Original article

The patterns of vascular plant discoveries in China

Muyang Lu[1, 2] and Fangliang He[1, 3]†

1. ECNU-Alberta Joint Lab for Biodiversity Study, Tiantong National Station for Forest Ecosystem Research, East China Normal University, Shanghai 200241, China
2. Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT, USA
3. Department of Renewable Resources, University of Alberta, Edmonton, Alberta T6G 2H1, Canada

Running head: patterns of plant discoveries

Words in the abstract: 295

Words in main body: 2846

References: 41

Table: 4

Figure: 4

† Corresponding author: Fangliang He (email: fhe@ualberta.ca)

ABSTRACT

In order to understand the patterns of plant discoveries in China and find out which taxa and which areas harbor most numbers of undiscovered species, we analyzed the discovery times of 25268 vascular plant species described between 1753 and 2000 from *Flora Reipublicae Popularis Sinicae*. We found that species with larger range size and distributed in northeastern part of China have a higher discovery probability. Species distributed on the coast were discovered earlier than inland species. Herbs have the highest discovery probability and ferns have the lowest discovery probability. Ferns and herbs have the largest number of undiscovered species in China. Most undiscovered species are found in southwestern part of China, one of the biodiversity hotspots in the world. Spatial patterns of mean discovery time and inventory completeness are mainly driven by the total number of species and human population density in an area and whether the area is coastal or not. We demonstrated that socio-economic factors shape the discovery patterns of vascular plants in China. Undiscovered species are mostly narrow-ranged, inconspicuous endemic species such as herbs and ferns, which are prone to extinctions and locate in biodiversity hotspots in southwestern China.

INTRODUCTION

Despite 270 years' discovery of species, our knowledge about the biological world in nature is still far from complete. According to a recent estimate, fewer than 20% of species on earth have been discovered since Linnaeus' invention of binomial nomenclature (Mora *et al.*, 2011). With the ongoing biodiversity crisis, this lack of knowledge has become a major obstacle to biodiversity conservation as many species could go extinct without ever being known to science (Costello, May, & Stork, 2013).

Early botanical and zoological discoveries are often biased towards large-sized, charismatic species with widespread geographic distributions (Gibbons *et al.*, 2005; Ferretti *et al.*, 2008; Stork *et al.*, 2008, 2015; Essl *et al.*, 2013; Randhawa, Poulin, & Krkošek, 2015). Positive correlations between species discovery probability and body size have been found in a variety of taxa including insects (Gaston & Hudson, 1994), birds (Blackburn & Gaston, 1995), mammals (Paxton, 1998; Medellín & Soberón, 1999; Collen, Purvis, & Gittleman, 2004), fishes (Zapata & Robertson, 2007), but not observed in taxa like reptiles (Reed & Roback, 2002) and marine holozooplanktons (Gibbons *et al.*, 2005). More recently discovered species are often of greater conservation interests because they are more likely to be narrow ranged and inconspicuous species are prone to extinctions (Diniz-Filho *et al.*, 2005; Tedesco *et al.*, 2014).

In addition to biological factors, species discovery is also influenced by human activities such as taxonomic efforts, technology innovations, and social-economic events. For example, it is commonly observed for many taxa that species discovery rates dropped to the bottoms during the two World Wars and peaked in the 1990s with the emergence of molecular

techniques (Gaston, 1995; Bebber *et al.*, 2007b; Joppa, Roberts, & Pimm, 2011b; Lu & He, 2017). Geographically speaking, Europe and North America have the most complete species inventories due to their long histories of exploration (Gaston, 1995; Essl *et al.*, 2013), while species discovery in less explored continents such as South America and Africa was affected by colonization histories and indigenous knowledge (Gaston, 1995; Diniz-Filho *et al.*, 2005; Rosenberg *et al.*, 2013; Ballesteros-Mejia *et al.*, 2013). As a result, biodiversity hotspots, most of which locate in less developed countries (Myers *et al.*, 2000), often harbor the largest number of undiscovered species (Joppa *et al.*, 2011a; Giam *et al.*, 2012). This imposes a more serious challenge for biodiversity conservation in underdeveloped regions where economic growth is often achieved at the expense of environmental degradation.

China harbors nearly one tenth of the plant species on earth (Joppa *et al.*, 2011b; Lu *et al.*, 2018). However, its rapid economic growth over the past three decades has resulted in the colossal loss of millions of hectares of natural habitats (He, 2009). The sustainable development of China depends on balancing economic growth and preservation of natural habitats. Knowing where the undiscovered species are is necessary for making decision on habitat protection.

The interest for cataloging species in China has long predated the invention of Linnaeus' binomial nomenclature, but in the light of modern taxonomy, much credit should be given to western naturalists who diligently collected specimens and described species since the first arrival of Jesuits in the 16th century, as reflected by the fact that nearly 70% of the type specimens of Chinese plants are kept in herbaria in Europe and America (Chen, 1994). Due to logistic constraints and political instability, most 18th and 19th naturalists made their botanical

collections in the coastal areas (Bretschneider, 1898; Fan, 2004), which should affect the patterns of discovery records.

There are around 31000 recorded vascular plant species in China, but this inventory is not complete and new species still await discovery (Lu & He, 2017). The objectives of this study are: (1) to model and map geographic variation in botanical discovery of vascular plant species in China, (2) to find out which taxa have the largest numbers of undiscovered species and where the undiscovered species are, and (3) to quantify what factors (e.g., species' range size and growth form) may influence the spatial patterns of plant discoveries in China. This study will contribute to understanding the pattern of species discoveries and their underlying factors. The identification of taxonomic and geographic gaps of undiscovered species will inform prioritizing our taxonomic and conservational efforts in the future.

