A Journal of Macroecology

Global Ecology and Biogeography

RESEARCH PAPER

Environmental and socio-economic factors shaping the geography of floristic collections in China

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, 20 Nanxincun, Xiangshan, 100093 Beijing, China, [2]Biodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences and Forest Ecology, University of Göttingen, Büsgenweg 1, 37077 Göttingen, Germany, [3]Ministry of Education Key Laboratory of Poyang Lake Wetland and Watershed Research, Jiangxi Normal University, Ziyang road 99, 330022 Nanchang, 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 Effort in collecting biodiversity information often varies strongly in space and may be driven by environmental, cultural and socio-economic factors. Understanding the constraints on collecting effort is crucial for identifying potential bias in distributional databases and for making future surveys more efficient. Here we test six competing hypotheses on drivers of geographical variation in collecting effort and identify the main factors shaping the geography of floristic collections in China.

Location China.

Methods We used the most comprehensive database of Chinese vascular plant distributions including 4,338,516 county-level occurrences derived from herbarium specimens and literature sources. Explanatory variables were assembled representing six different hypotheses: accessibility, human population density, the 'botanist effect', mountains, water availability and conservation priority. Ordinary least-squares models with eigenvector-based spatial filters were applied to investigate their effects on spatial patterns of two different facets of collecting effort, i.e. collection density and inventory incompleteness.

Results All hypotheses except accessibility and human population density received significant support. Elevational range was the strongest predictor with a positive effect on collection density. Inventory incompleteness in turn was best predicted by human population density, but unexpectedly showed a positive effect. In contrast to previous studies, collecting effort was only weakly and negatively related to road density. Counties with herbaria had significantly higher collecting effort, and the presence of herbaria had weakly positive effects on neighbouring counties.

Main conclusions Our results indicate that China's mountains are most intensively and completely collected, whereas densely populated areas are surprisingly under-sampled. Because densely populated areas are more seriously threatened by land-use change, our results show a need to increase biological collections in those areas for conservation assessment and monitoring. More generally, our study suggests that collecting effort and its environmental and socio-economic constraints have a strong region-specific component influenced by cultural context and by different botanical traditions.

Keywords

Biodiversity database, botanist effect, collecting effort, geographical sampling bias, specimen collection, vascular plants, Wallacean shortfall.

INTRODUCTION

Understanding spatial patterns of biodiversity, including the distribution of individual species and aggregated patterns like species richness and endemism, is crucial for biogeography, ecology and conservation management (Wilson, 1988; Gaston, 2000; Whittaker *et al.*, 2005). However, the uneven spatial distribution of collecting effort common in most databases may

DOI: 10.1111/geb.12225
http://wileyonlinelibrary.com/journal/geb

lead to a biased perception of patterns and drivers of biodiversity (Nelson *et al.*, 1990; Soberón & Peterson, 2004; Ballesteros-Mejia *et al.*, 2013; Yang *et al.*, 2013; Ficetola *et al.*, 2014). Previous studies have shown that the geography of biological collections is strongly related to environmental and socio-economic factors (Freitag *et al.*, 1998; Dennis & Thomas, 2000; Reddy & Dávalos, 2003; Sánchez-Fernández *et al.*, 2008). Understanding the determinants of spatial bias in distributional data should help to identify data limitations and to improve the efficiency of future field surveys (Dennis *et al.*, 1999; Romo *et al.*, 2006).

It has been reported that biological collections tend to be concentrated in areas near roads and navigable rivers (Reddy & Dávalos, 2003), suggesting that accessibility is an important factor influencing the selection of collecting sites – the 'highway effect' (Soberón *et al.*, 2000) or the 'road-map effect' (Crisp *et al.*, 2001). For example, plant collections in the inner parts of Australia have been largely restricted to the vicinity of a few main roads (Crisp *et al.*, 2001). This pattern has also been found for other taxa (e.g. frogs and passerine birds; Reddy & Dávalos, 2003; Botts *et al.*, 2011) and regions (e.g. South America and Thailand; Nelson *et al.*, 1990; Parnell *et al.*, 2003).

Furthermore, positive relationships have been found between human population density and the number of collections (Dennis *et al.*, 1999; Küper *et al.*, 2006; Tobler *et al.*, 2007; Ficetola *et al.*, 2014). Urban areas often show higher collecting effort than remote areas. For example, collecting sites for vascular plants in Thailand are concentrated in densely populated areas (Parnell *et al.*, 2003) – a pattern that may be driven by better infrastructure for transport and accommodation in densely populated areas.

Moerman & Estabrook (2006) found that species richness of flowering plants in the United States is higher in counties with universities, and argued that counties where botanists live are generally better sampled, leading to a higher number of documented species (the 'botanist effect'). Studies on butterflies (Dennis & Thomas, 2000) and water beetles (Sánchez-Fernández *et al.*, 2008) provided further support for the idea that collecting sites close to the homes of collectors are more frequently visited, suggesting that the distance to collectors' homes may affect collecting effort (Hortal *et al.*, 2004; Ahrends *et al.*, 2011).

