Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2013) **22**, 586–595

A Journal of Macroecology

Global Ecology and Biogeography

RESEARCH PAPER

Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths

Liliana Ballesteros-Mejia[1]*, Ian J. Kitching[2], Walter Jetz[3], Peter Nagel[1] and Jan Beck[1]

[1]*Department of Environmental Science (Biogeography), University of Basel, St Johanns-Vorstadt 10, 4056 Basel, Switzerland,* [2]*Department of Life Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK,* [3]*Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06520, USA*

*Correspondence: Liliana Ballesteros-Mejia, University of Basel, Environmental Sciences, St Johanns Vorstadt 10, 4056 Basel, Switzerland. E-mail: liliana.ballesteros@unibas.ch

ABSTRACT

Aim Many taxa, especially invertebrates, remain biogeographically highly understudied and even baseline assessments are missing, with too limited and heterogeneous sampling being key reasons. Here we set out to assess the human geographic and associated environmental factors behind inventory completeness for the hawkmoths of Africa. We aim to separate the causes of differential sampling from those affecting gradients of species richness to illustrate a potential general avenue for advancing knowledge about diversity in understudied groups.

Location Sub-Saharan Africa.

Methods Using a database of distributional records of hawkmoths, we computed rarefaction curves and estimated total species richness across 200 km × 200 km grid cells. We fitted multivariate models to identify environmental predictors of species richness and used environmental co-kriging to map region-wide diversity patterns. We estimated cell-wide inventory completeness from observed and estimated data, and related these to human geographic factors.

Results Observed patterns of hawkmoths species richness are strongly determined by the number of available records in grid cells. Both show spatially structured distributions. Variables describing vegetation type, emerge as important predictors of estimated total richness, and variables capturing heat, energy availability and topographic heterogeneity all show a strong positive relationship. Patterns of interpolated richness identify three centres of diversity: Cameroon coastal mountains, and the northern and southern East African montane areas. Inventory completeness is positively influenced by population density, accessibility, protected areas and colonial history. Species richness is still under-recorded in the western Congo Basin and southern Tanzania/Mozambique.

Main conclusions Sampling effort is highly biased and controlling for it in large-scale compilations of presence-only data is critical for drawing inferences from our still limited knowledge of invertebrate distributions. Our study shows that a baseline of estimate of broad-scale diversity patterns in understudied taxa can be derived from combining numerical estimators of richness, models of main environmental effects and spatial interpolation. Inventory completeness can be partly predicted from human geographic features and such models may offer fruitful guidance for prioritization of future sampling to further refine and validate estimated patterns of species richness.

Keywords

Co-kriging interpolation, hawkmoths, Lepidoptera, sampling effort, spatial pattern, Sphingidae, Sub-Saharan Africa.

DOI: 10.1111/geb.12039

INTRODUCTION

The compilation and mapping of species richness over large spatial extents have, over the past decade, considerably advanced our understanding of global gradients of diversity and underlying processes (e.g. Jetz & Rahbek, 2002; Currie *et al*., 2004; Kreft & Jetz, 2007; Field *et al*., 2009). Maps of species richness also offer an important first, if limited (Jetz & Rahbek, 2002), guide to identifying regions of potential conservation value (Beck *et al*., 2011, for a tropical insect example). However, broad-scale studies of diversity gradients are spatially biased (toward well-studied continents such as North America and Europe) and even more so taxonomically, with tropical invertebrates in particular receiving much less attention than their contribution to global biodiversity would dictate (Godfray *et al*., 1999; Boakes *et al*., 2010; Beck *et al*., 2012). Among recent studies on insects on continental to global extents, Jenkins *et al*. (2011) and Guénard *et al*. (2012) have investigated global ant diversity patterns, Beck *et al*. (2006a) have investigated Southeast Asian sphingid moths, and several additional taxa have been studied at a regional scale in the temperate zone (e.g. Hawkins & DeVries, 2009; Kumschick *et al*., 2009; Hortal *et al*., 2011; Kudrna *et al*., 2011).

Both geographic and taxonomic biases appear to be a direct function of sampling activity and data availability (Boakes *et al*., 2010; Beck *et al*., 2012; Jetz *et al*., 2012), which will depend to some degree on (and correlate with) factors of human geography. Incomplete knowledge of the spatial occurrence of taxa has thus usually prevented the reliable documentation of species richness patterns. Several techniques have been developed to make use of incomplete local inventory data (Colwell & Coddington, 1994) and successfully applied to provide estimates of species richness at larger extents (Beck & Kitching, 2007; Mora *et al*., 2008; Tittensor *et al*., 2010). These approaches (and further refinements) combined with increasingly mobilized and integrated distribution information (Beck *et al*., 2012; Jetz *et al*., 2012) open up new and exciting prospects for the use of natural history collections data.

Hawkmoths (Lepidoptera, family Sphingidae) are among the most well-known insects with regard to their taxonomy and distribution (Kitching & Cadiou, 2000), and therefore represent an ideal model taxon for studying insect macroecology at a global scale. Nevertheless, a shortage of distributional data for tropical species has so far prevented detailed, grid-based analyses of their broad-scale species richness patterns in relation to environmental factors (but see Beck *et al*., 2006a). Based on findings for other taxa (Field *et al*., 2009), we expect climatic variables and resulting patterns of habitat productivity to explain some variation in species richness. Given the herbivorous lifestyle of sphingid caterpillars, we also hypothesize that vegetation type affects their diversity.

However, inventory completeness (which may also affect observed species richness) is ultimately determined by collectors' decisions on where to engage in field sampling. While geographic patterns of sampling intensity will necessarily be partly idiosyncratic (e.g. a high record density near places of residence of particular collectors), we also expect some generalities to emerge (Reddy & Dávalos, 2003; Martin *et al*., 2012). For example, high human population density and dense infrastructure (i.e. accessibility to traffic, tourism) should have a positive effect on sampling effort, whereas regions of armed conflict have probably been avoided by collectors (Balmford *et al*., 2001). Given the impact of European colonialism on Sub-Saharan Africa even after the formal political independence of countries, we also expect effects of colonial history. This sort of political history may explain past sampling activity as well as mobilization and data access to date.

