Original article

The patterns of vascular plant discoveries in China

Muyang Lu[1, 2] , Lianming Gao[3], Hongtao Li[4], Fangliang He[1, 5]†

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. CAS Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China
4. Germplasm Bank of Wild Species in Southwest China, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China
5. Department of Renewable Resources, University of Alberta, Edmonton, Alberta T6G 2H1, Canada

Running head: Patterns of plant discoveries

Words in the abstract: 226

Words in main body: 4003

References: 51

Table: 4

Figure: 4

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

ABSTRACT

Botanical discovery has a long journey of revelation that contributes unparalleled knowledge to shape our understanding about nature. Plant discovery in China is an immanent part of that journey. To understand the patterns of plant discoveries in China and detect which taxa and areas harbor most numbers of undiscovered species, we analyzed the discovery times of 31093 vascular plant species described in *Flora of China*. 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. Trees and shrubs of seed plants have the highest discovery probability and ferns have the lowest discovery probability. Herbs hold the largest number of undiscovered species in China. Most undiscovered species are found in southwest China, where three global biodiversity hotspots locate. Spatial patterns of mean discovery year 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. Our results showed that socio-economic factors dictate the discovery patterns of vascular plants in China. Undiscovered species are mostly narrow-ranged, inconspicuous endemic such as herbs, which are prone to extinctions and locate in biodiversity hotspots in southwest China. We suggest that the future effort on plant discovery in China be prioritized in southwest China.

INTRODUCTION

Despite 270 years' discovery of species since Linnaeus, our knowledge about the biological diversity in nature is still far from complete. According to a recent estimate, fewer than 20% of species on earth have been so far discovered (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; Humphreys *et al.*, 2019).

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 in taxa like reptiles (Reed & Roback, 2002) and marine holozooplanktons (Gibbons *et al.*, 2005). More recently discovered species are often of greater conservation interest because they are more likely to be narrow ranged and rare, and thus are prone to extinctions (Diniz-Filho *et al.*, 2005; Bebber *et al.*, 2007a, 2010; Tedesco *et al.*, 2014; Xu *et al.*, 2019).

In addition to biological factors, species discovery is also influenced by human factors such as taxonomic effort, technology innovations, and social-economic events. For example, it is commonly observed for many taxa that species discovery rates dropped 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 developing regions where economic growth is often achieved at the expense of environmental degradation (He, 2009).

China harbors nearly one tenth of the plant species on earth (Joppa *et al.*, 2011b; Lu & He, 2017). 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 undiscovered species may locate is necessary for making decision on habitat protection and conservation managements.

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 naturalists in the 18th and 19th centuries

made their botanical collections in the coastal areas (Bretschneider, 1898; Fan, 2004), which likely had affected the patterns of discovery records.

There are more than 31000 vascular plant species documented in China, but this inventory is not complete and many new species still await discovery (Lu & He, 2017). Knowing what are those inconspicuous species and where they may locate is important to discover them. Therefore, 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 most likely are, and (3) to quantify what factors (e.g., species' range size and growth form) may influence the spatial distribution of plant discoveries in China. This study will contribute to understanding the pattern of species discoveries and their underlying factors and should be of significance to botanical discoveries of other regions beyond China. The identification of taxonomic and geographic gaps of undiscovered species will facilitate prioritizing our taxonomic and conservational efforts in the future.

MATERIALS AND METHODS

Data

Data including species names, authorities, discovery years, province-level biogeographic distributions, and genus-level growth forms were compiled from *Flora of China* (FOC) (Data available in Appendix S2) which has a totol number of ~31000 species (http://efloras.org). We treated discovery year as the time a species was first described in any name in a scientific publication. When estimating the number of undiscovered species, data after 2000 were

excluded as a routine to avoid the effect of delayed entrance of newly discovered 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. Genus-level growth forms were categorized as fern, herb, shrubs/trees and vine/liana. Shrubs and trees were categorized as one group because many species have both shrubs and trees as growthforms. Vines and liana include herbaceous vines, woody lianas and all other plants with climbing forms. When a genus has several different growth forms, we used the major growth form (which has the largest number of species within the genus) as the genus-level growth form. Range size, maximum/minimum latitude and maximum/minimum longitude 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 was found 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. Discovery time is the time it takes to discover a species, i.e., how long the species can "survive" not to be discovered. 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

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

The Cox PH regression model is:

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

where matrix βX is the linear component 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 generalized linear model. Compared to accelerated failure time model (AFT), Cox model is equivalent to a semi-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). 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 $h_0(t)$). Effect sizes are regression coefficients of the standardized predicting variables. All survival analyses were conducted using R package 'survival' (Therneau, 2015).

Estimating species richness

Because models that take into account of taxonomic effort (Joppa *et al.*, 2011a,b; Lu & He, 2017) did not converge at the province level, we used a modified logistic discovery model (Lu & He, 2017) to estimate species richness for different growth forms and for each province:

$$\Delta S_t = (a + bS_t)(S_{tot} - S_t) + \varepsilon_t ,$$

where ΔS_t is the number of species discovered per time interval (5 years in this study), S_{tot} is the total number of species in a region, S_t is the accumulative number of species discovered up to time *t* (=0, 5, 10, 15, 20,… years), *a* and *b* are fitting parameters and ε_t is the error term. Our goal is to estimate S_{tot}. Note that the logistic discovery model is different from the logistic regression model. The 'logistic' part of the discovery model derives from the logistic

shape of species discovery curve. 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 in a province.

