Journal of Biogeography (J. Biogeogr.) (2013) 40, 1415–1426

SPECIAL PAPER

Geographical sampling bias in a large distributional database and its effects on species richness–environment models

Wenjing Yang[1,2,3], Keping Ma[1]* and Holger Kreft[2]*

[1]State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, 100093, Beijing, China, [2]Biodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences and Forest Ecology, University of Göttingen, 37077, Göttingen, Germany, [3]Graduate University of Chinese Academy of Sciences, 100049, Beijing, China

*Correspondence: Keping Ma, State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, 20 Nanxincun, Xiangshan, 100093 Beijing, China.
E-mail: kpma@ibcas.ac.cn
Holger Kreft, Biodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences and Forest Ecology, University of Göttingen, Büsgenweg 1, 37077 Göttingen, Germany.
E-mail: hkreft@uni-goettingen.de

ABSTRACT

Aim Recent advances in the availability of species distributional and high-resolution environmental data have facilitated the investigation of species richness–environment relationships. However, even exhaustive distributional databases are prone to geographical sampling bias. We aim to quantify the inventory incompleteness of vascular plant data across 2377 Chinese counties and to test whether inventory incompleteness affects the analysis of richness–environment relationships and spatial predictions of species richness.

Location China.

Methods We used the most comprehensive database of Chinese vascular plants, which includes county-level occurrences for 29,012 native species derived from 4,236,768 specimen and literature records. For each county, we computed smoothed species accumulation curves and used the mean slope of the last 10% of the curves as a proxy for inventory incompleteness. We created a series of data subsets with different levels of inventory incompleteness by excluding successively more under-sampled counties from the full data set. We then applied spatial and non-spatial regression models to each of these subsets to investigate relationships between the species richness of subsets and environmental factors, and to predict spatial patterns of vascular plant species richness in China.

Results Log_{10}-transformed numbers of records and documented species were strongly correlated ($r = 0.97$). In total, 91% of Chinese counties were identified as under-sampled. After controlling for inventory incompleteness, the overall explanatory power of environmental factors markedly increased, and the strongest predictor of species richness switched from elevational range to annual wet days. Environmental models calibrated with more complete inventories yielded better spatial predictions of species richness.

Main conclusions Our results indicate that inventory incompleteness strongly affects the explanatory power of environmental factors, the main determinants of species richness obtained from regression analyses, and the reliability of environment-based spatial predictions of species richness. We conclude that even large distributional databases are prone to geographical sampling bias, with far-reaching implications for the perception of and inferences about macroecological patterns.

Keywords

Biodiversity patterns, Chao1, China, inventory incompleteness, richness–environment relationship, sampling effort, species accumulation curve, species richness prediction, vascular plants, Wallacean shortfall.

Journal of Biogeography

INTRODUCTION

Efforts to explain the geographical variation in species richness have attracted enormous interest in ecology and biogeography (Hawkins *et al.*, 2003). Particularly in the past two decades, a growing body of literature has documented spatial patterns in biodiversity and their relationships with contemporary and historical abiotic factors (Kreft & Jetz, 2007; Field *et al.*, 2008). This scientific interest is partly attributable to concerns about the current status of the world's species and their potential responses to climate and land use changes (Boakes *et al.*, 2010). The availability of vast amounts of distributional data from online databases, high-resolution environmental data (e.g. Hijmans *et al.*, 2005) and novel analytical techniques (e.g. Dormann *et al.*, 2007) has facilitated rapid development in the analysis of richness–environment relationships.

Many hypotheses have been proposed to explain geographical variation in species richness (Currie, 1991; Ricklefs, 2004; Mittelbach *et al.*, 2007). Several hypotheses postulate that species richness across broad scales is controlled mainly by contemporary climate, especially the availability of water and energy, by means of its effects on productivity or its interaction with the physiological tolerances of species (Francis & Currie, 2003; Kreft & Jetz, 2007). Another hypothesis is that environmental heterogeneity may lead to increased species richness by permitting more species to coexist in diverse habitats or by causing accelerated diversification rates in environmentally complex regions (Kerr & Packer, 1997; Rahbek & Graves, 2001). Other hypotheses suggest that geological or climatic histories play important roles in determining patterns of species richness by determining rates of evolution and extinction (Latham & Ricklefs, 1993; Qian & Ricklefs, 2000; Sandel *et al.*, 2011) or postglacial dispersal limitation (Svenning & Skov, 2005). Hypothesis testing in macroecology usually employs regression models to investigate relationships between species richness and potential determinants. However, model fit can be affected by many confounding factors, including the quality of species richness data and choice of explanatory variables.

Primary distributional databases are becoming publicly available at an unprecedented rate (Krishtalka & Humphrey, 2000; Soberón & Peterson, 2004) and are heavily used in macroecological studies. A common shortcoming in current databases is that sampling effort is not uniform in space (Hortal *et al.*, 2007; Soberón *et al.*, 2007). On the one hand, the ranges of many species are not fully documented owing to geographical sampling bias (known as the 'Wallacean shortfall'; Lomolino, 2004). On the other hand, regional species inventories are often incomplete, termed 'inventory incompleteness' in this study. Previous studies have shown that geographical sampling bias may lead to distorted spatial patterns of biodiversity (Hortal *et al.*, 2007; Boakes *et al.*, 2010; Ballesteros-Mejia *et al.*, 2013). It is thus important to investigate whether richness–environment relationships can be correctly perceived and whether spatial patterns of species richness can be reliably predicted from environmental models without considering the bias in species richness data.

China has one of the richest national floras in the world, harbouring four global biodiversity hotspots (Mittermeier *et al.*, 2005) and 31,847 native species of vascular plants (Wang *et al.*, 2011a). This enormous diversity is due to its large area and high environmental variability, which includes boreal, temperate, subtropical and tropical biomes, and complex topography and geological history (Axelrod *et al.*, 1996). Millions of specimens have been collected and stored in Chinese herbaria over the past *c.* 110 years. In 2005, a project of specimen digitization was started involving the majority of Chinese herbaria. We based our study on the specimen data that are now available through online databases as a result of this digitization effort. In addition, we integrated records from a wide range of literature to produce a working database, which to our knowledge represents the most extensive compilation of records for Chinese vascular plants and one of the largest regional plant databases in the world. Owing to its extensive geographical and taxonomic coverage, this database is suitable for studying data bias and its potential effects on macroecological analyses.