Collecting effort may also be related to environmental factors. For Iberian butterflies and water beetles, areas along prominent mountain ranges and with high precipitation have the highest collecting efforts (Romo *et al.*, 2006; Sánchez-Fernández *et al.*, 2008). This relationship might be due to the fact that moist, mountainous areas usually harbour more species (Kreft & Jetz, 2007; Ruggiero & Hawkins, 2008), and collectors may focus their effort on such areas to maximize the number of species collected ('diversity tracking'; Romo *et al.*, 2006).

Several studies have indicated that collectors tend to concentrate their activities within protected areas (e.g. nature reserves and national parks; Freitag *et al.*, 1998; Parnell *et al.*, 2003; Reddy & Dávalos, 2003). Protected areas are usually characterized by high levels of biodiversity, unique habitats, pristine ecosystems or protected (or threatened) species, and are thus particularly attractive to collectors.

Although a number of studies have related collecting effort to selected single factors, to our knowledge no study has tested the relative importance of all the above-mentioned factors simultaneously in a common framework. It is thus unclear which factors are the main determinants of the spatial distribution of collecting effort and whether the determinants and underlying causes can be generalized across taxa and regions.

With 31,847 currently recognized species of native vascular plants (Wang *et al.*, 2011), China harbours one of the most diverse floras in the world. This enormous diversity is due to its large area (*c.* 9.6 million km^2) and high environmental heterogeneity, encompassing boreal, temperate, subtropical and tropical biomes, and its complex topography and geological history (Axelrod *et al.*, 1996). The Chinese flora has been intensively surveyed over the last 100 years and millions of plant specimens have been collected and preserved (Fu, 1993). Nevertheless, collecting effort varies dramatically among regions and only 9% of Chinese counties can be considered to be well sampled (Yang *et al.*, 2013).

Here, we test six hypotheses for the spatial variation in collecting effort and identify the main factors shaping the geography of floristic collections in China using the most comprehensive collection of botanical data for the country. Based on the results of previous studies, we consider the following hypotheses.

Hypothesis 1: easily accessible areas receive higher collecting effort than less accessible, more remote areas (H_1).

Hypothesis 2: collecting effort is higher in more densely populated areas (H_2).

Hypothesis 3: counties with herbaria are better sampled (H_{3a}) and collecting effort is negatively related to the distance to herbaria (H_{3b}).

Hypothesis 4: mountainous areas have higher collecting effort than lowlands (H_4).

Hypothesis 5: areas with high water availability are better sampled (H_5).

Hypothesis 6: counties with a larger proportion of areas occupied by protected areas have higher collecting effort (H_6).

METHODS

Species distributional data

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

To improve the data quality, we cleaned the data according to the following criteria: (1) multiple entries referring to the same specimen that occurred during the digitization process were removed; (2) we excluded records collected outside China; (3) locality information was georeferenced to county level; (4) sci-