MATERIALS AND METHODS

Data

Data including species names, authorities, discovery dates, province-level biogeographic distributions, and growth forms were compiled from *Flora Reipublicae Popularis Sinicae* (http://frps.eflora.cn/), which contains 31142 species. Synonyms were excluded according to the updated list of *Flora of China* (http://foc.eflora.cn/) and only species discovered between 1753 and 2000 were retained for analysis, resulting in a total number of 25268 species (Data available in Appendix S2). We treated discovery date as the time a species was first described in any name in scientific publication. Data after 2000 were excluded as a routine to avoid the effect of delayed entrance of newly discovered species on estimating the number of

undiscovered species (Costello & Wilson, 2011; Costello, Wilson, & Houlding, 2012). Human population densities on province-level were obtained from *2010 Population Census of The People's Republic of China* (http://www.stats.gov.cn/), and *Monthly Bulletin of Interior Statistics* (http://sowf.moi.gov.tw/stat/month/elist.htm). Province areas were obtained from *National Fundamental Geographic Information System of China* (http://nfgis.nsdi.gov.cn/nfgis/). There are in total 28 provinces after merging independent municipalities such as Beijing and Shanghai to adjacent provinces. Range size, maximum/minimum latitude, maximum/minimum longitude and whether or not a species is distributed in coastal areas were obtained from province-level distributions. Growth forms were categorized as herb, shrub, arbor, fern and vine. Species with both herb and shrub forms are treated as herb while species with both shrub and arbor forms are treated as shrub (treating herb/shrub as shrub and shrub/arbor as arbor does not change our results). Range size, maximum/minimum latitude, maximum/minimum longitude and ordinal growth form were standardized to [0, 1] in order to calculate the effect size on discovery probability. Range size was log-transformed before standardization. Correlations between explanatory variables were checked prior to analysis. No collinearity existed among explanatory variables (the highest correlation was 0.7).

Cox proportional hazard model

We first modeled the discovery time using survival analysis in which the discovery time was treated as a random variable. Because only discovered species could be recorded, there are no censored data in our study. In this case, the empirical survival curve is just the inverse of the

accumulation curve (Steyskal, 1965; Essl *et al.*, 2013). Cox proportional hazard (Cox PH) model was used to model the instantaneous discovery probability which describes the probability that a species will be seen in the next step of time (t+Δt) if it remains unseen up to time t. This probability is expressed as a hazard function, $h(t)$, given as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr[(t \le T < t + \Delta t) \mid T > t]}{\Delta t}.$$

The Cox PH regression model is:

$$h_i(t) = h_0(t)\exp(\beta \mathrm{X}),$$

where matrix βX includes the linear components of the model (i.e., coefficients and dependent variables). $h_i(t)$ is the hazard function for species i (i.e., the "risk" to be discovered), and $h_0(t)$ is the baseline hazard function. Therefore, Cox PH model is a type of general linear model. Compared to accelerated failure time model (AFT), Cox model is equivalent to a non-parametric model which makes no assumption about the underlying distribution of survival time (i.e., discovery time), which is more appropriate for our data because discovery process is highly influenced by historical events (Lu & He, 2017) and the probability distribution of discovery time is unknown. Because it happened that multiple species were discovered in the same year, Efron approximation was used to break ties in discovery time (Hertz-Picciotto & Rockhill, 1997). A frailty term with Gaussian error distribution was added to account for genus-level random effects. The proportional hazard assumptions were examined by plotting the scaled Schoenfeld residuals (denoted as $\beta(t)$) against time. A horizontal trend of $\beta(t)$ implies that the time-independent coefficient assumption is met. Model selection was done by choosing the model with the minimum AIC value. Concordance statistic (C statistic) was used to show the discriminative ability of a

model. It is equivalent to the area under the Receiver Operating Characteristic Curve (AUC) in logistic regression, with the value of 0.5 indicating no discrimination power and the value of 1 indicating perfect discrimination. We also presented the fitted survival curves for different treatments (i.e. inland or coastal distributed species, and species with different growth forms) using strata models (in strata model each treatment has a different baseline function $h_0(t)$). All survival analyses were conducted using R package 'survival'.

Estimating species richness

We used a modified logistic discovery model (Lu & He, 2017) to estimate the species richness for different growth forms and for each province because models that take into account of taxonomic efforts (Joppa *et al*., 2011a,b; Lu & He, 2017) did not converge or failed to produce sensible estimates at the province level. Note that the logistic discovery model provides only conservative estimates in certain cases (Bebber *et al*., 2007b; Essl *et al*., 2013; Lu & He, 2017). Inventory completeness was calculated as the ratio of the number of discovered species to the estimated total number of species.

Spatial patterns of discovery time and inventory completeness

We further examined the relationship between mean discovery time and human population density, total number of species, whether a province is on the coast or inland, and province area using ordinary linear regression. The spatial autocorrelation of mean discovery time on the province scale was examined by Moran's *I*. The neighborhood structure of provincial polygons is defined by contiguity (only polygons with shared borders are counted as

neighbours). We proceeded with ordinary linear regression after no spatial autocorrelation was detected in the residuals of the model. Beta regression was used to model inventory completeness with the same set of covariates as modeling discovery time (population density, total number of species, whether coastal province or not, and area of each province). Beta regression was implemented using R package 'betareg' (Cribari-Neto & Zeileis, 2010). Beta regression assumes that inventory completeness (varying between 0 and 1) follows a beta distribution.