In this study, we aim to quantify the spatial pattern of inventory incompleteness across Chinese counties and to test whether inventory incompleteness affects the analysis of richness–environment relationships. Specifically, we ask the following questions. (1) What proportion of Chinese counties is under-sampled? (2) Are the explanatory power of richness–environment models, inferences about core determinants, and spatial predictions of species richness affected by incompleteness of sampling? (3) If so, how can we control for under-sampling to ensure the robustness and reliability of macroecological inferences?

MATERIALS AND METHODS

Species distributional data

We obtained information for *c.* 6.5 million specimens of Chinese vascular plants from the *Chinese Virtual Herbarium* (http://www.cvh.org.cn/cms/en/, accessed December 2008) and the *Chinese Educational Specimen Resource Centre* (http://mnh.scu.edu.cn/, accessed January 2009). These specimen data have been sourced from 42 major Chinese herbaria. In addition, we assembled *c.* 2.5 million species records from *c.* 500 national and provincial floras as well as from local survey reports.

To improve the quality of the database, we performed a data cleaning process involving the following steps. (1) All records not determined to species level or not geo-referenced to county level were excluded. (2) Scientific names were standardized according to the *Catalogue of Life: Higher Plants in China* (http://www.cnpc.ac.cn, accessed January 2009; Wang *et al.*, 2011a). (3) Infraspecific taxa were merged to species level. (4) Multiple entries referring to the same specimen were removed. (5) Exotic species, as defined according

to the classifications in national floras (Editorial Committee of Flora Reipublicae Popularis Sinicae, 1959–2004; Wu *et al.*, 1994–2011), were excluded from the analysis. (6) Records of species native to China but occurring outside their natural range were excluded from the database. Natural ranges at the province level were obtained from the national floras.

This process left 4,236,768 unique records (85% derived from specimens and 15% from the literature) with county-level distribution information in the database, including 29,012 native species (91% of all vascular plants in China), with a mean of 146 records per species (Appendix S1: Fig. S1a in the Supporting Information). China is divided politically into 2377 counties (Fig. 1), with a mean size of 4138 km^2. Our database covered 2315 counties, accounting for 97% of all counties and 98% of China's land area. We calculated the Pearson correlation coefficient between the $\log_{10}$-transformed number of records and documented species richness (i.e. the number of species documented in the database) per county. Statistical significance was examined based on geographically effective degrees of freedom (Dutilleul *et al.*, 1993). The same method was used in subsequent analyses.

Environmental factors

We focused on four core environmental factors that have previously been shown to be strongly correlated with plant species richness (Francis & Currie, 2003; Kreft & Jetz, 2007; Kreft *et al.*, 2010): area, elevational range, potential evapotranspiration (PET), and annual wet days. We also investigated other variables such as mean annual temperature, annual precipitation, and actual evapotranspiration (AET), but excluded them to avoid multicollinearity (Pearson's $r > 0.7$; Appendix S1: Table S1). Area sizes of counties were obtained from the *National Fundamental Geographic Information System of China* (http://nfgis.nsdi.gov.cn/nfgis/english/default.htm, accessed May 2007). Maximum elevational range within each county was used as a surrogate for topographic complexity and habitat diversity (Kerr & Packer, 1997; Rahbek & Graves, 2001) and calculated from the GTOPO-30 digital elevation model (US Geological Survey, 1996) at a spatial resolution of 30 arc-seconds. PET was used as a measure of ambient energy (Francis & Currie, 2003) and was derived from the Global Evapotranspiration and Water Balance Data Sets at a spatial resolution of 0.5° (Ahn & Tateishi, 1994). Annual wet days were extracted from a global high-resolution data set of climate at a spatial resolution of 10 arc-minutes (New *et al.*, 2002). For each county, spatial averages were calculated for climatic variables. Species richness and all environmental variables were $\log_{10}$-transformed to meet the assumption of normality of model residuals and to improve the linearity of models.

Species richness estimation using Chao1

Documented species richness was assumed to be consistently lower than actual richness. We employed Chao1 to estimate the actual species richness for each county. Chao1 is an abundance-based estimator that emphasizes the occurrences of 'rare' species in a sample, that is, the species represented

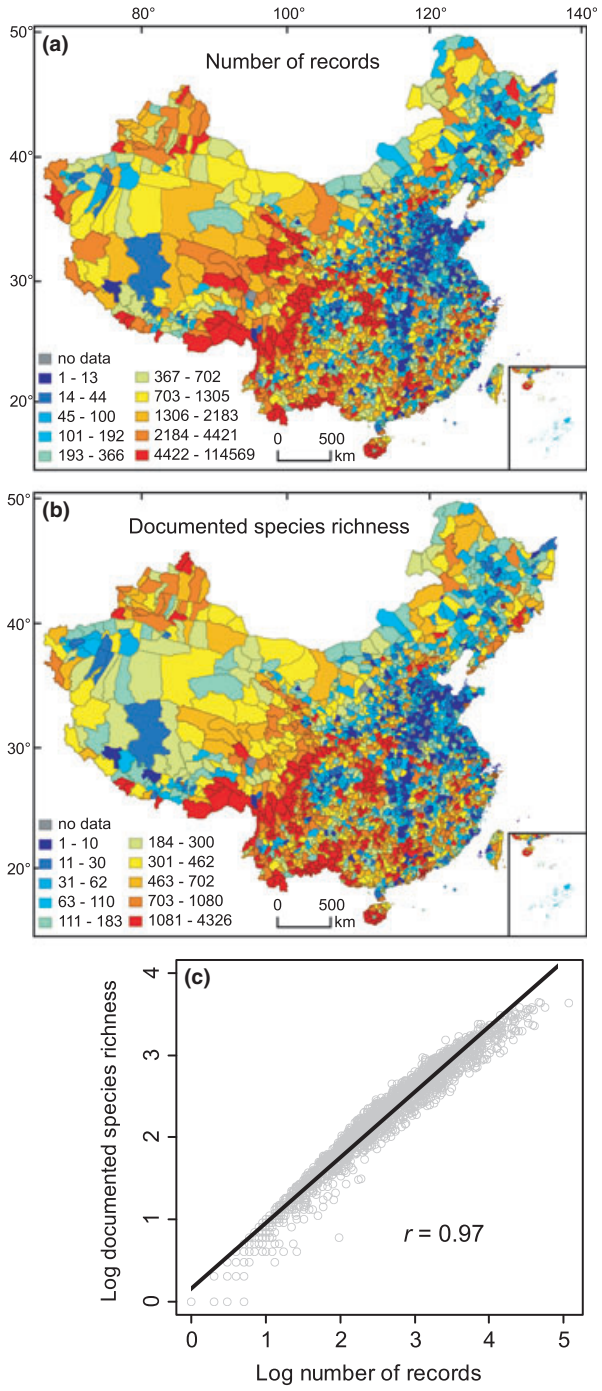

Figure 1 (a) Number of records and (b) documented species richness of vascular plants in 2315 counties of China, generated from 4,236,768 records referring to 29,012 species. Insets in the bottom right of figures show the south boundary of China, including all islands in the South China Sea. Legends are in quantile classification. Maps are in Albers projection. (c) Correlation between the $\log_{10}$-transformed number of records and documented species richness per county. The black line shows a linear fit.