Spatial patterns of discovery time and inventory completeness

We further examined the relationship between mean discovery time (i.e., the average number of years taken to discover the recorded species in an area) of a province 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 year 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 (i.e., population density, total number of species, whether coastal province or not, and area of each province) using R package 'betareg' (Cribari-Neto & Zeileis, 2010). Beta regression is preferred because it only assumes that the response variable ranges from 0 to 1 and is flexible to accommodate the shape of the distribution (symmetric or skewed). Logistic regression is for handling binary data, not for proportional data. If logistic regression is used for proportional data, the base counts that lead to the calculation of proportion must be known. Without that information, it is inappropriate to use logistic regression. In our data, the base population

which is the true total number of species is not known and hence logistic regression is not applicable.

RESULTS

Cox proportional hazard model for species discovery probability

The results in Table 1 show that ferns have the lowest discovery probability and trees and shrubs have the highest discovery probability among all taxonomic groups. For example, in year 1953 ferns have the highest proportion of undiscovered species and trees and shrubs have the lowest proportion of undiscovered species (Fig. 1A). Coastal species were discovered earlier 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, coastal distribution and growth form, suggesting the importance of geographic locations to discovery probability (Table 1).

Estimating species richness

The accumulation curve (i.e. the inverse of survival curve) of fern species has the steepest slope at recent years (Fig. 2A), suggesting that a considerable number of fern species remain to be discovered in the future. The logistic model estimates that the inventory completeness of ferns is 0.62. For seed plants, the estimated inventory completeness is 0.73 for herbs, 0.75 for shrubs and trees, and 0.68 for vines/lianas (Table 2). Herbs harbor the largest number of undiscovered species. The low ferns discovery rate is reflected by the fact that it took on average the longest time to discovery a fern species, while herbs were earlier to be discovered

($p < 0.05$ for all pairs except between tree/shrub and vine; Fig. 2E).

Spatial patterns of mean discovery year and inventory completeness

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

We estimate that in 18 of the 28 provinces, plant species discoveries are more than 90% complete. Provinces with the largest proportion of undiscovered species are in the southwest China (75.1% in Guangxi province and 82.7% in Yunnan province; Fig. 3B). Beta regression shows that inventory completeness is positively correlated with human population density, and negatively correlated with 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 (but see Collen *et al.*, 2004; Diniz-Filho *et al.*, 2005; Randhawa *et al.*, 2015). Worse still is that the data of species discoveries are not much used for answering questions important to

biogeography and biodiversity studies although there have been efforts in using such data for estimating regional and global species richness (Gaston,1995; Joppa *et al*. 2011a, b; Bebber *et al*. 2007a, b, 2014; Essle *et al*. 2013; Lu, & He 2017). The rich information provided by discovery history is especially valuable for filling the knowledge gap in biodiversity research (Meyer *et al.*, 2015; Meyer, Weigelt, & Kreft, 2016) because it provides guidance about when and where future discoveries are going to be made and what traits are important to future discovery (Collen *et al.*, 2004; Diniz-Filho *et al.*, 2005). Therefore, knowledge on species discovery is of great value for species conversation if we strive to describe all species before their extinction (Costello *et al.*, 2013; Tedesco *et al.*, 2014; Humphreys *et al*, 2019). In this study, we compiled data on vascular plant species discovered over 250 years 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 a larger range size is 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). Tree and shrub species of seed plants have the highest discovery probability while fern species have the lowest discovery probability. Herb and vine/liana species of seed plants all have similar discovery probabilities. 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), and is also shown by the result that province-level mean discovery time is negatively correlated with coastal distribution (Table

3). This is likely because coastal areas in China were most economically developed and much more accessible to western naturalists since the Opium War (Bretschneider, 1898; Fan, 2004).

Our results indicate that fern is likely to be the most under-discovered taxon in China because the species discovery curve for fern shows little sign of level-off (Fig. 2A, Table 2). Herbs have the largest number of undiscovered species (Table 2) and the second lowest discovery probability estimated from the Cox proportional hazard model (Table 1). Higher discovery probability usually leads to higher inventory completeness, which is shown by the concordance between ranks of discovery probability and ranks of inventory completeness among taxa in our results (Table 1, Table 2). Although the inventory completeness estimated from this study varies among the growth forms, the difference in inventory completeness is relatively small in magnitude (~ 2% between herb and tree/shrub). Herbs have the largest number of undiscovered species likely because the total number of herb species is larger than that of any other growth form of seed plants. We suspect that the effect of growth form on discovery probability and inventory completeness at least partly reflects the difference in the availability of taxonomic expertise, especially for ferns. The description of fern species started relatively late in China (~1920s, Chen, 1994) compared to other groups likely because of the difficulty in distinguishing subtle morphological characters, different species concept compared to seed plants and the lack of taxonomic expertise at that time. Given that herbs also contain the largest number of undiscovered species and that most specimens of undescribed species have already been preserved in herbaria (Bebber *et al.*, 2010), our study suggests that the limiting factor of discovering new species in China may be 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 human history did affect 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 efforts because the total number of species and number of discovered species are highly correlated (Joppa *et al*., 2011a; Giam *et al*., 2012; Fig. S3). We expect that new discoveries in the future will be most likely made in interior southwestern provinces with high species richness such as Xizang, 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). Yang *et al.* measured county-level inventory completeness using the slope of sample-based accumulation curves. Their assessment of inventory completeness might be biased by a sampling strategy that aims at collecting as many novel species as possible for a given amount of samples (Chen, 1994). We also argue that the slope of species accumulation curve is not a genuine measure of inventory completeness. Rather, it measures the variation in species composition in the samples used to construct the species-accumulation curve. Another possible reason for this discrepancy is that inventory completeness is scale-dependent. In a hypothetical scenario, even if the inventory completeness at the county level is on average 90%, the inventory completeness at the province level could be lower than 90% if most of the recorded species at

