Biological Conservation 231 (2019) 174–180

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/biocon

Spatial knowledge deficiencies drive taxonomic and geographic selectivity in data deficiency

Lina Zhao[a,b,1], Yuchang Yang[a,b,1], Huiyuan Liu[a,c], Zhangjian Shan[a,b], Dan Xie[a,b], Zheping Xu[d], Jinya Li[e,*]

[a] State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, 20 Nanxincun, Beijing 100093, China
[b] College of Life Sciences, University of Chinese Academy of Sciences, 19 Yuquan Road, Beijing 100049, China
[c] School of Life Sciences, Beijing Normal University, 19 Xinjiekouwai Street, Beijing 100875, China
[d] National Science Library, Chinese Academy of Sciences, 33 Beisihuan Xilu, Zhongguancun, Beijing 100190, China
[e] State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Beijing 100085, China

ARTICLE INFO

Keywords:
Data deficient
Non-randomness
Selectivity
Spatial patterns
Vulnerability

ABSTRACT

The uncertain threat status of species inevitably influences their focus on conservation. Just as in extinction risk, the non-randomness phenomenon related to uncertainty (also referred to as *selectivity*), which is a certain character cluster in some groupings, also exists in data deficiency of species' knowledge. In order to illustrate this kind of non-random phenomenon and explain the uncertainties it caused, we performed a hypergeometric test on taxonomic and geographic groupings of China's spermatophyte species and quantified two factors— *frequency of collections* and *spatial accessibility*— to indicate the primary causes of spatial knowledge deficiencies. We found that selectivity in data deficiency exists both taxonomically and geographically. Fifteen of the families were more deficient than expected, which included 30.0% of species and 56.3% ranked data deficient (DD). Among these, eight families were statistically highly significant with $p < 0.001$ and included 25.2% of species and 50.0% ranked DD. Forty-six families were less deficient than expected. With respect to floristic division, four of 29 floristic regions and subregions were more deficient than expected, and seven were less deficient than expected. Spatial autocorrelation analysis on DD species suggested an aggregated pattern of data deficiency in China (Moran's $I = 0.58$, z-score = 27.0, $p < 0.001$), and these areas that contained the highest numbers of DD species also contained the highest number of species (Spearman's $R^2 = 0.879$, $p < 0.001$). However, the largest DD ratio had a low correlation with the richest DD spatial diversity. Moreover, we found the larger the DD ratio was, the lower the frequency of collections and the poorer the spatial accessibility would be. In the research, we showed that the uncertainties associated with DD species would alter the non-randomness in the selectivity of data deficiency and further affect the focus of conservation. Only with a full understanding of the process and mechanisms of data deficiency can we determine where and what kind of actions are necessary to improve the knowledge of plant diversity.

1. Introduction

Effective conservation measures are needed urgently because of the unprecedented biodiversity loss today (Barnosky et al., 2011). Spatial assessments involving the identification of key areas or species that require conservation represent the first step in adequate conservation planning and implementation (Mittermeier et al., 1998; Knight et al., 2007; Langhammer et al., 2017). Furthermore, limited funding demands prioritization of those areas or species to ensure the efficient use of resources and effective conservation actions (Cowling et al., 2004; Wilson et al., 2007; Knight et al., 2008). In Target 11 of the Aichi biodiversity targets of the Convention on Biological Diversity (CBD 2010–2020) "by 2020, at least 17 percent of terrestrial and inland water, and 10 percent of coastal and marine areas are conserved through systems of protected areas…." As for conservation of endangered species, Target 12 of Aichi describes "By 2020 the extinction

* Corresponding author at: State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Beijing 100085, China.
E-mail address: jyli@rcees.ac.cn (J. Li).
[1] Equal contributors to this work.

https://doi.org/10.1016/j.biocon.2018.12.009
Received 23 July 2018; Received in revised form 30 November 2018; Accepted 7 December 2018
Available online 25 January 2019
0006-3207/

of known threatened species has been prevented and their conservation status, particularly of those most in decline…", and Target 2 of GSPC targets demanding "an assessment of the conservation status of all known plants as far as possible, to guide conservation action." There is a consensus that stopping declines in biodiversity is a critically important step in achieving more ambitious conservation goals.

The extinction risk assessment and its value of the threatened species in the IUCN Red List has been widely recognized (Gärdenfors et al., 2001; IUCN, 2001; Rodrigues et al., 2006; Hayward, 2009; Juslen et al., 2013; Saiz et al., 2015; Bennun et al., 2017). In the global assessment, all species on the Red List are listed according to 9 categories, 7 of which can indicate the extinction risk (except for Data Deficient (DD) and Not Evaluated (NE)). Avoiding species extinction can be seen as the fundamental goal of biodiversity conservation. The five quantitative Red List criteria explicitly defined as estimating extinction risk. Critically Endangered (CR), Endangered (EN), and Vulnerable (VU) are threatened categories and generally will be considered for prioritization (IUCN, 2016). However, data deficiencies of species in certain taxonomic groups and spatial regions may prevent them from being considered for protection. In IUCN Red List, the category DD highlights taxa for which information is insufficient to make a sound status assessment. Although the criteria are highly quantitative and defined, one can use projections, assumptions and inferences in order to place a taxon in the appropriate category. DD species introduce high uncertainty into the extinction risk assessment in the groups level due to their unknown risk status (Good et al., 2006; Butchart and Bird, 2010; Hoffmann et al., 2010; Sousa-Baena et al., 2014; IUCN, 2011; Jarić et al., 2016; Roberts et al., 2016; Bland et al., 2017). Although they are not classified within the threat categories, DD species may still face high extinction risks as those judged to be threatened (Howard et al., 2014; Bland et al., 2015; Jetz and Freckleton, 2015; Jarić et al., 2016; Roberts et al., 2016).