by only one (singletons) or two (doubletons) records (Chao, 1984; Colwell & Coddington, 1994). We also compared other estimators such as jackknife and bootstrap, which yielded qualitatively very similar results. To validate Chao1 estimates, we compared the estimated richness per county with species numbers derived from an independent data set of checklists for 86 nature reserves that were nested within single counties (Appendix S1: Table S2 & Fig. S2). Considering that reserves were smaller than counties, we expected Chao1 estimates to be equal to or higher than the values from reserve checklists.

Inventory incompleteness assessment

Two methods were used to assess the inventory incompleteness of Chinese counties. First, we calculated the ratio between the documented and the Chao1-estimated species richness for each county (Soberón *et al.*, 2007; Soria-Auza & Kessler, 2008). Second, we used the curvilinearity of smoothed species accumulation curves (SACs) (Tittensor *et al.*, 2010). This method is based on the fact that SACs of poorly sampled counties tend towards a straight line, while those of better sampled counties have a higher degree of curvature (Fig. 2b–e; Gotelli & Colwell, 2001). Smoothed SACs give the expected species richness for a certain number of records and were calculated with the method 'exact' of the function 'specaccum' in the R package VEGAN (Oksanen *et al.*, 2011). The application of SACs to our data is based on the assumptions that records are sampled randomly and that species occurrences are neither spatially nor temporally autocorrelated (Colwell & Coddington, 1994; Gotelli & Colwell, 2001). The mean slope of the last 10% of SACs reflects the degree of curvilinearity and was used as a proxy for inventory incompleteness (Fig. 2b–e). Shallow slopes (values close to zero) indicate saturation in the sampling and thus low levels of incompleteness, whereas steep slopes (values close to one) reflect high levels of incompleteness. Slope values can also be interpreted as the probability of discovering new species when sampling continues in the county (Dahl *et al.*, 2009). Here, we considered counties with slope values ≤ 0.05 as well sampled and those with slope values > 0.05 as under-sampled.

Data subsets with different level of inventory incompleteness

We created subsets of the data by successively excluding under-sampled counties (SAC slope > 0.05) from the full data set. The sequence of excluding counties was determined by SAC slope values; that is, less complete counties were excluded first. Each time we excluded one or several counties, we regarded the remaining counties as a subset (1485 subsets in total). Overall, the intention was to gradually minimize the inventory incompleteness of successive subsets.

Environmental representativeness of data subsets

To investigate how environmental factors varied by subset, we calculated summary statistics of environmental factors for each subset. We further investigated the environmental representativeness of each subset as compared with the full data set. Principal components analysis (PCA) was applied to three environmental factors (elevational range, PET and annual wet days). We created a kernel density surface on counties of each subset in a biplot space depicted by the first two principal components by using the function 'kde2d' in the R package MASS (Venables & Ripley, 2002). Kernel density captures both the extent and the spatial structure (i.e. aggregation and dispersion tendencies) of counties in the

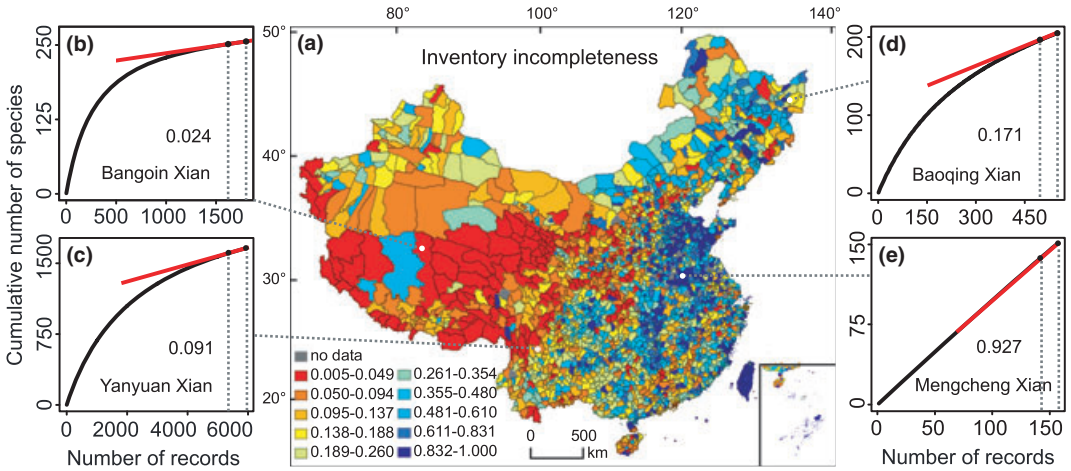

Figure 2 (a) Inventory incompleteness of vascular plants in 2315 counties of China. High values (in blue) indicate high levels of inventory incompleteness. The inset in the bottom right of the figure shows the south boundary of China, including all islands in the South China Sea. The legend is in quantile classification. The map is in Albers projection. (b)–(e) Species accumulation curves with different curvilinearity of four exemplary counties. The parts between the two black points indicate the last 10% of the species accumulation curves. Red lines and numbers indicate slopes. Confidence intervals of the curves are very narrow and are omitted for visual clarity.

space. The environmental representativeness of each subset was calculated as the Pearson correlation between kernel density surfaces of the subset and of the full data set within the PCA space.

Species richness–environment models

We first investigated the relationships between single environmental variables and species richness by applying simple ordinary least squares regressions (OLS) to all data subsets. We then performed multi-predictor OLS and explored the overall explanatory power of the four target predictors. Moran's I correlograms and global Moran's I values were used to evaluate the pattern and strength of spatial autocorrelation in model residuals (Dormann *et al.*, 2007). Strong and significant spatial autocorrelation was found among the residuals of OLS models (Appendix S2: Fig. S1). Spatial autocorrelation might inflate type I error rates and bias parameter estimates (Dormann *et al.*, 2007). We thus additionally employed spatial simultaneous autoregressive (SAR) models. Model selection was based on the Akaike information criterion (AIC) (Johnson & Omland, 2004). SAR models of the error type were chosen with a lag distance of 300 km and weighted neighbourhood structure. The selection of lag distance was based on the trade-off between AIC values and the number of counties having no neighbours within the distance class (Appendix S2: Fig. S2). The optimal lag distance varied slightly among subsets owing to different neighbourhood structures, but we opted to take a lag distance of 300 km for all models to make them comparable. The application of spatial models significantly reduced the spatial autocorrelation of SAR model residuals for both the full data set and the subset including only well-sampled counties (SAC slope $\leq$ 0.05; Appendix S2: Fig. S1). Pseudo-R^2 (hereafter R^2) values for SAR models were calculated as the squared Pearson correlation between predicted and observed values (Kissling & Carl, 2008). R^2 values were plotted against the inventory incompleteness (indicated by SAC slope values) of the most incomplete county in each subset.