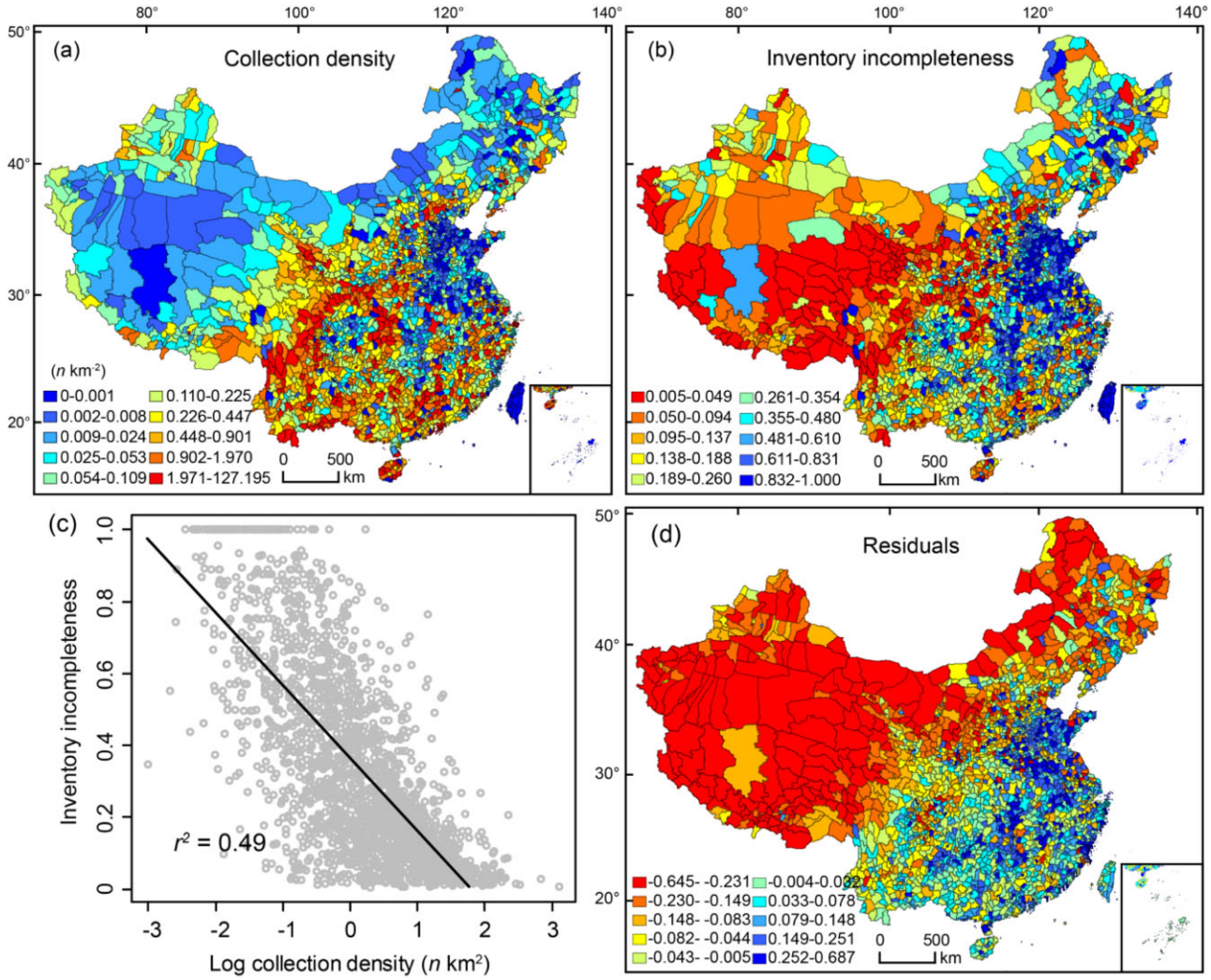

Figure 1 Maps of (a) collection density and (b) inventory incompleteness as indicators of collecting effort for vascular plants in 2377 Chinese counties. (c) Relationship between collection density and inventory incompleteness. (d) Map of absolute residuals from ordinary least-squares regression in (c). Collection density is calculated as the number of specimens per km^2. Inventory incompleteness is represented by the slope of the last 10% of species accumulation curves (see text for details). Negative values (red) in (d) indicate lower inventory incompleteness than expected from the collection density, whereas positive values (blue) indicate higher inventory incompleteness than expected. Legends use quantile classifications and maps are in Albers projections. Insets in the bottom right of maps show the south boundary of China, including all islands in the South China Sea.

entific names were standardized according to the *Catalogue of Life: Higher Plants in China* (http://www.etaxonomy.ac.cn/, accessed January 2009; Wang *et al.*, 2011); (5) infraspecific taxa were merged to species level. After this process, 4,338,516 records remained in the database for subsequent analysis. China is politically divided into 2377 counties with a mean and median size per county of 4138 and 1895 km^2, respectively.

Response variables

Collection density and inventory incompleteness of counties were entered as response variables in the regression analyses. Collection density and inventory incompleteness represent two different aspects of collecting effort, i.e. the number of collections per unit area and the degree of completeness reached with that number of collections. Collection density was calculated as the number of records per km^2 (Fig. 1a). Inventory incompleteness was represented by the slope of the last 10% of specimen accumulation curves (SACs; Fig. 1b; Yang *et al.*, 2013), based on the principle that the curvilinearity of species accumulation curves reflects the level of sampling incompleteness (Colwell & Coddington, 1994; Tittensor *et al.*, 2010).

Explanatory variables

We used road density to represent the accessibility of collecting sites (Table 1; see Appendix S1 in Supporting Information). Road data for the year 1999 were obtained from the Digital

Table 1 Predictor variables used to test different hypotheses on the spatial variation in collection density and inventory incompleteness of vascular plants in Chinese counties.

	Hypotheses	Predictor variables (units)	Data sources
H_1	Accessibility ('road-map effect' *sensu* Crisp *et al.*, 2001)	Road density (km km^{-2})	Digital Chart of the World (DCW)
H_2	Human population density	Human population density (n km^{-2})	China Historical GIS (CHGIS)
H_3	Home of collectors ('botanist effect' *sensu* Moerman & Estabrook, 2006)	Presence/absence of herbaria (H_{3a})	This study
		Distance to herbaria (H_{3b})	Represented by the reciprocal of accumulated kernel densities of herbaria calculated in ArcGIS 9.3
H_4	Mountains	Maximum elevational range (m)	GTOPO-30 digital elevation model (US Geological Survey, 1996)
H_5	Water availability	Annual wet days (number year^{-1})	Climatic Research Unit climatology (New *et al.*, 2002)
H_6	Conservation priority	Proportion of protected areas	World Database on Protected Areas (WDPA)

Chart of the World (DCW; http://www.fas.harvard.edu/~chgis/data/dcw/, accessed May 2012). Road density was calculated as the total length of roads (in km) divided by the area (in km^2) of each county. In preliminary analyses, we also considered the density of navigable rivers and railways as potential explanatory variables. However, these two variables were strongly collinear with each other and with road density (Pearson's $r \geq 0.7$) and were thus excluded from further analyses. County-level data of human population density for the year 1999 were obtained from the China Historical Geographic Information System (CHGIS; http://www.fas.harvard.edu/~chgis%20/work/downloads/, accessed May 2012; Table 1, Appendix S1).