However, there are contrasting attitudes about treating DD when setting conservation priorities. Some authors have suggested that DD species should be regarded as potentially threatened and deserve equal conservation as truly threatened, at least until there is more evidence that the DD rating can be designated clearly as "potentially threatened" (Mace et al., 2008; IUCN, 2011; Morais et al., 2013). Even these conservation measures may not work immediately, but they can reduce the speed of population decline in the near future. Conversely, others believe that if DD species are protected equally and unquestioningly, this will affect the utilization of protected resources for threatened species (Joaquim et al., 2012). Therefore, to solve such conflicts, additional methods and concepts directed to DD species are required to classify them into a more reliable extinction level. For example, many studies have shown that there is usually a significant tendency for extinction risk to be concentrated within certain large families or specific spatial areas, which is referred to as the selectivity of threat level (Bielby et al., 2006; Hoffmann et al., 2010; Böhm et al., 2013; Vaira et al., 2017). However, a similar phenomenon of selectivity also exists in data deficiency in that DD species often are taxonomically and spatially biased as well (Bielby et al., 2006; Fritz and Purvis, 2010; Hoffmann et al., 2010; Bland et al., 2012; Böhm et al., 2013; Vaira et al., 2017).

In recent years, China has paid more and more attention to research on biodiversity and conservation (Huang et al., 2011; Qin et al., 2017a, 2017b; Zhao et al., 2016). As one of the megadiversity countries in the world, China contains four of the 25 global biodiversity hotspots (Myers et al., 2000). China is rich in species diversity and endemic species diversity. There are about 30,000 spermatophyte species in China, which account for > 80% of all higher plants. China is also on a high threat level of biodiversity after decades of socio-economic development. Currently, > 10% of species are threatened. However, there are still few studies on biodiversity conservation related to the whole spermatophyte species in China, let alone studies on DD species. To fill this knowledge gap, we analyzed the selectivity of data deficiency at the taxonomic and geographic levels and used two quantifiable factors to explain the primary causes of spatial knowledge deficiency. Finally, we make recommendations for setting conservation priorities based on our results.

2. Materials and methods

2.1. IUCN Red List categories of spermatophytes in China

We divided non-extinct species into three types: threatened (CR, EN, and VU), Data Deficient (DD) and non-threatened (LC and NT). There are 30,319 spermatophyte species in China's Higher Plants Red List (RLCHP), of which 3511 are categorised as Threatened, and 3011 are categorised as DD. If DD species are included, the percentage of threatened species is 11.58% (3511/30319), whereas if DD species are omitted from the calculation, 12.86% are threatened (3511/27301). Two tags, DDP and DDT, were assigned to DD species by the RLCHP. DDP was defined as those with insufficient information on population size, trends, distribution and/or threats, while DDT was those with uncertain taxonomic status (IUCN, 2011). As DDT species have uncertainty in their taxonomy (IUCN, 2011), we excluded them from our spatial analysis in this paper.

2.2. Spatial information and spatial processing

Detailed spatial distribution information was important in our analysis. Some was taken from the RLCHP to which almost 300 experts contributed, while others were derived from large-scale specimen information platforms, such as the China Virtual Herbarium (CVH, http://www.cvh.ac.cn/en) and the National Science & Technology infrastructure (NSII, http://www.nsii.org.cn/2017/home-en.php). Some shortcomings are inevitable in our geographic distributions, such as duplication of records, invalid locations, erroneous coordinates, etc. We try to avoid these issues in the spatial pre-processing. First, we combined data from CVH and NSII into one database and then did some data pre-processing. In this process, we needed to confirm that one species had only one record in one collection locality. Then, we removed specimens with clearly invalid locations (i.e., outside the boundaries of China), specimens missing with georeferenced location data and erroneous coordinates, and specimens that were duplicates from the same collection date. Finally, 25% (5,002,566 original records, 1,252,047 retained) of the initial specimens were left for further analysis. Approximately one quarter of spermatophytes lack detailed distribution information, which limited the scale of the study. All of different forms of the datasets, except the layer of floristic regions, were unified to be spatial gridded layers with a resolution of one degree (~110 × 110 km at the equator). Some values of grid layers were normalized in the range of 0 to 100 (Chen, 2011).

We analyzed the spatial patterns of data deficiency species' selectivity after removing the protected areas in which DD species are distributed. Furthermore, we also analyzed the primary causes of spatial knowledge deficiency as in previous studies (Kier et al., 2005). Two quantifiable factors were used to explain the causes of spatial knowledge deficiency. We counted *frequency of collections* events based on the CVH to indicate the extent of scientific exploration. We used *spatial accessibility* to indicate the difficulty of reaching the research destination. We chose three density indices to synthesize a definition of spatial accessibility: length of river, of railway, and of road in each grid. Moreover, it should be noted that deviations might exist in some results in the Taiwan region as it lacks some suitable layers during the calculation of spatial accessibility. More details of spatial layers are showed in Table S1 of Appendix A. In addition, Wu (1979) also divided China floristically and partitioned the country into floristic regions based on the distributions of typical taxonomic groups and dominant vegetation. The details are presented in Fig. 1a.

In the spatial analysis, we used Moran's I as a measure of spatial autocorrelation on three layers: layers of DD Species distribution, layers