We investigated the relative importance of each variable in OLS and SAR multi-predictor models using the function 'calc.relimp' with metric 'pmvd' in the R package RELAIMPO (Grömping, 2006). The metric 'pmvd' calculates a weighted average of sequential R^2 values over all possible models. The adjusted R^2 (hereafter R^2) value of each OLS model was partitioned into relative proportions (proportional R^2) explained by each environmental variable. The relative proportions were then multiplied by the R^2 of the model to obtain the absolute fraction of R^2 value explained by a particular variable. To account for spatial autocorrelation, we first performed a standard SAR model, then removed the spatial component of the fitted values and entered richness excluding the spatial component as a new response variable in the R^2 partitioning procedure (Belmaker & Jetz, 2011).

To test whether the varying number of counties in subsets had an effect on model parameterization and fit, we created simulated subsets in which the number of counties was held equal to the numbers in actual subsets while the counties were randomly drawn from the pool of counties. Multi-predictor OLS was applied to the simulated subsets. Means and 95% confidence intervals of R^2 values for null models were obtained from 1000 permutations.

Spatial predictions of species richness

OLS models with all four environmental variables were used to predict the spatial pattern of vascular plant species richness in China. We parameterized the model separately with the documented species richness of the full data set and of the well-sampled counties. We also parameterized the model with the Chao1-estimated species richness of well-sampled counties to account for under-sampling in the counties. We excluded six island counties in the South China Sea from predictions owing to a lack of climate data and because islands show different richness–environment relationships from those on the mainland (Kreft *et al.*, 2008).

We conducted a cross-validation test to evaluate the predictive performance of the models calibrated by different data sets (Hortal *et al.*, 2007). A data set was randomly split 100 times into two subsets, with 85% of the counties used to calibrate the model and the remaining 15% to validate the result. The prediction error of models was calculated as

$$\frac{\sum_{k=1}^{n} \frac{(DR_k - PR_k)}{DR_k}}{n},$$

where DR and PR are the documented and predicted species richness in each county, respectively, and n is the number of counties.

RESULTS

Spatial patterns of records, richness and incompleteness

The number of records per county ranged from 0 to 114,569 with a mean of 1835 records (Fig. 1a, Appendix S1: Fig. S1b). The documented species richness ranged from 0 to 4328 species per county with a mean of 411 (Fig. 1b, Appendix S1: Fig. S1c). The $\log_{10}$-transformed numbers of records and documented species per county were strongly correlated (Pearson's $r = 0.97$, $P < 0.001$; Fig. 1c).

Inventory incompleteness of counties assessed by the slope at the last 10% of SACs was strongly and negatively correlated with the number of records (Pearson's $r = -0.80$, $P < 0.001$). Values of inventory incompleteness ranged from 0.005 to 1, with a mean of 0.35 (Fig. 2a, Appendix S1: Fig. S1d). A total of 216 counties (SAC slope $\leq$ 0.05), accounting for 9% of all counties and 21% of China's land area, were identified as well sampled. This pattern shows a strong correlation with the pattern of an alternative measure of inventory incompleteness expressed by the ratio between documented and Chao1-estimated richness (Pearson's

$r = 0.91$, $P < 0.001$). This suggests that the assessment of inventory incompleteness is not sensitive to the measure used.

Species richness estimation using Chao1

The Chao1-estimated species richnesses of all well-sampled counties with nature reserve checklist data ($n = 33$), except for two with highly even distributions of individuals among species in the samples, were higher than the species richnesses of the nature reserves nested within those counties (Fig. 3). However, the estimated richnesses of 37 under-sampled counties with checklist data ($n = 53$) were lower than the species richnesses of nature reserves, suggesting that the reliability of Chao1 estimation is affected by the completeness of sampling.

Species richness–environment models

The multi-predictor OLS model explained 23% of the variance in the species richness of the full data set. Elevational range contributed 55% to the explained variance and was the most important variable in the model (Fig. 4). When under-sampled counties were gradually excluded, the explanatory power of environmental factors strongly and consistently increased. A model including only the well-sampled counties explained 57% of the variance in species richness. The contribution of elevational range decreased to 16%, whereas annual wet days became the most important variable in the model, with a contribution increasing from 36% to 53%. Similar results with slightly better model fits were obtained from SAR models (Appendix S3: Fig. S1). The 95% confidence intervals of R^2 values of the null models became wider as county numbers in simulated subsets decreased, but the means did not vary. Elevational range was the strongest predictor for 87% of random subsets where the number of counties was the same as the number of well-sampled counties ($n = 216$), whereas annual wet days was the most important variable for the other 13%. Results from null models suggest that model fits and identification of core determinants were determined by the counties rather than by the number of counties in the analysis. Consistent with results from multi-predictor models, univariate analyses revealed that elevational range was the strongest predictor ($r^2 = 0.14$) for the species richness of the full data set (Appendix S3: Fig. S2), whereas annual wet days was the strongest predictor ($r^2 = 0.37$) for well-sampled counties.

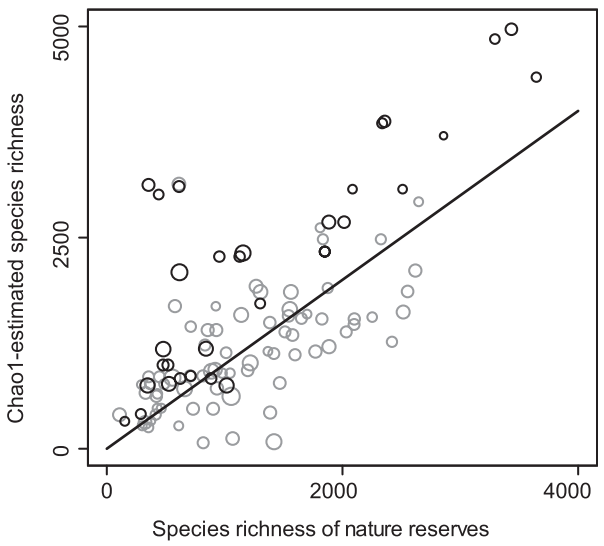

Figure 3 Comparison between the Chao1-estimated vascular plant species richness of 86 Chinese counties and species richness derived from an independent data set consisting of checklists for the same number of nature reserves each fully nested within a single county. Black circles represent 33 well-sampled counties (species accumulation curve slope ≤ 0.05), and grey circles represent 53 under-sampled counties (species accumulation curve slope > 0.05). Bigger circles indicate the nature reserves are closer in size to the respective counties. The black line shows the 1:1 fit.