Here, using an extensive, expert-validated data compilation, we provide a first quantitative assessment of species richness patterns of sphingid moths across Sub-Saharan Africa, using a variety of estimators and specifically addressing sampling effort. We use environmental predictors to identify and model the main correlates of spatial variation in sphingid richness and combine them with spatial interpolation techniques to provide a full subcontinental map of species richness. To assess the robustness of these findings, we specifically quantify patterns of survey completeness (see Moerman & Estabrook, 2006, Zagmajster *et al*., 2010, and Guénard *et al*., 2012, for relevance to biodiversity research and conservation) and model their potential human geographic determinants. Using African hawkmoths as a continental study system, we illustrate how separating the causes of species richness and its sampling facilitates a more rigorous documentation and understanding of the geographic diversity patterns of the many remaining understudied groups.

METHODS

Distribution data

We compiled distribution records for all Sphingidae of Sub-Saharan Africa (south of *ca*. 17° N latitude, including Madagascar), based on an extensive search of published literature and the internet (e.g. Lepidoptera blogs, specimen trading sites, the Barcode of Life Database (http://www.barcodinglife.org), the Global Biodiversity Information Facility (http://www.gbif.org)), as well as correspondence with a large number of professional and amateur collectors, our own field sampling, and through databasing several major natural history collections (e.g. museums in London, Berlin, Paris, Munich, Tervuren and Pittsburgh). We took the utmost care to exclude or correct confirmed or likely errors in locality and species identity. We georeferenced localities as precisely as feasible and applied a unified nomenclature of taxa (following Kitching & Cadiou, 2000, and recent, in parts as yet unpublished, updates; see also Boakes *et al*., 2010). For the purposes of this study, we ignored all locality records that could not be allocated with a precision of at least 1° latitude/longitude (*ca*. 110 km). We defined a record as a unique combination of species, locality, year and collector. Hence, a record may contain between one and many specimens caught at the same time, whereas temporal replicates (e.g. a species being caught repeatedly in different years at the same site) would be considered as separate records. While the oldest

data originated from the late 19th century, the vast majority of data were collected later than 1950 (and most from 1980 onward).

We mapped numbers of records (N) and observed species richness (S_{obs}) in an equal-area Mollweide projection, aggregated in raster grids with a cell size of 200 km × 200 km. Preliminary analyses identified this cell size as the best compromise between resolution and number of cells and data quality within cells.

Correcting for incomplete species inventories

We applied three approaches, all based on the distribution of records and species per grid cell, to attempt to control for variable sampling effort and ultimately provide an estimate of actual grid cell richness values.

1. We calculated rarefaction curves (i.e. randomized accumulation of species with records; S_{rar}) for each grid cell, which allows us to estimate how many species would have been observed in a cell if only a given number of specimens had been sampled (Gotelli & Colwell, 2001). Thus, rarefaction allows standardization of sampling effort across cells but outputs are not estimates of the complete species richness of cells (unlike the following methods). We used 25 records as a standard to compare estimated S_{rar}.

2. We fitted several asymptotic functions (i.e. Michaelis–Menten, negative exponential, asymptotic, Chapman–Richards, rational, Weibull; see Mora *et al*., 2008, for details) to the rarefaction curves to derive estimates of the total species richness expected with infinitely large sampling effort. Each of these functions was evaluated separately for each grid cell using Akaike's information criterion (AIC) for its fit with the rarefaction curves, and AIC-weighted average estimates of species richness (S_{asym}) were calculated. Asymptotic estimators have recently been used by Mora *et al*. (2008) in a similar context.

3. As an alternative estimator of 'true' species richness we calculated a nonparametric metric, *Chao1* (Chao, 1984), that makes use of the ratio of species recorded only once, or exactly twice, per cell (S_{Chao}). Because this method yields results similar to S_{asym} we only mention important data for S_{Chao} in the main text but present details in the Supporting Information (Table S2, Fig. S2).

Output from these three approaches varied in quality and reliability between cells, and we applied some 'pruning' rules to remove highly unreliable cell estimates, at the cost of reducing the number of cells available for analysis. We present here data for cells where at least 25 records were available, and where coefficients of variation (i.e. standard error of estimate/estimate) for species richness estimators were < 0.2. We also repeated our analyses using more (> 50 records per cell) and less rigorous (> 10, > 15 records) pruning rules (i.e. affecting the numbers of cells available versus the reliability of estimates), but this did not affect the main conclusions.

Environmental effects on species richness patterns

We investigated the effect of some environmental variables that have often been found to (or assumed to) affect species richness patterns at large extents and grain sizes on $\log_{10}$-transformed estimates of species richness. In particular, we investigated effects of potential evapotranspiration (PET; from http://edit.csic.es/Climate.html) as a measure of energy input into the ecosystem (Hawkins *et al*., 2003), actual evapotranspiration (AET; from http://edit.csic.es/Climate.html) as a measure of primary productivity (Currie *et al*., 2004), topographic heterogeneity (altitudinal range within cells) as a proxy of habitat variability and consequent beta diversity (Ruggiero & Hawkins, 2008) and vegetation structure (herb and tree cover from MODIS Vegetation Continuous Fields, http://glcf.umiacs.umd.edu/data/vcf/, means for 200-km cells). For sphingids, as herbivorous insects, we expected functional links with vegetation type although most species are not particularly host specific (i.e. specialization below plant family level is rare; Beck *et al*., 2006b). MODIS data are based on satellite imagery taken in 2000–01 and hence include aspects of human-induced changes to the landscape. Vegetation data are correlated with AET estimates (tree cover, $r^2 = 0.62$; herb cover, $r^2 = 0.30$), which may affect the interpretation of results (see below). Coastal raster cells may appear to harbour reduced species richness due to smaller area alone. However, in our data set coastal regions often contained well-sampled and species-rich cells, and given this sampling pattern the effect of land area on observed richness was weak (Spearman rank correlation, 405 cells: $r_s = -0.105$). We thus included in the analysis coastal cells down to 5.0% land area to avoid loss of critical information. We tested model residuals for spatial autocorrelation (software SAM, v.4; 999 permutations), finding significant Moran's $I > 0.1$ for lag distances up to *ca*. 500 km. We used spatially explicit multivariate generalized least square (GLS) models to account for spatial autocorrelation in the data (Beale *et al*., 2010; spherical variogram structure; software R.2.13.1, *nlme* package).