RESULTS

Cox proportional hazard model

The results show that ferns have the lowest discovery probability and herbs have the highest discovery probability among all taxonomic groups (Table 1; Fig. 1A). Coastal species were earlier discovered than inland species (Fig. 1B). The discovery probability of a species increases with its range size, maximum latitude and maximum longitude. The effect sizes of maximum latitude, minimum latitude, maximum longitude and minimum longitude are larger than the effect sizes of range size, coast and growth form, suggesting the importance of geographic locations on discovery probability (Table 1).

Estimating species richness

The accumulation curve (i.e. the inverse of survival curve) of fern species has not reached an asymptote (Fig. 2A), suggesting that there is a considerable number of fern species remained to be discovered in the future. We have tried but found no existing species discovery models

were able to make a sensible prediction about the total number of fern species. The most conservative logistic model estimates that there are 6960 fern species in total, but erroneously this estimate is not significantly different from 0! For other groups, the estimated inventory completeness is 0.873 for herb, 0.865 for shrub, 0.860 for arbor and 0.860 for vine (Table 2). Fern and herb harbor the largest number of undiscovered species (if we assume that at least 50% of fern species have not yet been discovered).

Spatial patterns of mean discovery time and inventory completeness

The mean discovery time increases from northeast towards southwest (Fig. 4A). Besides geographic variables (latitudes and coastal distribution), the best model shows that human population density and total number of species also have a significant effect on mean discovery time (Table 3). Human population density is negatively correlated with mean discovery time (Fig. 3A) and total number of species is positively correlated with mean discovery time (Fig. 3B). Mean discovery time decreases with latitude (Fig. 3C) and is shorter in coastal provinces than in inland provinces (Fig. 3D).

21 out of 28 provinces have more than 95% of their total number of species already being discovered according to the estimates of the modified logistic model (Lu & He, 2017). The most under-discovered provinces are in the southwestern China (85.1% in Guangxi province and 88.8% in Yunnan province; Fig. 4B). Beta regression shows that inventory completeness is positively correlated with human population density, and negatively correlated with total number of species, coastal distribution, area and latitude of a province (Table 4).

DISCUSSION

Today's knowledge about biodiversity is the result of arduous quest of generations of naturalists for botanical and zoological discoveries. Although the nomenclatures of species are universally binomial, the stories behind their discoveries are not and many of them are as colorful as the species that were discovered (Kilpatrick, 2014). It is however surprising that the literature in ecology is largely disconnected from this rich history of species discoveries. Worse still is that the data of species discoveries are not much used for answering questions important to biogeography and biodiversity study (but see Gaston,1995; Joppa *et al*. 2011a, b; Bebber *et al*. 2007a, b, 2014; Essle *et al*. 2013). In this study, we compiled data on vascular plant species discovered over 250 years' botanical exploration in China for understanding the geographic variation of discovery rates and for assessing the completeness of botanical inventory of the country.

Our analyses showed that vascular plant species with larger range size are more likely to be discovered in China, consistent with the previous findings that widespread species are discovered earlier in the history (Bebber *et al*., 2007a; Essl *et al*., 2013). But contrary to our expectation, the size of growth form does not have a consistent effect on discovery probability. Herb species have the highest discovery probability while fern species have the lowest discovery probability. Shrub, arbor and vine species all have similar discovery probabilities. This suggests that the effect of growth form on discovery probability reflects more of the difference in the number of taxonomic experts than in the size of the growth form. The unique history of botanical discovery in China is revealed by the fact that species distributed on the coast have a higher discovery probability than inland species even when

geographic information such as the maximum and minimum latitude range of a species is included in the model (Table 1). This is because coastal areas in China were most economically developed and much more accessible to western naturalists since the Opium War (Bretschneider, 1898; Fan, 2004).

We show that fern is the most under-discovered taxon in China (Fig. 2A), while herb has the largest number of undiscovered species (Table 2). Fern and herb have the largest number of undiscovered species because (1) they are more inconspicuous and difficult to identify and (2) the enterprise of cataloging ferns in China started only at the beginning of the 20th century (Chen, 1994). Given that fern and herb together contain the largest number of undiscovered species, and that most specimens of undiscovered species have already been preserved in herbaria (Bebber *et al.*, 2010), our study suggests that the limiting factor of discovering new species in China is the lack of taxonomic expertise, which resonates with the call to address the challenge of "taxonomic impediments" (Ebach, Valdecasas, & Wheeler, 2011; Bebber *et al.*, 2014).

The spatial patterns of mean species discovery time and inventory completeness are largely driven by human population density and species diversity in an area. Although colonization history did shape the spatial patterns of species discovery (Diniz-Filho *et al.*, 2005; Rosenberg *et al.*, 2013; Ballesteros-Mejia *et al.*, 2013), geographic sampling bias does not change the prioritization of the current conservation designs because the total number of species and number of discovered species are highly correlated (Joppa *et al.*, 2011a; Giam *et al.*, 2012). We expect that future discoveries will continue to be made in provinces with high species richness such as Sichuan, Guangxi and Yunnan.