Environmental representativeness of subsets

The environmental characteristics of subsets shifted when undersampled counties were successively excluded (Fig. 5a–d), and the variation was particularly significant for elevational range and PET. The mean elevational range increased from 1431 to 2252 m, while the average PET decreased from 764 to 496 mm year^{-1}, indicating that well-sampled counties have a larger elevational range and lower PET than other counties. Kernel density surfaces of well-sampled counties and the full data set were strongly correlated (Pearson's $r = 0.73$, $P < 0.001$; Fig. 5e, f).

Spatial predictions of species richness

The predicted species richness based on the full data set ranged from 42 to 992 species per county, with a mean of 351 (Fig. 6a, Appendix S3: Table S1), while the prediction based on well-sampled counties varied from 45 to 4782 species per county, with an average of 1067 (Fig. 6b, Appendix S3: Table S1). With prediction errors of 0.35 as compared with 0.69, the prediction based on well-sampled counties was much more plausible than that based on the full data set. The prediction from the model calibrated by the Chao1-estimated richness of well-sampled counties ranged from 31 to 5207 species per county (mean = 1206; Fig. 6c, Appendix S3: Table S1). A prediction error of 0.33 indicated that this prediction was the best of the three.

DISCUSSION

Species distribution information is becoming increasingly available through online databases (Graham *et al.*, 2004; Soberón *et al.*, 2007; Jetz *et al.*, 2012). However, these data often suffer from significant bias in the spatial distribution of sampling effort (Meier & Dikow, 2004; Soberón & Peter-

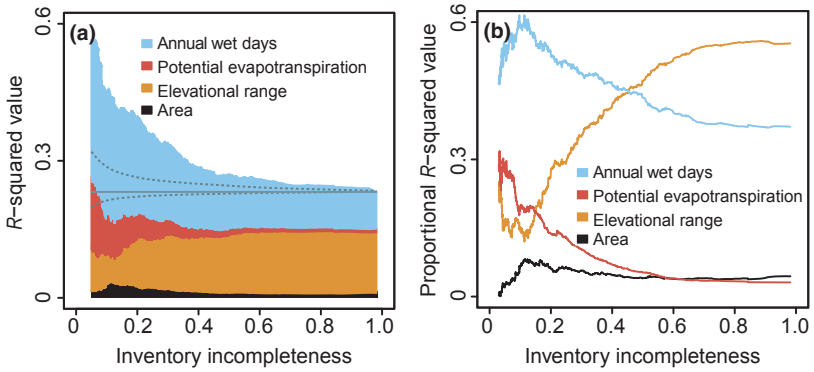

Figure 4 (a) Explanatory power of multi-predictor ordinary least squares (OLS) models and (b) proportional R^2 of environmental variables on vascular plant species richness of 1485 subsets of Chinese counties with different levels of inventory incompleteness. (a) The coloured area represents the total variance in species richness explained for each subset. Areas with specific colours indicate the hierarchical partitioning of total variance to the relative contribution of each environmental factor. Grey solid and dotted lines indicate means and 95% confidence intervals of R^2 values of null models, respectively. Null models represent results from 1000 permutations while keeping the number of counties constant as in the respective subset.

