

Estimating regional species richness of tropical insects from museum data: a comparison of a geography-based and sample-based methods

JAN BECK^{*‡} and IAN J. KITCHING[†]

^{*}Kuala Belalong Field Studies Centre, Universiti Brunei Darussalam, Tungku Link, Gadong BE1410, Brunei; and

[†]Department of Entomology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

Summary

1. The identification of spatial patterns of species richness at regional scales, such as biodiversity hotspots, is complicated in invertebrate taxa (particularly those from tropical regions) by incomplete and biased inventory data. Estimation techniques of regional species richness from incompletely sampled landscapes have recently become available, but their applicability to data from museum collections and local taxonomic checklists has not been investigated.

2. We use records of sphingid moths in grid cells of 1° latitude and longitude on 14 Malesian islands to estimate the total species richness of these islands, as well as that of the whole archipelago, by five parametric and nonparametric estimation techniques. We compare these values to figures based on GIS-supported estimates of geographical ranges of each species.

3. Our analyses suggest that the F3 estimator of regional species richness leads to least deviation from the GIS-based estimate, followed by Chao2 and ICE, which are often employed in local comparisons. Overestimation occurred more often than underestimation and estimates of well-sampled islands (with many available grid cells) are less deviant than those of poorly sampled islands. We did not obtain conclusive results as to whether strongly undersampled grid cells are better excluded from an analysis (at the cost of reduced grid cell number) or not.

4. *Synthesis and applications.* In agreement with some previously published assessments, we conclude that the F3 estimator has the greatest potential for predicting regional species richness in partially sampled landscapes. Sample-based methods for estimating regional species richness can provide an alternative to the much more work-intensive geographical modelling of species distributions, which may facilitate the inclusion of tropical invertebrate groups in documenting global diversity patterns. However, under conditions of incomplete, non-systematic sampling, which is typical for museum and checklist data, errors can still be large, particularly if the number of sampling units (e.g. grid cells) is low. Estimation values should not be interpreted uncritically when the data conditions that lead to biased values are not precisely defined.

Key-words: biodiversity, Lepidoptera, Malesia, Sphingidae, South-east Asia, WS2M.

Journal of Applied Ecology (2007) **44**, 672–681

doi: 10.1111/j.1365-2664.2007.01291.x

Introduction

Describing and understanding species diversity patterns is a major research challenge in biodiversity, biogeo-

graphy and conservation (e.g. Rosenzweig 1995, 2003) that might generate the much-needed knowledge necessary for ecosystem-wide ‘diversity-management’ in the wake of the current anthropogenic extinction wave (e.g. Sodhi *et al.* 2004 for South-east Asia). The shortage of reliable and complete sampling data has limited most larger-scale analyses to temperate regions or to well-sampled taxa such as plants or vertebrates (e.g. Barthlott *et al.* 1996; Kier & Barthlott 2001; Jetz &

Rahbek 2002; Hawkins *et al.* 2003; Gaston & Evans 2004), whereas the incompleteness of invertebrate sampling, particularly in hyper-diverse tropical rainforest regions, has mostly prevented such efforts in other taxonomic groups.

The analysis of local species diversity on the basis of incomplete samples has led to well-established methods such as diversity indices, rarefaction and local species richness estimators (see, e.g. Southwood & Henderson 2000; pp. 462–506; Ulrich & Ollik 2005 review recent developments). Methods to estimate regional species richness on the basis of incomplete local samples and incomplete coverage of local habitat differences (β -diversity) have only recently begun to be applied (Mawdsley 1996). The software WS2M (Turner *et al.* 2000) can be used to compute a number of parametric and nonparametric estimators, some of which are also used for local estimates (e.g. incidence-based estimator (ICE), asymptotic Michaelis-Menten (MM) functions; Chazdon *et al.* 1998). It also includes new methods specifically designed to fit the shape of species accumulation curves, and thus the spatial species turnover between sites (see Turner *et al.* 2000; Rosenzweig *et al.* 2003; for details; we use their acronyms for methods throughout this paper). Rosenzweig *et al.* (2003) showed that some methods (MM, F3, F5) performed very well in estimating continental species richness for North American butterflies on the basis of a relatively small number of included ecoregions, providing these were not spatially clustered. Other studies (e.g. Oertli *et al.* 2002; Dangles *et al.* 2004) have also applied estimation methods from WS2M, though mostly on relatively small spatial scales. It is unknown, however, to what degree these promising results might be dependent on local completeness of samples, which is unusual for most invertebrate samples. Most recently, Hortal *et al.* (2006) reported rather poor performance of these techniques using arthropod data from the Azorean islands.

Museum collections and taxonomic publications with local species checklists contain a huge and largely untapped amount of information on the large-scale distribution of biodiversity (Graham *et al.* 2004). Data that has accumulated over decades or centuries are far more comprehensive than those that can be collected in modern field studies. These data might, if properly analysed, allow comparisons of the spatial distribution of biodiversity (e.g. 'hotspots'; Reid 1998; Myers *et al.* 2000) of less investigated taxa with the well-established patterns found in plants or vertebrates, similar to that attempted at local scales in the context of biodiversity indication (Lawton *et al.* 1998; Moritz *et al.* 2001; Schulze *et al.* 2004; Hilt & Fiedler 2006). Such knowledge would considerably substantiate management and conservation decisions, as well as allowing investigations of the coincidence, or divergence, of patterns between different groups. 'Specimen label' data, as provided by museums and checklists, however, have a number of biases (as discussed in Graham *et al.* 2004) with as yet unexplored effects on species richness estimates.

Here we analyse the species richness of hawkmoths (Lepidoptera: Sphingidae) in Malesia, a highly fragmented archipelago with distinct biogeographical sub-regions (Whitmore 1987; Hall & Holloway 1998; Beck *et al.* 2006a). Although sphingid moths are among the best-sampled tropical insect taxa, our data still feature all the problems typical for collection-based, presence-only data, such as incomplete coverage of local samples, and habitat heterogeneity, and a clustering of sampling effort in logistically favourable spots – many of which violate the inherent assumptions of most estimation techniques (e.g. random sampling). We compare sample-based estimator techniques with figures based on geographical distribution estimates for each species on two spatial scales, estimating the sphingid diversity of several large islands from records in grid cells, as well as estimating the diversity across the whole archipelago from the records on those islands. Furthermore, we assess how results depend on the utilization of more, but less thoroughly sampled, locations, as the number of sites vs. sampling intensity per site is a common trade-off in biodiversity surveys (e.g. Novotny & Missa 2000).