The presence or absence of herbaria in each county was incorporated as a binary variable into the regression analyses (Table 1, Appendix S1). The distance of each county to herbaria was represented by the reciprocal of accumulated kernel densities of herbaria. Kernel densities for each herbarium were calculated in ArcGIS 9.3 (ESRI, 2008) and weighted by the number of specimens in the herbarium (Table 1, Appendix S1). The proportion of area occupied by protected areas in each county was calculated to represent conservation priority (Table 1, Appendix S1). Polygon layers of nature reserves (i.e. data category Ia), the main type of protected area in China, were obtained from the World Database on Protected Areas (WDPA, http://www.wdpa.org/), accessed May 2012.

Maximum elevational range was used as an indicator of topographical complexity and calculated for each county based on the GTOPO-30 digital elevation model (US Geological Survey, 1996) at a spatial resolution of 30 arcsec (Table 1, Appendix S1). Annual wet days were extracted from a global high-resolution climatology at a spatial resolution of 10 arcmin (Table 1, Appendix S1; New *et al.*, 2002) to represent the amount and seasonality of precipitation (Kreft & Jetz, 2007). Spatial averages of annual wet days were calculated for each county.

Statistical analyses

Ordinary least-squares (OLS) models were applied to investigate the relationships between explanatory variables and response variables. We used the Akaike information criterion (AIC; Johnson & Omland, 2004) and step-wise backward selection to identify the most parsimonious multipredictor models as minimum adequate models. All explanatory variables except the presence of herbaria, annual wet days and the proportion of protected areas were $\log_{10}$-transformed to achieve the best model fits and approximate normally distributed residuals.

Spatial correlograms and global Moran's I-tests showed that spatial autocorrelation in the residuals of minimum adequate models was relatively weak but significant (Appendix S2). Spatial autocorrelation leads to overestimated degrees of freedom and results in inflated Type I errors in our OLS models (Dormann *et al.*, 2007). Eigenfunction-based procedures are able to address complex spatial patterns and are widely used for taking spatial autocorrelation in regression analyses into account (Diniz-Filho & Bini, 2005; Griffith & Peres-Neto, 2006). Geographical coordinates of the centre of each county were used to calculate pairwise Euclidean distances of all counties. We truncated the distance matrix at a distance of 300 km, i.e. we replaced distances over 300 km with 1200 km (four times the truncation distance), while keeping distances of 300 km or less as they were calculated (Borcard & Legendre, 2002). We chose a distance of 300 km because spatial autocorrelation in residuals of non-spatial OLS models was very weak beyond this distance (Appendix S2). Principal coordinates of neighbour matrices (PCNM; Borcard & Legendre, 2002) were applied to decompose the spatial structures among counties using the function 'pcnm' in the R package 'vegan' (R Development Core Team, 2011; Oksanen *et al.*, 2012). One thousand and seventy-four eigenvectors with positive eigenvalues were obtained that captured the spatial relationships among counties at different scales. The first eigenvectors represent broad-scale variation of spatial structures, whereas those with smaller eigenvalues represent finer-scale variation (Diniz-Filho & Bini, 2005; Appendix S3).

Eigenvector-based spatial filters were then added as predictors in multipredictor models (mostly following the procedure by Diniz-Filho & Bini, 2005). A forward model selection procedure was performed to select the most significant spatial filters. The selection started with the first filter (in descending order based on eigenvalues) and a filter was selected only when it was

Table 2 Results from minimum adequate ordinary least-squares models with eigenvector-based spatial filters and either collection density or inventory incompleteness of vascular plants in Chinese counties as a dependent variable.

Explanatory variables	Estimate	*t*-value	*P*-value	R^2_{adj}	Partial R^2
Collection density*				0.43	
Human population density	0.23	6.05	< 0.01		0.02
Presence of herbaria	0.73	7.06	< 0.001		0.03
Distance to herbaria	−1.43	−6.79	< 0.001		0.02
Elevational range	0.61	16.90	< 0.001		0.15
Annual wet days	0.01	11.14	< 0.001		0.10
Protected areas	2.32	7.29	< 0.001		0.03
23 spatial filters			< 0.05		0.07
Inventory incompleteness†				0.45	
Road density	0.10	3.03	< 0.01		0.01
Human population density	0.22	14.43	< 0.001		0.13
Presence of herbaria	−0.43	−9.25	< 0.001		0.04
Distance to herbaria	0.24	5.37	< 0.001		0.02
Elevational range	−0.17	−12.31	< 0.001		0.09
Protected areas	−0.40	−8.68	< 0.001		0.03
25 spatial filters			< 0.05		0.08

*Collection density is calculated as the number of specimens per km^2.
†Inventory incompleteness is represented by the slope of the last 10% of species accumulation curves (see text for details). Protected areas, proportion of area in a county occupied by protected areas; R^2_{adj}, adjusted R^2 of multiple-predictor models; partial R^2, partial R^2 of each predictor in the models.

significant in the model. This procedure was stopped when spatial autocorrelation in model residuals disappeared.