For all three response variables (i.e. S_{rar}, S_{asym}, S_{Chao}), we evaluated full models (all listed variables, no interactions) and various simplified models using the AIC; we used only the best (lowest AIC) for further analyses. We calculated the pseudo-R^2 of models as a correlation of predicted versus observed values. GLS model coefficients were used to extrapolate species richness estimates across Sub-Saharan Africa, allowing intuitive evaluation of the consistency of patterns derived from the three estimation methods. We also applied co-kriging (i.e. spatial interpolation of raw estimates based on their autocorrelation, the autocorrelation of environmental model predictions and the cross-correlation between them) for mapping (Kreft & Jetz, 2007). Co-kriging was carried out in ArcGIS 10 software, assuming anisotropic variogram structures. Estimates were optimized by cross-validation, and we report final root mean square errors (RMSEs) of interpolation predictions.

Quantifying and analysing inventory completeness

Using co-kriging estimates of 'true' species richness (S_{asym}), we determined cell-wide species inventory completeness as S_{obs}/S_{asym} and as yet unrecorded species richness as $S_{asym} - S_{obs}$. Cells without any data consequently had an inventory completeness

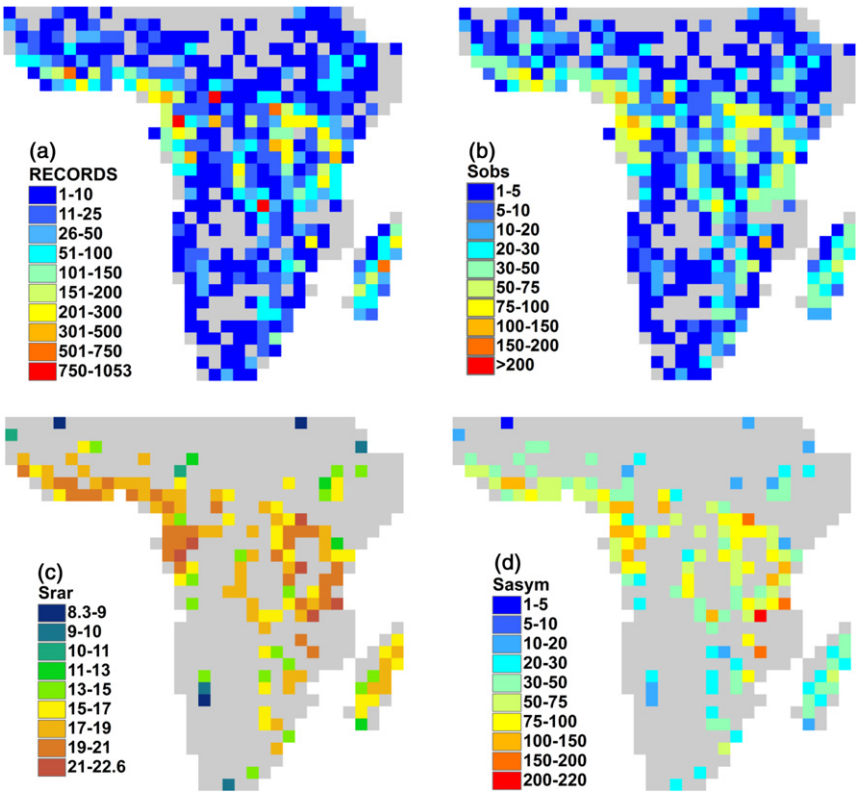

Figure 1 (a) Number of records (N). (b) Observed species richness (S_{obs}). (c) Rarefied species richness at 25 records (S_{rar}). (d) Asymptotic estimate of total species richness (S_{asym}). Note that S_{obs} and S_{asym} are shown on the same colour scale. Grid cells without data are shown in grey.

of zero. In some cells estimates of S_{asym} were lower than S_{obs} due to imperfect function fitting; for these we defined inventories as complete (i.e. $S_{obs}/S_{asym} = 1$) and set the number of unrecorded species to zero ($S_{asym} - S_{obs} = 0$).

We related the geographic patterns of inventory completeness to human factors such as road and tourism infrastructure, habitat encroachment, population density, armed conflict and colonial history (see Table S1 for details and sources). We hypothesized that each of these factors may play a role in affecting collectors' inclination to be active in a region. Inventory completeness is a zero-inflated response variable (i.e. there are many cells without records; Zuur *et al.*, 2010) and we used $\log_{10}(x + 1)$-transformation to reduce extreme deviations from normality. We first carried out AIC-based model selection of ordinary least square (OLS) models to identify important effects of these variables on inventory completeness. For the best model (lowest AIC), we found significant positive autocorrelation in residuals for lag distances up to *ca*. 870 km. Using the variables in the best OLS model, we reanalysed effects in a spatially explicit GLS model, and we repeated this analysis without the zero cells to avoid spurious conclusions due to zero inflation.