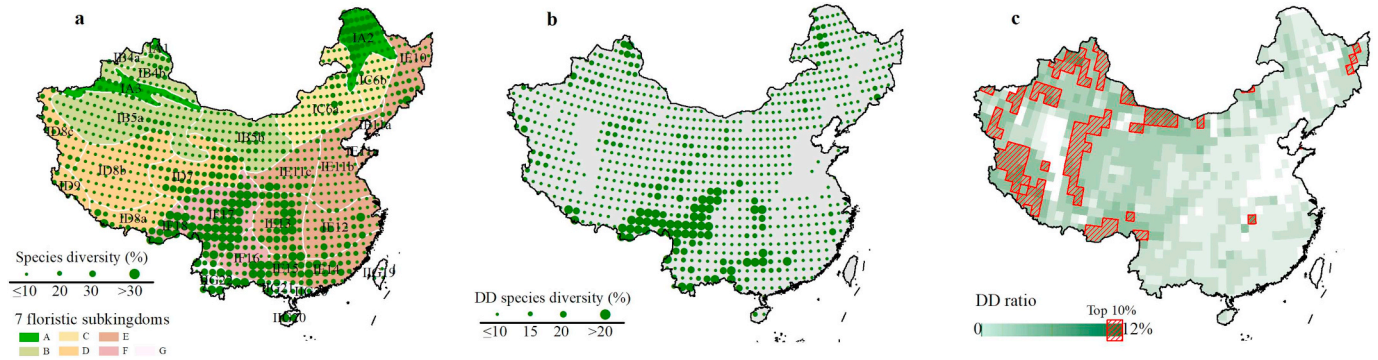

Fig. 1. Species diversity pattern and floristic divisions of China. The spatial patterns of species diversity and 29 floristic regions (a), DD species diversity (b), DD ratio (c). Wu (1979)'s floristic hierarchical divisions are marked with codes, where the leading roman numeral represents kingdom, the following upper-case roman letter represents subkingdoms, the arabic numeral represents region, and the possible lower-case roman letter at the end represents subregion. I. Holarctic Kingdom; II. Paleotropic Kingdom. A. Eurasia Forest Subkingdom; B. Central Asia Desert Subkingdom; C. Eurasia Grassland Subkingdom; D. Qinghai-Tibet Plateau Subkingdom; E. Sino-Japan Forest Subkingdom; F. Sino-Himalaya Forest Subkingdom; G. Southeast Asia Subtropics Subkingdom. 1. Altay Region; 2. Daxing'anling Region; 3. Tianshan Region; 4. West Central Asia Region (a. Tacheng-Yili Subregion; b. Dzungaria Subregion); 5. East Central Asia Region (a. Kashgar Subregion; b. West-South Mongolia Subregion); 6. Mongolian Grassland Region (a. East Mongolia Subregion; b. Northeast China Plain Subregion); 7. Tangut Region; 8. Pamir-Kunlun-Tibet Region (a. Ü-Tsang Subregion; b. Changtang Subregion; c. Pamir-Kunlun Subregion); 9. West Himalaya Region; 10. Northeast China Region; 11. North China Region (a. Liaodong Peninsula - Shandong Penisula Subregion; b. North China Plain and Montane Subregion; c. Loess Plateau Subregion); 12. East China Region; 13. Central China Region; 14. South China Region; 15. Yunnan-Guizhou-Guangxi Region; 16. Yunnan Plateau Region; 17. Hengduan Mountain Region; 18. East Himalaya Region; 19. Taiwan Region; 20. South China Sea Region; 21. Tonkin Gulf Region; 22. Yunnan-Myanmar-Thailand Region (Wu, 1979).

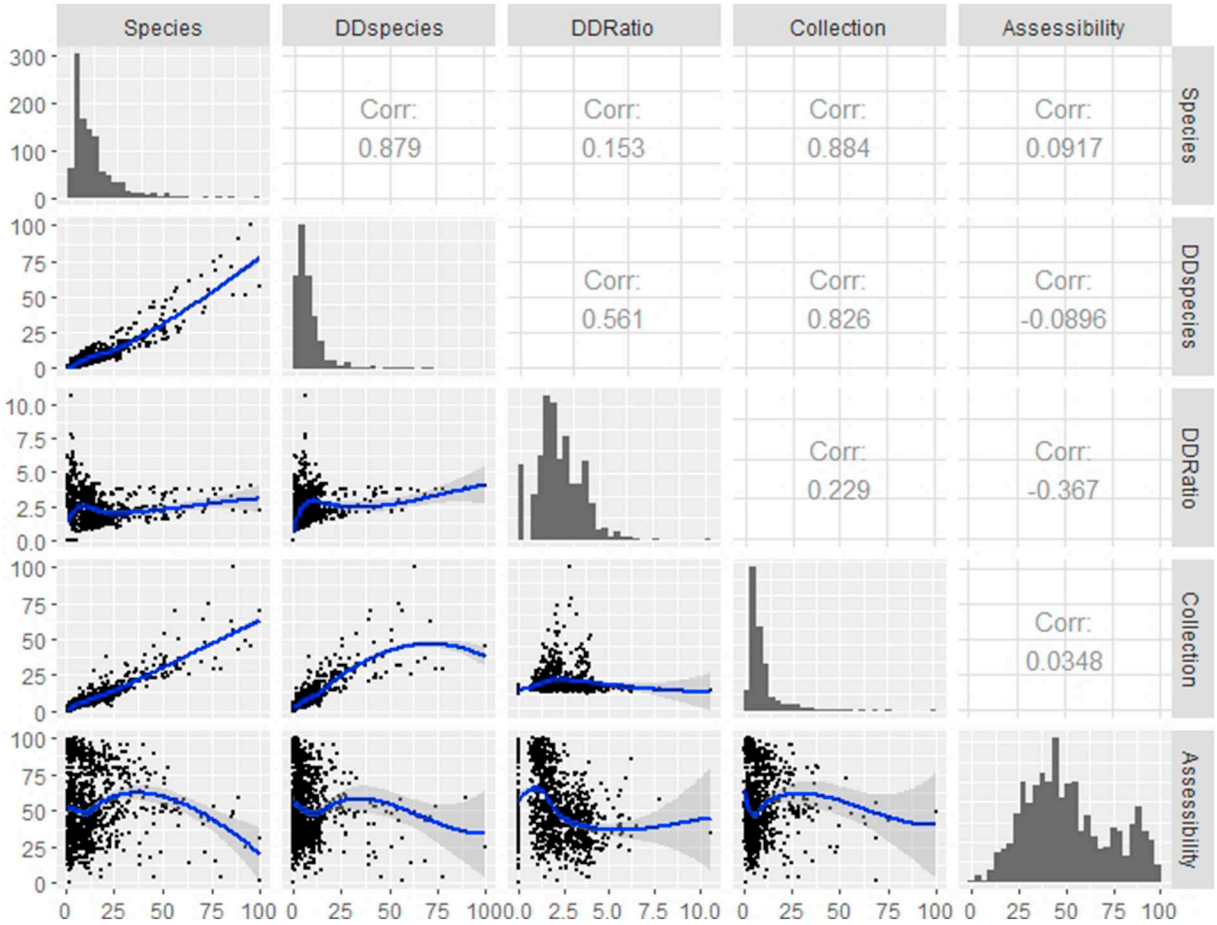

Fig. 2. The matrix of Spearman's R^2s between the two primary causes of spatial knowledge deficiency (frequency of collections and spatial accessibility) and three biodiversity measures (species, DD species, and DD Ratio).

of frequency of collections, layers of spatial accessibility. The values of Moran's I range from -1 (indicating a perfect dispersion) to 1 (indicating a perfect correlation). A Moran's $I \in (0,1)$, indicates a positive spatial autocorrelation; $I \in (-1,0)$ indicates a negative spatial autocorrelation, and $I = 0$ indicates a random spatial pattern (Li et al., 2007). We also utilized Spearman's rank correlations (Spearman's R^2) and spatial coverage ratio (SCR) to explain the correlation between two primary causes of spatial knowledge deficiency (frequency of

Table 1
Taxonomic selectivity of data deficiency and extinction risk of spermatophytes with different treatments of DD (DD excluded, DD considered non-threatened, and DD considered threatened).