We used partial R^2 for each predictor from minimum adequate models with spatial filters to assess the explanatory power of the predictor for collecting effort while controlling for the effects of other variables. We obtained residuals from OLS regressions between the predictor of interest and the other predictors (including spatial filters) in the minimum adequate model, and between the response variable and the other predictors, respectively. Partial R^2 was the calculated as the R^2 of the OLS regression between these two sets of residuals (Legendre & Legendre, 2012).

RESULTS

The collection density of vascular plants in Chinese counties (Fig. 1a) ranged from 0 to 127.2 specimens per km^2 (mean = 0.9). The inventory incompleteness (Fig. 1b) varied between 0.005 and 1 (mean = 0.35). As expected, collection density and inventory incompleteness were negatively related, but the relationship was not very strong ($r^2 = 0.49$; Fig. 1c), indicating that counties with a high density of collections are not necessarily more completely sampled. For example, some counties in the Tibetan Plateau had fewer collections but more complete inventories than average, whereas many counties in south China had a much higher collecting density but less complete inventories than average (Fig. 1a,b,d), which would be expected based on differences in species richness between the two areas.

Twenty-three and 25 spatial filters from the first 51 were selected into the minimum adequate OLS models for collection density and inventory incompleteness, respectively. Global Moran's *I*-tests and correlograms indicated that spatial autocorrelation was successfully removed in model residuals by adding spatial filters (Appendices S2 & S4).

The minimum adequate models including all predictors except road density and 23 spatial filters explained 43% of the variance in collection density, with 7% explained by spatial filters alone (Table 2). Elevational range emerged as the strongest predictor (Appendix S5), explaining 15% (indicated by partial R^2) of the variance in collection density, whereas annual wet days was the second strongest predictor (partial $R^2 = 0.10$). The effects of other variables were relatively weak, with partial R^2 values between 0.02 and 0.03 (Table 2). All predictors except distance to herbaria showed positive relationships with collection density.

For inventory incompleteness, all predictors except annual wet days and 25 spatial filters were included in the minimum adequate model, explaining 45% of the total variance, with 8% exclusively explained by the spatial filters (Table 2). Human population density was the strongest predictor (Appendix S6) and explained 13% of the variance in inventory incompleteness, followed by elevational range (partial $R^2 = 0.09$). All other predictors were of minor importance (partial R^2 between 0.01 and 0.04; Table 2). Human population density, distance to herbaria and road density had positive effects on inventory incompleteness, whereas the other variables had negative effects.

DISCUSSION

Using a large Chinese plant distributional database, this study is the first to systematically test six proposed hypotheses on the origin of geographical variation in collecting effort at regional scales. We found that elevational range was the strongest predictor of collection density and had a positive effect on it, whereas

human population density had a strong positive effect and best predicted inventory incompleteness of vascular plants in China. Road density only showed a surprisingly weak correlation with collecting effort. These results are in contrast to previous studies for other regions or taxa reporting strong positive effects of human population density and accessibility on collection patterns (Soberón *et al.*, 2000; Reddy & Dávalos, 2003; Tobler *et al.*, 2007; Ficetola *et al.*, 2014).

Road density was not selected in the multiple-predictor model for collection density, and was weakly and positively correlated with inventory incompleteness, indicating a non-significant even negative effect of this predictor on collecting effort (Table 2). This finding does not support our H_1 and previous studies that counties with easy access and good transportation infrastructure have higher collecting effort (Crisp *et al.*, 2001; Parnell *et al.*, 2003; Küper *et al.*, 2006). The Chinese flora is mostly collected by trained taxonomists who might be specifically interested in surveying pristine vegetation in remote areas (Chen, 1994). This could explain our result because high road density not only implies easy access but also higher levels of human disturbance (Li *et al.*, 2010; Marcantonio *et al.*, 2013). However, the effect of accessibility is likely to be strongly scale dependent. At finer spatial resolutions, specimen collections may still cluster along roads or paths even if such collecting behaviour is not detectable at the scale of counties or large grid cells. Fine-scale collection data and cross-scale analyses would be needed to test this.

Human population density was the strongest and a positive predictor of inventory incompleteness, but only had a weak effect on collection density (Table 2). This indicates that species inventories in densely populated areas tend to be more incomplete, sharply contrasting with previous studies (Parnell *et al.*, 2003; Luck, 2010; Botts *et al.*, 2011; Ficetola *et al.*, 2014) and not supporting our H_2. Moreover, collection density and human population density for different decades (from the 1970s to the 2000s) in 87 counties of south central China were negatively related (Appendix S7), suggesting that the overall negative effect of human population density on collecting effort also holds at smaller spatial scales and for different time periods. One possible explanation for this difference is that taxonomists may expect densely populated areas (e.g. urban and intensive agricultural areas) to have few native or narrow endemic species and thus regard these areas as less interesting (Redford & Richter, 1999; Angermeier, 2000).