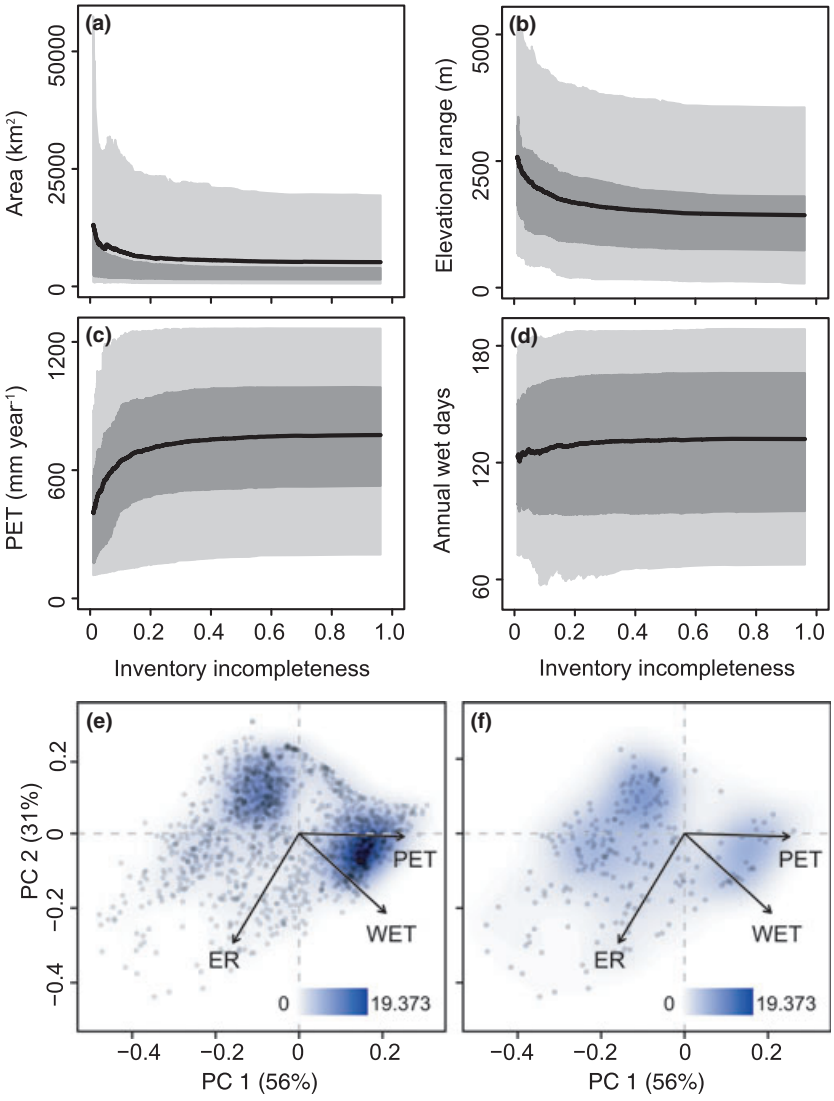

Figure 5 (a)–(d) Environmental characteristics of 1485 subsets of Chinese counties with different levels of inventory incompleteness of vascular plants. Black lines indicate means. Dark grey areas indicate the 25% and 75% percentiles, and light grey areas indicate the 5% and 95% percentiles. (a) Area (km^2); (b) elevational range (m); (c) potential evapotranspiration (mm year^{-1}); (d) annual wet days (number year^{-1}). Biplots showing the distribution of (e) all counties and (f) well-sampled counties in a space depicted by the first two principal components of a principal component analysis (PCA) applied to three environmental factors: ER, elevational range; PET, potential evapotranspiration; WET, annual wet days. Grey points represent counties. The blue cloud shows the kernel density surface on the counties of each data set. The intensity of blue colour indicates the sum of kernel densities contributed by each data point.

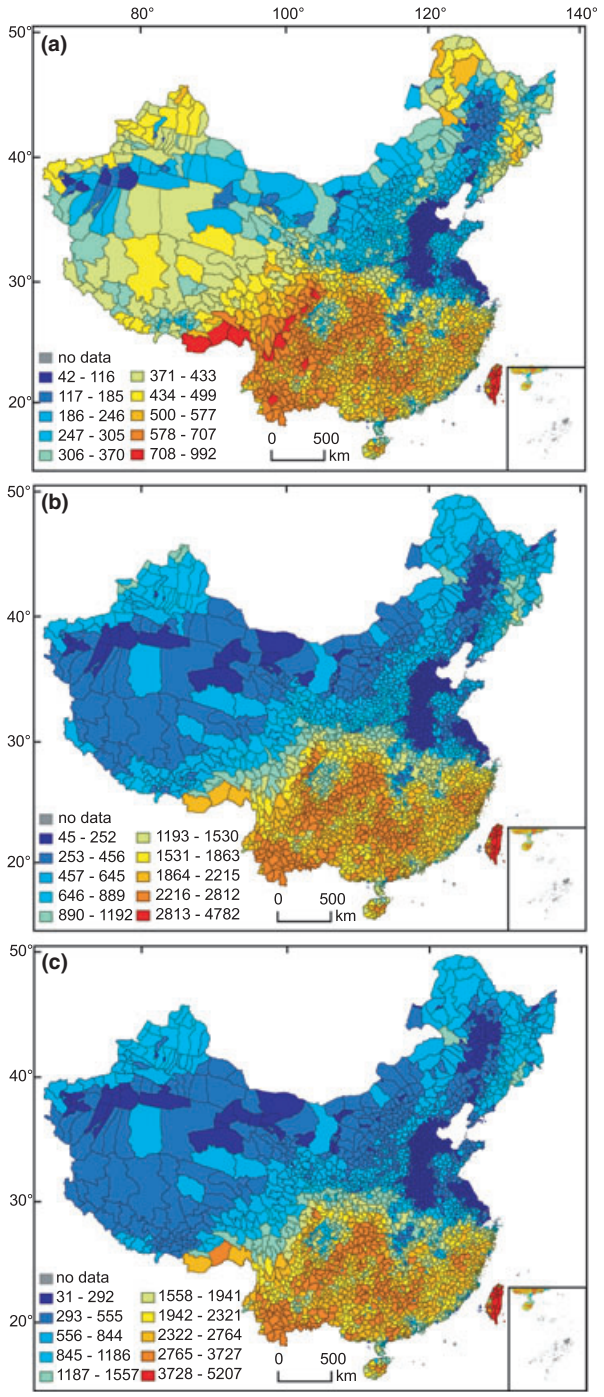

Figure 6 Spatial predictions of vascular plant species richness in 2371 counties of China derived from multi-predictor ordinary least squares (OLS) models calibrated by (a) the documented species richness of all counties, (b) the documented, and (c) Chao1-estimated species richness of well-sampled counties. Insets in the bottom right of figures show the south boundary of China, including all islands in the South China Sea. Legends are in quantile classification. Maps are in Albers projection.

son, 2004; Hortal *et al.*, 2007), which may affect macroecological analyses and inferences. Using a large distributional database of Chinese vascular plants, we demonstrate that geographical sampling bias has strong effects on the perception of species richness patterns in commonly applied analyses of richness–environment relationships and on the spatial predictions of species richness.

What proportion of Chinese counties is under-sampled?

According to our assessment, 91% ($n = 2161$) of Chinese counties are under-sampled (Fig. 2a). The high incompleteness of sampling at the county level may be due to the relatively short history of intensive floristic surveys in China and to the limited number of collections given the country's large size and enormous floristic diversity (Wang *et al.*, 2011a). To our knowledge, few surveys have been conducted specifically for the purpose of completing county inventories. Rather, collecting effort has been concentrated on regions with the highest species richness or number of endemics (e.g. Hengduan Mountains), with less diverse regions having a considerably lower collection intensity (e.g. North China Plain; Fig. 1). However, complete distribution information at finer scales could provide deeper insights into the processes that create and maintain biodiversity patterns (Beck *et al.*, 2012). Under-sampled counties in south-eastern China and the North China Plain should be particularly emphasized in the future. However, it should be noted that reaching a reasonable degree of completeness requires a tremendous effort. Assuming that a minimum of 5914 specimens should be collected for each county (i.e. the average number of specimens of well-sampled counties), a roughly estimated total of 10 million specimens are still required to accomplish a near-complete survey of all counties.

Does geographical sampling bias impair macroecological inferences?

We found that the R^2 values of multi-predictor models strongly and consistently increased when under-sampled counties were successively excluded from the analysis (Fig. 4, Appendix S3: Fig. S1). The overall effect of environmental predictors was severely underestimated when all counties were included, ignoring the bias in species richness data. Importantly, the strongest single predictor of species richness switched from elevational range to annual wet days when excluding under-sampled counties (Fig. 4, Appendix S3: Figs S1 & S2). The correlation between elevational range and species richness of the full data set may be inflated by the geographical sampling bias that mountainous regions have more collections, resulting in higher documented numbers of species (Fig. 1). Annual wet days and PET appeared to be the most important predictors for the species richness of well-sampled counties and together accounted for 44% of the variance (Fig. 4). These results underscore how difficult it may be in hypothesis testing to distinguish between the importance of water, energy and environmental heterogeneity when geographical sampling

bias in species richness data varies non-randomly with environment and is not accounted for.