Methods

We compiled approximately 10 000 sphingid records from Malesia (i.e. Malaysia, Indonesia, East Timor, Brunei, Philippines and Papua New Guinea excluding Bougainville), where 302 sphingid species (four as yet unnamed) are currently known. Data originate from museum collections, publications and unpublished lists (see Beck & Kitching 2004 for details; note that one 'record' refers to a unique combination of species, place, altitude, year and source, but may involve from one to several hundreds of specimens). Sample sites were allocated with a precision of at least one degree latitude and longitude. More precise localization of records that can be up to 185 years old was not feasible in most cases. A small number of records (*c.* 0.2%) were excluded from all analyses due to doubts about their credibility (possible misidentifications or mislabelling, see Beck & Kitching 2004). Fewer than 5% of records had no precise location information available. These records were set to the most likely grid cell (given logistic conditions at the time of sampling) and included in analyses. Analyses of 1°-grid cells (see below) are slightly biased as their actual surface area changes with latitude as well as topography, but with the 'inner-tropical' extent of our analyses (11°S to 20°N) this error can be neglected.

GEOGRAPHY-BASED ESTIMATES

Records were entered into a Geographic Information System (GIS; ArcView 3.2, 2000). We estimated the geographical ranges of each species as described elsewhere (Beck & Kitching 2004). In short, we applied a mixture of a niche-based approach, underlying records with habitat maps (such as altitudinal relief, vegetation zones, precipitation or minimum winter temperature,

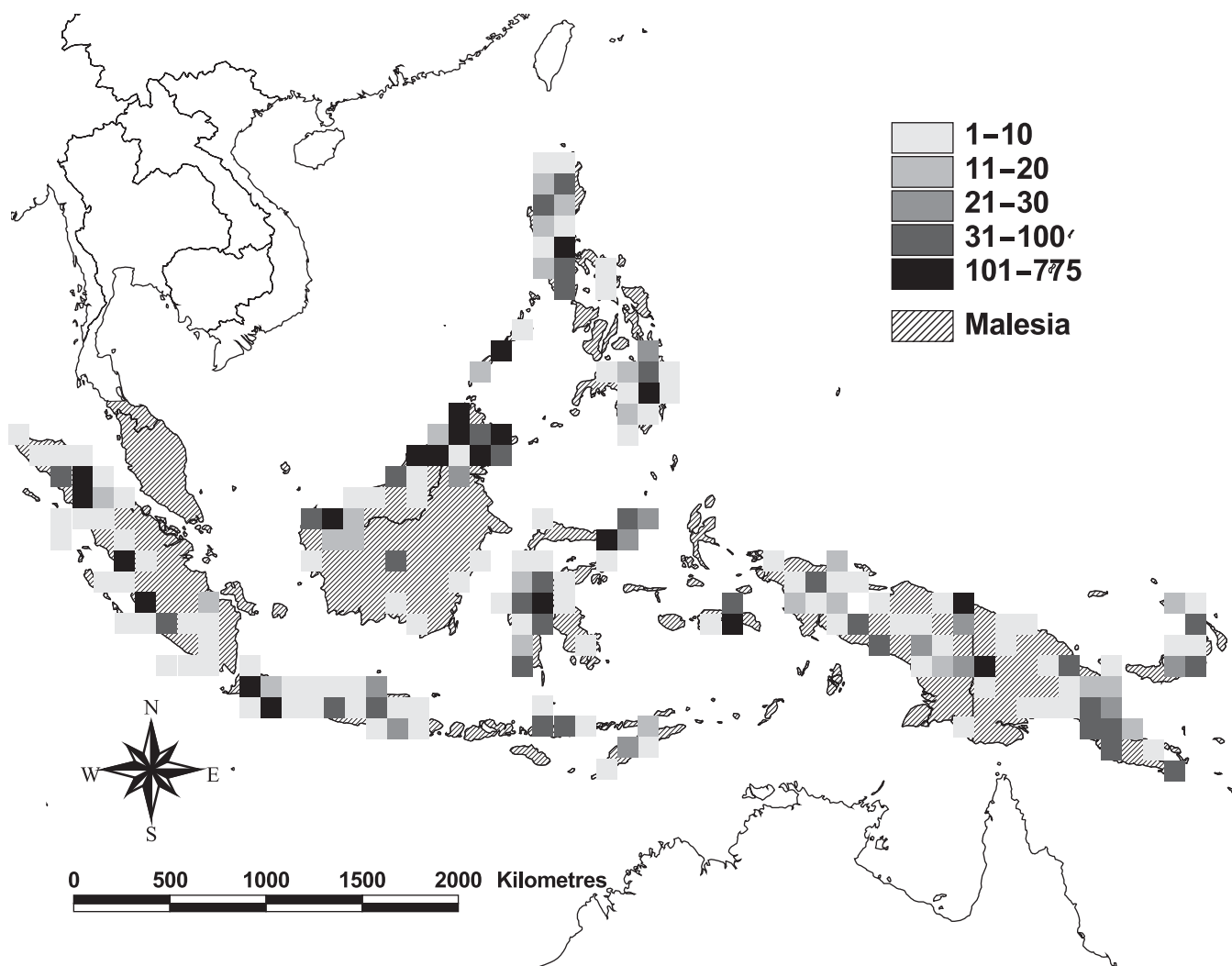


Fig. 1. Location and sampling intensity (number of records) of grid cells in analysis. Grid cells have a size of 1-degree latitude and longitude, which translates to an area of approximately 12 000 km² near the equator. Sampling intensity is probably determined mostly by logistic and administrative accessibility.

cf. Cowley *et al.* 2000) to find potentially suitable habitats for each species, and a consideration of historical and species-specific constraints. The limits of many species within Malesia are determined to a large extent by historical dispersal limits rather than by habitat alone, so we did not extend species ranges beyond actual records across such known borders. Uneven sampling efforts in different regions (see also Fagan & Kareiva 1997; Soberón *et al.* 2000; Graham *et al.* 2004) and the likelihood of finding, correctly identifying or reporting different species were carefully assessed on a species-by-species basis. Taking all these factors into consideration, the best estimate of each species' range was digitized. We did not use an explicit computer model to estimate ranges, as the analysis of presence-only data is still problematic for statistical habitat models (e.g. Zaniwski *et al.* 2002), particularly for rare species (Engler *et al.* 2004), despite successful application on smaller geographical scales (e.g. Iverson & Prasad 1998; Ray *et al.* 2002; Raxworthy *et al.* 2003). Overlaying species' ranges produced estimated checklists for islands

of the Malesian region. Original records, range maps and island checklists are presented by Beck & Kitching (2004).