It is worth noting that the spatial pattern of species inventory completeness at the province level is opposite to the spatial pattern at the county level which shows that eastern part of China has lower inventory completeness (Yang, Ma, & Kreft, 2013, 2014). This might result from a collecting strategy in Yang *et al*'s studies that aimed at collecting as many novel species as possible in a certain amount of samples (Chen, 1994) because their county-level inventory completeness is derived from the steepness of sample-based accumulation curves. The other explanation for this discrepancy is that inventory completeness is scale-dependent. For example, even if the inventory completeness at county level is on average 90%, the inventory completeness at province level could be lower than 90% if most of the recorded species at county level are common species. While Yang *et al.* (2014) argued that more efforts should be devoted to increasing botanical collections in eastern densely populated areas where county-level inventories are less complete.

There are two major limitations in this study. The first one is that species discovery date is not the time when the species was first collected in the field but is the time when the species was first described in a scientific publication. This could lead to inflating discovery probability for widespread species if their type specimens were collected outside of China. To address this problem, we reran our Cox proportional hazard model using species only endemic to China (12464 species). The results show that the key drivers of discovery probability do not change substantially (Table S1), which suggests that the observed patterns are not artifacts of including species collected outside China in our analysis. The effect size of coastal distribution became negative because the majority of these endemic species were discovered after 1900, which removed the effect of early western naturalists (Fig. S1).

The second limitation is that our dataset does not distinguish the process of discovering a new species from resolving a synonym. To model the transition rates from synonyms to valid names, more data including all synonyms are required (Alroy, 2002). Our treatment of the data could be seen as an approximation of species discovery process, which minimizes the effect of synonyms by using only valid names scrutinized from the most authoritative database (http://foc.eflora.cn/).

In summary, our results demonstrate that most of undiscovered vascular plant species in China are ferns and herbs, which are likely to be narrowly distributed endemic species in the Southwest biodiversity hotspot of China. Given the "taxonomic impediments" we are facing, more resources should be channeled to the training and recruitments of taxonomic expertise in these two particular groups. There is an urgency of cataloging undiscovered species in southwest mountainous areas for future conservation designs and taxonomic activities.

ACKNOWLEDGEMENTS

This work was funded by Sun Yat-sen University, East China Normal University and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Author contributions: F.H. and M.L. conceived the study; M.L. collected and analyzed the data; M.L. and F.H. wrote the paper.

REFERENCES

Alroy J. **2002**. How many named species are valid? *Proceedings of the National Academy of Sciences of the United States of America* **99**: 3706–3711.

Ballesteros-Mejia L, Kitching IJ, Jetz W, *et al.* 2013. Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. *Global Ecology and Biogeography* **22**: 586–595.

Bebber DP, Carine M a, Wood JRI, *et al.* 2010. Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 22169–22171.

Bebber DP, Harris SA, Gaston KJ, *et al.* 2007a. Ethnobotany and the first printed records of British flowering plants. *Global Ecology and Biogeography* **16**: 103–108.

Bebber DP, Marriott FHC, Gaston KJ, *et al.* 2007b. Predicting unknown species numbers using discovery curves. *Proceedings of the Royal Society B: Biological Sciences* **274**: 1651–8.

Bebber DP, Wood JRI, Barker C, *et al.* 2014. Author inflation masks global capacity for species discovery in flowering plants. *New Phytologist* **201**: 700–706.

Blackburn TM & Gaston KJ. **1995**. What determines the probability of discovering a species - a study of south-American oscine passerine birds. *Journal of Biogeography* **22**: 7–14.

Bretschneider E. **1898**. *History of European botanical discoveries in China*. St. Petersburg.

Chen C. **1994**. History of plant taxonomy in China. In: *Flora Reipublicae Popularis Sinicae*. Beijing: Science Press, 121–144.

Collen B, Purvis A & Gittleman JL. **2004**. Biological correlates of description date in carnivores and primates. *Global Ecology and Biogeography* **13**: 459–467.

Costello MJ, May RM & Stork NE. **2013**. Can we name Earth's species before they go extinct? *Science* **339**: 413–416.

Costello MJ & Wilson SP. **2011**. Predicting the number of known and unknown species in European seas using rates of description. *Global Ecology and Biogeography* **20**: 319–330.

Costello MJ, Wilson S & Houlding B. **2012**. Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Systematic Biology* **61**: 871–883.

Cribari-Neto F & Zeileis A. **2010**. Beta regression in R. *Journal of Statistical Software* **34**: 24.

Diniz-Filho JAF, Bastos RP, Rangel TFLVB, *et al.* 2005. Macroecological correlates and spatial patterns of anuran description dates in the Brazilian Cerrado. *Global Ecology and Biogeography* **14**: 469–477.

Ebach MC, Valdecasas AG & Wheeler QD. **2011**. Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics* **27**: 550–557.

Essl F, Rabitsch W, Dullinger S, *et al.* 2013. How well do we know species richness in a well-known continent? Temporal patterns of endemic and widespread species descriptions in the European fauna. *Global Ecology and Biogeography* **22**: 29–39.

Fan FT. **2004**. *British naturalists in Qing China*. Cambridge, Massachusetts: Harvard University Press.

Ferretti F, Myers RA, Serena F, *et al.* 2008. Loss of large predatory sharks from the Mediterranean Sea. *Conservation Biology* **22**: 952–964.

Gaston K. **1995**. Patterns in species description: a case study using the Geometridae

(Lepidoptera). *Biological Journal of the Linnean Society* **55**: 225–237.

