribution, with almost 80% of the localities having 6 or less species registered, as displayed in Figure 5. Therefore, procedures are needed to decide if small numbers of species are due to bad sampling or to the particular historical and ecological factors of the locality.

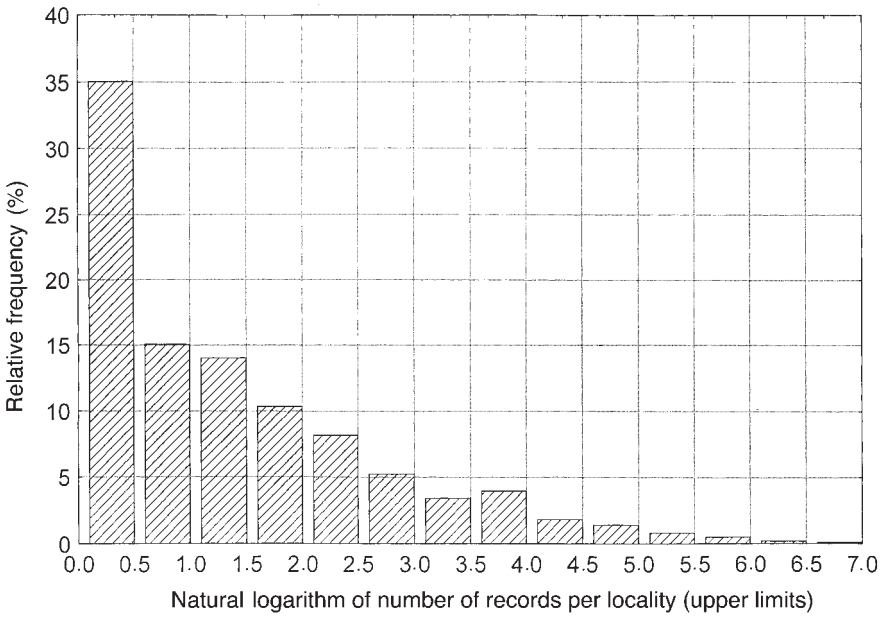

Figure 4. Distribution of the natural logarithm of the number of records among localities.

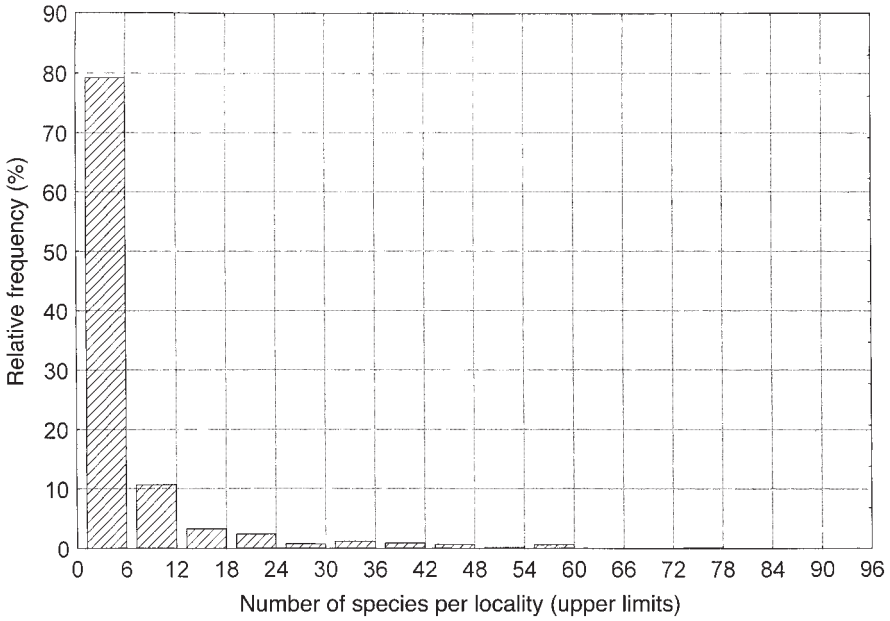

Figure 5. Distribution of the number of species among localities.

The first and obvious thing to do is to establish a threshold for the minimal number of records or species required to regard localities as acceptable. For example, localities with less than ten records might be regarded as poorly sampled. However, in some cases a few records may be enough to characterize a locality for certain groups.

One way out of this problem might be to use methods like those described by Prendergast et al. (1993a), Soberon and Llorente (1993), Colwell and Coddington (1994) and Leon-Cortes et al. (1998) in which one extrapolates the sampling history to obtain estimates of the 'total' or 'true' number of species in a site (i.e., a number representing the size of the list, assuming good and thorough collecting methods was applied to all relevant ecological and temporal conditions for a reasonably long period of time). To do this, we selected those localities with at least 10 records. This reduced the number of localities from 2261 to 480, i.e., a reduction to 21% of the original. This also reduced the number of species found in the subsample from 176 to 174 species. In order to maintain a full representation of species, we augmented the set of 480 remaining localities to include all those in which the 2 missing species were found. We ended with a set of 486 localities that either contained at least 10 different records or contained one of the species required to maintain a total count of 176 species. Data 'cleaning' and preprocessing, which might include disregarding subsets of a database is an integral part of knowledge discovery in databases (Fayyad et al. 1996).

The set of 486 localities was then used to construct a table of records, containing 31,413 records (i.e., all the records differing either in date or collector or species for each one of the selected localities). Notice that a reduction to 21% of the localities

caused only a reduction to 85% of the original number of records. Finally, a cross-tabulation was made of species per sampling date. Sampling dates are simply dates differing in year, month or day. In this table there are 486 (one for each site) sub-tables with an average of 8.9 different sampling dates in each. Each sub-table is just a species per samples' presence–absence matrix that can be used as the input for the non-parametric, incidence-based extrapolation estimators described by Colwell (1997) and Chazdon et al. (1998). This yields an estimate of the total number of species present in a locality based on the way the species number grows with each new sample (in our case, distinct date in each locality).

A Qbasic program was constructed to calculate from each sub-table its 'ICE' (Incidence Coverage Estimation) estimator, which estimates the 'true' species number for each locality (see Colwell 1997, Chazdon et al. 1998 and Lee and Chao 1994). The ICE estimator was chosen among many possibilities because it is incidence-based (it does not require estimates of the abundance of the species in each sample, only of its presence), and was designed to take into account explicitly the right-biased nature of many biological datasets (Colwell 1997).

The program rejected those subtables that were too small to allow a sensible application of the method (less than 2 species or 4 sampling dates). Therefore, after application of the algorithm, only 291 localities had fulfilled all the requisites. In Figure 6 we report the results of the above procedure as an scattergram of values of the ratio Observed Species/Predicted Species, henceforth called the completeness ratio, or *C*, versus the number of species reported in the database for that locality. The *C* ratio can be interpreted as a measure of how complete is the inventory in a given locality. If all localities were well sampled, we would expect all the points to have a *C* value near to one, perhaps with most of them located in a band around 0.9 or 0.8, regardless of the observed number of species.

Unfortunately, the Figure depicts a set of localities that still appear to be poorly sampled. About 75% of the localities have completeness ratios of less than 0.65. However, the picture we get now is in a sense clearer than in the previous ones. It is now possible at a glance to distinguish those localities with small number of species because probably they are poorly sampled, from those that the test regards as well sampled and for which a small number of species should be a consequence of ecological, historical or human factors. One can set a threshold, like 0.8 or 0.9, and decide that only localities with less than that amount need to have more intense sampling in order to finish their butterflies' lists.

Another interesting feature of Figure 6 is that most localities with smaller *C* ratios are also those with few species. To increase the ratio, most work would have to concentrate in localities with few reported species.