SAMPLE-BASED ESTIMATES

We used the software WS2M (Turner *et al.* 2000) to estimate species richness of the larger islands of Malesia (we avoided small islands because of our relatively large grid cells) on the basis of recorded presences in 1°-grid cells (Fig. 1). Furthermore, we estimated the species richness of Malesia as a whole from checklists of these islands as well as from the grid cells within them. For some well-sampled islands, as well as Malesia as a whole, we repeated analyses using only grid cells with at least 10, 20, 30 or 100 records, respectively, to exclude data from severely undersampled grid cells. Sampling units were randomized 1000 times to remove entry order effects. We focused our investigation on three methods that are commonly used in estimations of local habitat diversity (Chao2, ICE, MM), but which

Table 1. Number of 1°-grid cells, records, recorded species (S_{obs}) and various estimates of species richness for 14 islands. Scores for the estimator F6, which was recommended against in Rosenzweig *et al.* (2003), are also presented, but were not considered in analysis. Estimator scores (except GIS) were calculated with WS2M, applying a 1000-fold randomization of sampling order, otherwise using recommended settings (see Appendix S1, Supplementary material, for details)

Island	Cells	Records	S_{obs}	GIS	Chao2	ICE	MM	F3	F5	F6
New Guinea	47	1032	78	78	81.6	82.0	86.3	81.1	81.3	84.5
Sumatra	30	843	127	127	152.6	144.3	181.8	157.0	165.3	153.6
Borneo	27	3013	114	114	127.8	127.0	126.8	121.2	122.0	132.2
Java	20	676	108	108	133.9	130.1	156.7	133.6	142.2	140.3
Sulawesi	20	673	84	84	94.6	93.6	118.1	99.9	104.3	98.8
Luzon	13	480	86	86	103.4	102.3	135.1	108.7	117.5	110.7
Mindanao	11	259	78	82	80.6	84.4	131.4	99.0	106.1	92.4
New Britain	5	77	34	42	39.0	43.2	60.7	40.3	45.4	44.4
New Ireland	5	72	36	39	47.0	53.4	59.9	43.8	49.2	57.3
Timor	4	49	30	40	60.3	74.0	110.5	46.2	68.4	72.6
Palawan	3	250	84	89	300.0	426.0	195.6	424.1	421.9	333.2
Seram	3	186	39	53	55.5	69.5	92.8	55.1	64.9	64.7
Flores	3	135	41	48	49.6	59.0	74.3	53.3	57.5	48.3
Mindoro	3	127	65	84	69.6	77.5	102.0	80.1	81.9	

are also applied occasionally to estimations at regional scales (e.g. Folgarait *et al.* 2005), and two methods (F3, F5) that were specifically designed and recommended for estimating regional diversity in unsampled habitats (Rosenzweig *et al.* 2003). These methods represent variations of a family of asymptotic species-accumulation curves that are all forced through the point (1,1) (one species at a sample size of one individual), but vary in their curvature (see Rosenzweig *et al.* 2003 for details; see Appendix S1, Supplementary material, for formulae of all estimators and details on program settings).

The software EstimateS (version 7.5, Colwell 2005) also allows calculation of Chao2, ICE and MM, using an analytically derived species-accumulation curve rather than randomized data ('Mao Tau' function; Colwell *et al.* 2004; Mao *et al.* 2005). As expected, scores from the two programs are identical for ICE and very similar for Chao2 and MM (median divergence < 3 species). Large deviations were only observed on an island that generally produced unreasonable estimates (Palawan, see below). For consistency we choose to present only WS2M-scores. Many more nonparametric or curve-fitting estimation techniques exist (e.g. Colwell & Coddington 1994; Colwell *et al.* 2004; Ulrich & Ollik 2005; Jiménez-Valdere *et al.* 2006), and some have been used in similar comparisons (e.g. Chazdon *et al.* 1998; Hortal *et al.* 2006). We choose to concentrate here on a handful of easily applied measures that appeared promising for applications at large geographical scales.

Results

We used 7867 records from 194 grid cells on 14 islands to produce sample-based richness estimates. Grid cells and islands differ substantially in sampling effort, i.e. in the number of records (Fig. 1), which correlates with the number of recorded species ($N = 194$, Spearman's $R = 0.98$, $P < 0.0001$).

Table 1 gives results of the various estimation techniques for the islands. Most estimators are considerably

higher than recorded or GIS-estimated species richness (median across all estimators and islands: 17.4, respectively, 16.1 species). Ranks of estimation methods correlate ($N = 14$, $P < 0.001$, Spearman's $R > 0.90$ for all pairs except GIS-MM, $R = 0.84$). Linear regressions of GIS-estimates with other methods are significant only for F3 (Pearson's $r^2 = 0.69$, $P < 0.0001$), Chao2 ($r^2 = 0.36$, $P = 0.025$) and MM ($r^2 = 0.29$, $P = 0.048$).

For further comparisons, we calculated residuals from the GIS-estimate for the sample-based estimates and normalized their absolute values by the expected number of species according to the GIS-estimate. Negative residuals are relatively small, whereas some very large positive deviations occur (Fig. 2). Estimates are closer to GIS-derived measures on islands with a large number of grid cells or records, but rank correlations with residuals are not significant except for grid cells and MM-scores ($N = 14$, Spearman's $R = -0.62$, $P < 0.02$). Surprisingly, normalizing grid cell count and records for island size (i.e. grid cells or records per km²) did not yield clear relations with estimate residuals (not shown) and produced non-significantly positive rank correlations.