Family	Total species	Threatened species	DD species	DD ratio	Family	Total species	Threatened species	DD species	DD ratio
Schisandraceae***	31	2	12	38.7	Moraceae††	160	28	7	4.4
Ericaceae***	1012	142	291	28.8	Euphorbiaceae†††	356	26	14	3.9
Piperaceae***	66	5	18	27.2	Theaceae†††	346	91	13	3.8
Elaeagnaceae***	70	8	18	25.7	Hamamelidaceae†	77	38	3	3.8
Asteraceae***	2139	20	507	23.7	Asclepiadaceae†††	272	19	10	3.7
Lardizabalaceae*	38	3	8	21.0	Urticaceae†††	388	13	14	3.6
Berberidaceae***	308	48	61	19.8	Aristolochiaceae†	82	40	3	3.6
Celastraceae**	178	34	30	16.9	Lamiaceae†††	864	42	27	3.1
Poaceae***	1881	53	308	16.4	Styracaceae†	65	22	2	3.1
Boraginaceae***	291	9	47	16.1	Rhamnaceae†††	171	11	5	3.0
Caryophyllaceae***	364	10	56	15.4	Clusiaceae††	100	18	3	3.0
Zingiberaceae*	202	22	29	14.3	Polygonaceae†††	223	13	6	2.7
Rosaceae***	1206	62	158	13.1	Cucurbitaceae†††	157	28	4	2.6
Papaveraceae*	451	19	59	13.1	Primulaceae†††	571	45	15	2.6
Gesneriaceae*	496	73	62	12.5	Lauraceae†††	465	99	12	2.6
Ranunculaceae†	1081	87	91	8.5	Tiliaceae†	79	11	2	2.6
Fabaceae††	1463	126	114	7.8	Vitaceae†††	175	19	4	2.2
Orchidaceae†††	1502	653	102	6.8	Melastomataceae†††	135	10	3	2.2
Rubiaceae††	713	48	48	6.8	Commelinaceae†	47	1	1	2.2
Liliaceae†††	760	103	50	6.5	Apocynaceae†††	114	17	2	1.8
Brassicaceae††	432	16	28	6.5	Aceraceae†††	130	47	2	1.5
Gentianaceae††	449	15	26	5.8	Caprifoliaceae†††	115	9	1	0.9
Crassulaceae†	255	57	15	5.8	Thymelaeaceae†††	123	15	1	0.8
Oleaceae†	162	15	9	5.5	Juncaceae†††	88	0	0	0
Acanthaceae††	298	0	16	5.4	Anacardiaceae†††	71	12	0	0
Saxifragaceae†††	638	9	33	5.2	Menispermaceae†††	81	23	0	0
Rutaceae†	139	17	7	5.1	Onagraceae††	53	4	0	0
Araceae†	181	44	9	5.0	Capparaceae†	42	9	0	0
Salicaceae†††	443	45	22	4.9	Cornaceae†	35	2	0	0
Myrsinaceae†	122	5	6	4.9	Bignoniaceae†	35	8	0	0
Apiaceae†††	614	33	28	4.5					

Number of species described is based on Qin et al. (2017a, 2017b). Superscripts of families denote the significance level of the selectivity of data deficiency. Significantly more deficient than expected: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Significantly less deficient than expected: ††† $p < 0.001$; †† $p < 0.01$; † $p < 0.05$. Only the families with significant difference in data deficiency are listed in this table. See Table S2 in Appendix A for full list of families.

collections and spatial accessibility) and three biodiversity measures (species richness, DD species richness, and DD ratio). In this paper, the DD ratio was defined as the proportion of DD species to total species and SCR was defined as the degree of overlap between the top 10% of the grids of two primary causes of spatial knowledge deficiency and three biodiversity measures.

2.3. Examine and quantify taxonomic and geographic selectivity of data deficiency

In order to minimize the effect of families with few species and balance the conflicting aims of obtaining a reasonable number of data sets and obtaining reasonable statistical power within each, we selected suitable families that contained > 30 species. Families with an insufficient number of species were excluded. Chi-square test was used to determine the non-randomness in data deficiency of all the families. Then, we focused on the remaining taxa. We followed Bielby et al.'s (2006) null hypothesis that threatened species are distributed randomly within and without a certain subset. We shuffled the threat labels across subsets by a random permutation without changing the number of times the threatened and non-threatened labels occur, and then counted the threatened labels in a certain subset, what was the probability that the real number would reoccur? To answer this question, we have to know the distribution function of the number of threatened labels after such a random shuffling of that particular subset. Bielby et al. (2006) preferred to estimate the distribution function empirically by repeating the random shuffling process 10,000 times. As in other Monte Carlo methods, such simulation experiments suffer from high time complexity and low accuracy. In fact, the statistic X, defined as the number of threatened labels in a subset, obeys a hypergeometric distribution, the probability density function of which can be expressed explicitly:

$$f(k;n,N,M) = P(X = k \mid n,N,M) = \frac{C_k^M C_{n-k}^{N-M}}{C_n^N},$$

where N is the total number of both threatened and non-threatened labels; M is the number of threatened labels; n is the size of the subset; C_a^b denotes the binomial coefficient, defined as b!/a!(a-b)!; and $k = 0, 1, 2, \ldots$, max that runs through all probable numbers of threatened labels in the subset. Given a significance level α, we reject the null hypothesis if the actual number of threatened labels falls into either α tail of the hypergeometric distribution. We refer to such a procedure as a *hypergeometric test* (on selectivity) hereinafter. In the DD related section, we used the same hypergeometric test to test the randomization of DD species in each subset.