RESULTS

Observed and estimated species richness

A total of 21,194 records provide occurrence data for 322 species over 405 grid cells (of size 200 km × 200 km) covering all of Sub-Saharan Africa (145 additional grid cells had no data available; Fig. 1). After applying 'pruning' rules of data inclusion for estimating species richness (see Methods), 146 cells were left for analysis.

Most occurrence samples (N) come from the coastal parts of western and central Africa, from the Great Lakes regions of eastern Africa and from Madagascar. Few records are available for the drier parts of southern Africa. This geographically highly uneven availability of occurrence data was strongly reflected in the patterns of observed species richness (S_{obs}). N and S_{obs} are strongly positively correlated (linear correlation of $\log N \sim \log S_{obs}$: $r = 0.95$, $n = 405$ grid cells), suggesting pervasive effects of sampling effort even at this coarse spatial resolution. Restricting the test to the 146 cells with ≥ 25 records confirms this relationship ($r = 0.90$).

To overcome these effects of sampling on richness we calculated rarefied species richness (S_{rar}) and estimated expected full species richness given sufficient sampling using parametric asymptotic (S_{asym}) and nonparametric (S_{Chao}) methods. S_{asym} showed similar geographic patterns (Fig. 1c). These measures accordingly exhibited much weaker relationships with N (i.e. $\log N \sim \log S_{asym}$, $r = 0.65$; $\log N \sim \log S_{Chao}$, $r = 0.64$) and there was barely any association with rarefied species richness ($\log N \sim \log S_{rar}$, $r = 0.45$). Relationships with observed species richness were also weak ($S_{obs} \sim S_{asym}$, $r = 0.78$; $S_{obs} \sim S_{Chao}$, $r = 0.70$; $\log_{10} S_{obs} \sim S_{rar}$, $r = 0.67$). Estimates of full species richness agree with each other (i.e. $S_{asym} \sim S_{Chao}$, $r = 0.94$) while showing some deviation from rarefied data ($S_{rar} \sim \log S_{asym}$, $r = 0.88$; $S_{rar} \sim \log S_{Chao}$, $r = 0.84$).

Table 1 Generalized least squares model details for rarefied species richness (S_{rar}) and asymptotic estimates of species richness (S_{asym}). Pseudo-R^2 = 0.405 for S_{rar}, 0.138 for S_{asym} (n = 146 grid cells).

	$\log_{10}S_{rar}$			$\log_{10}S_{asym}$		
Variable	Coefficient	t	P	Coefficient	t	P
Intercept	0.62548	7.307	< 0.001	0.52334	1.258	0.211
Topo. het.	0.00001	1.839	0.068	0.00003	1.348	0.180
AET	0.00004	1.255	0.212	0.00032	2.273	0.025
PET	0.00010	2.253	0.026	0.00016	0.760	0.448
Tree cover (%)	0.00430	6.074	< 0.001	0.00544	1.893	0.061
Herb cover (%)	0.00404	6.817	< 0.001	0.00562	2.286	0.024

Topo. het., topographic heterogeneity; AET, actual evapotranspiration; PET, potential evapotranspiration.

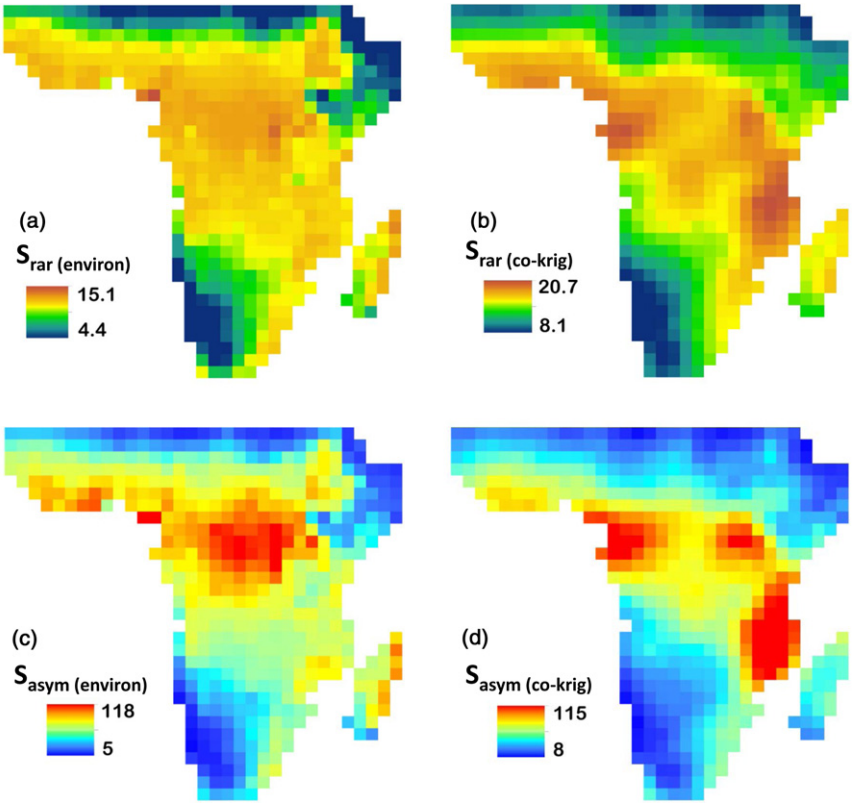

Figure 2 (a) Estimates of species richness based on the environmental model of rarefied species richness at 25 records ($S_{rar\ (environ)}$, upper left). (b) Co-kriging interpolation of rarefied species richness ($S_{rar\ (co\text{-}krig)}$, upper right). (c) Environmental model of asymptotic estimate of total species richness ($S_{asym\ (environ)}$, lower left). (d) Co-kriging interpolation of asymptotic estimate of total species richness ($S_{asym\ (co\text{-}krig)}$, lower right). See Table 1 for details on environmental models. Root mean square errors for co-kriging interpolations are 1.95 for S_{rar} and 29.05 for S_{asym}.