Medians of absolute residuals (in per cent of GIS species richness) lead to a rank order of deviation from the GIS-estimate (Table 2), and island-specific differences in estimator performance are significant (Friedman-ANOVA: $N = 14$, $\chi^2_{df=4} = 26.11$, $P < 0.0001$). F3 is judged closest to GIS-scores, with Chao2 and ICE closely following. In contrast, F5, and particularly MM, deviate considerably from GIS estimates. A consideration of mean deviations of the non-normal distributed data (not shown) confirms this rank order, but differences in estimator performance become evident even within the 'best' three methods, as gross overestimations occurred least commonly in F3. However, even after excluding data for Palawan (which did not produce reasonable results with any sample-based technique, Table 1) deviations from the GIS-estimate ranged from 11 (F3) to > 30 (MM) species, or 13 to > 50% (means across islands).

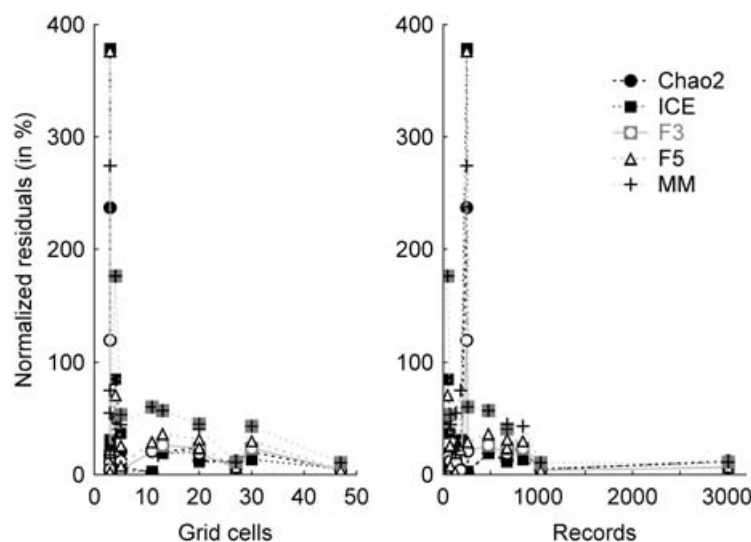


Fig. 2. Deviations of sample-based estimators from GIS-estimate on islands of different sampling intensity. Sampling intensity (x-axis) is measured as number of grid cells (left graph) and numbers of records (right graph). Estimator performance (y-axis, log-scale) is measured as the residual from the GIS estimate, normalized by the GIS-estimated species richness (as percentages).

Table 2. Species richness estimators, sorted by median residuals from GIS-estimates of 14 islands. Absolute residuals were normalized by each island's species richness (according to GIS-estimate). Medians were also calculated separately for nine well-sampled ($S_{\text{obs}}/\text{GIS} > 0.9$) and five poorly sampled ($S_{\text{obs}}/\text{GIS} = 0.7\text{--}0.9$) islands

Estimator	Median (%) [all islands]	Median (%) [well-sampled]	Median (%) [poorly sampled]
F3	13.9	20.8	4.7
Chao2	14.9	20.2	7.1
ICE	16.3	13.6	22.9
F5	25.1	29.4	19.8
MM	49.3	45.1	54.9

On the basis of normalized residuals of the estimators, the species richness of New Guinea has been assessed relatively well (a counter-intuitive result, as this island contains huge stretches of virtually unexplored forests), whereas the data from Palawan led to particularly bad estimates. Small numbers of grid cells lead to erratic estimator values, which are least extreme in F3 (see Fig. 3 for some examples). The general assessment of estimator performance did not change if only the well sampled or only the poorly sampled islands were considered, although scores for some estimators varied (Table 2).

Estimates of the species richness of the whole Malaysian archipelago, based on checklists for 14 large islands as well as the grid cells within them, are compared with the total number of currently known species from the region, including islands and biogeographical subregions that are not part of our data (Fig. 4). Among those based on grid cells, F3 and F5 produce stable estimates that are very close to the number of known species after about 100 grid cells, whereas the other estimators remained below that value. However, estimators based on island checklists are still rising after inclusion of all 14 islands and lead to estimates considerably higher than the number of known species.

The exclusion of severely undersampled sampling units (grid cells or islands) does not lead to unequivocal changes in estimator scores (Fig. 5). There are no consistent patterns that hold across estimator type or island. Even on Borneo, where, after exclusion of all grid cells with fewer than 30 records the data still contain all recorded species, estimators do not follow a common trend toward higher or lower scores.

Discussion

Our analyses indicate that of the sample-based methods of estimating regional species diversity, F3 performed best overall, thereby corroborating the assessment of Rosenzweig *et al.* (2003). The estimates still have considerable deviations of up to more than 25% from supposedly true values (see below for discussion), even if scores for Palawan with its extremely poor estimations are excluded (Table 1). The Michaelis-Menten estimator performed worst in our comparison, never reaching values within 10% of the GIS assessment.

Estimation techniques that were designed for application to local samples did not perform any worse than those specifically formulated for assessments in a regional context, contrary to results of Rosenzweig *et al.* (2003). Furthermore, we did not find a general distinction in estimation quality between curve-fitting techniques (F3, F5, MM) and those that do not depend directly on the shape of the species accumulation curve (Chao2, ICE). On theoretical grounds, nonparametric estimators might be expected to be more robust to different relative abundance distributions (RADs) in samples (which would be advantageous, all else being equal), but a general sensitivity to RADs was nevertheless evident for all estimators in simulation studies (Brose *et al.* 2003).

However, a large number of randomizations were necessary to obtain smooth accumulation curves close to the analytically derived curves computed using EstimateS (not shown; Colwell 2005; an application of the 'Mao Tau' function would probably benefit all estimation techniques, but software is not yet available). Rosenzweig *et al.* (2003) presented the estimator F6 as theoretically suitable, but obtained poor results in their empirical test on North American butterflies (see below), and hence did not recommend it. However, we did not find any indication of an outstandingly bad performance except for its failure in one case only to compute a meaningful value (see Table 1 for scores, analysis not shown).

Although overestimation was common among all estimation techniques applied to the 14 islands in this analysis, the application of local estimating techniques such as Chao2 or MM to a substantial number of quantitatively sampled sites on Borneo led to an underestimation of the island's totally known species inventory (Beck *et al.* 2006b).