Unfortunately, although the test is informative, it is still not conclusive because the extrapolation algorithm depends on the assumption that the sampling was well performed. A careless collector (or collectors) might keep accumulating samples for a locality without ever getting the rarer species due to the lack of skill or thoroughness

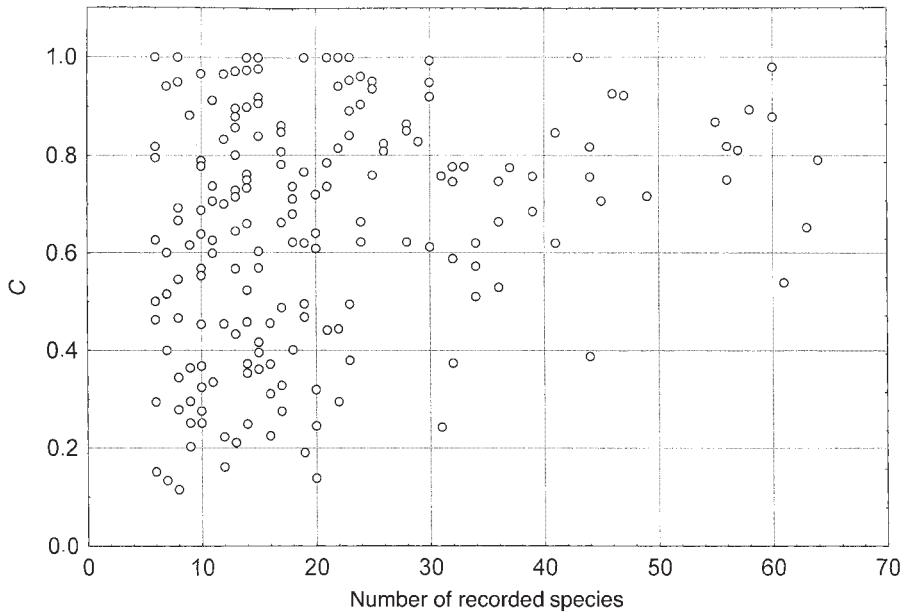

Figure 6. Completeness ratio for 291 localities (see text) as a function of the recorded number of species. The *C* ratio represents the proportion of the number of species actually observed in a locality relative to the number predicted by the ICE estimator described in Colwell (1997).

and thus making *C* to converge to unity without having obtained a near-complete list. An algorithm such ICE cannot distinguish this situation from one in which the true species value has been reached. A summary of the statistics of the analysis appears in Table 1.

The scale of the 1/2 degree grid

The above patterns apply to scales that are measured in a few square kilometers. Perhaps pooling data in units of decreased resolution will lead to improved knowledge of the two families under consideration. There are several ways of aggregating data to increase the scale of observation. A very common one is to impose a grid and to obtain lists of species within each cell.

To do this, we used a grid pattern utilized by Arita et al. (1998) for their analyses of the diversity of Mexican mammals. This is a grid of 704 cells of 1/2 degree of side (roughly 2800 km^2 each). We avoided a number of nonsensical cells (those located almost entirely on the sea or beyond the borders of Mexico) and thus we allowed a total of 94 out of the 2261 localities to remain disassociated from a cell. The localities in this situation were removed from the analysis.

In Figure 7 we present the distribution of the logarithm of the number of records per cell. The distribution of the number of records is so right-skewed that the use of logarithms is needed. It is noticeable however, that the distribution is much less

Table 1. Descriptive statistics for the completeness ratio analysis of the number of species in the localities and cell spatial scales.

	Scale of the localities			Scale of the 1/2 degree cells		
	Samples	Observed species	*C*	Samples	Observed species	*C*
N	224			190		
Mean	15.36	26.29	0.60	19.07	26.22	0.62
Median	11	22	0.62	14	20	0.65
Minimum	5	6	0.08	5	2	0.07
Maximum	95	78	1	102	79	1
s^2	12.52	15.13	0.23	17.66	18.66	0.23
25th Percentile	8	14	0.42	7	11	0.43
75th Percentile	17	35	0.80	22	38	0.82

skewed than in the case of the localities. The shift in the scale has produced a noticeable (although still negligible from a practical perspective) increase in the degree of knowledge about the distributions of butterflies that can be obtained from the database.

This can be also appreciated in Figure 8, where we display the distribution of number of cells with a given number of species. Notice how now about 80% of the cells have 20 or less species reported, in comparison of 80% localities with 6 or less species reported before.

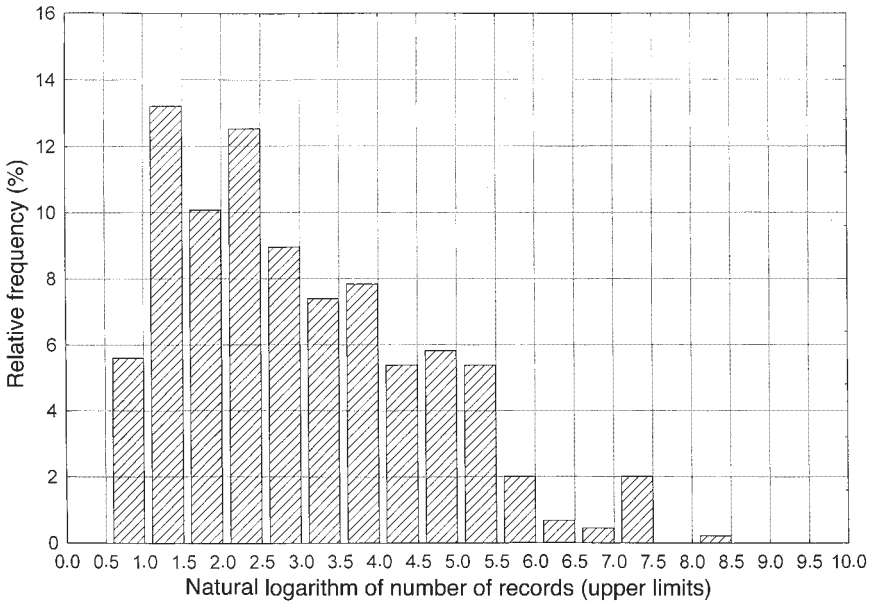

Figure 7. Distribution of the natural logarithm of the number of records among cells. The cells are 1/2 degree of side squares superimposed on the localities and with the data of localities aggregated in each cell.

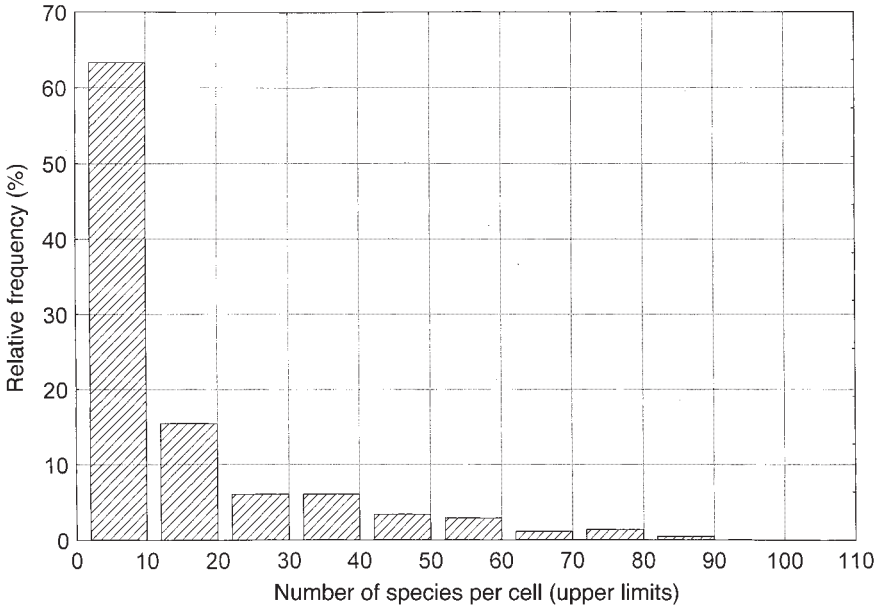

Figure 8. Distribution of the number of species among cells. The cells are 1/2 degree of side squares superimposed on the localities and with the data of localities aggregated in each cell.