Gaston KJ & Hudson E. **1994**. Regional patterns of diversity and estimates of global insect species richness. *Biodiversity & Conservation* **3**: 493–500.

Giam X, Scheffers BR, Sodhi NS, *et al.* 2012. Reservoirs of richness: least disturbed tropical forests are centres of undescribed species diversity. *Proceedings of the Royal Society B: Biological Sciences* **279**: 67–76.

Gibbons MJ, Richardson AJ, Angel M V., *et al.* 2005. What determines the likelihood of species discovery in marine holozooplankton: is size, range or depth important? *Oikos* **109**: 567–576.

He F. **2009**. Price of prosperity: economic development and biological conservation in China. *Journal of Applied Ecology* **46**: 511–515.

Hertz-Picciotto I & Rockhill B. **1997**. Validity and Efficiency of Approximation Methods for Tied Survival Times in Cox Regression. *Biometrics* **53**: 1151.

Joppa LN, Roberts DL, Myers N, *et al.* 2011a. Biodiversity hotspots house most undiscovered plant species. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 13171–6.

Joppa LN, Roberts DL & Pimm SL. **2011b**. How many species of flowering plants are there? *Proceedings of the Royal Society B: Biological Sciences* **278**: 554–559.

Kilpatrick J. **2014**. *Fathers of botany: the discovery of Chinese plants by European missionaries*. Chicago: University of Chicago Press.

Lu M & He F. **2017**. Estimating regional species richness: The case of China's vascular plant species. *Global Ecology and Biogeography* **26**: 835–845.

Lu L min, Mao L feng, Yang T, *et al.* 2018. Evolutionary history of the angiosperm flora of China. *Nature* **554**: 234–238.

Medellín RA & Soberón J. **1999**. Predictions of mammal diversity on four land masses. *Conservation Biology* **13**: 143–149.

Mora C, Tittensor DP, Adl S, *et al.* 2011. How many species are there on earth and in the ocean? *PLoS Biology* **9**: 1–8.

Myers N, Mittermeier RA, Mittermeier CG, *et al.* 2000. Biodiversity hotspots for conservation priorities. *Nature* **403**: 853–858.

Paxton CGM. **1998**. A cumulative species description curve for large open water marine animals. *Journal of the Marine Biological Association of the United Kingdom* **78**: 1389–1391.

Randhawa HS, Poulin R & Krkošek M. **2015**. Increasing rate of species discovery in sharks coincides with sharp population declines: implications for biodiversity. *Ecography* **38**: 96–107.

Reed RN & Roback SM. **2002**. Does body size predict dates od species description among north Amercian and Australian reptiles. *Global Ecology and Biogeography* **11**: 41–47.

Rosenberg R, Johansson MA, Powers AM, *et al.* 2013. Search strategy has influenced the discovery rate of human viruses. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 13961–4.

Steyskal GC. **1965**. Trend Curves of the Rate of Species Description in Zoology. *Science* **149**: 880–882.

Stork NE, Grimbacher PS, Storey R, *et al.* 2008. What determines whether a species of

insect is described? Evidence from a study of tropical forest beetles. *Insect Conservation and Diversity* **1**: 114–119.

Stork NE, McBroom J, Gely C, *et al.* 2015. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 7519–23.

Tedesco PA, Bigorne R, Bogan AE, *et al.* 2014. Estimating how many undescribed species have gone extinct. *Conservation Biology* **28**: 1360–1370.

Yang W, Ma K & Kreft H. **2013**. Geographical sampling bias in a large distributional database and its effects on species richness-environment models (W Daniel Kissling, Ed.). *Journal of Biogeography* **40**: 1415–1426.

Yang W, Ma K & Kreft H. **2014**. Environmental and socio-economic factors shaping the geography of floristic collections in China. *Global Ecology and Biogeography* **23**: 1284–1292.

Zapata FA & Robertson DR. **2007**. How many species of shore fishes are there in the tropical Eastern Pacific? *Journal of Biogeography* **34**: 38–51.

Table 1. Cox proportional hazard model with all variables (growth form as categorical data and arbor was used as the baseline variable).

	Effect size	Standard error	χ^2	d.f.	*P*-value
Growth form.fern	-0.311	0.08	16.60	1	<0.001
Growth form.herb	0.111	0.04	6.93	1	<0.05
Growth form.shrub	-0.063	0.04	3.11	1	0.08
Growth form.vine	-0.093	0.05	3.16	1	0.08
Range size	0.70	0.08	81.10	1	<0.001
Coast	0.23	0.02	109.76	1	<0.001
Maximum longitude	1.43	0.05	713.17	1	<0.001
Minimum longitude	-1.05	0.05	378.87	1	<0.001
Maximum latitude	0.83	0.07	153.16	1	<0.001
Minimum latitude	-0.77	0.04	410.62	1	<0.001
Frailty(genus)			13511.15	1329	<0.001
N = 25268					
Variance of random effect = 0.35					
Concordance = 0.735 (se = 0.002)					

Table 2. Estimated species richness for different growth forms.

	Number of discovered species	Estimated total number of species	Lower 95% bound	Upper 95% bound	Completeness
Fern	1977	--	--	--	--
Vine	1776	2035	1691	2380	0.873
Herb	13501	15611	13809	17412	0.865
Shrub	5210	6055	5023	7087	0.860
Arbor	2799	3254	2750	3759	0.860

Table 3. Linear regression of mean discovery time.