Brose *et al.* (2003) and Brose & Martinez (2004) have shown that the accuracy of various local estimators is dependent on the sample coverage. Indeed, F3 and Chao2 scored much more closely to the GIS estimate

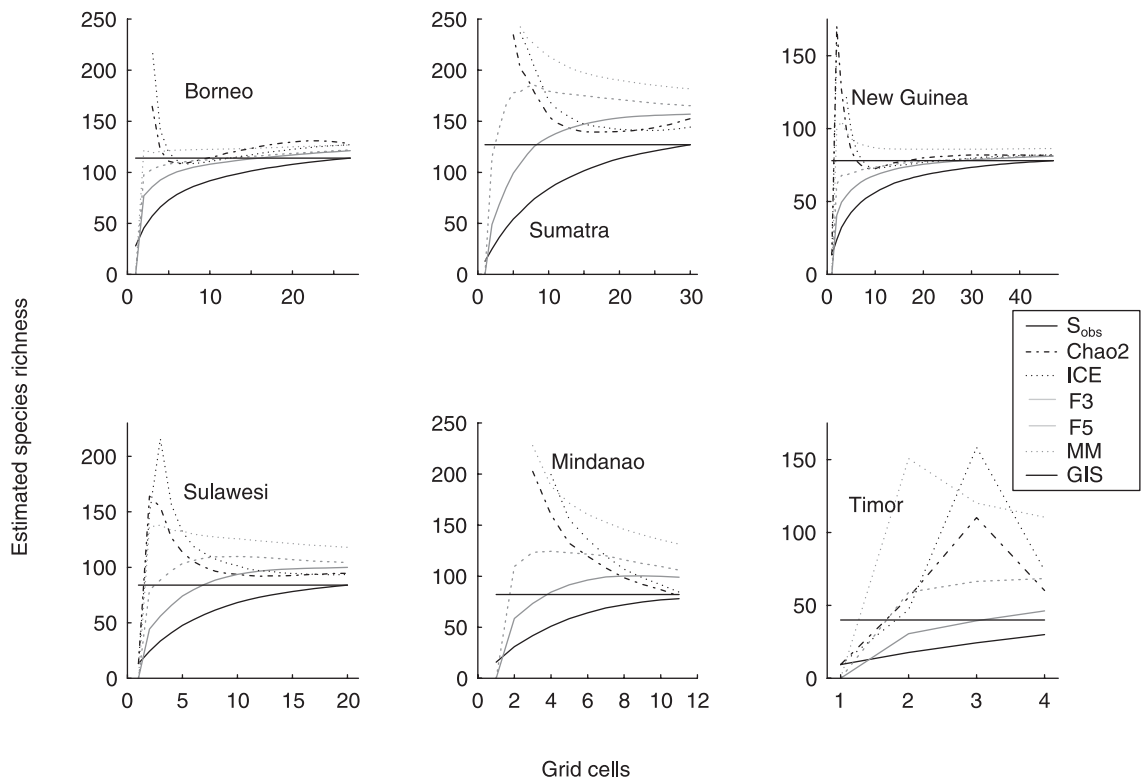


Fig. 3. Species accumulation and various estimators as a function of the number of included grid cells on exemplarily chosen islands. Values are averages from randomizations of grid cell orders (see main text for details) to make data entry irrelevant. The graphs show islands with different degree of sampling effort and homogeneity in estimator values. Note that y-axis values were cut at a maximum of 250 species.

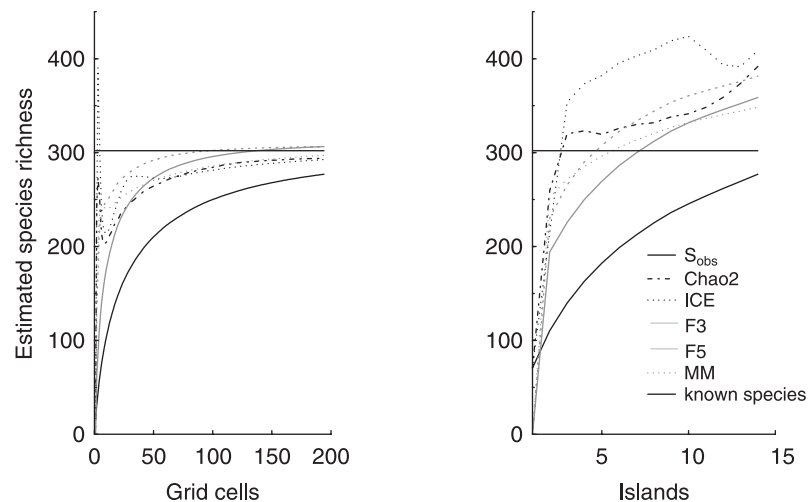


Fig. 4. Species accumulation and various estimators for the species richness of Malesia, based on grid cells from 14 major islands (left graph) or 14 island-wide checklists (right graph). 'Known species' are the total number of recorded species from the Malay Archipelago, including islands other than the 14 used for estimator calculation.

on five poorly sampled islands (Table 2), whereas other techniques performed better on well-sampled islands. However, the number of islands might be too small for the meaningful documentation of a pattern.

THE RELIABILITY OF GEOGRAPHY-BASED ESTIMATES

Rosenzweig *et al.* (2003) tested butterfly data from an artificially rarefied number of North American ecore-

gions, which they reasonably assumed to have been completely sampled, against the total known species richness of the continent. Thus, although perfect to test the precision of estimation methods, their data are unrealistically good for the proposed purpose of producing estimates from incompletely sampled habitats. Furthermore, North American species richness is probably mostly determined by climate-mediated habitat differences (Kerr *et al.* 2001). In contrast, our data are strongly affected by historical biogeography (Beck