In order to obtain an estimate of the true number of species per cell we used again the ICE algorithm. First we reduced the number of cells removing those with 10 or less records. This left us with a subset of 235 of the original 448 cells with at least one register. The 235 cells have associated 35,180 records of an original 36,685. That is, a database with 53% of the original number of cells still maintains 96% of the records.

The database was queried to produce a table with the cell identity code, the years and months of collections, and the register and species identity codes. The data were arranged in a cross-tabulation in such a way that for each cell, all sampling events differing in either the year or the month were regarded as different and every species had a zero or a one depending on its presence on that site and sampling event. A Qbasic program calculated ICE estimates for each cell containing at least 4 samples. The results appear in Figure 9.

As previously discussed for localities, by defining a threshold for *C* this procedure can be used to differentiate those cells regarded as poorly sampled. It is very interesting to note that the statistics for the two different scales do not differ much, nor the general aspect of the scatterplot of observed vs. *C* ratio. One would expect that by aggregating data from several localities better lists might be obtained, but the data do not support this. The general aspect in both Figures is one in which the majority of the sites can be regarded as poorly sampled. The descriptive statistics appear in Table 1.

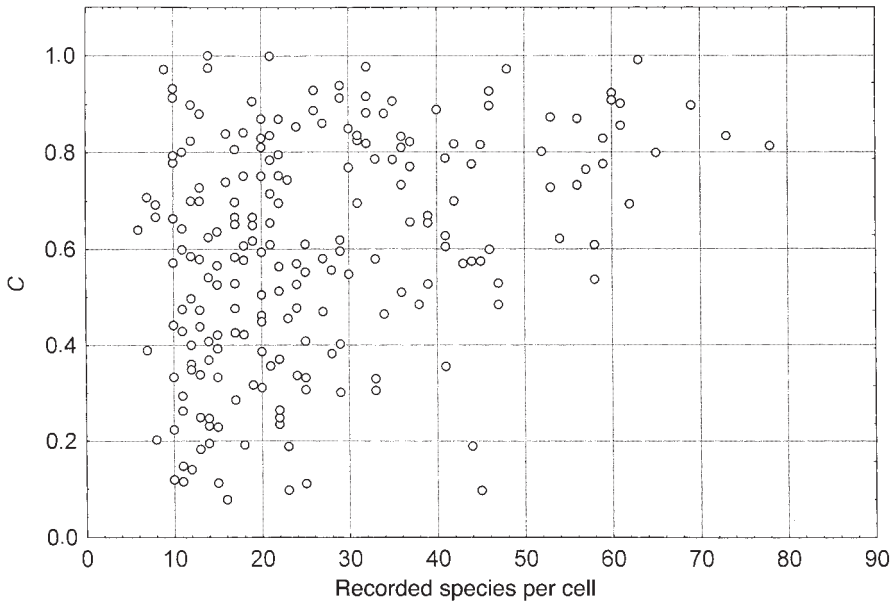

Figure 9. Completeness ratio for 190 1/2 degree of side cells (see text) as a function of the observed number of species. The *C* ratio represents the proportion of the number of species actually observed in a cell relative to the number predicted by the ICE estimator described in Colwell (1997).

The scale of vegetation types of Mexico

The last scale we shall explore in this work is that of the vegetation type. We used the Potential Vegetation of Mexico, according to Rzedowsky (1978) in an electronic version of the map published by UNAM, the National University of Mexico (scale 1:4,000, 000) and provided by the National Comission for Biodiversity (CONABIO). We used eight of the subdivisions of Rzedowsky, excluding only the wetland vegetation. By using GIS software, subsets of the records table were obtained, containing all the records located within each one of Rzedowsky's main vegetation types.

A further subset was obtained for each vegetation type, with only those records with the data for the year available. These subsets were then used to obtain extrapolations of the likely number of species using the ICE algorithm as implemented in the EstimateS software package (Colwell 1997). In this case, each sample is the number of species registered in a given decade. The pooled results for both families appear in Table 2.

The data used to extrapolate excluded almost 30% of records without a date. However, ICE is still very good at predicting the total number of species reported, as depicted in Figure 10. The line is the identity function. The open circles show the number of species observed at each vegetation type, ignoring data without a date. The ICE estimator predicts increases in the number of species in all points. However, the prediction is always very close to the number of species observed when

Table 2. Results of the analysis of the database at the scale of vegetation regions of Mexico.

	Area (km^2)	Fraction[a]	Localities	Density[b]	Sps.[c]	Sps.[d]	Year[e]	ICE
Bce	374395.21	0.1940	439	1.173	146	125	1880	136.38
Be	114806.01	0.0595	113	0.984	72	54	1900	73.04
Bmm	17158.66	0.0089	80	4.662	106	99	1880	108.85
Btc	268976.00	0.1393	453	1.684	128	105	1890	126.91
Btp	186277.99	0.0965	291	1.562	113	109	1890	114.15
Bts	54133.89	0.0280	104	1.921	104	83	1920	102.66
Mx	751571.71	0.3893	589	0.784	93	82	1890	90.4
P	163085.77	0.0845	87	0.533	49	35	1900	44.9

The vegetation types used are the following: Bce = Oak and conifer forest; Be = Thorny forest; Bmm = Cloud forest; Btc = Deciduous tropical forest; Btp = Evergreen tropical forest; Bts = Semi-deciduous tropical forest; Mx = xerophitic shrub; P = grasslands.
[a]Fraction of area relative to the total surface of Mexico.
[b]Number of localities per 1000 km^2.
[c]All records, included those without a year.
[d]Only records with an associated year were used to obtain the ICE estimate.
[e]Starting decade for the Ice estimation.

the undated records are included (closed circles). Therefore, this graph support the idea that probably very few (in relative terms) new species will be added to the lists for the main Rzedowsky's (1978) vegetation types in Mexico. However, the main

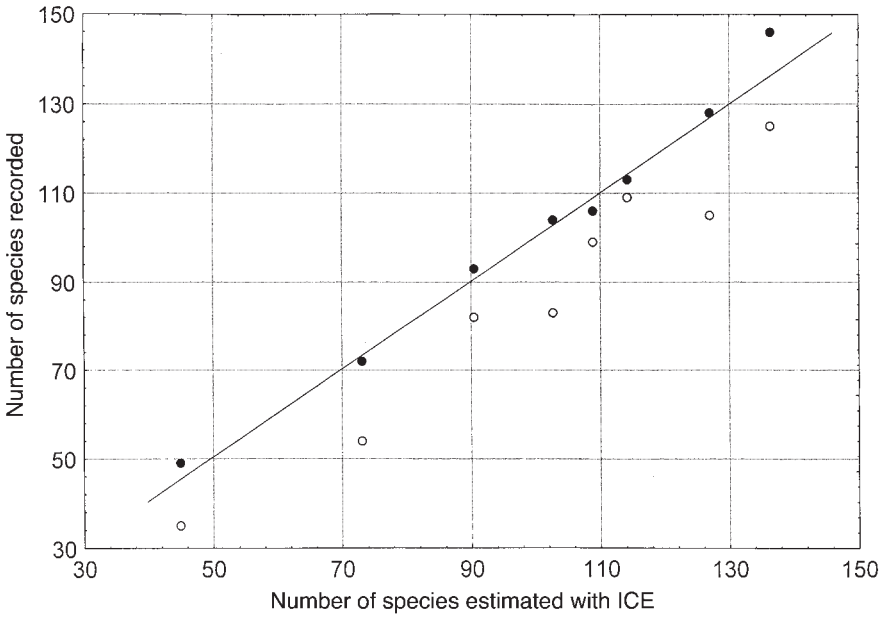

Figure 10. Observed vs. predicted number of species in each one of eight major vegetation types of Mexico. The predicted value was obtained using only data with an associated year-of-capture in the database. ○ Represent the recorded number of species with an associated year-of-capture and ● represent all records of species, with and without associated year-of-capture.

point we want to make is that it is only at a gross scale that the database appears to provide enough data as to make complete lists.