Environmental models and interpolation

For both rarefied species richness (S_{rar}) and asymptotic estimators (S_{asym}) the strongest environmental models included all predictors according to AIC (see Table 1). Notably, environment explains considerably more of the variation in S_{rar} (pseudo-R^2 = 0.41) than S_{asym} (pseudo-R^2 = 0.14), and the models agree only partly in the importance of variables. For both models, positive coefficients of similar magnitude were found for tree and herb cover. The model for S_{rar} is additionally driven by PET and, more weakly, topographic heterogeneity, whereas that for S_{asym} is mainly affected by AET. Figure 2 illustrates these differences by extrapolating the models (note, for example, different prediction for the Congo Basin, a region of very high AET). The model for S_{Chao} shows performance similar to that of S_{asym} (pseudo-R^2 = 0.15; Table S2).

Environmental models based on different species richness estimates lead to broadly similar predicted patterns of diversity (Fig. 2; $S_{rar} \sim \log_{10}S_{asym}$, r = 0.97; N = 550), and so did co-kriging interpolations ($S_{rar} \sim \log_{10}S_{asym}$, r = 0.95). Estimates of S_{asym} and S_{Chao} (see Fig. S2) are correlated for the environmental model (r = 0.92) and even stronger for co-kriging (r = 0.97). Surprisingly, predictions from environmental models and co-kriging interpolations within the same metrics deviate considerably from each other (S_{rar}, r = 0.83; S_{asym}, r = 0.75; S_{Chao}, r = 0.68). Deviations are particularly strong in the Ethiopian Highlands and the Horn of Africa (the environmental model predicts more species in the former and fewer in the latter than co-kriging).

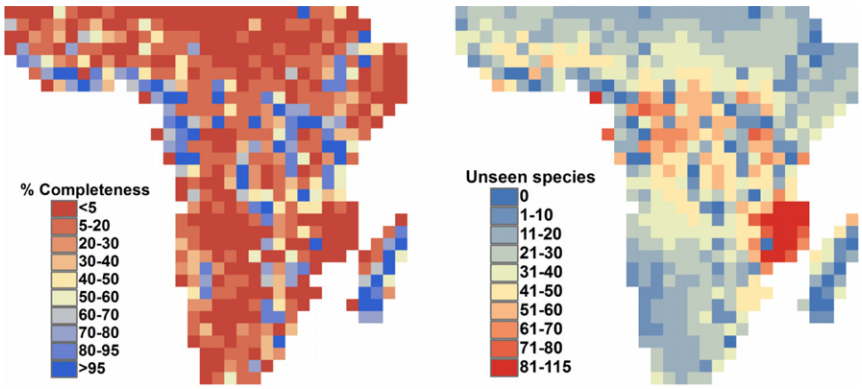

Figure 3 (Left) Cell-wide inventory completeness (based on co-kriging estimate of S_{asym}; these data were used to investigate effects of human geographic factors, Table 2). (Right) Estimated number of unrecorded species in each grid cell measured as the difference between S_{obs} and S_{asym}.

Table 2 Ordinary least squares (OLS) and generalized least squares (GLS) models explaining estimated inventory completeness [Fig. 3; $\log_{10}(x + 1)$-transformed] of cells by human geographic factors. Country names refer to colonial powers in 1919 (see Table S1 for details on predictor variables; $n = 502$ grid cells). Data are zero-inflated, but a GLS model without the 145 zero cells recovered all results except the marginal effect of protected areas (Table S3).

	OLS; $R^2_{adj} = 0.21$			GLS; pseudo-$R^2 = 0.22$		
	Coefficient	t	P	Coefficient	t	P
Intercept	0.03285	2.356	0.019	0.03286	2.340	0.020
Britain	−0.02825	−2.837	0.005	−0.02828	−2.815	0.005
Belgium	0.01243	0.838	0.402	0.01190	0.794	0.428
Portugal	−0.06021	−4.015	< 0.001	−0.06045	−3.997	< 0.001
France	0*			0*		
$\log_{10}$(population + 1)	0.02954	4.427	< 0.001	0.02980	4.433	< 0.001
Airports	0.04446	3.378	0.001	0.04387	3.350	0.001
Railways	0.00015	3.064	0.002	0.00015	3.050	0.002
Tourism	0.04163	3.968	< 0.001	0.04115	3.936	< 0.001
Protected	0.01691	1.982	0.048	0.01652	1.942	0.053

*Zero by default.

For both estimates of total species richness (S_{asym} and S_{Chao}), environmental models predict more species in the Congo Basin and fewer in Tanzania/Mozambique than co-kriging (residual data not shown).

Inventory completeness

We used the predictions of the best-performing model of total species richness, co-kriging of S_{asym}, to estimate the geographic variation in inventory completeness (S_{obs}/S_{asym}) and undetected species richness ($S_{asym} - S_{obs}$; Fig. 3). Model selection based on AIC (see Methods) led to a model including population density, railway lines, airports, tourist hotspots, protected areas and colonial history as the most important variables for predicting inventory completeness (explaining *ca.* 21% of data variability). Coefficients (Table 2) reveal the expected positive effects of infrastructure (traffic access, tourism) and of protected areas, but also effects of colonial history (although the large majority of data stemmed from the post-colonial era). In particular, the formerly Portuguese (i.e. Mozambique) and British regions were less well-sampled or well-mobilized than formerly French and Belgian regions. A univariate model (not shown) that does not account for differences in infrastructure (which may itself be an outcome of colonial history) confirms the effect of past Portuguese occupation but no other effects of colonial history. Models with slightly higher AIC (ΔAIC < 2; data not shown) contain additional positive effects of road density and negative effects of pristine regions. A map of residuals from the OLS model (not shown) indicates only weak spatial structure, with particularly positive residuals (i.e. better sampling than predicted) in Madagascar and Cameroon, and negative residuals in the Sahel, western Congo Basin and Zimbabwe.