the 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, our study here does not support this advocacy. Instead, we suggest that future botanical collection efforts should be more allotted to the provinces of southwest China where there is high species diversity and the botanical inventory is least complete. Recent findings support our conlusion by showing that the majority of newly discovered species in China after the completion of *Flora of China* in 2013 came from the global biodiversity hotspots of Indo-Burma and Mountains of Southwest China (Cai *et al*., 2019), and southwest China (Yunnan, Guangxi, Sichuan, Xizang) and Taiwan are places where most new species were discovered during 2000 - 2019 (Du *et al.*, 2020).

There are two major limitations in this study. The first one is that species discovery year 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 (12917 species). The results show that the key drivers of discovery probability are mostly consistent but with some noticeable differences (Table S1): the effect size of coastal distribution became negative (Table S1) because the majority of these endemic species were discovered after 1900 (Fig. S1A), which removed the effect of early discoveries made by western naturalists. For the province-level analysis, excluding non-endemic species in China does not significantly change the spatial pattern of mean discovery year (Fig. S2A).

Human population density became negatively correlated with the mean discovery year of endemic species; area became positively correlated with mean discovery year; the effect of coastal distribution and number of species are no longer significant (Table S2). For province-level inventory completeness, the effects of predicting variables are mostly consistent with the model that includes all species except that coastial distribution and longitude also became significant variables (Table S3).

The second limitation is that our data do not distinguish the discovery of a new species from a species renamed from a known synonym. As data including a full list of all synonyms at each time step are not available, we are not able to model the transition rate from synonyms to valid names (Alroy, 2002).

In summary, our study shows that most underdiscovered vascular plant species in China are ferns and herbs of seed plants, which are mostly narrowly distributed endemic species in the southwest biodiversity hotspots 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 botanical study.

ACKNOWLEDGEMENTS

This work was funded by Sun Yat-sen University, East China Normal University, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB31000000) and the Science and Technology Basic Resources Investigation Program of China (Grant No. 2019FY100900).

We thank Xingli Giam and an anonymous reviewer for the critical comments that greatly improved this study.

Author contributions: FH and ML conceived the study; ML collected and analyzed the data; LG and HL provided the province-level checklist of *Flora of China*; ML led the writing with input from all authors.

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.

Cai J, Yu W Bin, Zhang T, *et al.* 2019. China's biodiversity hotspots revisited: A treasure chest for plants. *PhytoKeys* **130**: 1–24.

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.

Du C, Liao S, Boufford DE, *et al.* 2020. Twenty years of Chinese vascular plant novelties, 2000 through 2019. *Plant Diversity* **42**: 393–398.

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.

Humphreys AM, Govaerts R, Ficinski SZ, *et al.* 2019. Global dataset shows geography and life form predict modern plant extinction and rediscovery. *Nature Ecology and Evolution* **3**: 1043–1047.

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.

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

Meyer C, Kreft H, Guralnick R, *et al.* 2015. Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6**: 8221.

Meyer C, Weigelt P & Kreft H. **2016**. Multidimensional biases, gaps and uncertainties in

global plant occurrence information (JHR Lambers, Ed.). *Ecology Letters* **19**: 992–1006.

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.

Therneau T. **2015**. A Package for Survival Analysis in S. version 2.38.

Xu WB, Svenning JC, Chen GK, *et al.* 2019. Human activities have opposing effects on distributions of narrow-ranged and widespread plant species in China. *Proceedings of the National Academy of Sciences of the United States of America* **116**: 26674–26681.

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 tree/shrub used as the baseline variable).

	Effect size	Standard error	*P*-value
Growth form.fern	-0.82	0.03	<0.001
Growth form.herb	-0.13	0.01	<0.001
Growth form.vine.liana	-0.13	0.03	<0.001
Range size	0.51	0.06	<0.001
Coast	0.13	0.02	<0.001
Maximum longitude	1.43	0.05	<0.001
Minimum longitude	-1.05	0.05	<0.001
Maximum latitude	0.83	0.07	<0.001
Minimum latitude	-0.77	0.04	<0.001
N = 30944			
Concordance = 0.653 (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 (lower 95% -)
Fern	1921	2712	749	4676	0.622 (0.41-)
Vine/Liana	1202	1710	856	2563	0.685 (0.47-)
Herb	18370	23867	16982	30752	0.737 (0.59-)
Tree/Shrub	9451	12998	8855	15741	0.754 (0.60-)

Table 3. Linear regression of mean discovery year.

	Coefficient	Standard Error	*P*-value
Intercept	1941	1.5	<0.001
Number of species	0.0029	0.0008	<0.01
Coast	-13.8	4.49	<0.01
Mean latitude	-2.03	0.36	<0.001
Multiple R^2 = 0.82, Adjusted R^2 = 0.80			

Table 4. Beta regression of province level inventory completeness.