The 'beta diversity' point of view

Another way of exploring the database is from the perspective of how much information it provides about the range of presence of the species. In the limit of very high resolution the range will be a precise area, but in this paper we regard the location of sites or cells in which a species has been registered as an approximation of the 'true' distribution area.

The scale of the localities

In Figure 11 we display the frequency distribution of the number of localities for which each species is reported in the database. If the geographic range of every species were well known, the mean of the distribution would be a direct measure of the beta component of papilionid and pierid diversity of the country.

The fact that the number of sites per species is right-skewed has been noted very often (Preston 1948, 1962; Rapoport 1982; Gaston 1994; Brown 1995) and again one is left with the problem of not knowing what part of the log-normal pattern is due to biology (or the law of large numbers. See May 1973) and what part is due to a very poor and incomplete sampling regime of the country.

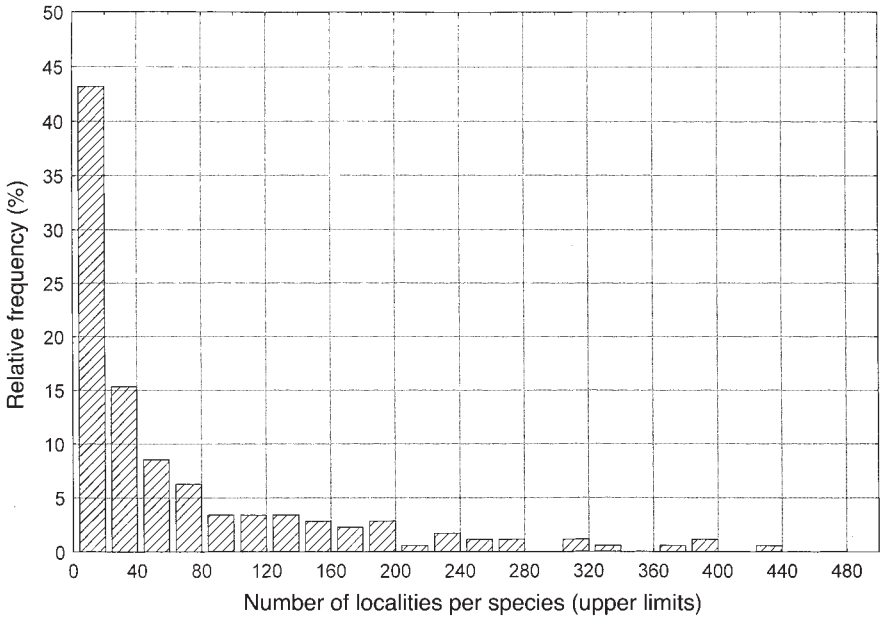

Figure 11. Distribution of the number of localities for which each species has been recorded.

In the ‘alpha diversity’ perspective, we approached the problem of separating ignorance from biology by performing the superficially drastic exclusion of those localities with very few species or sampling dates. This allowed us to proceed further in distinguishing between localities with few species due to bad sampling and those with few species due to biological reasons. The essence of the procedure was first to reduce the database to meaningful localities and then apply an extrapolative procedure to obtain estimates of the unknown number ‘true species richness’.

However, in the ‘beta diversity’ view of the database at the locality scale applying the same methodology is not feasible for two reasons: (1) Many species reported in few localities might well be legitimately rare (in the sense that their range is restricted (Gaston 1994), or they might be in the limit of a wider range and thus present in very few localities. This is the case for several species with very few (5 or less) localities like *Papilio indra pergamus, P. zelicaon, Pterourus glaucus alexiares, P. rutulus rutulus, Anthocharis sara inghami* and *A. cethura*, which in Mexico are at the southernmost part of their ranges. Other species like *Pterourus esperanza* appear to be indeed very restricted in their distributions. Therefore, by rejecting species with few localities one would be disposing off valuable data. (2) There is not an accepted procedure to extrapolate samples to an estimated ‘true’ area or number of localities. One idea that comes to mind is to apply the ICE algorithm to the ‘beta’ view of the database and, for each species, obtain an estimate of the number of localities where it will be found. However at the scale of localities this is not possible because their size and shape are not defined and therefore a discrete and finite universe for the predictions is lacking. In the next section where localities are replaced by a subdivision of Mexico in discrete cells, this ‘inverse’ application of the ICE algorithm will be attempted.

The scale of cells

The grid of cells used before supply a well-defined and finite number of classes that cover the surface of the country. It should be possible then to apply the ICE algorithm to the beta view of the database in order to obtain estimates of how many cells a species should be found in. This was done by querying the database to obtain a cross-tabulation having, for each species, all years and months where it was registered (the samples) and a zero or a one for every cell. This generates presence–absence matrices for each species but now the ‘species’ (for the purpose of applying ICE) are the different cells, and the samples are all the records of the particular butterfly that differ in year and month. Notice that this is not a simple transpose of the matrix used to estimate the true number of species in the cells.

A Qbasic program was run on all the species with at least four different samples to obtain the completeness ratio (for localities) for each species with enough samples. The results are displayed in Figure 12, which is a scatterplot of the number of localities in which a species has been observed versus the ICE-predicted proportion.

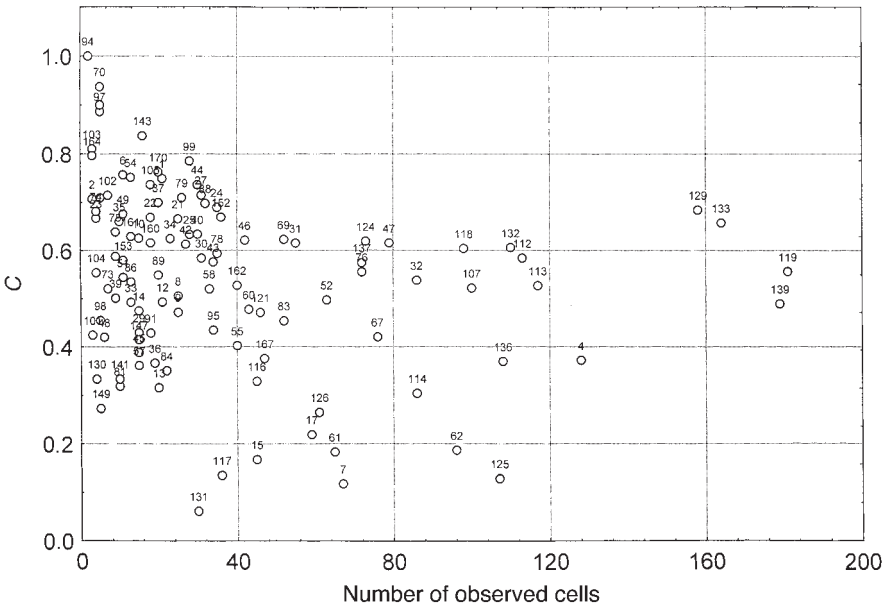

Figure 12. Completeness ratio for 104 species as a function of the observed number of localities in which they have been reported. In this figure the *C* ratio was obtained by using an ICE estimate to predict the number of cells in which a given species should be located given its pattern of sampling in all cells where the species has been reported (see text). The numbers correspond to the Id-Species column in the Appendix.