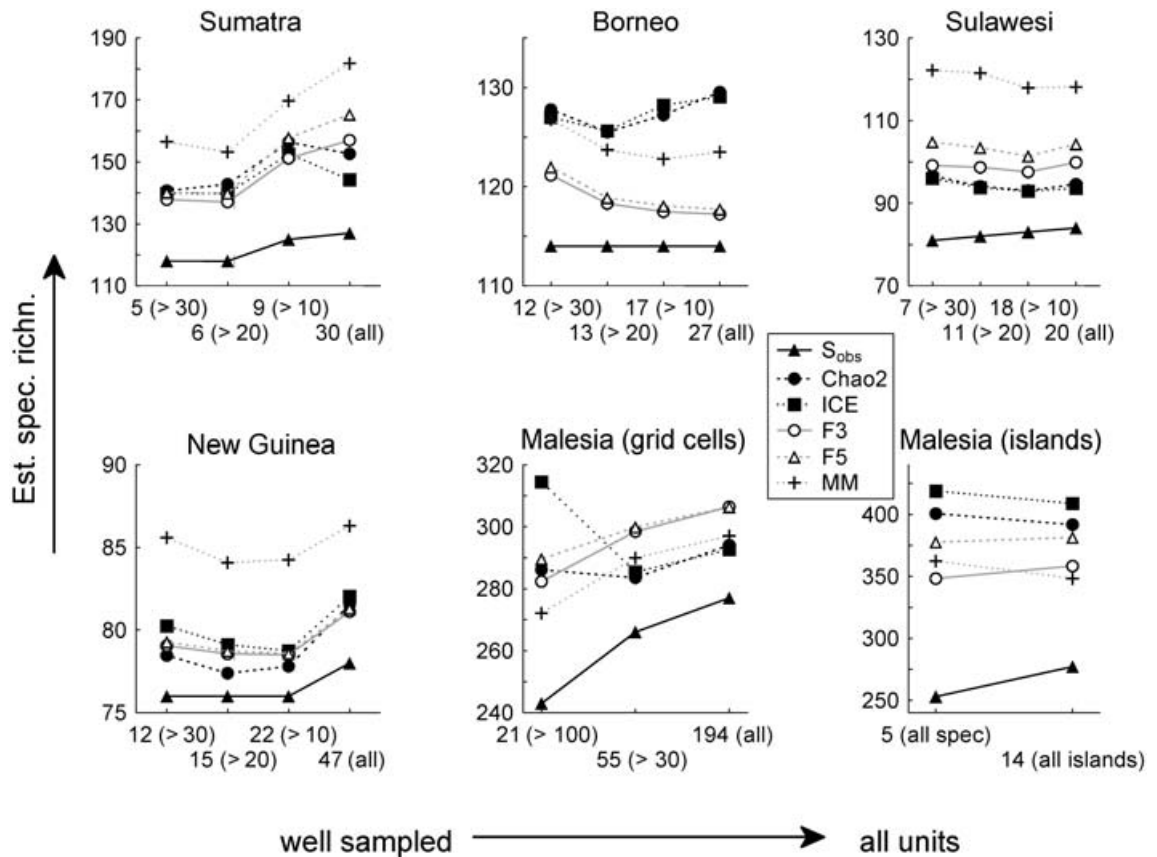


Fig. 5. Observed species richness and various estimators with or without the inclusion of poorly sampled units (grid cells or islands) for four well-sampled islands as well as for Malesia (based on grid cells and islands). Captions on the x-axis indicate the number of sampling units and the inclusion criterion (in brackets). For example, '11 (> 20)' means 11 grid cells with more than 20 records (each) were used for estimator computation; in the lower right graph (islands as sampling units) the five islands with completely recorded species inventory (according to GIS-estimate) were used.

et al. 2006a) and contain all the biases that may be found in distributional data of tropical invertebrates in a large spatial context: we cannot test estimates against values of known species richness. Rather, we have compared results from sample-based methods against our far more time-intensive GIS estimates, which we consider to be more reliable. Hence, the value of our analyses depends on the correctness of that assessment, which cannot be empirically confirmed or refuted, as large-scale, systematic sampling across broad areas of Malesia is unlikely to be undertaken in the foreseeable future. However, it appears very unlikely to us that a truly large number of unexpected taxa are present on the islands used in this analysis, although new records will certainly prove some GIS estimates to be incorrect by a few species.

With a large number of sampling units, sample-based methods estimated the number of species at or below that currently known across Malesia, whereas they produced considerably higher estimates for the same data using a spatially more coarse resolution (Fig. 4). This shows the dependency of estimator scores on the number of sampling units, which we also noticed in our comparison between islands (Fig. 3).

As an independent measure of estimator quality we correlated log-transformed island sizes with species

richness estimates, expecting a positive relation (the 'species-area relationship', e.g. Rosenzweig 1995) despite additional historical and environmental impacts on species richness (Beck *et al.* 2006a). A significant correlation was found only for the GIS-estimate ($N = 14$, Pearson's $r^2 = 0.44$, $P < 0.01$), and not for any of the sample-based methods (all $r^2 < 0.01$ except $F3$, $r^2 = 0.12$), although after exclusion of the Palawan data all methods produced significant ($P < 0.05$), positive species-area curves, with $r^2 > 0.50$ for Chao2, GIS, $F3$ and ICE. This makes it seem likely that the high estimates for islands with a low number of grid cells, most particularly on Palawan, are indeed errors, rather than the result of the GIS method greatly underestimating the islands' species richness. Furthermore, GIS estimates led to easily interpretable patterns of regional species richness (Beck *et al.* 2006a), and the frequency distribution of species' range sizes follows the same pattern found in many other taxa (Beck *et al.* 2006c).

APPLICATION OF ESTIMATORS

Estimators often performed very well on well-sampled islands, but the most serious problems arose in just those situations where estimates would be most needed, on islands where observed species richness is certainly

incomplete. Erroneous estimates occur in many estimators of local or regional species richness at small sample sizes (Ulrich & Olrik 2005 and references therein). Underestimations are commonly described in the literature, whereas here we report mostly overestimation errors. This point warns particularly against the uncritical application of species richness estimators from specimen-label data to heavily under-sampled locations or taxa, especially considering that every island in our present analysis had at least 70% of their GIS-expected species inventory recorded. Only a better understanding of what data biases lead to erroneous estimates may allow *a priori* excluding such samples in the future. Similar problems are known from local species richness estimates, where estimates based on a small number of sampling units may be quite unreliable before reaching more or less asymptotic values at higher sample sizes (see, e.g. Brehm *et al.* 2005 for data). From our analyses we tentatively conclude that a large number of sampling units (grid cells) is a good guard against such effects. Many fine-scaled sampling units (grid cells) produced more stable, and probably also more reliable, estimates than few large-scaled samples (island checklists), but may lead to underestimations if biogeographical regions (or other sources of β -diversity) are excluded from sampling.

Unfortunately, our analyses allow no general conclusions to be drawn regarding the effects of very poorly sampled grid cells on species richness estimates. We tentatively suggest that if data are available, sample-based estimates may benefit from using many small rather than few large grid cells. When the necessary data are available, estimates should be checked for a realistic range of values by comparing them to independent (i.e. not sample-based) methods, such as rank–abundance distributions (Ulrich & Olrik 2005), species–area curves (Krishnamani *et al.* 2004), GIS-supported habitat models (e.g. Iversen & Prasad 1998; Wiley *et al.* 2003) or species description rates (e.g. Gaston *et al.* 1995; Scoble *et al.* 1995; Solow & Smith 2005).