	Coefficient	Standard Error	*P*-value
Intercept	-0.39	0.23	<0.05
Human population density	8.70e-04	3.17e-04	<0.001
Area	-5.23e-07	1.33e-07	<0.001
Latitude	0.07	6.8e-03	<0.001
Pseudo R^2 = 0.85			

Figure Legends

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

Figure 2. (A) – (D) Species accumulation curves for four different taxonomic groups (based on 5 years-interval data). (E) Boxplot for discovery time of the four growth forms. “a” labels the groups with no significant difference in the TurkeyHSD test.

Figure 3. (A) Distribution of mean province-level species discovery year in China. (B) Distribution of species inventory completeness. Mollweide projection is used for the distributions.

Figure 4. (A) – (C) Mean species discovery year against human population density, number of discovered species, and latitude. (D) Boxplot of mean discovery year 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 1

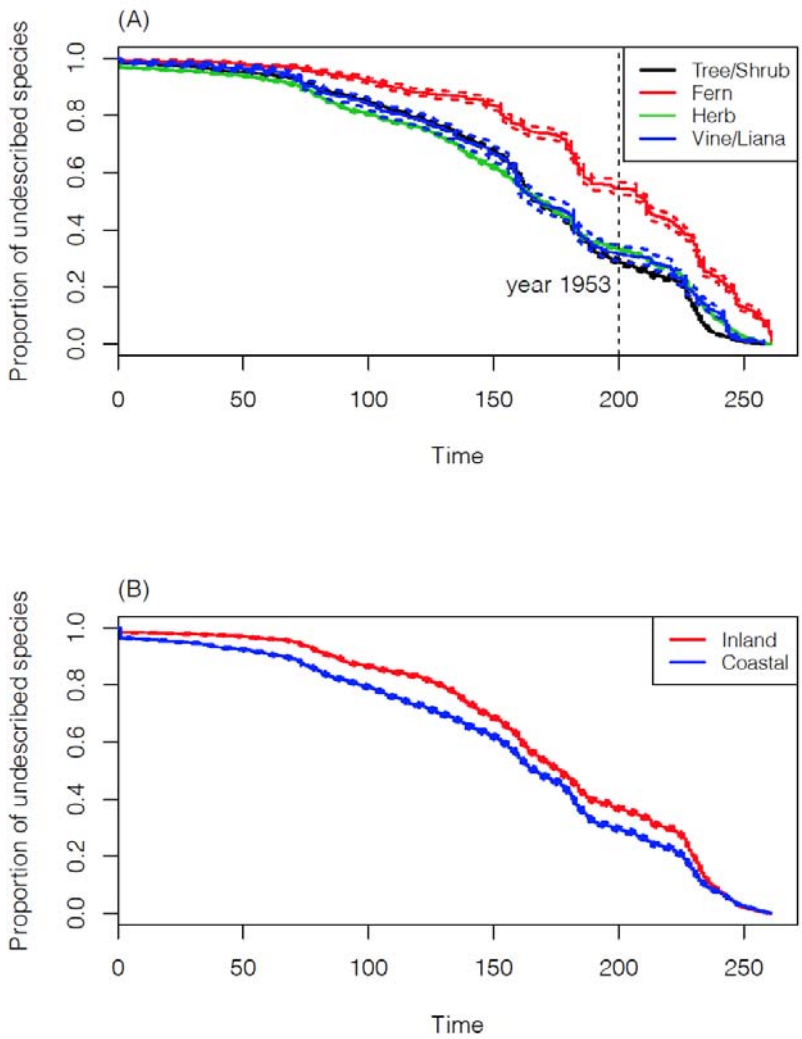

(A)
Proportion of undescribed species
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
Time
Tree/Shrub
Fern
Herb
Vine/Liana
year 1953
(B)
Proportion of undescribed species
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
Time
Inland
Coastal