There is a marked inverse relationship that, without any attempt to formality, can be explained by noting that the ICE estimator will converge to the observed number of cells (and therefore *C* to the value of one) when the frequency of uncommon observations diminish (Colwell and Coddington 1994). In the extreme, $C = 1$ when all observations of the species are composed of at least two dates in the observed cells and in no cell the species has been observed only once.

The inverse shape of the cloud of points in the graph means that those species with a large number of samples tend to have many cells with a single observation, whereas for species with few samples many have been sampled several times in their cells. Biologically speaking, for a species with a very restricted area, it is likely that new observations will occur in those cells already sampled, which will lead to a decrease in the numbers of cells sampled only once. This makes the predicted value to converge to the observed number of cells (values of the completeness ratio near one).

On the contrary, for those species with wide distribution areas, it is easy to add new observations in new cells, making the number of cells sampled only once to increase and thus lower the completeness ratio.

Indeed, in the region of high number of observed cells and low value of *C* we get species like *Battus philenor philenor*, *Protographium philolaus philolaus*, *Phoebis sennae marcellina*, *Eurema mexicana mexicana* and *Krycogonia lyside*, which are

widely distributed in Mexico. On the other hand, in the region of the species registered in a few cells, but with a high value of *C* we see *Lienix neblina*, *Lienix nemesis atthis*, two new subspecies of *Catastica ochracea* and *Catasticta teutila*, *Dismorphia amhiona lupita* and *Hesperocharis crocea jaliscana*, all of which have restricted distributions in our database.

In view of the above, to interpret correctly the graph one has to remember that for a given value of number of cells where the species has been observed, high values of *C* correspond to a high number of observations of the species being repeated in the same cells, and low values of *C* to a low number of observations of the species in the observed cells. New sampling will increase the values of *C* only if no new cells were added to the coverage of a species. Adding new cells will decrease or maintain the value of *C*.

The application of the methodology then creates a picture in which a very high percentage of the species are expected to expand their registered range when non-sampled cells are added to the database. However, we get the interesting prediction that this effect will be much more marked in the common than in the rare species. The descriptive statistics for this analysis appear in Table 3.

In a way similar to the alpha view of the database, we believe that the application of the extrapolation methodology is illuminating but still leaves too many avenues for interpretation open. Perhaps the way out of the problem is the application of other extrapolation tools, like bioclimatic or artificial intelligence models (Soto and Gomez-Pompa 1990; Stockwell and Noble 1992; Carpenter et al. 1993; Butterfield et al. 1994) that generate inferred predictions of niche variables like temperature, precipitation, altitude etc. with a clearcut geographical counterpart. For this purpose, the methods outlined here can be useful to select those species likely to have enough distributional data to apply the niche-extrapolation models.

Table 3. Descriptive statistics for the completeness ratio analysis of the number of cells for each species at the cell spatial scale.

	Scale of the 1/2 degree cells		
	Samples	Observed cells	*C*
N	124		
Mean	48.66	36.17	0.55
Median	33	20	0.57
Minimum	5	2	0.06
Maximum	249	181	1
s^2	50.21	39.04	0.19
25th Percentile	13.5	10	0.42
75th Percentile	65.5	45.5	0.67

Discussion

The compiling and maintenance of large databases of label information is a trend in which most large museums of the world are taking part. In the near future, many if not all of those databases will be interconnected via the Internet, allowing access to a totally unprecedented amount of locality information. Since in all likelihood such databases, centralized or otherwise, will be used by many users with little or no taxonomic expertise, it is very important that the main weaknesses associated with such novel and complex structures are explored.

There are daunting problems related with the taxonomic and nomenclatural aspects of such large and heterogeneous databases, for example to identify synonyms, which in certain cases might compose as much as 50% of the names (Gaston and Mound 1993). However in this paper we decided to focus on the other side of the problem, namely, the gaps and biases of the databases when they are used to obtain listings of species or provide the basis for area of distribution extrapolations.

Despite the fact that the database we studied contains data from 55,000 specimens deposited in all the major collections of Mexican papilionids and pierids, two of the better known families of butterflies, the pattern that emerges is one of extreme bias in the temporal and spatial distribution of the information. In the alpha view of the database the scales of spatial units of 10^1 km^2 (the localities) or 10^3 km^2 (the grid cells) display very similar patterns of completeness ratios. It is only after the degree of resolution has decreased to the scale of units of 10^5 km^2 (Table 2) that the lists of species for most units appear to be complete or nearly complete, as suggested by the *C* ratio.

The *C* ratio can be interpreted as a prediction of how much the species list will increase in a given locality after more samples are obtained. For example, $C = 0.8$ means that 20% of increase in the species list should be expected after a thorough sampling. However, we should be cautious about how seriously to take these predictions until a substantial amount of empirical tests of the ratio have been obtained. It would be probably unsafe to expect *C* ratios much lower than 0.5 to represent reliable predictions. Leon-Cortes et al. (1998) found that a number of extrapolative methods did predict within 20% the number of sphingid moths in three Mexican localities, but there is an urgent need to subject the methods presented here to extensive empirical testing.

Of course the relevant question is not whether the data are well distributed in an abstract sense, but if they are well distributed in relation to the important ecological and biogeographical factors for the taxonomic groups in question. After all, the apparent gaps in knowledge might correspond to ecological conditions containing localities in which very good collections have already been made, and therefore perhaps that ecological condition is well sampled despite an apparent sparseness of data points. Indeed, an experienced field biologist can predict with a reasonable degree of confidence the species likely to be found in a place, or the distribution of many

species, without having access to large databases or complicated software. However there are several important points here:

1. In the first place, the reliability of such 'expert-based' predictions is very dependent on the scale used. A country as ecologically complex as Mexico is still producing surprises (both of presence and absence) to the most expert biologist, whenever new zones are explored. Even in small, developed, and low-diversity countries like Switzerland the quality of inventories varies spatially and affects modeling efforts (Wohlgemuth 1998). This means that even if the relevant ecological and biogeographical factors that affect the distributions of butterflies are reasonably well known at rough scales, the myriad of details that influence observations at finer scales still require extensive research. Our analysis of the database shows that at the scales of localities and 1/2 degree cells the country is basically unexplored. At this scale Mexico still displays an enormous amount of the topographical, vegetational, edaphic and climatic variations that are known to affect butterfly distributions (Weiss and Murphy 1993).
2. The needs of biodiversity information for management and conservation in Mexico and in many other high-diversity countries are already beyond the gross 'hot spots' predictions that have spatial detail of tens of thousands of square kilometers or more (Myers 1988; Dinerstein et al. 1995). More and more conservationists, NGOs or government officers require data at the scale of units of tens to hundreds of square kilometers, and in the absence of the human and economical resources, as well as the time to mount full-fledged research expeditions, they are resorting to modeling based on presence databases (Nelson et al. 1990; Prendergast et al. 1993b; Bojorquez et al. 1995; Soberon et al. 1996; Austin 1998; Scott and Jennings 1998).
3. The essence of natural science is the need to check predictions, informal or formal, with empirical data. The analysis presented here shows that at this point in time, and accepting that the database is a reasonable expression of the available information about the Papilionids and Pierids of Mexico, there is not enough empirical data available to check predictions at any but the roughest spatial scales, without resorting to further field work.

Since the pattern of very poor collections at the majority of localities appears to be very common (Peterson et al. 1998 for Mexican birds and Sanchez Cordero, personal communication, for the Mexican mammals) and probably almost universal, a pressing question is what can one do with such biased and incomplete information? Despite the overtly negative tone of the previous discussion, our experience working with museum labels databases leads us to believe that they can provide a powerful and very useful instrument for conservation planning as the raw material for extrapolative techniques with high potential in conservation (Soto and Gomez-Pompa 1990; Stockwell and Noble 1992; Carpenter et al. 1993; Chapman and Busby 1994; Bojorquez-Tapia et al. 1995; Mourelle and Ezcurra 1996; Scott et al. 1996; Wohlgemuth 1998, and many others). Moreover, it is probably true to say that in the near future we will

witness a massive movement towards computerization and Internet-linkage of zoological and botanical collections. This will be driven both by the intrinsic value to biological research and by the powerful tool such databases will provide to governments, NGOs and others interested in conservation and sustainable management.