Efforts in biodiversity conservation in China have so far largely focused on pristine and remote areas and little there has been little emphasis on integrated conservation in densely populated or agricultural landscapes (Tang, 2005). Other studies have reported that urban areas even have higher native species richness including a large proportion of rare and endangered species (Kühn *et al.*, 2004; Pautasso & McKinney, 2007). Rapid land-use change (especially urbanization as well as agricultural expansion and intensification) is one of the major threats to global biodiversity, and might be particularly detrimental in densely populated areas (McKinney, 2002). In the light of these threats, we see an urgent need to increase collections in and close to human settlements as an initial step to effectively increase the knowledge base for conservation assessments and environmental monitoring in these areas.

Although the explanatory power of the presence of herbaria was relatively low, it still indicates that counties with herbaria had a significantly higher collecting effort, i.e. a higher collection density and lower inventory incompleteness (Table 2). This finding supports the 'botanist effect' hypothesis (H_{3a}) stating that homes of botanists tend to be better sampled (Moerman & Estabrook, 2006). For example, collecting effort is very high in Beijing and Kunming (Fig. 1a,b) where China's largest herbaria are located and many taxonomists work. The presence of herbaria also has a weak but positive impact on the collecting effort in neighbouring counties (Table 2), supporting H_{3b} that collecting effort is negatively related to the distance to herbaria (Dennis & Thomas, 2000).

Elevational range had a positive effect on and emerged as the strongest predictor of collection density, and had a negative effect on inventory incompleteness (Table 2). This supports H_4 that mountainous areas have a higher collecting effort than lowlands. This may be because expedition teams of trained taxonomists have conducted organized surveys in China since the 1950s and have regularly been sent to remote and mountainous areas anticipated to have particularly high levels of biodiversity, especially the Hengduan Mountains, where the Himalayan uplift caused a great complexity of different habitats, stimulated allopatric speciation and ultimately gave rise to high levels of species richness and endemism (Chen, 1994). Collecting in such areas is thought to maximize the number of species in the collections and to result in higher probabilities of finding species that are rare or new to science (Romo *et al.*, 2006; Tang *et al.*, 2006; Soria-Auza & Kessler, 2008; Sastre & Lobo, 2009). However, the northern part of the Tibetan Plateau shows high levels of completeness because both species richness and broad-scale turnover are comparatively low (Barthlott *et al.*, 2007), and it is thus easy to reach fair levels of completeness with only a few collections (Fig. 1b,d).

We tested whether areas with high water availability are better sampled (H_5). This hypothesis receives support from our findings that annual wet days had a positive effect, and it was the second strongest predictor of collection density (Table 2). Previous studies demonstrated that water availability is the most important environmental constraint on plant species richness in China (Wang, 1992; Yang *et al.*, 2013). Collectors have probably been guided by their experience that moister areas usually harbour more plant species, leading to a higher number of collections in such areas. However, annual wet days had little explanatory power on inventory incompleteness, indicating that areas with high water availability are not more completely sampled.

The proportion of protected areas had a positive yet relatively weak effect on collecting effort (Table 2), i.e. collection density was higher and incompleteness lower in counties with large proportions of protected areas. The effect of protected areas might be weakened by the fact that some areas are protected not primarily because of their biological importance (e.g. high

diversity or endemism) but for their cultural and geological importance that does not necessarily attract botanists. Nevertheless, this result supports H_6 that protected areas receive higher collecting effort. This is consistent with previous reports that collecting effort tends to be concentrated in areas designated as conservation priorities (Freitag *et al.*, 1998; Parnell *et al.*, 2003; Reddy & Dávalos, 2003). Reddy & Dávalos (2003) demonstrated that conservation prioritization based on species richness and endemism might be affected by biased knowledge about species distributions. It is thus important to take collecting effort into account when using currently available distributional data for conservation assessments and priority setting.

A potential drawback of this as well as of similar previous analyses is that specimen data and explanatory variables do not stem from exactly the same time period. While specimen data have accumulated over more than 100 years, population and road data are from a recent decade, due to the difficulty of obtaining historical data for these variables in China. However, spatial patterns of number of collections in different decades are similar (pairwise Spearman's rank $r_s = 0.54$–0.73; Appendices S8 & S9), suggesting a strong temporal autocorrelation in collecting effort. Furthermore, human population density has not shifted much spatially in China during the last century (Wang *et al.*, 1996), because the spatial distribution of human population is largely controlled by environmental factors such as natural productivity and soil fertility (Waide *et al.*, 1999). In the light of the strong temporal autocorrelations of collection patterns and human activities, we consider that the reliability of our results is not affected by these drawbacks.