	Coefficient	Standard Error	*t*-value	*P*-value
Intercept	1903	1.1	170.50	<0.001
Human population density	-0.021	0.0084	-2.45	<0.05
Number of species	0.0033	0.0007	4.34	<0.001
Coast	-9.10	3.56	-2.56	<0.05
Mean latitude	-1.52	0.25	-4.81	<0.001
Multiple R^2: 0.88, Adjusted R^2: 0.86				

Table 4. Beta regression of province level inventory completeness.

	Coefficient	Standard Error	*z*-value	*P*-value
Intercept	0.89	0.53	1.69	0.09
Human population density	1.59e-03	4.71e-04	3.37	<0.001
Number of species	-9.67e-05	2.66e-05	-3.64	<0.001
Coast	-0.40	0.18	-2.18	<0.05
Area	-4.93e-07	2.16e-07	-2.29	<0.05
Mean latitude	0.085	0.014	6.02	<0.001
Pseudo R^2: 0.86				

Figure Legends

Figure 1. Fitted survival curves of the Cox proportional hazard models (A) stratified on variable ‘growth form’; (B) stratified on variable ‘coast’. Dashed lines show 95% confidence intervals.

Figure 2. (A) – (E) Species accumulation curves for five different taxonomic groups (based on 5 years-interval data). (F) Boxplot for discovery time of five different taxonomic groups.

Figure 3. (A) – (C) Mean species discovery time against human population density, number of discovered species, and latitude. (D) Boxplot of mean discovery time for inland and coastal provinces. (E) – (G) Species inventory completeness against human population density, number of discovered species, and latitude. (H) Boxplot of inventory completeness for inland and coastal provinces. Solid curves show the fitted smooth spline curves.

Figure 4. (A) Mean province-level species discovery date in China. (B) Species inventory completeness. Mollweide projection is used.

Figure 1

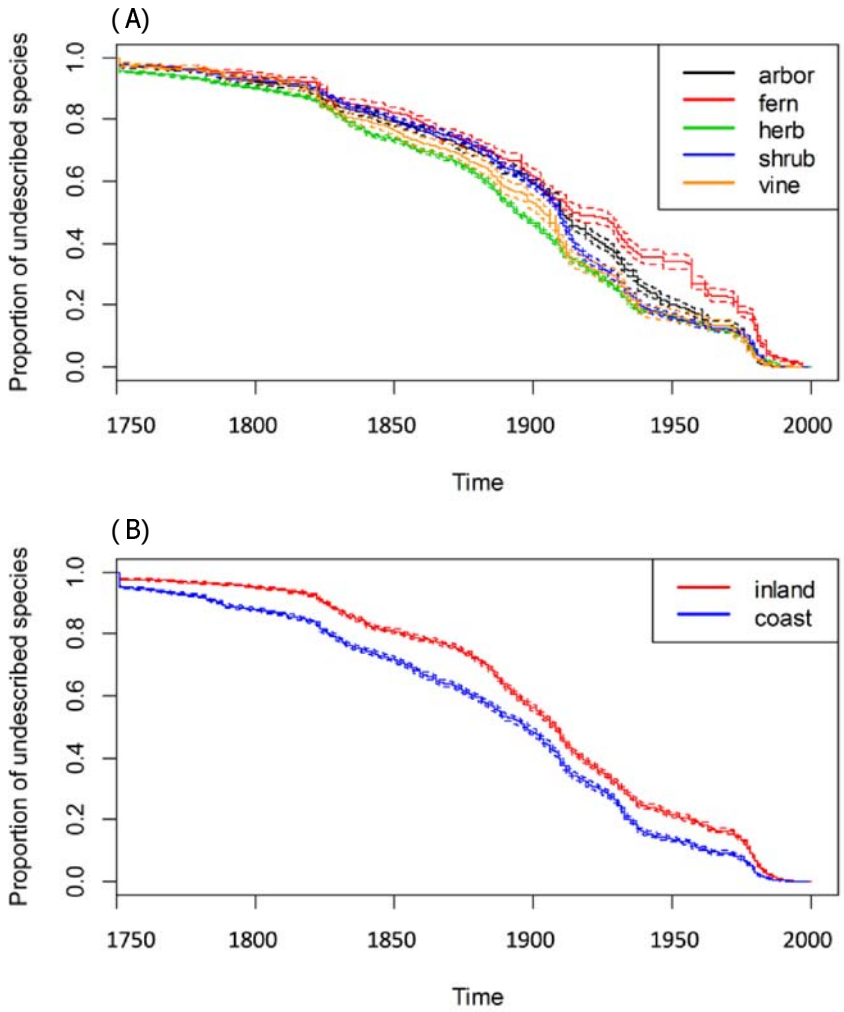

(A)
Proportion of undescribed species
0.0
0.2
0.4
0.6
0.8
1.0
1750
1800
1850
1900
1950
2000
Time
arbor
fern
herb
shrub
vine
(B)
Proportion of undescribed species
0.0
0.2
0.4
0.6
0.8
1.0
1750
1800
1850
1900
1950
2000
Time
inland
coast