In this process, care should be taken at all steps of the assembling and use of large, mixed-origin databases, from the checking of specimens by experts, use of accepted and updated taxonomic authority files, and proper and careful georeferencing of the data, but also to the statistical and graphical analysis of the main gaps in the databases that we illustrated with the methods presented in this work. We would like to stress that there is still ample space for developing and perfecting methods to describe and assess gaps of collection labels databases. Generally speaking, the field of 'data mining' in computer sciences is a new one (Imielinski and Mannila 1996). Conservation biologists, systematists and biogeographers will have to develop the concepts and statistical tools and algorithms required extracting valid, relevant knowledge from the rapidly growing collection of biodiversity databases available even now.

The inescapable conclusion of the analysis we presented is that much more work is needed in the future along three lines: (1) Increase the efforts to collect in poorly explored areas and to house, curate and study the resulting specimens. (2) Increase the efforts to computerize museum information and to incorporate the data in properly designed and maintained databases. (3) Develop the software tools and analytical methods that will allow knowledge extraction and predictive modeling on the basis of large, composite databases.

Acknowledgements

We are grateful to Armando Luis and Isabel Vargas, without whose help the database could never have been completed. We also would like to thank Townsend Peterson and Victor Sanchez-Cordero for comments, suggestions and illuminating discussions on the subject of specimen databases. Patricia Koleff and Raul Jimenez of CONABIO, Mexico, generously gave us technical advice and useful comments on relational databases. Rob Colwell also provided us with many useful suggestions and with the use of his program EstimateS, which was used to obtain the ICE values at the vegetation-type scale. Jorge Soberon acknowledges the support of CONACyT (Grant 981057) and UNAM for a sabbatical period that allowed completion of the research, of the National Science Foundation for the acquisition of computing equipment and of the Natural History Museum of Kansas University for providing space and general facilities for research. Jorge Llorente acknowledges support of grants DGAPA IN-211397 and CONACyT 32002 and the hospitality of the Instituto de Ciencias Naturales, Universidad Nacional de Colombia.

Appendix. List of identity codes for the 176 species of Papilionidae and Pieridae used in the analysis.

Id_Species	Family	Genus	Species	Subspecies
1	Papilionidae	*Baronia*	*brevicornis*	*brevicornis*
2	Papilionidae	*Baronia*	*brevicornis*	*rufodiscalis*
4	Papilionidae	*Battus*	*philenor*	*philenor*
5	Papilionidae	*Battus*	*philenor*	*orsua*
6	Papilionidae	*Battus*	*philenor*	*acauda*
7	Papilionidae	*Battus*	*polydamas*	*polydamas*
8	Papilionidae	*Battus*	*laodamas*	*iopas*
9	Papilionidae	*Battus*	*laodamas*	*copanae*
10	Papilionidae	*Battus*	*eracon*	
12	Papilionidae	*Battus*	*ingenuus*	
13	Papilionidae	*Battus*	*lycidas*	
14	Papilionidae	*Parides*	*alopius*	
15	Papilionidae	*Parides*	*photinus*	*photinus*
17	Papilionidae	*Parides*	*montezuma*	*montezuma*
19	Papilionidae	*Parides*	*eurymedes*	*mylotes*
21	Papilionidae	*Parides*	*sesostris*	*zestos*
22	Papilionidae	*Parides*	*panares*	*panares*
23	Papilionidae	*Parides*	*panares*	*lycimenes*
24	Papilionidae	*Parides*	*erithalion*	*polyzelus*
25	Papilionidae	*Parides*	*erithalion*	*trichopus*
27	Papilionidae	*Parides*	*iphidamas*	*iphidamas*
29	Papilionidae	*Protographium*	*epidaus*	*tepicus*
30	Papilionidae	*Protographium*	*epidaus*	*fenochionis*
31	Papilionidae	*Protographium*	*epidaus*	*epidaus*
32	Papilionidae	*Protographium*	*philolaus*	*philolaus*
33	Papilionidae	*Protographium*	*agesilaus*	*fortis*
34	Papilionidae	*Protographium*	*agesilaus*	*neosilaus*
35	Papilionidae	*Protographium*	*dioxippus*	*lacandones*
36	Papilionidae	*Protographium*	*calliste*	*calliste*
37	Papilionidae	*Protographium*	*thyastes*	*marchandi*
38	Papilionidae	*Protographium*	*thyastes*	*occidentalis*
39	Papilionidae	*Eurytides*	*salvini*	
40	Papilionidae	*Protesilaus*	*macrosilaus*	
42	Papilionidae	*Mimoides*	*thymbraeus*	*thymbraeus*
43	Papilionidae	*Mimoides*	*thymbraeus*	*aconophos*
44	Papilionidae	*Mimoides*	*ilus*	*branchus*
45	Papilionidae	*Mimoides*	*ilus*	*occiduus*
46	Papilionidae	*Mimoides*	*phaon*	*phaon*
47	Papilionidae	*Priamides*	*pharnaces*	
48	Papilionidae	*Priamides*	*rogeri*	
49	Papilionidae	*Priamides*	*erostratus*	*erostratinus*
50	Papilionidae	*Priamides*	*erostratus*	*vazquezae*
51	Papilionidae	*Priamides*	*erostratus*	*erostratus*
52	Papilionidae	*Priamides*	*anchisiades*	*idaeus*
53	Papilionidae	*Troilides*	*torquatus*	*mazai*
54	Papilionidae	*Troilides*	*torquatus*	*tolus*
55	Papilionidae	*Calaides*	*ornythion*	*ornythion*
57	Papilionidae	*Calaides*	*astyalus*	*bajaensis*

Appendix. Continued.

Id_Species	Family	Genus	Species	Subspecies
58	Papilionidae	*Calaides*	*astyalus*	*pallas*
60	Papilionidae	*Calaides*	*androgeus*	*epidaurus*
61	Papilionidae	*Heraclides*	*thoas*	*autocles*
62	Papilionidae	*Heraclides*	*cresphontes*	
63	Papilionidae	*Papilio*	*indra*	*pergamus*
65	Papilionidae	*Papilio*	*zelicaon*	*zelicaon*
66	Papilionidae	*Papilio*	*polyxenes*	*coloro*
67	Papilionidae	*Papilio*	*polyxenes*	*asterius*
68	Papilionidae	*Pterourus*	*esperanza*	
69	Papilionidae	*Pterourus*	*pilumnus*	
70	Papilionidae	*Pterourus*	*palamedes*	*leontis*
72	Papilionidae	*Pterourus*	*glaucus*	*alexiares*
73	Papilionidae	*Pterourus*	*glaucus*	*garcia*
74	Papilionidae	*Pterourus*	*rutulus*	*rutulus*
75	Papilionidae	*Pterourus*	*eurymedon*	
76	Papilionidae	*Pterourus*	*multicaudatus*	
78	Papilionidae	*Pyrrhosticta*	*garamas*	*garamas*
79	Papilionidae	*Pyrrhosticta*	*abderus*	*abderus*
80	Papilionidae	*Pyrrhosticta*	*abderus*	*baroni*
81	Papilionidae	*Pyrrhosticta*	*abderus*	*electryon*
83	Papilionidae	*Pyrrhosticta*	*victorinus*	*victorinus*
84	Papilionidae	*Pyrrhosticta*	*victorinus*	*morelius*
85	Pieridae	*Pseudopieris*	*nehemia*	*irma*
86	Pieridae	*Enantia*	*lina*	*marion*
88	Pieridae	*Enantia*	*albania*	*albania*
89	Pieridae	*Enantia*	*jethys*	
90	Pieridae	*Enantia*	*mazai*	*mazai*
91	Pieridae	*Enantia*	*mazai*	*diazi*
92	Pieridae	*Lieinix*	*lala*	*lala*
93	Pieridae	*Lieinix*	*lala*	*turrenti*
94	Pieridae	*Lieinix*	*neblina*	
95	Pieridae	*Lieinix*	*nemesis*	*atthis*
96	Pieridae	*Lieinix*	*nemesis*	*nayaritensis*
97	Pieridae	*Dismorphia*	*amphiona*	*lupita*
98	Pieridae	*Dismorphia*	*amphiona*	*isolda*
99	Pieridae	*Dismorphia*	*amphiona*	*praxinoe*
100	Pieridae	*Dismorphia*	*crisia*	*virgo*
101	Pieridae	*Dismorphia*	*crisia*	*alvarezi*
102	Pieridae	*Dismorphia*	*eunoe*	*eunoe*
103	Pieridae	*Dismorphia*	*eunoe*	*popoluca*
104	Pieridae	*Dismorphia*	*eunoe*	*chamula*
105	Pieridae	*Dismorphia*	*theucharila*	*fortunata*
106	Pieridae	*Colias*	*alexandra*	*harfordii*
107	Pieridae	*Colias*	*eurytheme*	
108	Pieridae	*Colias*	*philodice*	*philodice*
109	Pieridae	*Colias*	*philodice*	*guatemalena*
110	Pieridae	*Zerene*	*cesonia*	*cesonia*
111	Pieridae	*Zerene*	*eurydice*	
112	Pieridae	*Anteos*	*clorinde*	*nivifera*
113	Pieridae	*Anteos*	*maerula*	*lacordairei*