Our results show that different hypotheses on collecting effort are not mutually exclusive but that many factors can act synergistically as drivers of biological collection patterns. Mountains and human population density are the two main determinants of the spatial distribution of collecting effort for vascular plants in China. Whereas the former have a positive effect, the latter has a negative effect on collecting effort. Differences from studies from other parts of the world are probably caused by the fact that the Chinese flora has been mostly surveyed by experienced taxonomists. Expedition teams have been centrally organized and coordinated by governments and research institutes, and have been regularly sent to mountainous and remote areas where high levels of biodiversity and pristine vegetation have been anticipated (Chen, 1994). As the differences between China and other regions demonstrate, the determinants and causes of the uneven collecting effort may therefore be strongly dependent on the regional context (Vale & Jenkins, 2012).

In conclusion, our results indicate that mountainous areas are most intensively collected in China, whereas densely populated areas tend to be neglected by plant collectors. This sampling bias leads to woefully incomplete inventories, particular in urban and agricultural areas, and thus to a pronounced 'Wallacean shortfall', i.e. an incomplete documentation of species ranges. Our study highlights the need to increase biological collections in areas with specific environmental and socio-economic characteristics (e.g. densely populated areas) to improve the quality and representativeness of distributional data as well as the knowledge base for biological conservation.

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 and J. L. Zhang for help with data preparation. We thank Z. H. Wang, Y. Kisel and C. Meyer for valuable comments on this study. We are particularly grateful for helpful comments provided by the handling editor J. A. Diniz-Filho and anonymous 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

Ahrends, A., Rahbek, C., Bulling, M.T., Burgess, N.D., Platts, P.J., Lovett, J.C., Kindemba, V.W., Owen, N., Sallu, A.N., Marshall, A.R., Mhoro, B.E., Fanning, E. & Marchant, R. (2011) Conservation and the botanist effect. *Biological Conservation*, **144**, 131–140.

Angermeier, P.L. (2000) The natural imperative for biological conservation. *Conservation Biology*, **14**, 373–381.

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*, **22**, 586–595.

Barthlott, W., Hostert, A., Kier, G., Küper, W., Kreft, H., Mutke, J., Rafiqpoor, M.D. & Sommer, J.H. (2007) Geographic patterns of vascular plant diversity at continental to global scales. *Erdkunde*, **61**, 305–315.

Borcard, D. & Legendre, P. (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.

Botts, E.A., Erasmus, B.F.N. & Alexander, G.J. (2011) Geographic sampling bias in the South African Frog Atlas Project: implications for conservation planning. *Biodiversity and Conservation*, **20**, 119–139.

Chen, C. (1994) History of plant taxonomy in China. *History of Chinese botany* (ed. by Z. Wang), pp. 121–144. Science Press, Beijing.

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

Crisp, M.D., Laffan, S., Linder, H.P. & Monro, A. (2001) Endemism in the Australian flora. *Journal of Biogeography*, **28**, 183–198.

Dennis, R.L.H. & Thomas, C.D. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, **4**, 73–77.

Dennis, R.L.H., Sparks, T.H. & Hardy, P.B. (1999) Bias in butterfly distribution maps: the effects of sampling effort. *Journal of Insect Conservation*, **3**, 33–42.

Diniz-Filho, J.A.F. & Bini, L.M. (2005) Modelling geographical patterns in species richness using eigenvector-based spatial filters. *Global Ecology and Biogeography*, **14**, 177–185.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kuhn, I., Ohlemuller, R., Peres-Neto, P.R., Reineking, B., Schroder, 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.

ESRI (2008) *ArcMap version 9.3*. ESRI Press, Redlands, CA.

Ficetola, G.F., Cagnetta, M., Padoa-Schioppa, E., Quas, A., Razzetti, E., Sindaco, R. & Bonardi, A. (2014) Sampling bias inverts ecogeographical relationships in island reptiles. *Global Ecology and Biogeography*, doi: 10.1111/geb.12201.

Freitag, S., Hobson, C., Biggs, H.C. & Van Jaarsveld, A.S. (1998) Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation*, **1**, 119–127.

Fu, L. (1993) *Index herbariorum Sinicorum*. China Science and Technology Press, Beijing.

Gaston, K.J. (2000) Global patterns in biodiversity. *Nature*, **405**, 220–227.

Griffith, D.A. & Peres-Neto, P.R. (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology*, **87**, 2603–2613.

Hortal, J., Garcia-Pereira, P. & García-Barros, E. (2004) Butterfly species richness in mainland Portugal: predictive models of geographic distribution patterns. *Ecography*, **27**, 68–82.

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

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

Kühn, I., Brandl, R. & Klotz, S. (2004) The flora of German cities is naturally species rich. *Evolutionary Ecology Research*, **6**, 749–764.

Küper, W., Sommer, J.H., Lovett, J.C. & Barthlott, W. (2006) Deficiency in African plant distribution data – missing pieces of the puzzle. *Botanical Journal of the Linnean Society*, **150**, 355–368.

Legendre, P. & Legendre, L. (2012) *Numerical ecology*, 3rd edn Elsevier, Amsterdam.

Li, T., Shilling, F., Thorne, J., Li, F., Schott, H., Boynton, R. & Berry, A.M. (2010) Fragmentation of China's landscape by roads and urban areas. *Landscape Ecology*, **25**, 839–853.