Figure 2

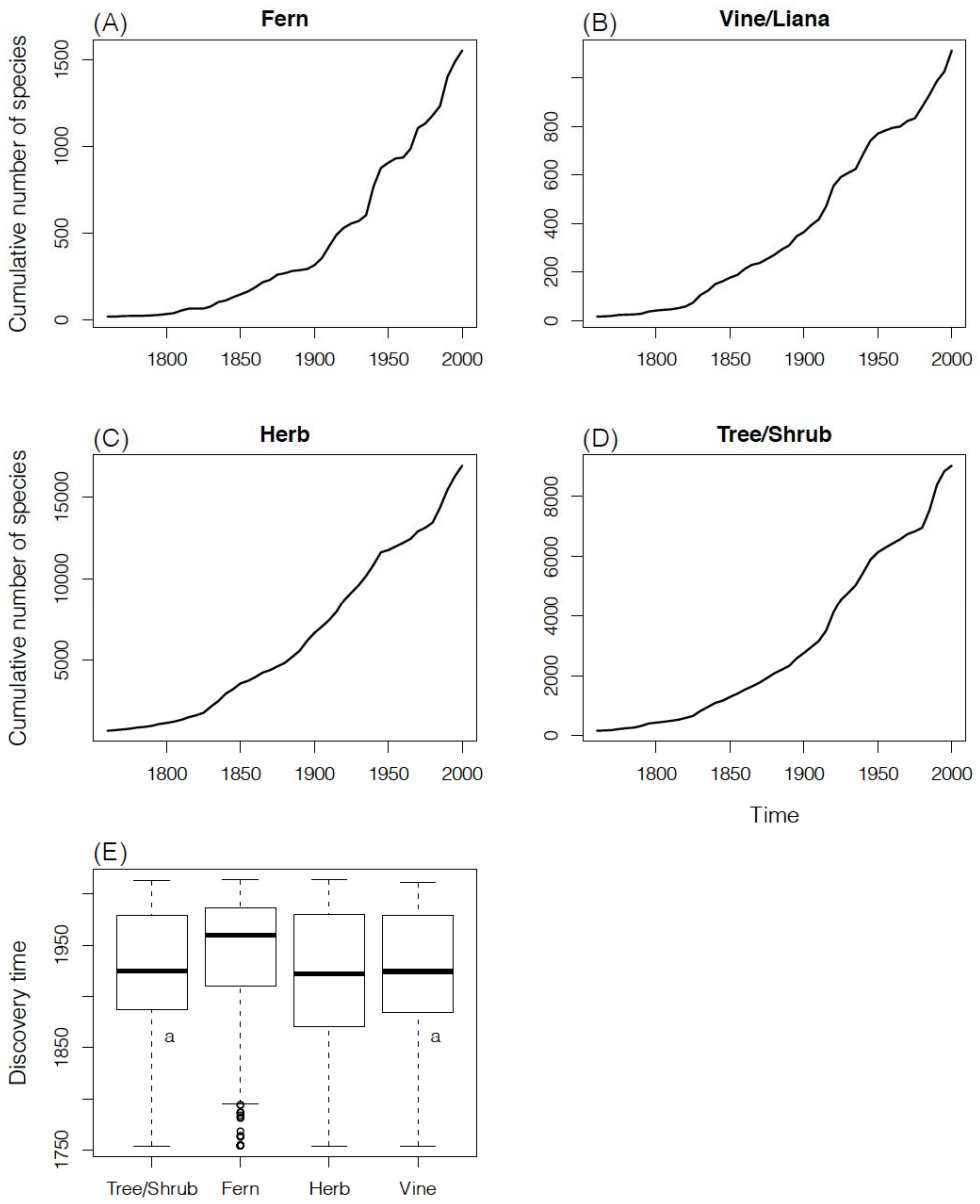

(A)
Fern
Cumulative number of species
0
500
1000
1500
1800
1850
1900
1950
2000
(B)
Vine/Liana
0
200
400
600
800
(C)
Herb
Cumulative number of species
5000
10000
15000
(D)
Tree/Shrub
0
2000
4000
6000
8000
Time
(E)
Discovery time
1750
1850
1950
a
a
Tree/Shrub
Fern
Herb
Vine

Figure 3

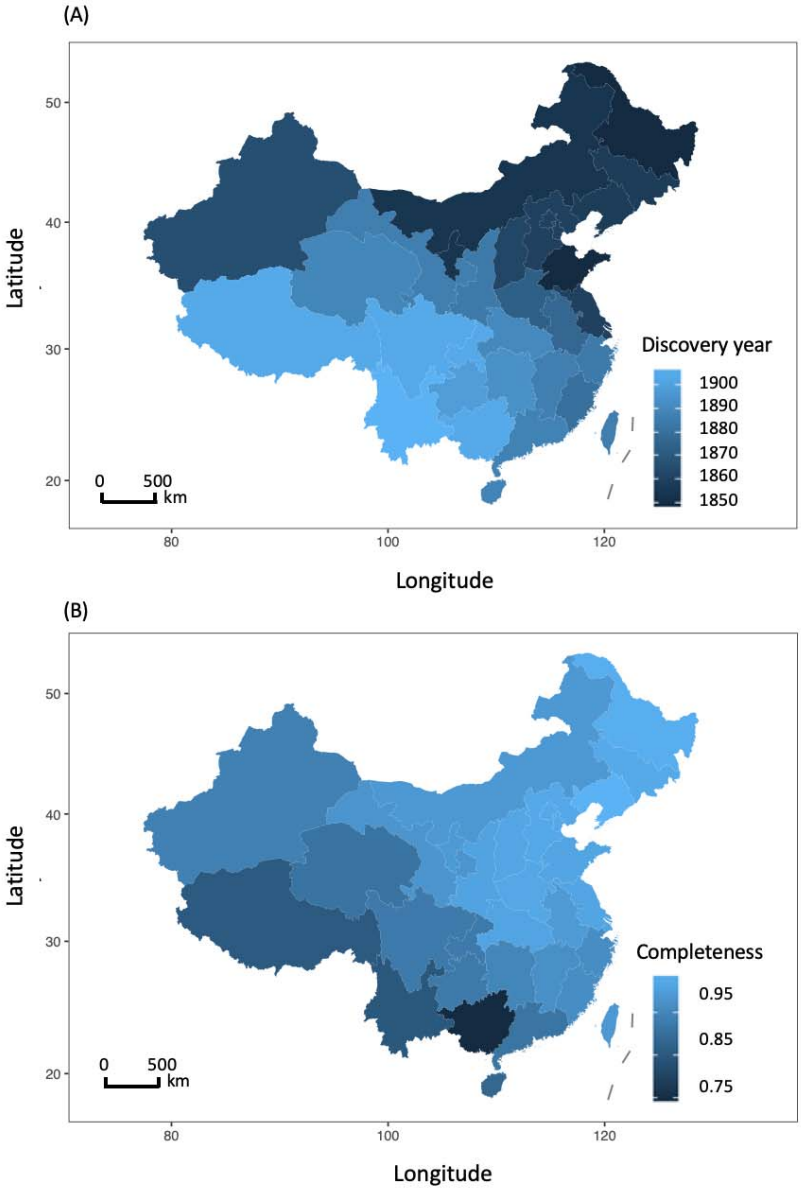
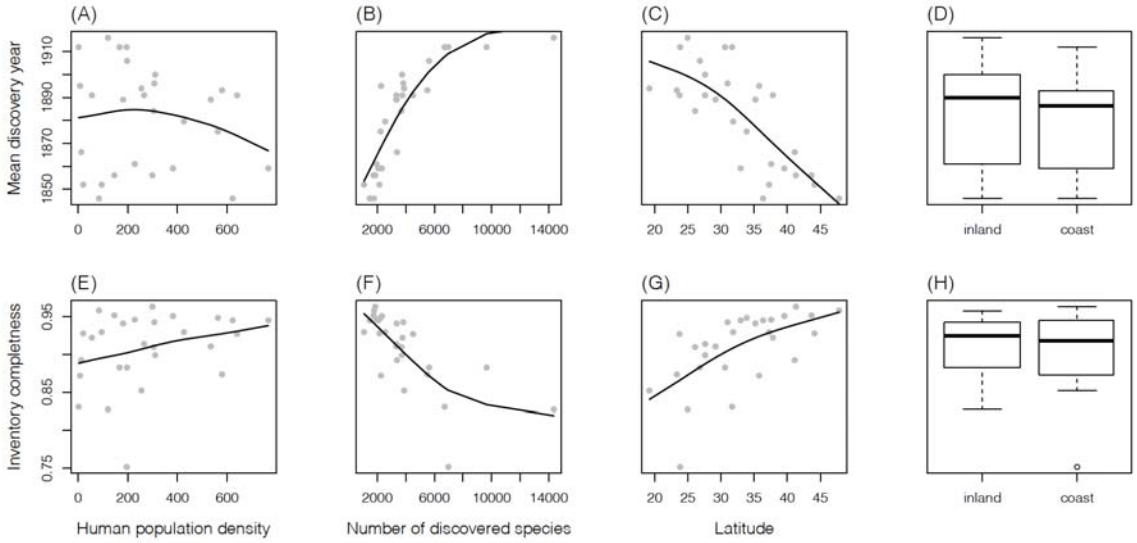