Appendix. Continued.

Id_Species	Family	Genus	Species	Subspecies
114	Pieridae	*Phoebis*	*agarithe*	*agarithe*
115	Pieridae	*Phoebis*	*agarithe*	*fisheri*
116	Pieridae	*Phoebis*	*argante*	*argante*
117	Pieridae	*Phoebis*	*neocypris*	*virgo*
118	Pieridae	*Phoebis*	*philea*	*philea*
119	Pieridae	*Phoebis*	*sennae*	*marcellina*
120	Pieridae	*Prestonia*	*clarki*	
121	Pieridae	*Rhabdodryas*	*trite*	*trite*
124	Pieridae	*Aphrissa*	*statira*	*jada*
125	Pieridae	*Abaeis*	*nicippe*	
126	Pieridae	*Pyrisitia*	*dina*	*westwoodi*
127	Pieridae	*Pyrisitia*	*lisa*	*centralis*
128	Pieridae	*Pyrisitia*	*nise*	*nelphe*
129	Pieridae	*Pyrisitia*	*proterpia*	*proterpia*
130	Pieridae	*Eurema*	*agave*	*millerorum*
131	Pieridae	*Eurema*	*albula*	*celata*
132	Pieridae	*Eurema*	*boisduvaliana*	
133	Pieridae	*Eurema*	*daira*	
136	Pieridae	*Eurema*	*mexicana*	*mexicana*
137	Pieridae	*Eurema*	*salome*	*jamapa*
138	Pieridae	*Eurema*	*xantochlora*	*xantochlora*
139	Pieridae	*Nathalis*	*iole*	*iole*
140	Pieridae	*Kricogonia*	*lyside*	
141	Pieridae	*Anthocharis*	*cethura*	*cethura*
142	Pieridae	*Anthocharis*	*cethura*	*pima*
143	Pieridae	*Anthocharis*	*sara*	*sara*
144	Pieridae	*Anthocharis*	*sara*	*inghami*
145	Pieridae	*Paramidea*	*lanceolata*	
147	Pieridae	*Paramidea*	*limonea*	
148	Pieridae	*Euchloe*	*guaymasensis*	
149	Pieridae	*Euchloe*	*hyantis*	*hyantis*
150	Pieridae	*Euchloe*	*hyantis*	*lotta*
152	Pieridae	*Hesperocharis*	*costaricensis*	*pasion*
153	Pieridae	*Hesperocharis*	*crocea*	*crocea*
154	Pieridae	*Hesperocharis*	*crocea*	*jaliscana*
155	Pieridae	*Hesperocharis*	*graphites*	*graphites*
156	Pieridae	*Hesperocharis*	*graphites*	*avivolans*
157	Pieridae	*Eucheira*	*socialis*	*socialis*
158	Pieridae	*Eucheira*	*socialis*	*westwoodi*
159	Pieridae	*Neophasia*	*terlooii*	
160	Pieridae	*Archonias*	*brassolis*	*aproximata*
161	Pieridae	*Charonias*	*theano*	*nigrescens*
162	Pieridae	*Catasticta*	*flisa*	*flisa*
164	Pieridae	*Catasticta*	*flisa*	*oaxaca*
165	Pieridae	*Catasticta*	*flisella*	
166	Pieridae	*Catasticta*	sp1.	
167	Pieridae	*Catasticta*	*nimbice*	*nimbice*
168	Pieridae	*Catasticta*	*ochracea*	*ochracea*
169	Pieridae	*Catasticta*	*ochracea*	ssp..
170	Pieridae	*Catasticta*	*teutila*	*teutila*

Appendix. Continued.

Id_Species	Family	Genus	Species	Subspecies
171	Pieridae	*Catasticta*	*teutila*	ssp1.
172	Pieridae	*Catasticta*	*teutila*	*flavifaciata*
173	Pieridae	*Catasticta*	*teutila*	ssp2.
175	Pieridae	*Pereute*	*charops*	*charops*
176	Pieridae	*Pereute*	*charops*	*leonilae*
177	Pieridae	*Pereute*	*charops*	*nigricans*
178	Pieridae	*Pereute*	*charops*	*sphocra*
180	Pieridae	*Melete*	*lycimnia*	*isandra*
181	Pieridae	*Melete*	*polyhymnia*	*florinda*
182	Pieridae	*Melete*	*polyhymnia*	*serrana*
183	Pieridae	*Glutophrissa*	*drusilla*	*tenuis*
184	Pieridae	*Pieris*	*rapae*	*rapae*
187	Pieridae	*Pontia*	*beckeri*	
188	Pieridae	*Pontia*	*protodice*	
189	Pieridae	*Pontia*	*sisymbrii*	*sisymbrii*
190	Pieridae	*Leptophobia*	*aripa*	*elodia*
191	Pieridae	*Itaballia*	*demophile*	*centralis*
192	Pieridae	*Itaballia*	*pandosia*	*kicaha*
193	Pieridae	*Pieriballia*	*viardi*	*viardi*
194	Pieridae	*Pieriballia*	*viardi*	*laogore*
195	Pieridae	*Perrhybris*	*pamela*	*chajulensis*
196	Pieridae	*Perrhybris*	*pamela*	*mapa*
197	Pieridae	*Ascia*	*monuste*	*monuste*
198	Pieridae	*Ascia*	*monuste*	*raza*
199	Pieridae	*Ganyra*	*howarthi*	*howarthi*
200	Pieridae	*Ganyra*	*howarthi*	*kuschei*
201	Pieridae	*Ganyra*	*josephina*	*josepha*
202	Pieridae	*Ganyra*	*phaloe*	*tiburtia*

References

Arita H, Figueroa F, Frisch A, Rodriguez P and Santos del Prado K (1998) Geographical range size and the conservation of Mexican mammals. Conservation Biology 11: 92–100

Austin M (1998) An ecological perspective on biodiversity investigations: examples from Australian Eucalypt forests. Annals of the Missouri Botanical Garden 85: 2–17

Bojorquez-Tapia L, Balvanera P and Cuaron AD (1994) Biological inventories and computer databases: their role in environmental assessments. Environmental Management 18: 545–551