3. Results

3.1. Spatial patterns of DD species

Spatial autocorrelation analysis on relative species richness in grid cells suggests an aggregated biodiversity pattern in China (Moran's $I = 0.61$, z-score = 28.0; $p < 0.001$). In general, species' richness increases from north to south along latitude, as shown in Fig. 1a. DD species in China were distributed unevenly (Moran's $I = 0.58$, z-score = 27.0; $p < 0.001$). Fig. 1b shows only a few centers of plant diversity in which there were relatively many DD species (Spearman's $R^2 = 0.879$ in Fig. 2). These concentrations of DD species that generally were disproportionately high occurred in the Hengduan Mountain Region (IF17), East Himalaya Region (IF18), and were scattered in a minority of certain regions (IA2, IE13, IE16, IE25, and G22). Fig. 1c shows that the DD ratio was distributed primarily in Northwestern China. Larger clusters of areas were located largely in Northwest China

Table 2
Geographic selectivity of data deficiency and extinction risk of spermatophytes under different treatments of DD (DD excluded, DD considered non-threatened, and DD considered threatened).

Code of floristic regions	Total species	Threatened species	DD species	DD ratio
IA1 ^{ns}	304	11	14	4.6
IA2 ^{ns}	791	20	34	4.3
IA3 ^{ns}	996	30	55	5.5
IB4a ^{ns}	373	11	19	5.1
IB4b ^{ns}	1222	38	64	5.2
IB5a ^{ns}	829	25	41	5.o
IB5b ^{ns}	1397	37	64	4.6
IC6a ^{ns}	1465	49	61	4.2
IC6b ^{ns}	684	11	25	3.7
ID7 ^{ns}	2130	74	102	4.8
ID8a***	1596	63	97	6.1
ID8b ^{ns}	815	18	45	5.5
ID8c*	620	24	40	6.4
ID9 ^{ns}	529	16	26	4.9
IE10 ^{ns}	953	40	37	3.9
IE11a†	490	22	13	2.6
IE11b††	2052	106	72	3.5
IE11c†	3155	153	122	3.9
IE12 ^{ns}	3522	252	166	4.7
IE13 ^{ns}	5776	475	290	5.o
IE14†	4053	338	160	4.0
IE15 ^{ns}	4989	479	219	4.4
IF16 ^{ns}	6864	630	333	4.8
IF17***	9144	726	569	6.3
IF18***	3088	219	175	5.7
IIG19†††	1408	103	32	2.3
IIG20†††	2663	291	71	2.7
IIG21††	3006	303	109	3.6
IIG22 ^{ns}	5566	694	250	4.5

Codes of floristic regions and species diversity patterns of China. The details are presented in Fig. 1a.
Significantly more deficient than expected: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Significantly less deficient than expected: ††† $p < 0.001$; †† $p < 0.01$; † $p < 0.05$.

(IA 1, IB4, north IB5a, north IB 5b, IF18, and the junction of ID8b, ID8c, and ID9) and some grid cells were scattered in other regions. We marked the areas with the largest DD ratio, which resulted in a selection of 105 cells (9.3% of the study area in which the DD ratio was < 3.8%). A high DD ratio had a low correlation with the DD species richness (compare Fig. 1b and c: Spearman's rank correlation $R^2 = 0.11$).

3.2. Taxonomic and geographic selectivity of data deficiency

Of the 259 families in the data set, 108 had enough species to depart significantly from the overall average extinction risk prevalence. The Chi-squared test showed that data deficiency was distributed non-randomly among families ($\chi^2_{107} = 1232.75$, $p < 0.001$). 15 of the families remaining were more deficient than expected, which included 30.0% of species and 56.3% of DD species. Among these, 8 families were statistically highly significant at $p < 0.001$ (Schisandraceae, Ericaceae, Piperaceae, Elaeagnaceae, Asteraceae, Berberidaceae, Poaceae, Rosaceae), which included 25.2% of species and 50.0% of DD species. 46 families were less deficient than expected and included 50.8% of species, 25.7% of DD species (Table 1).

Data deficiency also was distributed non-randomly among floristic regions ($\chi^2_{28} = 148.55$, $p < 0.001$). 4 floristic regions were more deficient than expected, among which 3 were statistically highly significant at $p < 0.001$ (Ü-Tsang Subregion (ID8a), Hengduan Mountain Region (IF17), East Himalaya Region (IF18)). 7 floristic regions were less deficient than expected. The Pamir-Kunlun Subregion (code ID8a) in the Pamir-Kunlun-Tibet Region (ID8) had the largest ratio (6.4%) in data deficiency (Table 2).

3.3. Conservation prioritization based on data deficiency selectivity

Fig. 3 showed the new overly-deficient areas were significantly larger after some DD species were removed (Fig. 3b, c) and many were statistically highly significant ($p < 0.001$). Spatially, these new overly-deficient areas were clustered largely in IA1, IA3, IB4a, IB4b, IB5a, IB5b, and the juncture of ID8b, ID8c, and ID9. Many were located in the hotspots of China (Fig. 3a, c) (Myers et al., 2000; Mittermeier et al., 2005). In addition, differences also existed before and after analysis of less deficient areas (Fig. 3c). Most of the new under-deficient areas were distributed around protected areas in the ID7 region except for a small number of grids that are scattered in E12, E13, and other areas (Fig. 3c).

3.4. Primary causes of spatial knowledge deficiency

Spatial aggregations were found in both frequency of collections and spatial accessibility. However, the two causes were in low correlation (Spearman's $R^2 = 0.035$). They show significantly different in patterns. Fig. 2 shows the matrix of Spearman's rank correlations between the two primary causes of spatial knowledge deficiency and the three biodiversity measures. Fig. 4a shows that the frequency of collections was distributed unevenly in China (Moran's $I = 0.47$, z-score = 48.70; $p < 0.001$), and was concentrated primarily in Southwest China (IE13, IE15, IF16, IF17, IF18, IIG22) that are rich in plant diversity (R^2 between Figs. 4a and 1a $= 0.88$, $p < 0.001$) and DD species (R^2 between Figs. 4a and 1b $= 0.83$, $p < 0.001$). However, the frequency of collections was correlated very weakly with the DD ratio ($R^2 = 0.23$, $p < 0.001$), and the spatial coverage ratio was low