Environmental models calibrated by only the least biased counties yielded more reliable spatial predictions of species richness (Fig. 6). In the prediction based on the full data set, the species richness of many counties was significantly underestimated, particularly in south-western China, where it is expected to be highest (Fig. 6a). Spatial predictions based on well-sampled counties appeared to be much more plausible, with species richness high in mountainous regions and decreasing significantly from south to north (Fig. 6b, c). This is consistent with previous studies on spatial patterns of Chinese woody plants (Wang *et al.*, 2011b) as well as with global predictions (Kreft & Jetz, 2007). Nevertheless, species richness in glacial refugia or centres of diversification (López-Pujol *et al.*, 2011; Huang *et al.*, 2012) is likely to be underestimated by a model considering only contemporary environment unless historical patterns co-vary with contemporary conditions. The prediction is likely to be improved by including further factors representing climatic and geological histories (Qian & Ricklefs, 2000; Sandel *et al.*, 2011).

How to control for geographical sampling bias to ensure the robustness and reliability of macroecological inferences

Macroecological analyses are likely to be affected when strong geographical bias exists in distributional data. Our study offers an approach to observe potential shifts in environmental predictors along an inventory incompleteness gradient and to identify well-sampled sampling units. To ensure the robustness of results, we suggest including in analyses only well-sampled units. However, whether the inferences can be scaled up may depend on how well the well-sampled units represent the whole study area. In our study, PCA and kernel density analyses indicated that *c.* 73% of the environmental variance (expressed by elevational range, PET and annual wet days) in China is still covered by the well-sampled counties (Fig. 5e, f). The conclusions based on this subset are consistent with findings from previous studies that water availability and ambient energy rather than topographic complexity are the main determinants for vascular plant diversity in China (e.g. Wang *et al.*, 2011b). It has been widely recognized that the latitudinal gradient, namely plant diversity decreasing from south to north, is determined mainly by ambient energy, while plant diversity from east to west is controlled largely by water availability (Wang, 1992). Our study provides an example of how to utilize an imperfect distributional database for fundamental macroecological research (cf. Ballesteros-Mejia *et al.*, 2013).

Comparison of two methods for assessing inventory incompleteness

A number of previous studies used the ratio between documented and 'true' species richness as a proxy for inventory incompleteness (Soberón *et al.*, 2007; Mora *et al.*, 2008; Soria-Auza & Kessler, 2008). To this end, the 'true' species richness of each sampling unit is commonly estimated based on the existing sampling and application of species richness estimators (Chao, 1984; Colwell & Coddington, 1994) or statistical models (Bebber *et al.*, 2007). However, this estimation might not be consistently reliable and accurate, and is limited by spatial and temporal variation in sampling effort (Fig. 3). In contrast, our method does not estimate the 'true' species richness. Instead, we compute the slope at the tail of SACs to represent the rate at which new species are added to the inventory with continuing sampling. This method appears to be more robust for counties with few records, where commonly applied species richness estimators produce unrealistic results. However, this method is not without drawbacks. For instance, the probability of finding new species depends not only on the number of undiscovered species, but also on factors such as species abundance distributions that might vary considerably between sampling units (Mora *et al.*, 2008). Despite the different principles of the two methods, the patterns of inventory incompleteness of Chinese counties obtained from them are strongly correlated ($r = 0.91$, $P < 0.001$), suggesting that the pattern and rankings of inventory incompleteness are relatively robust to the choice of method (Fig. 2).

In conclusion, strong geographical sampling bias is found in this database, which hampers an unbiased perception of spatial patterns of vascular plant species richness in China. Macroecological analyses based on databases with strong geographical bias are likely to yield an underestimation of environmental effects, a misidentification of the main determinants of species richness patterns, and unreliable spatial predictions of species richness. Our study highlights the importance of carefully evaluating data quality before using such information for theoretical and practical applications. In addition, our results clearly show the urgent need for continuing intensive and targeted field surveys, particularly in poorly explored regions.

ACKNOWLEDGEMENTS

Data compilation was carried out within the project 'China National Specimen Information Infrastructure of the National Science and Technology Resource Platform' (2005DKA21401), funded by the Ministry of Science and Technology of China. We are grateful to B. Chen, T.M. Chen, J.L. Zhang and L.S. Wang for help with data preparation. We thank Z.H. Wang and W. Jetz for valuable comments on preliminary results. We also thank P. Weigelt, Y. Kisel, R. Soria-Auza and J.H. Huang for technical assistance and valuable suggestions. We are grateful for helpful comments provided by the handling editor W. Daniel Kissling and three referees. W.Y. was financially supported by the China Scholarship Council as a visiting PhD student in H.K.'s lab at the University of Göttingen. H.K. acknowledges funding from the German Initiative of Excellence of the German Research Foundation (DFG).

REFERENCES

Ahn, C.H. & Tateishi, R. (1994) Development of a global 30-minute grid potential evapotranspiration data set. *Journal of the Japanese Society of Photogrammetry and Remote Sensing*, **33**, 12–21.

Axelrod, D.I., Al-Shehbaz, I. & Raven, P.H. (1996) History of the modern flora of China. *Floristic characteristics and diversity of East Asian plants* (ed. by A. Zhang and S. Wu), pp. 43–55. China Higher Education Press, Beijing.

Ballesteros-Mejia, L., Kitching, I.J., Jetz, W., Nagel, P. & Beck, J. (2013) Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. *Global Ecology and Biogeography*, doi:10.1111/geb.12039.

Bebber, D.P., Marriott, F.H.C., Gaston, K.J., Harris, S.A. & Scotland, R.W. (2007) Predicting unknown species numbers using discovery curves. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1651–1658.

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

Belmaker, J. & Jetz, W. (2011) Cross-scale variation in species richness–environment associations. *Global Ecology and Biogeography*, **20**, 464–474.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Ding, C., 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.

Chao, A. (1984) Nonparametric 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. (1991) Energy and large-scale patterns of animal- and plant-species richness. *The American Naturalist*, **137**, 27.

Dahl, C., Novotny, V., Moravec, J. & Richards, S.J. (2009) Beta diversity of frogs in the forests of New Guinea, Amazonia and Europe: contrasting tropical and temperate communities. *Journal of Biogeography*, **36**, 896–904.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M. & Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Dutilleul, P., Clifford, P., Richardson, S. & Hemon, D. (1993) Modifying the *t* test for assessing the correlation between two spatial processes. *Biometrics*, **49**, 305–314.