SYNTHESIS AND APPLICATION

The use of sample-based methods for estimating regional species richness can provide an alternative to the much more work-intensive geographical modelling of species distribution, if the aim is a comparison of species richness irrespective of species identity. This may facilitate the inclusion of tropical invertebrate groups in documenting global diversity patterns. For many taxa, the necessary distribution records may already be available in museums (e.g. Graham *et al.* 2004) and (often locally) published checklists (see, e.g. literature in Beck & Kitching 2004), but these data have to be processed to correct for their incompleteness. We concur with previously published assessments that the F3 estimator (Rosenzweig *et al.* 2003) has the greatest potential of predicting regional species richness in incompletely sampled landscapes. However,

under conditions of locally incomplete, spatially non-systematic sampling, which is typical for museum and checklist data, estimate errors were still moderate to large (> 20% for many islands), particularly if the number of sampling units (e.g. grid cells) is low. Estimation values should not be used uncritically as long as the data conditions that lead to errors are not precisely defined.

Acknowledgements

We are very grateful to all those professional and non-professional collectors who supplied unpublished distribution records of species (full details are listed in Beck & Kitching 2004). We particularly thank Konrad Fiedler and Robert Colwell for helpful comments on earlier drafts of the manuscript. Parts of data preparation were carried out during J.B.'s PhD work at Würzburg University, supported financially by scholarships of the German Research Council (DFG), the German Academic Exchange Service (DAAD) and the Sys-Resource Program of the EU.

References

- ArcView GIS 3.2a (2000) *Computer Program and Manual*. Environmental Systems Research Institute (ESRI), Redlands, California. <http://www.esri.com>.
- Barthlott, W., Lauer, W. & Placke, A. (1996) Global distribution of species diversity in vascular plants: towards a world map of phytodiversity. *Erdkunde*, **50**, 317–327.
- Beck, J. & Kitching, I.J. (2004) *The Sphingidae of Southeast Asia (incl. New Guinea, Bismarck & Solomon Islands)*. <http://www.sphingidae-sea.biozentrum.uni-wuerzburg.de>.
- Beck, J., Kitching, I.J. & Linsenmair, K.E. (2006a) Determinants of regional species richness: an empirical analysis of the number of hawkmoth species (Lepidoptera: Sphingidae) on the Malaysian archipelago. *Journal of Biogeography*, **33**, 694–706.
- Beck, J., Kitching, I.J. & Linsenmair, K.E. (2006b) Effects of habitat disturbance can be subtle yet significant: biodiversity of hawkmoth-assemblages (Lepidoptera: Sphingidae) in Southeast-Asia. *Biodiversity and Conservation*, **15**, 465–486.
- Beck, J., Kitching, I.J. & Linsenmair, K.E. (2006c) Measuring range sizes of Southeast-Asian hawkmoths (Lepidoptera, Sphingidae): effects of scale, resolution and phylogeny. *Global Ecology and Biogeography*, **15**, 339–348.
- Brehm, G., Pitkin, L.M., Hilt, N. & Fiedler, K. (2005) Montane Andean rain forests are a global diversity hotspot of geometrid moths. *Journal of Biogeography*, **32**, 1621–1627.
- Brose, U. & Martinez, N.D. (2004) Estimating the richness of species with variable mobility. *Oikos*, **105**, 292–300.
- Brose, U., Martinez, N.D. & Williams, R.J. (2003) Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, **84**, 2364–2377.
- Chazdon, R.L., Colwell, R.K., Denslow, J.S. & Guariguata, M.R. (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern Costa Rica. *Forest Biodiversity Research and Modeling – Conceptual Backgrounds and Old World Case Studies* (eds F. Dallmeier & J.A. Corniskey), pp. 285–309. Smithsonian Institution, Washington DC, USA. Parthenon Publishing Group and UNESCO, Paris.
- Colwell, R.K. (2005) *Estimates: Statistical Estimation of Species Richness and Shared Species from Samples*, Version 7.50. <http://purl.oclc.org/estimates>.

- Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B*, **345**, 101–118.
- Colwell, R.K., Mao, C.X. & Chang, J. (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, **85**, 1717–1727.
- Cowley, M.J.R., Wilson, R.J., Leon-Cortez, J.L., Gutierrez, D., Bulman, C.R. & Thomas, C.D. (2000) Habitat-based statistical models for predicting the spatial distribution of butterflies and day-flying moths in a fragmented landscape. *Journal of Applied Ecology*, **37**, 60–72.
- Dangles, O., Malmqvist, B. & Laudon, H. (2004) Naturally acid freshwater ecosystems are diverse and functional: evidence from boreal streams. *Oikos*, **104**, 149–155.
- Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrences and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.
- Fagan, W.F. & Kareiva, P.M. (1997) Using compiled species lists to make biodiversity comparisons among regions: a test case using Oregon butterflies. *Biological Conservation*, **80**, 249–259.
- Folgarait, P.J., Bruzzone, O., Porter, S.D., Pesquero, M.A. & Gilbert, L.E. (2005) Biogeography and macroecology of phorid flies that attack fire ants in south-eastern Brazil and Argentina. *Journal of Biogeography*, **32**, 353–367.
- Gaston, K.J. & Evans, K.L. (2004) Birds and people in Europe. *Proceedings of the Royal Society (London) B*, **271**, 1649–1655.
- Gaston, K.J., Scoble, M.J. & Crook, A. (1995) Patterns in species description: a case study using the *Geometridae* (*Lepidoptera*). *Biological Journal of the Linnean Society*, **55**, 225–237.
- Graham, C.H., Ferrier, S., Huetmann, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.
- Hall, R. & Holloway, J.D., eds. (1998) *Biogeography and Geological Evolution of South East Asia*. Backhuys, Leiden.
- Hawkins, B.A., Field, R., Cornell, H.V., Currie, D.J., Guegan, J.F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M., Porter, E.E. & Turner, J.R.G. (2003) Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, **84**, 3105–3117.
- Hilt, N. & Fielder, K. (2006) Arctiid moth ensembles along a successional gradient in the Ecuadorian montane rain forest zone: how different are subfamilies and tribes? *Journal of Biogeography*, **33**, 108–120.
- Hortal, J., Borges, P.A.V. & Gaspar, C. (2006) Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology*, **75**, 274–287.
- Iverson, L.R. & Prasad, A. (1998) Estimating regional plant biodiversity with GIS modelling. *Diversity and Distributions*, **4**, 49–61.
- Jetz, W. & Rahbek, C. (2002) Geographic range size and determinants of avian species richness. *Science*, **297**, 1548–1551.
- Jiménez-Valdere, A., Mednoza, S.J., Cano, J.M. & Munguira, M.L. (2006) Comparing relative model fit of several species-accumulation functions to local Papilionoidea and Hesperioidea butterfly inventories of Mediterranean habitats. *Biodiv. Conservation*, **15**, 177–190.
- Kerr, J.T., Southwood, T.R.E. & Cihlar, J. (2001) Remotely sensed habitat diversity predicts butterfly species richness and community similarity in Canada. *Proceedings of the National Academy of Sciences (USA)*, **98**, 11365–11370.
- Kier, G. & Barthlott, W. (2001) Measuring and mapping endemism and species richness: a new methodological approach and its application on the flora of Africa. *Biodiversity and Conservation*, **10**, 1513–1529.
- Krishnamani, R., Kumar, A. & Harte, J. (2004) Estimating species richness at large spatial scales using data from small discrete plots. *Ecography*, **27**, 637–642.
- Lawton, J.H., Bignell, D.E., Bolton, B., Bloemers, G.F., Eggleton, P., Hammond, P.M., Hodda, M., Holt, R.D., Larsoen, T.B., Mawdsley, N.A., Stork, N.E., Srivastava, D.S. & Watt, A.D. (1998) Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nature*, **391**, 72–76.
- Mao, C.X., Colwell, R.K. & Chang, J. (2005) Estimating the species accumulation curve using mixtures. *Biometrics*, **61**, 433–441.
- Mawdsley, N. (1996) The theory and practice of estimating regional species richness from local samples. *Tropical Rain-forest Research – Current Issues* (eds D.S. Edwards, W.E. Booth & S.C. Choy), pp. 193–212. Kluwer, Dordrecht, the Netherlands.
- Moritz, C., Richardson, K.S., Ferrier, S., Monteith, G.B., Stanisci, J., Williams, S.E. & Whiffin, T. (2001) Biogeographical concordance and efficiency of taxon indicators for establishing conservation priority in a tropical rainforest biota. *Proceedings of the Royal Society*, **268**, 1875–1881.
- Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A.B. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.
- Novotny, V. & Missa, O. (2000) Local versus regional species richness in tropical insects: one lowland site compared with the island of New Guinea. *Ecological Entomology*, **25**, 445–451.
- Oertli, B., Joye, D.A., Castella, E., Juge, R., Cambin, D. & Lachavanne, J.-B. (2002) Does size matter? The relationship between pond size and biodiversity. *Biological Conservation*, **104**, 59–70.
- Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A. & Peterson, A.T. (2003) Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, **426**, 837–841.
- Ray, N., Lehmann, A. & Joly, P. (2002) Modelling spatial distribution of amphibian populations: a GIS approach based on habitat matrix permeability. *Biodiversity and Conservation*, **11**, 2143–2165.
- Reid, W.V. (1998) Biodiversity hotspots. *Trends in Evolution and Ecology*, **13**, 275–279.
- Rosenzweig, M.L. (1995) *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, Massachusetts, USA.
- Rosenzweig, M.L. (2003) Reconciliation ecology and the future of species diversity. *Oryx*, **37**, 194–205.
- Rosenzweig, M.L., Turner, W.R., Cox, J.G. & Ricketts, T.H. (2003) Estimating diversity in unsampled habitats of a biogeographical province. *Conservation Biology*, **17**, 864–874.
- Schulze, C.H., Waltert, M., Kessler, P.J.A., Pitopang, R., Shahbuddin, Veddeler, D., Mühlenberg, M., Gradstein, S.R., Leuschner, C., Steffan-Dewenter, I. & Tschamtkke, T. (2004) Biodiversity indicator groups of tropical land-use systems: comparing plants, birds and insects. *Ecological Applications*, **14**, 1321–1333.
- Scoble, M.J., Gaston, K.J. & Crook, A. (1995) Using taxonomic data to estimate species richness in Geometridae. *Journal of the Lepidopterists' Society*, **49**, 136–147.
- Soberón, J.M., Llorente, J.B. & Oñate, L. (2000) The use of specimen-label databases for conservation purposes: an example using Mexican papilionid and pierid butterflies. *Biodiversity and Conservation*, **9**, 1441–1466.
- Sodhi, N.S., Lian, P.K., Brook, B.W. & Ng, P.K.L. (2004) Southeast-Asian biodiversity: an impending disaster. *Trends in Ecology and Evolution*, **19**, 654–660.
- Solow, A.R. & Smith, W.K. (2005) On estimating the number of species from the discovery record. *Proceedings of the Royal Society B*, **272**, 285–287.
- Southwood, T.R.E. & Henderson, P.A. (2000) *Ecological Methods*, 3rd edn. Blackwell Science, Oxford.
- Turner, W., Leitner, W. & Rosenzweig, M. (2000) *WS2M. Software for the measurement and analysis of species diversity*.

- Software and User's Manual. <http://eebweb.arizona.edu/diversity> (accessed March 2005).
- Ulrich, W. & Ollik, M. (2005) Limits to the estimation of species richness: The use of relative abundance distributions. *Diversity and Distributions*, **11**, 265–273.
- Whitmore, T.C. (ed.) (1987) *Biogeographical Evolution of the Malay Archipelago*. Clarendon Press, Oxford.
- Wiley, E.O., McNyset, K.M., Peterson, A.T., Robins, C.R. & Steward, A.M. (2003) Niche modelling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography*, **16**, 120–127.
- Zaniewski, A.E., Lehmann, A. & Overton, J.McC. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

Received 1 April 2006; final copy received 10 January 2007
Editor: Paul Giller

Supplementary material

The following supplementary material is available for this article:

Appendix S1. Formulas and software settings.

This material is available as part of the online article from: <http://www.blackwell-synergy.com/doi/full/10.1111/j.1365-2664.2007.01291.x>.

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.