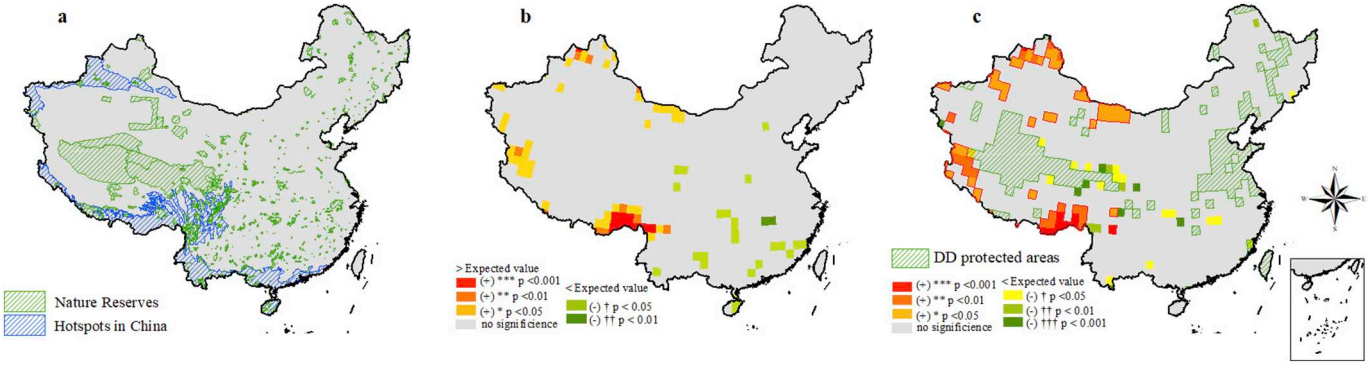

Fig. 3. Selectivity of data deficiency and conservation prioritization. **a** shows the nature reserves from WCPA (https://www.iucn.org/theme/protected-areas/wcpa) and hotspots from CI (Myers et al., 2000; Mittermeier et al., 2005); **b** shows the spatial pattern of data deficiency selectivity when all DD species are considered and **c** shows the spatial patterns of data deficiency after omitting the areas in which DD were under protected. More deficient than expected: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; less deficient than expected: ††† $p < 0.001$; †† $p < 0.01$; † $p < 0.05$. (ns)–non-significant.

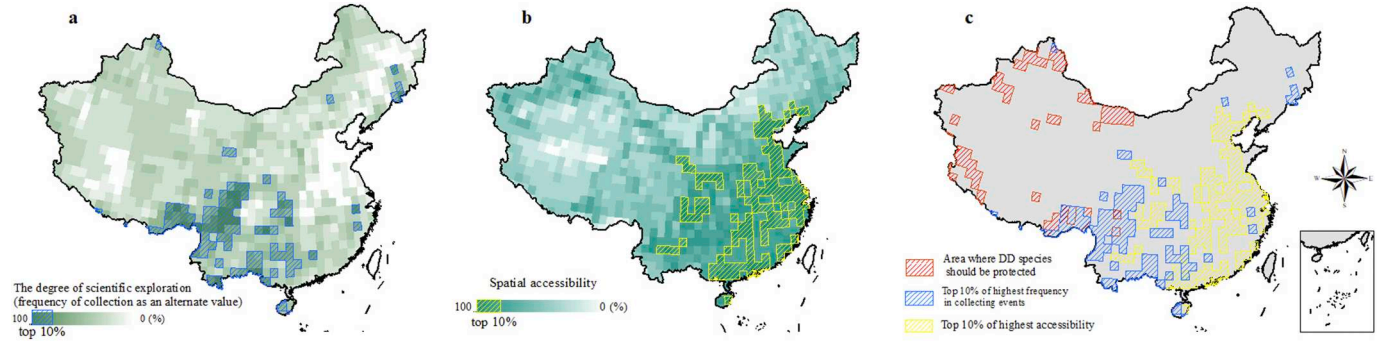

Fig. 4. The overlying effects of the two primary causes of spatial knowledge deficiency (a. frequency and collections; b. accessibility), and regions in which DD species should be protected (c).

(SCR = 3.7%, Fig. 4). Spatial accessibility also was distributed non-randomly (Moran's $I = 0.59$, z-score = 60.1, $p < 0.001$, Fig. 4b) and concentrated in central China and south China (IE11b, IE12, IE14, IIG20, IIG21, and some other fragmented areas). The areas with best accessibility were low in DD species diversity (Spearman's $R^2 = -0.09$, $p = 0.005$; SCR = ~0%) and DD ratio (Spearman's $R^2 = -0.37$, $p < 0.001$; SCR = 0%). In addition, Fig. 4 showed the new overly-deficient areas were correlated marginally both with the importance (SCR = 16.0%) and difficulty of scientific research (SCR = 0%).

4. Discussion

As we know, conservation managers tend to prioritize those species that are classified as threatened (Bland et al., 2015). Determining the conservation status of DD species is essential to achieve an accurate understanding of global biodiversity that provides adequate protection for threatened species. The absence of a random distribution of data deficiency suggests that there are biological or geographic drivers of deficiency that can help focus conservation activity indirectly (Cardillo and Meijaard, 2012). DD species, which constitute such a large proportion of spermatophytes in China, are associated with different degrees of uncertainty in the estimates of the threat level both for taxonomic groups known poorly and even those known well.

Our results show national patterns of geographical selectivity were inconsistent with one another in given different treatments of DD species. That was because the number of grids and the percentage of DD species changed in the process of the randomization test, as did the patterns. Moreover, the differentiations of endangerment also may be exaggerated artificially by the subjectivity of the assessors. Based on these uncertainties, we have reason to believe that there might be significant differences in the selectivity of extinction risk when DD species are included or omitted. Given the significant impact of DD species on the understanding of different patterns, DD species should be given high research priority to determine their true status. In addition, when setting priority protection, DD species, especially DDP species, should receive special attention rather than neglect. DDP species are hotspot of scientific and conservation researches, as they represent an identifiable gap in knowledge (i.e., a species is either DD or not).

Similarly, the uncertainties attributable to DD species also might alter the patterns of selectivity of extinction risk. These differentiations in patterns indicated that taxonomists and conservationists have a degree of knowledge deficiency about China's plant species. Although we cannot avoid the uncertainties raised by DD, what we can do is address the knowledge gap and develop an accurate picture of biodiversity of China. This uncertainty affects not only the monitoring of progress to achieve global biodiversity targets (i.e., CBD, 2012), but also conservation priorities that rely on threatened species lists, such as Key Biodiversity Areas (IUCN, 2016), biodiversity hotspots (Myers et al., 2000; Mittermeier et al., 2005), Areas of Zero Extinction (Ricketts et al., 2005), and many others (Brooks et al., 2006). Therefore, the first thing necessary to address the knowledge gap is to identify the primary causes of spatial knowledge deficiency.