(A)
Latitude
Longitude
Discovery year
1900
1890
1880
1870
1860
1850
0 500 km
(B)
Latitude
Longitude
Completeness
0.95
0.85
0.75
0 500 km

Figure 4


(A)
(B)
(C)
(D)
Mean discovery year
1850
1870
1890
1910
0
200
400
600
2000
6000
10000
14000
20
25
30
35
40
45
inland
coast
(E)
(F)
(G)
(H)
Inventory completness
0.75
0.85
0.95
Human population density
Number of discovered species
Latitude

Appendices: **The patterns of vascular plant discoveries in China**

Muyang Lu, Lianming Gao, Hongtao Li, Fangliang He

Table S1. Cox proportional hazard model with all variables for endemic species in China (growth forms were coded as a categorical variable in which tree/shrub was used as the baseline variable).

	Effect size	Standard error	*P*-value
Growth form.fern	-0.83	0.04	<0.001
Growth form.herb	-0.082	0.02	<0.001
Growth form.vine	-0.23	0.10	<0.001
Range size	1.21	0.03	<0.001
Coast	-0.12	0.03	<0.001
Maximum longitude	1.30	0.10	<0.001
Minimum longitude	0.09	0.09	0.32
Maximum latitude	0.30	0.11	<0.01
Minimum latitude	-0.70	0.08	<0.001
N = 12917			
Concordance = 0634 (se = 0.003)			

Table S2. Linear regression of mean discovery time for endemic species.

	Coefficient	Standard Error	*P*-value
Intercept	1964	6.8	<0.001
Human population density	-0.013	0.0071	0.074
Area	2.30e-05	3.91e-06	<0.001
Mean latitude	-1.46	0.18	<0.001
Multiple R^2: 0.80, Adjusted R^2: 0.77			

Table S3. Beta regression of province-level inventory completeness for endemic species.

	Coefficient	Standard Error	*P*-value
Intercept	-1.49	5.18	<0.01
Human population density	2.06e-03	5.53e-04	<0.001
Coast	-0.49	0.17	<0.01
Area	-8.58e-07	3.16e-07	<0.01
Mean longitude	-0.13	0.047	<0.01
Mean latitude	0.68	0.14	<0.001
Longitude:latitude	-5.45e-03	1.23e-03	<0.001
Pseudo R^2: 0.76			