Figure 2

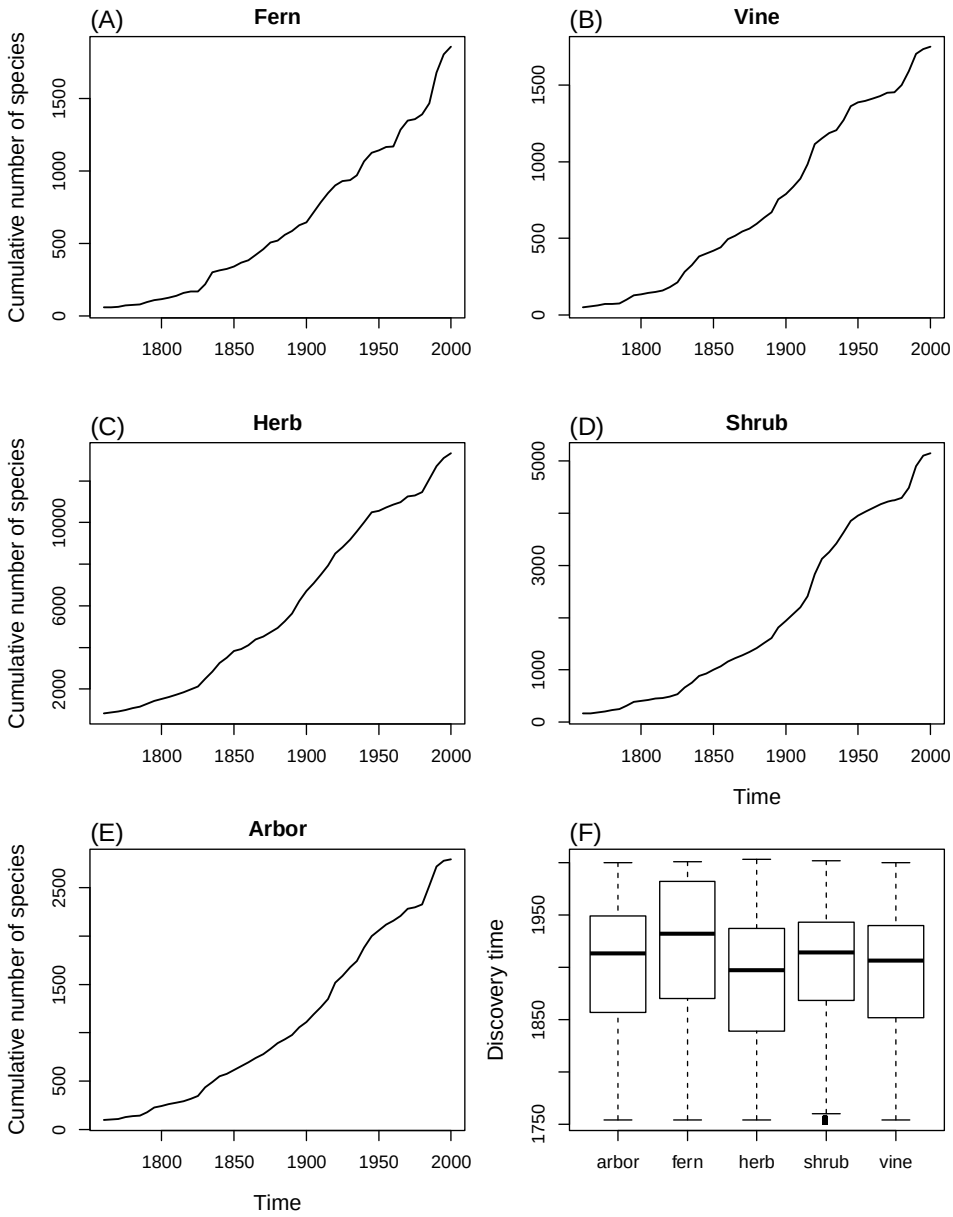

(A)
Fern
Cumulative number of species
0
500
1000
1500
1800
1850
1900
1950
2000
(B)
Vine
0
500
1000
1500
1800
1850
1900
1950
2000
(C)
Herb
Cumulative number of species
2000
6000
10000
1800
1850
1900
1950
2000
(D)
Shrub
0
1000
3000
5000
1800
1850
1900
1950
2000
Time
(E)
Arbor
Cumulative number of species
0
500
1500
2500
1800
1850
1900
1950
2000
Time
(F)
Discovery time
1750
1850
1950
arbor
fern
herb
shrub
vine