Luck, G.W. (2010) Why is species richness often higher in more densely populated regions? *Animal Conservation*, **13**, 442–443.

McKinney, M.L. (2002) Urbanization, biodiversity, and conservation. *BioScience*, **52**, 883–890.

Marcantonio, M., Rocchini, D., Geri, F., Bacaro, G. & Amici, V. (2013) Biodiversity, roads, and landscape fragmentation: two Mediterranean cases. *Applied Geography*, **42**, 63–72.

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.

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

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., Minchin, P.R., O'Hara, R., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2012) Vegan: community ecology package. R package version 2.0-5. Available at: http://vegan.r-forge.r-project.org/ (accessed March 2012).

Parnell, J.A.N., Simpson, D.A., Moat, J., Kirkup, D.W., Chantaranothai, P., Boyce, P.C., Bygrave, P., Dransfield, S., Jebb, M.H.P., Macklin, J., Meade, C., Middleton, D.J., Muasya, A.M., Prajaksood, A., Pendry, C.A., Pooma, R., Suddee, S. & Wilkin, P. (2003) Plant collecting spread and densities: their potential impact on biogeographical studies in Thailand. *Journal of Biogeography*, **30**, 193–209.

Pautasso, M. & McKinney, M.L. (2007) The botanist effect revisited: plant species richness, county area, and human population size in the United States. *Conservation Biology*, **21**, 1333–1340.

R Development Core Team (2011) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. Available at: http://cran.r-project.org.

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

Redford, K.H. & Richter, B.D. (1999) Conservation of biodiversity in a world of use. *Conservation Biology*, **13**, 1246–1256.

Romo, H., García-Barros, E. & Lobo, J.M. (2006) Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography*, **29**, 873–885.

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

Sánchez-Fernández, D., Lobo, J.M., Abellán, P., Ribera, I. & Millán, A. (2008) Bias in freshwater biodiversity sampling: the case of Iberian water beetles. *Diversity and Distributions*, **14**, 754–762.

Sastre, P. & Lobo, J.M. (2009) Taxonomist survey biases and the unveiling of biodiversity patterns. *Biological Conservation*, **142**, 462–467.

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

Soberón, J.M., Llorente, J.B. & Oñate, L. (2000) The use of specimen-label databases for conservation purposes: an

example using Mexican Papilionid and Pierid butterflies. *Biodiversity and Conservation*, **9**, 1441–1466.

Soria-Auza, R.W. & 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.

Tang, X. (2005) Analysis of the current situation of China's nature reserve network and a draft plan for its optimization. *Biodiversity Science*, **13**, 81–88.

Tang, Z., Wang, Z., Zheng, C. & Fang, J. (2006) Biodiversity in China's mountains. *Frontiers in Ecology and the Environment*, **4**, 347–352.

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.

Tobler, M., Honorio, E., Janovec, J. & Reynel, C. (2007) Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two Neotropical plant families (Moraceae and Myristicaceae) in Peru. *Biodiversity and Conservation*, **16**, 659–677.

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

Vale, M.M. & Jenkins, C.N. (2012) Across-taxa incongruence in patterns of collecting bias. *Journal of Biogeography*, **39**, 1743–1748.

Waide, R.B., Willig, M.R., Steiner, C.F., Mittelbach, G., Gough, L., Dodson, S.I., Juday, G.P. & Parmenter, R. (1999) The relationship between productivity and species richness. *Annual Review of Ecology and Systematics*, **30**, 257–300.

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

Wang, Z., Zhang, P. & Zhou, Q. (1996) The impacts of climate on the society of China during historical times. *Acta Geographica Sinica*, **51**, 329–339.

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

Whittaker, R.J., Araújo, M.B., Jepson, P., Ladle, R.J., Watson, J.E.M. & Willis, K.J. (2005) Conservation biogeography: assessment and prospect. *Diversity and Distributions*, **11**, 3–23.

Wilson, E.O. (1988) *Biodiversity*. National Academy Press, Washington, DC.

Yang, W., Ma, K. & Kreft, H. (2013) Geographical sampling bias in a large distributional database and its effects on species richness–environment models. *Journal of Biogeography*, **40**, 1315–1426.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1 Explanatory variables for collecting effort.
Appendix S2 Moran's *I* correlograms for response variables and residuals from regression models.
Appendix S3 Examples of spatial patterns depicted by eigenvector maps.
Appendix S4 Moran's *I* and *P*-values of Moran's *I* in residuals of ordinary least-squares models.
Appendix S5 Partial residual plots for collection density.
Appendix S6 Partial residual plots for inventory incompleteness.
Appendix S7 Relationships between collection density and human population density of different decades in 87 counties of south central China.
Appendix S8 Collection patterns of vascular plants in two-decade time periods.
Appendix S9 Pairwise Spearman's rank correlation coefficients between collection patterns of two-decade time periods.

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. wrote the manuscript.

Editor: José Alexandre Diniz-Filho