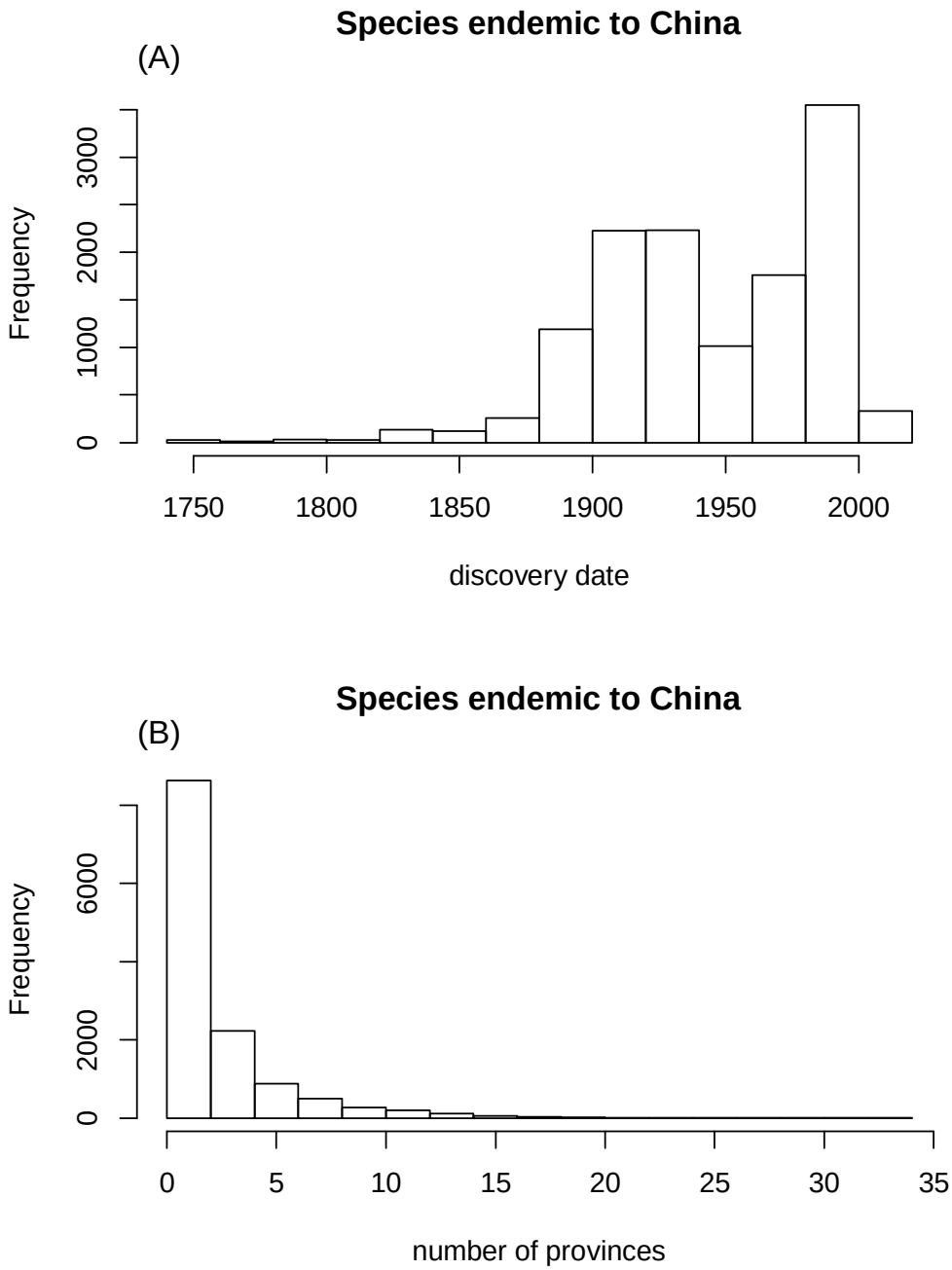


Figure S1. Histograms of discovery date (A) and the number of distributed provinces of endemic species to China (B).