Bojorquez-Tapia L, Azuara I, Ezcurra E and Flores O (1995) Identifying conservation priorities in Mexico through GIS and modelling. Ecological Applications 5: 215–231

Brown J (1995) Macroecology. The University of Chicago Press. Chicago

Butterfield BR, Csuti B and Scott J (1994) Modeling vertebrate distributions for GAP analysis. In: Miller R (ed) Mapping the Diversity of Nature, pp 53–68. Chapman & Hall, London

Carpenter G, Gillison AN and Winter J (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation 2: 667–680

Chapman A and Busby J (1994) Linking plant species information to continental biodiversity inventory, climate modeling and environmental monitoring. In: Miller R (ed) Mapping the Diversity of Nature, pp 179–194. Chapman & Hall, London

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

Colwell RK (1997) EstimateS: Statistical estimation of species richness and shared species from samples. Version 5. User's Guide and application published at http://viceroy.eeb.unconn.edu/estimates

Colwell RK and Coddington J (1994) Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B 345: 101–118

Dinerstein E, Olson D, Graham D, Webster A, Primm S, Bookbinder M and Ledec G (1995) A Conservation Assessment of the Terrestrial Ecoregions of Latin America and the Caribbean. The World Wildlife Fund and The World Bank. Washington, DC, 129 pp

Fayyad U, Piatetsky-Shapiro G and Smyth P (1996) The KDD process of extracting useful knowledge from volumes of data. Communications of the ACM 39(11): 27–34

Gaston K (1994) Rarity. Chapman & Hall, London, 205 pp

Gaston K and Mound LH (1993) Taxonomy, hypothesis testing and the biodiversity crisis. Proceedings of the Royal Society of London B 251: 139–142

Holt R and Gaines M (1992) Analysis of adaptation in heterogeneous landscapes: implications for the evolution of fundamental niches. Evolutionary Ecology 6: 433–447

Hutchinson GE (1987) An Introduction to Population Ecology. Yale University Press, New Haven, Connecticut

ICBP (International Council for Bird Preservation) (1992) Putting Biodiversity on the Map: Priority Areas for Global Conservation. International Council for Bird Preservation, Cambridge, UK

Imielinski T and Mannila H (1996) A database perspective on knowledge discovery. Communications of the ACM 39(11): 58–64

Jones PG, Beebe S, Tohme J and Galwey NW (1997) The use of geographical information systems in biodiversity exploration and conservation. Biodiversity and Conservation 6: 947–958

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

Leon-Cortes JL, Soberon J and Llorente J (1998) Assessing completeness of Mexican sphinx moth inventories through species accumulation functions. Diversity and Distributions 4: 37–44

Llorente J and Luis A (1993) Conservation-oriented analysis of Mexican butterflies: Papilionidae (Lepidoptera, Papilionoidea). In: Ramamoorthy T, Bye R, Lot A and Fa J. Biological Diversity of Mexico. Origins and Distribution. Oxford University Press, Oxford, 812 pp

Llorente J, Luna IV, Soberon J and Bojorquez L (1994) Biodiversidad, su inventario y teoría y práctica en conservación: la taxonomía alfa contemporánea. In: Llorente J (ed) Taxonomía Biológica, pp 507–520. Fondo de Cultura Económica. México

Llorente J, Onate L, Luis A and Vargas I (1997) Papilionidae y Pieridae de Mexico: Distribucion Geografica e Ilustracion. UNAM, Mexico, 227 pp

Margules CR and Austin MP (1995) Biological models for monitoring species decline: the construction and use of databases. In: Lawton J and May RM (eds) Extinction Rates, pp 183–196. Oxford University Press, Oxford

May RM (1973) Patterns of species abundance and diversity. In: Cody ML and Diamond J (eds) Ecology and Evolution of Communities, pp 81–120. Belknap, Cambridge, Massachussetts

Miller R (1994) Mapping the Diversity of Nature. Chapman & Hall, London

Mourelle C and Ezcurra E (1996) Species richness of Argentine cacti: a test of biogeographic hypothesis. Journal of Vegetation Science 7: 667–680

Myers N (1988) Threatened biotas: 'hot spots' in tropical forests. The Environmentalist 8(3): 187–208

Nelson B, Ferreira C, da Silva M and Kawasaki M (1990) Endemism centres, refugia and botanical collection density in Brazilian Amazonia. Nature 345: 714–716

Pankhurst R (1991) Practical Taxonomy Computing. Cambridge University Press, Cambridge

Peterson T, Navarro A and Benitez H (1998) The need for continued scientific collecting; a geographic analysis of Mexican bird specimens. IBIS 140: 288–294

Peterson T, Soberón J and Sanchez-Cordero V (1999) Conservatism of ecological niches in evolutionary time. Science 285: 1265–1267

Prendergast S, Wood N, Lawton J and Eversham BC (1993a) Correcting for variation in recording effort in analyses of diversity hotspots. Biodiversity Letters 1: 39–53

Prendergast JR, Quinn R, Lawton J, Eversham B and Gibbons D (1993b) Rare species, the coincidence of diversity hotsposts and conservation strategies. Nature 365: 335–337

Preston FW (1948) The commonness, and rarity of species. Ecology 29: 254–283

Preston FW (1962) The canonical distribution of commonness and rarity. Ecology 43: 185–215

Rapoport E (1982) Areography: Geographical Strategies of Species. Pergamon, Oxford

Roman S (1997) Access Database. Design and Programming. O'Reilly & Associates, Cambridge, 251 pp

Rzedowsky J (1978) Vegetacion de Mexico. Editorial Limusa, Mexico, 432 pp

Schluter D and Ricklefs RE (1993) Species diversity. An introduction to the problem. In: Ricklefs RE and Schluter D (eds) Species Diversity in Ecological Communities, pp 1–12. The University of Chicago Press, Chicago

Scott M, Tear T and Davies F (1996) Gap Analysis. A Landscape Approach to Biodiversity Planning. The American Society for Photogrammetry and Remote Sensing, Maryland, 320 pp

Scott M and Jennings MD (1998) Large-area mapping of biodiversity. Annals of the Missouri Botanical Garden 85: 34–47

Soberon J and Llorente J (1993) The use of species accumulation functions for the prediction of species richness. Conservation Biology 7: 480–488

Soberon J, Llorente J and Benítez H (1996) An International view of National Biological Surveys. Annals of the Missouri Botanical Gardens 83: 562–573

Soto M and Gomez-Pompa A (1990) Bioclimatologia de la Flora de Veracruz, No. 1 Instituto de Ecologia A.C., Xalapa, Mexico

Stockwell D and Noble I (1992) Induction of sets of rules from animal distribution data: a robust and informative method for data analysis. Mathematics and Computers in Simulation 33: 385–390

Tyler H, Brown KS and Wilson K (1994) Swallowtail Butterflies of the Americas. A Study in Biological Dynamics, Ecological Diversity, Biosystematics and Conservation. Gainesville Scientific Publications

Umminger B and Young S (1997) Information Management for Biodiversity: a proposed US National Biodiversity Information Center. In: Reaka-Kudla M, Wilson DE and Wilson EO (eds) Biodiversity II. Understanding and Protecting our Biological Resources, pp 491–504. Joseph Henry Press, Washington, DC

Weiss S and Murphy D (1993) Climatic considerations in reserve design and ecological restoration. In: Saunders D, Hobbs R and Ehrlich P (eds) Nature Conservation 3: Reconstruction of Fragmented Ecosystems, pp 89–107. Beatty & Sons, London

Whittaker RH (1972) Evolution and measurement of species diversity. Taxon 21: 213–251

Wohlgemuth T (1998) Modelling floristic species richness on a regional scale: a case study in Switzerland. Biodiversity and Conservation 7: 159–177