Based on this research, we found a high congruence between the spatial distribution of DD species and species richness, which is consistent with the results of previous studies (e.g., Zhao et al., 2016). Possible causes may be that the richness, nature, and community aspects of mountainous Southwest China, and their geographic distribution patterns are controlled largely by the complex and diversified eco-environment (Huang et al., 2011; Zhao et al., 2016). These complex and highly heterogeneous regions can create certain microclimates that are suited for narrow niche species, particularly endemic species. Another possible explanation of the spatial clustering of the two causes may be that researchers always focus on certain taxonomic groups or research areas (Kier et al., 2005; Küper et al., 2006; Yang et al., 2014; Sofef et al., 2017). Moreover, they may be prone to invest more effort in areas with rich biodiversity where they can get what they need and observe what they expect more easily and achieve more progress in a short time. A full understanding of the causes of these data deficiencies can determine where and what kind of actions are required to improve the status of knowledge on plant diversity. However, when we constructed the map of the spatial pattern of selectivity in data deficiency and then plotted the layer of frequency of collections and spatial accessibility, we found that the clustering centers of DD with ratios significantly larger than expected are located in regions where there was less collections and more difficult to reach (Figs. 2 and 4). This is also a challenge for biodiversity conservation that generally has been given little attention. The more deficient than expected areas (Fig. 4c), which are less accessible spatially, always have worse infrastructures that increase the difficulty of scientific research and prevent further surveys from being easy. Thus, for these areas, strategic field collections are required urgently. In addition, because of the low spatial accessibility of these areas, more research costs are required to conduct research activities. We suggest that these regions should receive more financial support.

Therefore, a clearer understanding of the patterns of taxonomic selectivity and geographical selectivity in DD is essential to developing a more representative picture of biodiversity. These results help to allocate restricted conservation resources. Based on the above analysis, we will have new thoughts and ideas when we set priority areas on conservation or set priority objects on research. Selectivity, whether that of extinction risk or of data deficiency, inevitably must be considered when establishing conservation and research priorities, while comprehensive and definite conservation strategies also should be implemented in some sensitive areas that are related to DD, such as the DD-driven significantly more threatened than expected areas or particular significantly more deficient than expected areas that always have fragile habitats or instable populations. Moreover, these areas should be given the highest priority in conservation. Whereas the protection measures for others with taxonomical uncertainty (named DDT) are vigorously implemented before they can be distinguished into threatened and non-threatened categories, it is likely to waste resources.

DDT species should be emphasized in taxonomic and molecular identification studies and funding should be increased for taxonomists' work (Callmander et al., 2005).

Acknowledgements

This work was jointly supported by the National Key R&D Program of China (Grant No. 2016YFC0500406 and 2017YFC0505804), the National Natural Science Foundation of China (NSFC, 41601439), and the Plant Specimen Sub-Platform, National Specimen Information Infrastructure (2005DKA21401).

We are grateful to Dr. Haining Qin, Bojian Bao, Yi Li, Guoxia Han, Naxin Xue for data collecting and preparation. We also thank all other people having commented on earlier drafts of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biocon.2018.12.009.