Inventory completeness (Fig. 3, left) is related to [$\log_{10}(x + 1)$-transformed] number of records (Fig. 1; $r^2 = 0.85$), and repeating the OLS analysis with records as a response variable (a proxy of sampling effort) leads to identical conclusions (not shown).

When looking at absolute numbers of yet-to-be-recorded species, Mozambique and southern Tanzania, as well as the Congo Basin, stand out as containing much unrecorded (at grid cell level, not necessarily undescribed) biodiversity (Fig. 3, right).

DISCUSSION

Controlling species richness for sampling effort

Our analyses demonstrate that for incompletely sampled taxa (i.e. the great majority of taxa in most regions), numerical estimates of cell-wide species richness based on the relative

distributions of records and species can provide data that enable first large-scale mapping and analysis of diversity. These sorts of assessments are urgently needed to put global biodiversity research on a broader taxonomic basis. As observed data are often heavily affected by sampling effort (e.g. Palmer *et al*., 2002; Boakes *et al*., 2010), such estimates may currently be the only alternative to overlaying estimated range maps of individual species (based on expert knowledge or distribution modelling). Expert range maps for individual species are, for tropical regions, currently only available for vertebrates, and they can have spatial characteristics different from cell-based estimates, with consequences for further analysis and inference (McPherson & Jetz, 2007).

Although estimates of total species richness are easiest to understand and interpret, our data suggest that rarefaction may currently be the more reliable method of controlling for sampling effort in diversity patterns. We found S_{rar} to be less dependent on record numbers than S_{asym} or S_{Chao}, and the environmental model based on S_{rar} explained considerably more variability. This indicates that rarefaction introduces less random error than extrapolation. Based on similar arguments, Fiedler & Truxa (2012) recently came to the same conclusions for finer-scale data.

Environmental effects and spatial interpolation

We found positive effects of energy-related variables in environmental models, but there was inconsistency between models as to whether links with AET (a proxy of primary productivity) or PET (a proxy of solar energy input) are more important. Plausible mechanisms have been postulated for both variables (see Evans *et al*., 2005, for review), and published analyses leave uncertainty similar to that revealed here (e.g. Mittelbach *et al*., 2001; Currie *et al*., 2004; Buckley & Jetz, 2007). Energy availability was found to have a large effect on regional and local richness (Jetz & Fine, 2012). Additionally temperature was found to be positively associated with ectotherm richness whereas primary productivity is correlated with endotherm richness (Buckley *et al*., 2012). At a much smaller scale temperature was found to be negatively associated with butterfly richness (Stefanescu *et al*., 2004). Possibly the coarse-scale, imprecise measurement of currently available AET data prevents a clear distinction between these effects. The high similarity of coefficients for tree and herb cover (i.e. forest versus savannah) after controlling for energy and productivity is interesting, suggesting that other differences between those habitat types (such as three-dimensional structure) are not very important at this spatial scale of analysis.

We found relatively weak correspondence of patterns recovered from the environmental models and from co-kriging. Spatial interpolation can yield equal or better estimates than environmental models (e.g. Bahn & McGill, 2007; Lin *et al*., 2008), although they are less informative with regard to the causes of patterns. By being closer to observed data patterns, interpolation can also map historical effects that are undetectable by correlation with the current environment. Our co-kriging estimates (Fig. 2) clearly identify three areas of high diversity, i.e. the coastal mountains of western central Africa, and the northern and southern mountain ranges of East Africa. Notably, this pattern is not explained by topographic heterogeneity (which was included in environmental models). All three regions were identified as regions of complex biogeographic history and high endemism in other taxa (e.g. Jetz *et al*., 2004; Linder *et al*., 2012), suggesting potential effects of geographic history. Also, co-kriging interpolations yield patterns of species richness broadly similar to those published for birds (Jetz & Rahbek, 2002; Lin *et al*., 2008), amphibians (Buckley & Jetz, 2007) and plants (Kreft & Jetz, 2007). Scale dependency of species richness hinders quantitative comparison across studies (Rahbek, 2005; Beever *et al*., 2006). However, the prediction of high diversity not only in the montane areas of southern Tanzania and Mozambique but also in their coastal lowlands appears novel; it is unwarranted by actual data (Fig. 1) and requires further data collection for confirmation.

Sampling effort and the large-scale evaluation of biodiversity

Our data showed clearly that cell-wide inventory completeness was not equally distributed in space (Fig. 3). However, the causality of the relationship between sampling effort (i.e. number of records, *N*) and observed species richness is not entirely clear. Collectors may be particularly drawn to places known or presumed to be high in species diversity (which often also feature high human population density; Balmford *et al*., 2001). Alternatively, more comprehensive sampling may lead to the finding of more species.

Inventory completeness was substantially related to accessibility and infrastructure. Modelling cannot infer causality directly, but it is plausible to conclude that collectors make conscious decisions to visit those places that are easy to access. Protected areas had a positive effect on inventory completeness, although it cannot be concluded from our data whether this is caused by specific conservation interest in surveys or by the infrastructure allowing access to, for example, national parks. 'Pristine' landscapes, on the other hand, had a (weak) negative effect, which is most likely due to lack of access. This (non-significant) effect is somewhat in contradiction to the assumption that such places often have a more complete inventory, and also to Guénard *et al*. (2012), who estimated many unrecorded ant genera in regions of high anthropogenic habitat destruction.