Editorial Committee of Flora Reipublicae Popularis Sinicae (1959–2004) *Flora Reipublicae Popularis Sinicae*. Science Press, Beijing.

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. (2008) Spatial species-richness gradients across scales: a meta-analysis. *Journal of Biogeography*, **36**, 132–147.

Francis, A.P. & Currie, D.J. (2003) A globally consistent richness-climate relationship for angiosperms. *The American Naturalist*, **161**, 523–536.

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

Graham, C.H., Ferrier, S., Huettman, 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.

Grömping, U. (2006) Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, **17**, 1–27.

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.M., Porter, E.E. & Turner, J.R.G. (2003) Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, **84**, 3105–3117.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hortal, J., Lobo, J.M. & Jiménez-Valverde, A. (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology*, **21**, 853–863.

Huang, J., Chen, B., Liu, C., Lai, J., Zhang, J. & Ma, K. (2012) Identifying hotspots of endemic woody seed plant diversity in China. *Diversity and Distributions*, **18**, 673–688.

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.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

Kerr, J.T. & Packer, L. (1997) Habitat heterogeneity as a determinant of mammal species richness in high-energy regions. *Nature*, **385**, 252–254.

Kissling, W.D. & Carl, G. (2008) Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography*, **17**, 59–71.

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

Kreft, H., Jetz, W., Mutke, J., Kier, G. & Barthlott, W. (2008) Global diversity of island floras from a macroecological perspective. *Ecology Letters*, **11**, 116–127.

Kreft, H., Jetz, W., Mutke, J. & Barthlott, W. (2010) Contrasting environmental and regional effects on global pteridophyte and seed plant diversity. *Ecography*, **33**, 408–419.

Krishtalka, L. & Humphrey, P.S. (2000) Can natural history museums capture the future? *BioScience*, **50**, 611–617.

Latham, R.E. & Ricklefs, R.E. (1993) Continental comparisons of temperate-zone tree species diversity. *Species diversity in ecological communities* (ed. by R.E. Ricklefs and D. Schluter), pp. 294–314. University of Chicago Press, Chicago.

Lomolino, M.V. (2004) Conservation biogeography. *Frontiers of biogeography: new directions in the geography of nature* (ed. by M.V. Lomolino and L.R. Heaney), pp. 293–296. Sinauer Associates, Sunderland, MA.

López-Pujol, J., Zhang, F., Sun, H., Ying, T. & Ge, S. (2011) Centres of plant endemism in China: places for survival or for speciation? *Journal of Biogeography*, **38**, 1267–1280.

Meier, R. & Dikow, T. (2004) Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conservation Biology*, **18**, 478–488.

Mittelbach, G.G., Schemske, D.W., Cornell, H.V. *et al.* (2007) Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecology Letters*, **10**, 315–331.

Mittermeier, R.A., Gil, P.R., Hoffmann, M., Pilgrim, J., Brooks, T., Mittermeier, C.G., Lamoreux, J. & Fonseca, G.A.B. (2005) *Hotspots revisited: Earth's biologically richest and most endangered terrestrial ecoregions*. University of Chicago Press, Chicago.

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.

New, M., Lister, D., Hulme, M. & Makin, I. (2002) A high-resolution data set of surface climate over global land areas. *Climate Research*, **21**, 1–25.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2011) *vegan: community ecology package*. R package version 1.17-8. Available at: http://vegan.r-forge.r-project.org/.

Qian, H. & Ricklefs, R. (2000) Large-scale processes and the Asian bias in species diversity of temperate plants. *Nature*, **407**, 180–182.

Rahbek, C. & Graves, G.R. (2001) Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences USA*, **98**, 4534–4539.

Ricklefs, R.E. (2004) A comprehensive framework for global patterns in biodiversity. *Ecology Letters*, **7**, 1–15.

Sandel, B., Arge, L., Dalsgaard, B., Davies, R.G., Gaston, K.J., Sutherland, W.J. & Svenning, J.C. (2011) The influence of Late Quaternary climate-change velocity on species endemism. *Science*, **334**, 660–664.

Soberón, J. & Peterson, T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **359**, 689–698.

Soberón, J., Jiménez, R., Golubov, J. & Koleff, P. (2007) Assessing completeness of biodiversity databases at different spatial scales. *Ecography*, **30**, 152–160.

Soria-Auza, R. & Kessler, M. (2008) The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: a case study of ferns from Bolivia. *Diversity and Distributions*, **14**, 123–130.

Svenning, J.-C. & Skov, F. (2005) The relative roles of environment and history as controls of tree species composition and richness in Europe. *Journal of Biogeography*, **32**, 1019–1033.

Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Vanden Berghe, E. & Worm, B. (2010) Global patterns and predictors of marine biodiversity across taxa. *Nature*, **466**, 1098–1103.

US Geological Survey (1996) *GTOPO30*. Available at: http://www1.gsi.go.jp/geowww/globalmap-gsi/gtopo30/gtopo30.html (accessed November 2010).

Venables, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*. Springer, New York.

Wang, H. (1992) *Floristic geography*. Science Press, Beijing.

Wang, L., Zhang, Y., Xue, N. & Qin, H. (2011a) Floristics of higher plants in China – report from Catalogue of Life: Higher Plants in China Database. *Plant Diversity and Resources*, **33**, 69–74.

Wang, Z., Fang, J., Tang, Z. & Lin, X. (2011b) Patterns, determinants and models of woody plant diversity in China. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 2122–2132.

Wu, Z., Raven, P.H. & Hong, D. (1994–2011) *Flora of China*. Missouri Botanical Garden Press, St. Louis, MO.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Detailed information of species distributional data, environmental variables and nature reserves used in this analysis.

Appendix S2 Moran's *I* correlograms and the selection of lag distances for simultaneous autoregressive models.

Appendix S3 R^2 values and coefficients of regression models.

BIOSKETCHES

Wenjing Yang is a PhD student with particular interests in the fields of biodiversity, biogeography and plant taxonomy. She specifically focuses on data bias in species distributional databases and its potential impact on biodiversity research.

Keping Ma is interested in biodiversity conservation, biogeography and biodiversity informatics. He is particularly interested in understanding mechanisms of species coexistence in forest communities, as well as biodiversity patterns and the underlying processes at broader scales.

Holger Kreft is interested in biogeographical and ecological patterns from local to global scales, particularly gradients of species richness and endemism. His research includes analyses of plant and vertebrate diversity, and island and conservation biogeography.

Author contributions: H.K. and W.Y. conceived the ideas; K.M. contributed the data; W.Y. and H.K. analysed the data; and W.Y., H.K. and K.M. led the writing.

Editor: W. Daniel Kissling