References

Barnosky, A.D., et al., 2011. Has the Earth's sixth mass extinction already arrived? Nature 471, 51–57.
Bennun, L., et al., 2017. The value of the IUCN Red List for business decision-making. Conserv. Lett. 2017 (00), 1–8.
Bielby, J., Cunningham, A.A., Purvis, A., 2006. Taxonomic selectivity in amphibians: ignorance, geography or biology? Anim. Conserv. 9, 135–143.
Bland, L.M., Collen, B., Orme, C.D.L., Bielby, J., 2012. Data uncertainty and the selectivity of extinction risk in freshwater invertebrates. Divers. Distrib. 18, 1211–1220.
Bland, L.M., Collen, B., Orme, C.D.L., Bielby, J., 2015. Predicting the conservation status of data-deficient species. Conserv. Biol. 29, 250–259.
Bland, L.M., Bielby, J., Kearney, S., Orme, C.D.L., Watson, J.E.M., Collen, B., 2017. Toward reassessing data-deficient species. Conserv. Biol. 31, 531–539.
Böhm, M., et al., 2013. The conservation status of the world's reptiles. Biol. Conserv. 157, 372–385.
Brooks, T.M., Mittermeier, R.A., da Fonseca, G.A., Gerlach, J., Hoffmann, M., Lamoreux, J.F., Mittermeier, C.G., Pilgrim, J.D., Rodrigues, A.S., 2006. Global biodiversity conservation priorities. Science 313, 58–61.
Butchart, S.H.M., Bird, J.P., 2010. Data deficient birds on the IUCN Red List: what don't we know and why does it matter? Biol. Conserv. 143, 239–247.
Callmander, M.W., Schatz, G.E., Lowry, P.P., 2005. IUCN Red List assessment and the global strategy for plant conservation: taxonomists must act now. Taxon 54, 1047–1050.
Cardillo, M., Meijaard, E., 2012. Are comparative studies of extinction risk useful for conservation? Trends Ecol. Evol. 27, 167–171.
CBD, 2012. Global Strategy for Plant Conservation: 2011–2020. Botanic Gardens Conservation International, Richmond, UK.
Chen, Y., 2011. Mathematical Methods for Geography: Foundations and Applications. Science press, Beijing.
Cowling, R.M., Knight, A.T., Faith, D.P., Ferrier, S., Lombard, A.T., Driver, A., Rouget, M., Maze, K., Desmet, P.G., 2004. Nature conservation requires more than a passion for species. Conserv. Biol. 18, 1674–1676.
Fritz, S.A., Purvis, A., 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. Conserv. Biol. 24, 1042–1051.
Gärdenfors, U., Hilton-Taylor, C., Mace, G.M., Rodríguez, J.P., 2001. The application of IUCN Red List criteria at regional levels. Conserv. Biol. 15, 1206–1212.
Good, T.C., Zjhra, M.L., Kremen, C., 2006. Addressing Data Deficiency in Classifying Extinction Risk: A Case Study of a Radiation of Bignoniaceae From Madagascar.
Hayward, M.W., 2009. The need to rationalize and prioritize threatening processes used to determine threat status in the IUCN Red List. Conserv. Biol. 23, 1568–1576.
Hoffmann, M., et al., 2010. The impact of conservation on the status of the world's vertebrates. Science 330, 1503–1509.
Howard, S.D., Bickford, D.P., Ferrier, S., 2014. Amphibians over the edge: silent extinction risk of data deficient species. Divers. Distrib. 20, 837–846.
Huang, J., Chen, J., Ying, J., Ma, K., 2011. Features and distribution patterns of Chinese endemic seed plant species. J. Syst. Evol. 49, 81–94.
IUCN, 2001. IUCN Red List Categories and Criteria: Version 3.1. IUCN Species Survival Commission, Gland, Switzerland.
IUCN, 2011. Guidelines for using the IUCN Red List categories and criteria. Version 9.0. In: Prepared by the Standards and Petitions Working Group of the IUCN SSC Biodiversity Assessments Sub-Committee in August 2008.
IUCN, 2016. A Global Standard for the Identification of Key Biodiversity Areas, Version 1.0, First edition. IUCN, Gland,Switzerland.
Jarić, I., Courchamp, F., Gessner, J., Roberts, D.L., 2016. Potentially threatened: a data deficient flag for conservation management. Biodivers. Conserv. 25, 1995–2000.
Jetz, W., Freckleton, R.P., 2015. Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. Philos. Trans. R. Soc. B 370.
Joaquim, T.-F., de Carvalho, R.A., Brito, D., Loyola, R.D., 2012. How does the inclusion of data deficient species change conservation priorities for amphibians in the Atlantic Forest? Biodivers. Conserv. 21, 2709–2718.
Juslen, A., Hyvarinen, E., Virtanen, L.K., 2013. Application of the red-list index at a national level for multiple species groups. Conserv. Biol. 27, 398–406.
Kier, G., Mutke, J., Dinerstein, E., Ricketts, T.H., Küper, W., Kreft, H., Barthlott, W., 2005. Global patterns of plant diversity and floristic knowledge. J. Biogeogr. 32, 1107–1116.
Knight, A.T., et al., 2007. Improving the key biodiversity areas approach for effective conservation planning. Bioscience 57, 256–261.
Knight, A.T., Cowling, R.M., Rouget, M., Balmford, A., Lombard, A.T., Campbell, B.M., 2008. Knowing but not doing: selecting priority conservation areas and the research-implementation gap. Conserv. Biol. 22, 610–617.
Küper, W., Sommer, J.H., Lovett, J.C., Barthlott, W., 2006. Deficiency in African plant distribution data – missing pieces of the puzzle. Bot. J. Linn. Soc. 150, 355–368.
Langhammer, P., et al., 2017. Identification and Gap Analysis of Key Biodiversity Areas: Targets for Comprehensive Protected Area Systems. IUCN, Gland, Switzerland.
Li, H., Calder, C.A., Cressie, N., 2007. Beyond Moran's *I*: testing for spatial dependence based on the spatial autoregressive model. Geogr. Anal. 39, 357–375.
Mace, G.M., Collar, N.J., Gaston, K.J., Hilton-Taylor, C., AkÇAkaya, H.R., Leader-Williams, N., Milner-Gulland, E.J., Stuart, S.N., 2008. Quantification of extinction risk: IUCN's system for classifying threatened species. Conserv. Biol. 22, 1424–1442.
Mittermeier, R.A., Myers, N., Thomsen, J.B., da Fonseca, G.A.B., Olivieri, S., 1998. Biodiversity hotspots and major tropical wilderness areas: approaches to setting conservation priorities. Conserv. Biol. 12, 516–520.
Mittermeier, R.A., Gil, P.R., Hoffman, M., Pilgrim, J., Brooks, T., Mittermeier, C.G., Lamoreux, J., Fonseca, G.A.B.D., 2005. Hotspots Revisited. The University of Chicago Press, Chicago.
Morais, A.R., Siqueira, M.N., Lemes, P., Maciel, N.M., De Marco, P., Brito, D., 2013. Unraveling the conservation status of data deficient species. Biol. Conserv. 166, 98–102.
Myers, N., Mittermeier, R.A., Mittermeier, C.G., Fonseca, G.A.B.D., Kent, J., 2000. Biodiversity Hotspots for Conservation Priorities.
Qin, H., et al., 2017a. Threatened species list of China's higher plants. Biodivers. Sci. 25, 696–744.
Qin, H., et al., 2017b. Evaluating the endangerment status of China's angiosperms through the red list assessment. Biodivers. Sci. 25, 745–757.
Ricketts, T.H., et al., 2005. Pinpointing and preventing imminent extinctions. Proc. Natl. Acad. Sci. U. S. A. 102, 18497–18501.
Roberts, D.L., Taylor, L., Joppa, L.N., Beggs, J., 2016. Threatened or data deficient: assessing the conservation status of poorly known species. Divers. Distrib. 22, 558–565.
Rodrigues, A.S.L., Pilgrim, J.D., Lamoreux, J.F., Hoffmann, M., Brooks, T.M., 2006. The value of the IUCN Red List for conservation. Trends Ecol. Evol. 21, 71–76.
Saiz, J.C.M., Lozano, F.D., Gómez, M.M., Baudet, Á.B., 2015. Application of the Red List Index for conservation assessment of Spanish vascular plants. Conserv. Biol. 29, 910–919.
Sofef, et al., 2017. Exploring the floristic diversity of tropical Africa. BMC Biol. 15, 15.
Sousa-Baena, M.S., Garcia, L.C., Townsend Peterson, A., 2014. Knowledge behind conservation status decisions: data basis for “data deficient” Brazilian plant species. Biol. Conserv. 173, 80–89.
Vaira, M., Pereyra, L.C., Akmentins, M.S., Bielby, J., 2017. Conservation status of amphibians of Argentina: an update and evaluation of national assessments. Amphib. Reptile Conserv. 11, 36–44.
Wilson, K.A., et al., 2007. Conserving biodiversity efficiently: what to do, where, and when. PLoS Biol. 5, e223.
Wu, C., 1979. The regionalization of Chinese flora. Acta Bot. Yunnanica 1, 1–24.
Yang, W.J., et al., 2014. Environmental and socio-economic factors shaping the geography of floristic collections in China. Glob. Ecol. Biogeogr. 23, 1284–1292.
Zhao, L., Li, J., Liu, H., Qin, H., 2016. Distribution, congruence, and hotspots of higher plants in China. Sci. Rep. 6, 19080.