Even though some patterns of model residuals match the preconceived expectations of most Africa researchers on collection intensity (e.g. poor knowledge of the Congo Basin, well-sampled Madagascar), we mostly observed only idiosyncratic deviations from model expectations of inventory completeness. Some large positive residuals (more complete data than modelled) seemed to be associated with single places with large quantities of data collected over a few years, suggesting intense activity by a single collector or a particular survey programme. Additionally, georeferencing issues could also cause such effects. Records saying nothing but 'Kivu', for example, were referenced to the same coordinates whereas they could sometimes relate to

a much wider geographic interpretation, i.e. the Kivu provinces or the entire region around Lake Kivu. Furthermore, some well-sampled places did not stand out in the population or traffic network, but they may nevertheless have drawn collectors (such as expatriate workers, missionaries) because of their administrative (e.g. Yaoundé, Cameroon's capital) or economic importance (e.g. Lubumbashi, a mining town in the Congo). We did not include locations of universities (cf. Moerman & Estabrook, 2006) in our analyses as very few of our data stemmed from African collectors or collections (e.g. databasing the collection of the Natural History Museum of Addis Ababa yielded < 10% of available records for Ethiopia). Exploratory analysis of inventory completeness and human geographic variables, using geographically weighted regression (not shown), suggested some spatial patterns that deserve further study, such as increasing predictability of inventory completeness from west to east.

CONCLUSIONS

Sampling effort is a crucial variable when assessing large-scale species richness patterns, and ignoring this would probably lead to flawed perceptions of patterns. Numerical estimates based on the accumulation of species with records, in combination with environmental models and spatial interpolation, can help us to estimate broad-scale richness patterns. However, they necessarily contain estimation error, and important patterns should be backed up by future field surveys. Similar to what has been found for the much better studied vertebrates, vegetation cover, energy-related variables and topographic heterogeneity are also important environmental correlates for sphingid moth species richness, while leaving considerable variation unexplained – possibly due to historical component in the patterns of species richness. Inventory completeness can be predicted to a certain degree from human population density, infrastructure and colonial history. Our approach and results expose areas of extensive and poor sampling given expected discoverable species richness, thus highlighting regions where future sampling efforts should be directed.

ACKNOWLEDGEMENTS

We thank all the professional and amateur collectors (too numerous to mention here) who made their data available for our project. S. P. Loader, W. Schwanghart and two anonymous referees provided critical comments on an earlier version of the manuscript. M. Curran, M. Kopp, R. Hagmann, S. Widler and S. Lang helped to process the distributional data. The study received financial support from the Swiss National Science Foundation (project 3100AO_119879) and the Synthesys programme of the European Union.

REFERENCES

Bahn, V. & McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16**, 733–742.

Balmford, A., Moore, J.L., Brooks, T., Burgess, N., Hansen, L.A., Williams, P. & Rahbek, C. (2001) Conservation conflicts across Africa. *Science*, **291**, 2616–2619.

Beale, C., Lennon, J., Yearsley, J. & Brewer, M. (2010) Regression analysis of spatial data. *Ecology Letters*, **13**, 246–264.

Beck, J. & Kitching, I.J. (2007) Estimating regional species richness of tropical insects from museum data: a comparison of a geography-based and sample-based methods. *Journal of Applied Ecology*, **44**, 672–681.

Beck, J., Kitching, I.J. & Eduard Linsenmair, K. (2006a) Determinants of regional species richness: an empirical analysis of the number of hawkmoth species (Lepidoptera: Sphingidae) on the Malesian Archipelago. *Journal of Biogeography*, **33**, 694–706.

Beck, J., Kitching, I.J. & Linsenmair, K.E. (2006b) Diet breadth and host plant relationships of Southeast-Asian sphingid caterpillars. *Ecotropica*, **12**, 1–13.

Beck, J., Schwanghart, W., Chey, V.K. & Holloway, J.D. (2011) Predicting geometrid moth diversity in the heart of Borneo. *Insect Conservation and Diversity*, **4**, 173–183.

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.-K., Winter, M. & Dormann, C. (2012) What's on the horizon for macroecology? *Ecography*, **35**, 673–683.

Beever, E.A., Swihart, R.K. & Bestelmeyer, B.T. (2006) Linking the concept of scale to studies of biological diversity: evolving approaches and tools. *Diversity and Distributions*, **12**, 229–235.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Ding, C.Q., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, **8**, e1000385.

Buckley, L.B. & Jetz, W. (2007) Environmental and historical constraints on global patterns of amphibian richness. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1167–1173.

Buckley, L.B., Hurlbert, A.H. & Jetz, W. (2012) Broad-scale ecological implications of ectothermy and endothermy in changing environments. *Global Ecology and Biogeography*, **21**, 873–885.

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

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

Currie, D.J., Mittelbach, G.G., Cornell, H.V., Field, R., Guégan, J.-F., Hawkins, B., Kaufman, D.M., Kerr, J.T., Oberdorff, T., O'Brien, E. & Turner, J. (2004) Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecology Letters*, **7**, 1121–1134.

Evans, K.L., Warren, P.H. & Gaston, K.J. (2005) Species–energy relationships at the macroecological scale: a review of the mechanisms. *Biological Reviews*, **80**, 1–25.

Fiedler, K. & Truxa, C. (2012) Species richness measures fail in resolving diversity patterns of speciose forest moth assemblages. *Biodiversity and Conservation*, **21**, 2499–2508.

Field, R., Hawkins, B.A., Cornell, H.V., Currie, D.J., Diniz-Filho, J.A.F., Guégan, J.-F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M. & Turner, J.R.G. (2009) Spatial species-richness gradients across scales: a meta-analysis. *Journal of Biogeography*, **36**, 132–147.

Godfray, H.C.J., Lewis, O.T. & Memmot, J. (1999) Studying insect diversity in the tropics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **354**, 1811–1824.

Gotelli, N. & Colwell, R. (2001) Quantifying biodiversity?: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Guénard, B., Weiser, M.D. & Dunn, R.R. (2012) Global models of ant diversity suggest regions where new discoveries are most likely are under disproportionate deforestation threat. *Proceedings of the National Academy of Sciences USA*, **109**, 7368–7373.