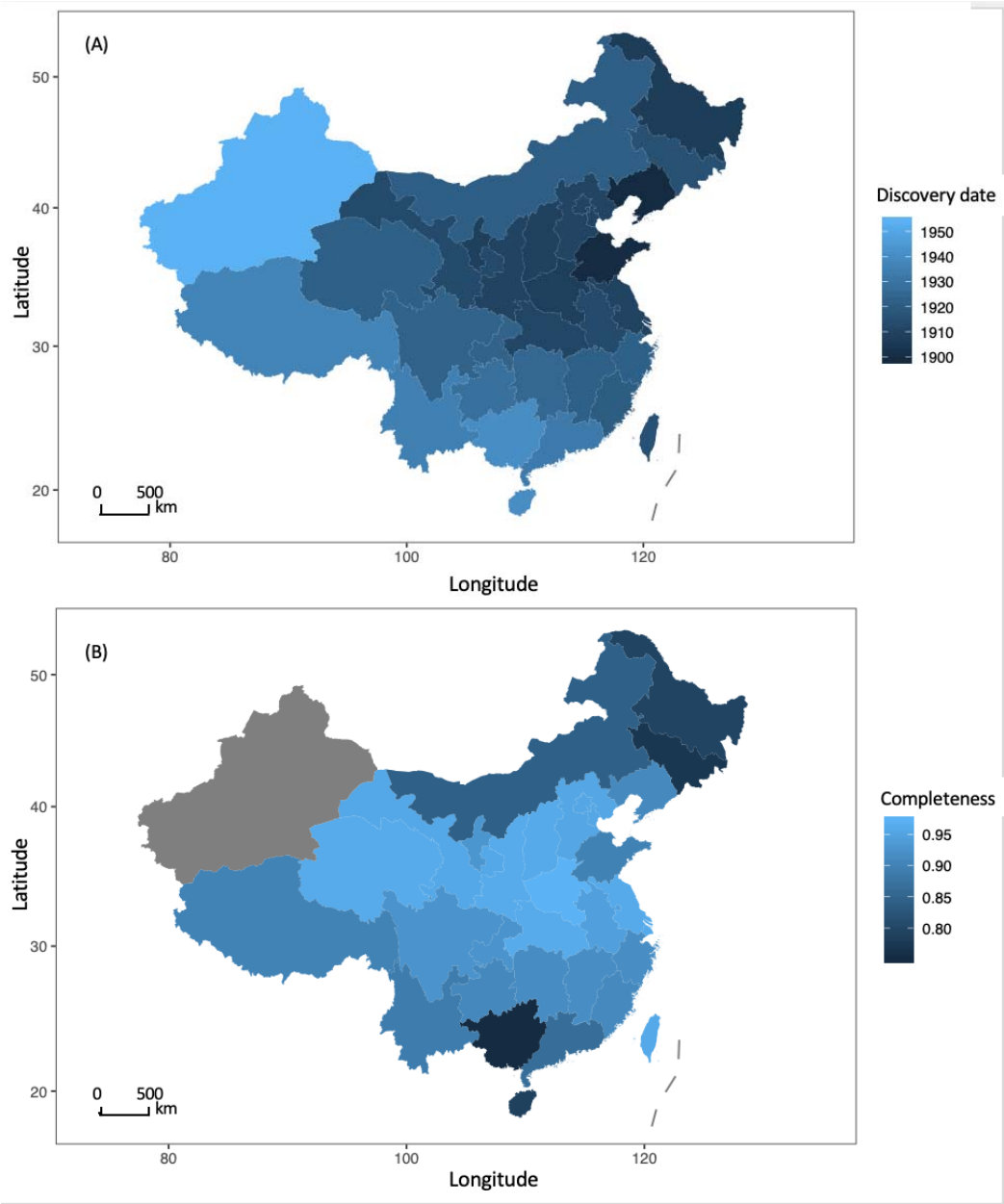


Figure S2. (A) Mean discovery time for province-level endemic species in China. (B) Species inventory completeness for province-level endemic species. Mollweide projection is used.

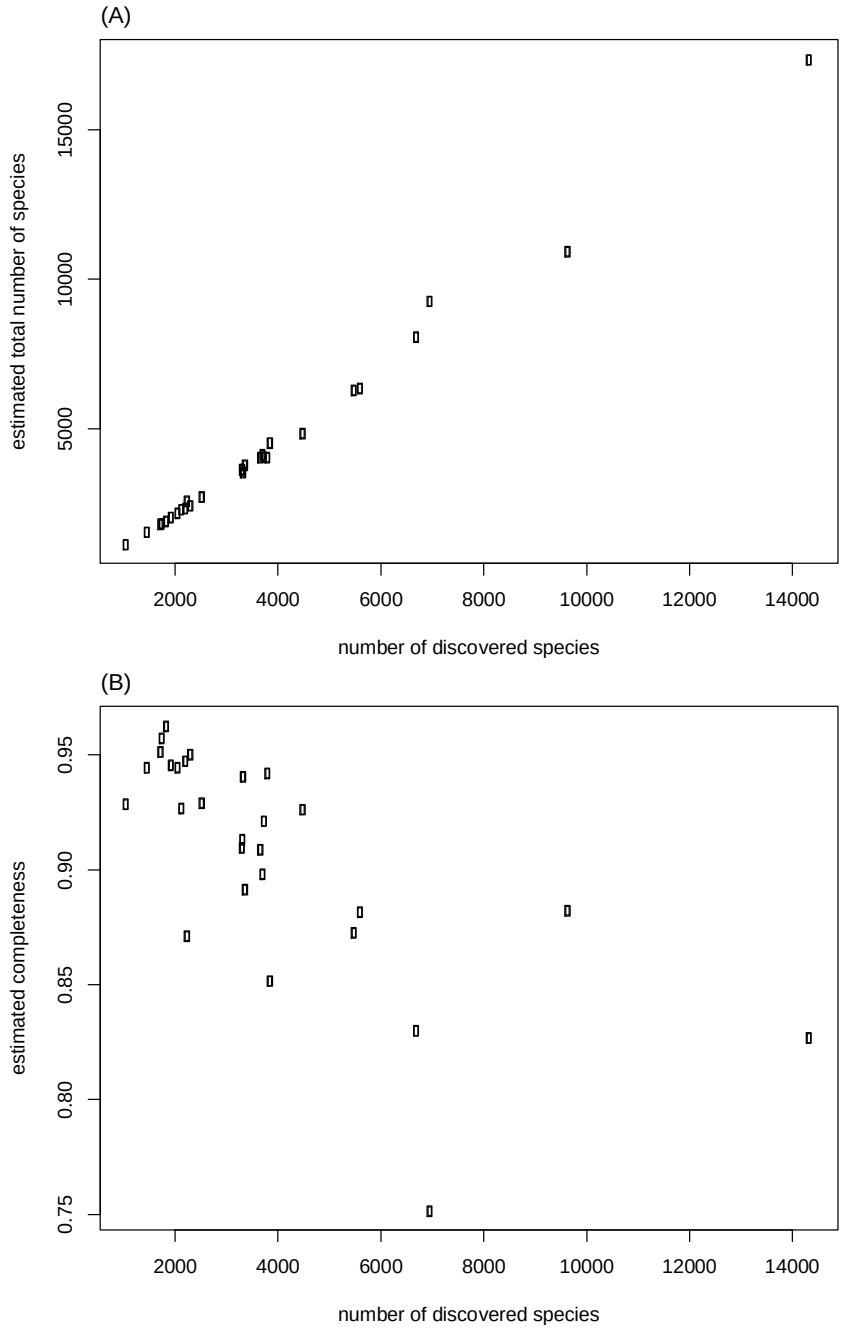


Figure S3. (A) Estimated total number of species against the number of discovered species at the province-level. (B) Estimated inventory completeness against the number of discovered species at the province-level.