Figure 3

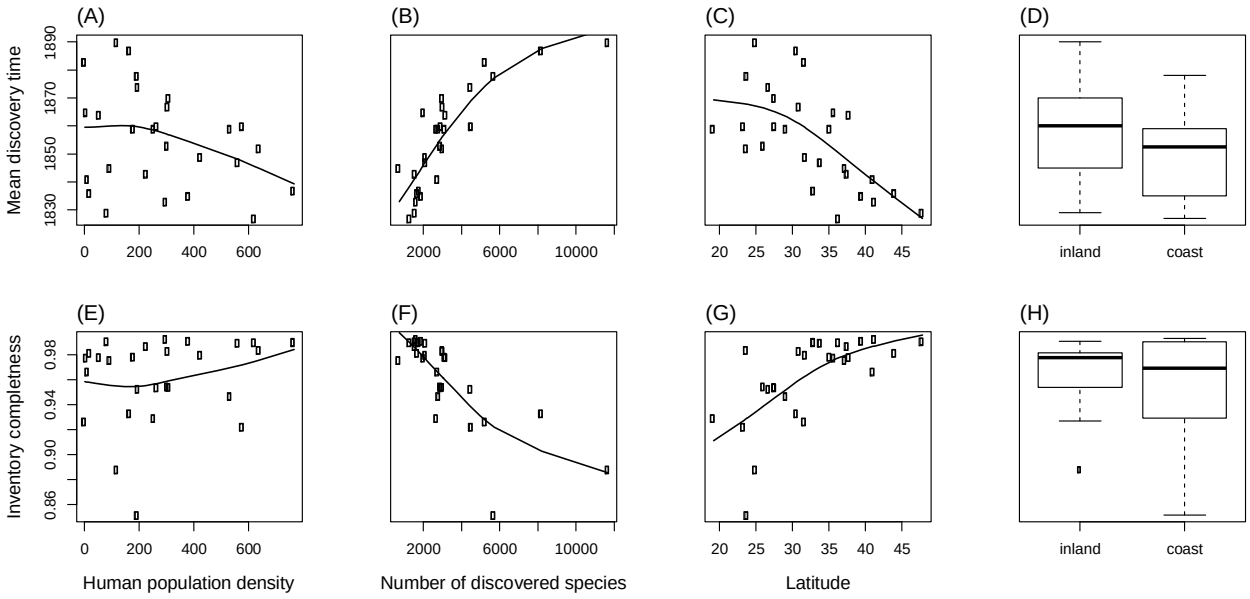
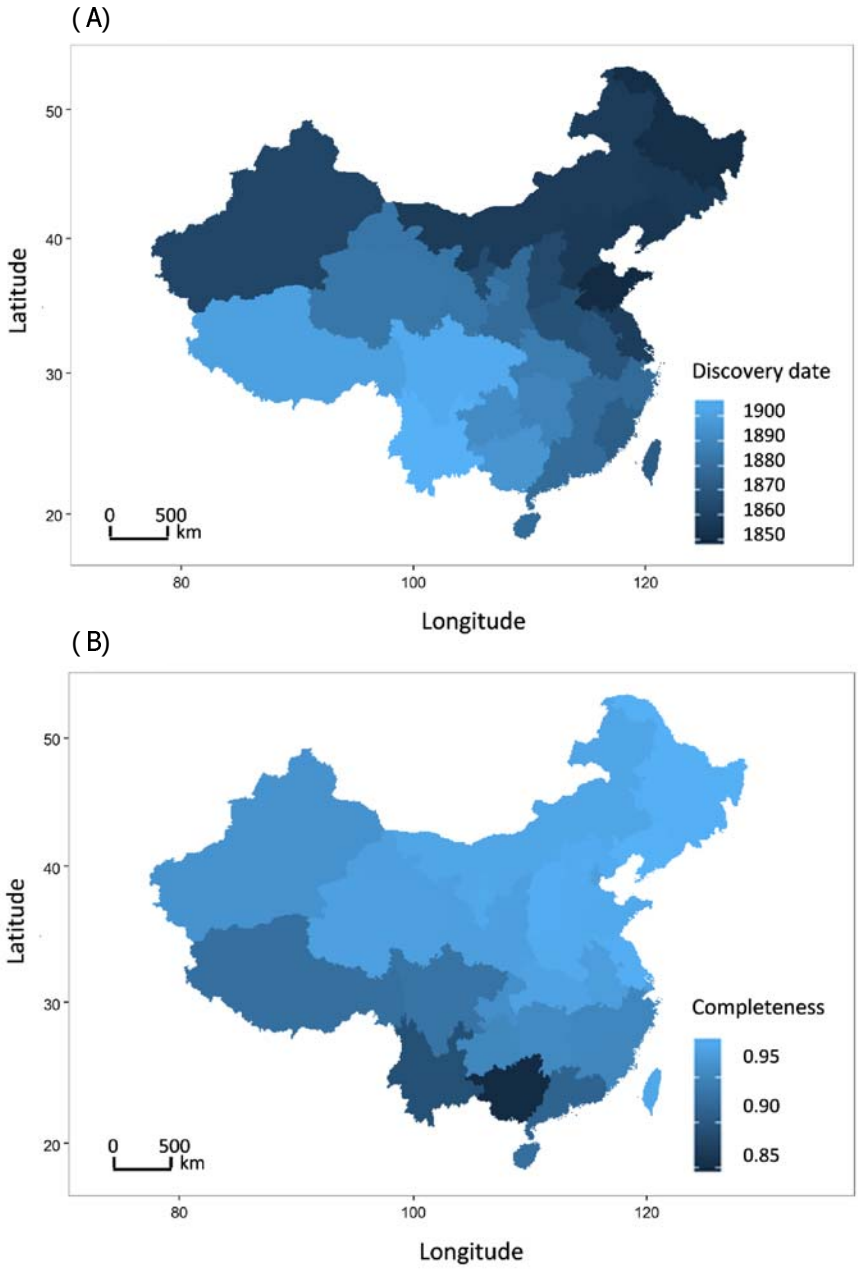

(A)
(B)
(C)
(D)
Mean discovery time
1830
1850
1870
1890
0
200
400
600
2000
6000
10000
20
25
30
35
40
45
inland
coast
(E)
(F)
(G)
(H)
Inventory completness
0.86
0.90
0.94
0.98
Human population density
Number of discovered species
Latitude

Figure 4


(A)
Latitude
Longitude
Discovery date
1900
1890
1880
1870
1860
1850
0 500 km
(B)
Latitude
Longitude
Completeness
0.95
0.90
0.85
0 500 km