Hawkins, B. & DeVries, P.J. (2009) Tropical niche conservatism and the species richness gradient of North American butterflies. *Journal of Biogeography*, **36**, 1698–1711.

Hawkins, B.A., Field, R., Cornell, H.V., Currie, D.J., Guégan, J.-F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E., Porter, E.E. & Turner, J.R.G. (2003) Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, **84**, 3105–3117.

Hortal, J., Diniz-Filho, J.A.F., Bini, L.M., Rodríguez, M.A., Baselga, A., Nogués-Bravo, D., Rangel, T.F., Hawkins, B.A. & Lobo, J.M. (2011) Ice age climate, evolutionary constraints and diversity patterns of European dung beetles. *Ecology Letters*, **14**, 741–748.

Jenkins, C.N., Sanders, N.J., Andersen, A.N. *et al*. (2011) Global diversity in light of climate change: the case of ants. *Diversity and Distributions*, **17**, 652–662.

Jetz, W. & Fine, P.V. (2012) Global gradients in vertebrate diversity predicted by historical area–productivity dynamics and contemporary environment. *PloS Biology*, **10**, e1001292.

Jetz, W. & Rahbek, C. (2002) Geographic range size and determinants of avian species richness. *Science*, **297**, 1548–1551.

Jetz, W., Rahbek, C. & Colwell, R.K. (2004) The coincidence of rarity and richness and the potential signature of history in centres of endemism. *Ecology Letters*, **7**, 1180–1191.

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution*, **27**, 151–159.

Kitching, I.J. & Cadiou, J.-M. (2000) *Hawkmoths of the world*. The Natural History Museum, London and Cornell University Press, London.

Kreft, H. & Jetz, W. (2007) Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences USA*, **104**, 5925–5930.

Kudrna, O., Harpke, A., Lux, K., Pennersdorfer, J., Schweiger, O., Settele, J. & Wiemers, M. (2011) *Distribution atlas of butterflies in Europe*. Gesellschaft für Schmetterlingsschutz, Halle, Germany.

Kumschick, S., Schmidt-Entling, M.H., Bacher, S., Hickler, T., Espadaler, X. & Nentwig, W. (2009) Determinants of local ant (Hymenoptera: Formicidae) species richness and activity density across Europe. *Ecological Entomology*, **34**, 748–754.

Lin, Y., Yeh, M.-S., Deng, D. & Wang, Y.-C. (2008) Geostatistical approaches and optimal additional sampling schemes for spatial patterns and future sampling of bird diversity. *Global Ecology and Biogeography*, **17**, 175–188.

Linder, H.P., de Klerk, H.M., Born, J., Burgess, N.D., Fjeldså, J. & Rahbek, C. (2012) The partitioning of Africa: statistically defined biogeographical regions in sub-Saharan Africa. *Journal of Biogeography*, **39**, 1189–1205.

McPherson, J.M. & Jetz, W. (2007) Type and spatial structure of distribution data and the perceived determinants of geographical gradients in ecology: the species richness of African birds. *Global Ecology and Biogeography*, **16**, 657–667.

Martin, L.J., Blossey, B. & Ellis, E. (2012) Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment*, **10**, 195–201.

Mittelbach, G.G., Steiner, C.F., Scheiner, S.M., Gross, K.L., Reynolds, H.L., Waide, R.B., Willig, M.R., Dodson, S.I. & Gough, L. (2001) What is the observed relationship between species richness and productivity? *Ecology*, **82**, 2381–2396.

Moerman, D.E. & Estabrook, G.F. (2006) The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography*, **33**, 1969–1974.

Mora, C., Tittensor, D.P. & Myers, R.A. (2008) The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 149–155.

Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S. & Wohlgemuth, T. (2002) Quantitative tools for perfecting species lists. *Environmetrics*, **13**, 121–137.

Rahbek, C. (2005) The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology Letters*, **8**, 224–239.

Reddy, S. & Dávalos, L. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719–1727.

Ruggiero, A. & Hawkins, B.A. (2008) Why do mountains support so many species of birds? *Ecography*, **31**, 306–315.

Stefanescu, C., Herrando, S. & Páramo, F. (2004) Butterfly species richness in the north-west Mediterranean Basin: the role of natural and human-induced factors. *Journal of Biogeography*, **31**, 905–915.

Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E.V. & Worm, B. (2010) Global patterns and predictors of marine biodiversity across taxa. *Nature*, **466**, 1098–1101.

Zagmajster, M., Culver, D., Christman, M. & Sket, B. (2010) Evaluating the sampling bias in pattern of subterranean species richness: combining approaches. *Biodiversity and Conservation*, **19**, 3035–3048.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Figure S1 Map of *Chao1* estimated species richness.
Figure S2 Maps of extrapolation of the environmental model of species richness and co-kriging for *Chao1*.
Table S1 Predictors of inventory completeness.
Table S2 Details of *Chao1* estimates of species richness model.
Table S3 Details of *Chao1* estimates of inventory completeness model.

BIOSKETCHES

Liliana Ballesteros-Mejia recently finished her PhD research on the analysis of geographical distributions of Old World sphingid moths.

Ian J. Kitching's main research interests are the taxonomy, systematics and phylogeny of Sphingidae (and other bombycoids), their distribution, host plants and parasitoids.

Walter Jetz is interested in understanding and predicting biodiversity.

Peter Nagel is interested in insect biogeography and African ecology.

Jan Beck has a research focus on the ecology, diversity and distribution patterns of insects.

L.B.M., J.B. and W.J. conceived and designed the study, L.B.M. and J.B. carried out the statistical analyses. I.J.K. and J.B. provided data. All authors contributed to writing the paper.

Editor: José Alexandre Diniz-Filho