Environmental and geographical biases in plant specimen data from the Colombian Andes

Vargas, Carlos A; Bottin, Marius; Särkinen, Tiina; Richardson, James E; Raz, Lauren; Garzon-Lopez, Carol X; Sanchez, Adriana

Published in:
Botanical Journal of the Linnean Society

DOI:
10.1093/botlinnean/boac035

Botanical Journal of the Linnean Society, 2022, **200**, 451–464. With 5 figures.

Environmental and geographical biases in plant specimen data from the Colombian Andes

CARLOS A. VARGAS[1,2,*], MARIUS BOTTIN[3], TIINA SÄRKINEN[4], JAMES E. RICHARDSON[1,4,5], LAUREN RAZ[6], CAROL X. GARZON-LOPEZ[7] and ADRIANA SANCHEZ[1]

[1]*Departamento de Biología, Facultad de Ciencias Naturales, Universidad del Rosario, Bogotá, D.C., Colombia*
[2]*Subdirección Científica, Jardín Botánico de Bogotá 'José Celestino Mutis', Bogotá, D.C., Colombia*
[3]*Independent Researcher, Bogotá, D.C., Colombia*
[4]*Tropical Diversity Section, Royal Botanic Garden, Edinburgh, UK*
[5]*School of Biological, Earth and Environmental Sciences, University College Cork, Cork, Ireland*
[6]*Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Bogotá, D.C., Colombia*
[7]*Knowledge Infrastructures, Campus Fryslân, University of Groningen, Leeuwarden, The Netherlands*

Received 6 January 2022; revised 27 April 2022; accepted for publication 22 July 2022

Specimen records are a major source of species information for biodiversity research. However, specimen records currently available may be geographically or environmentally biased. Detailed knowledge of biases is useful for understanding and accounting for errors they introduce into analyses of biodiversity patterns. Here we study geographical and environmental biases in online records representing the flora of the Colombian Andes and explore their effect on sample completeness at different spatial scales. We found a strong geographical and environmental sampling bias. Plant records were concentrated close to cities where herbaria and researchers are located. The highlands > 2000 m are better sampled, whereas mid- and lowlands remain poorly sampled (i.e. montane and lowland forest). Sampling completeness (SC) median across the Colombian Andes is < 75% at the scales studied. We explore possible causes of sampling bias, identify critical gaps and priority areas for plant sampling and make recommendations for strategies to increase SC and reduce biases.

ADDITIONAL KEYWORDS: collecting bias – Colombia – flora – herbarium specimens – northern Andes – sampling completeness

INTRODUCTION

Primary specimen records are a major source of information on species occurrence in space and time. Many of the specimens have been deposited in museums and biological collections through the work of scientists and explorers through time, back to the 14th century (Thiers, 2020). Today, these biological data have become available through online data aggregators, such as the Global Biodiversity Information Facility (GBIF) that includes ≥ 333 million plant occurrence records (www.gbif.org, accessed 22th February 2021) and herbaria that have digitized their collections. These digitally available specimen records may be used to study biodiversity patterns and to inform management and conservation policy decisions.

Despite the increasing amount of digitally available specimen data, gaps and biases have been detected in datasets, particularly in temporal, geographical and taxonomical dimensions (Meyer *et al.*, 2016). Geographical bias includes uneven distribution of records concentrated along roads and near cities where scientific infrastructure is available (Sousa-Baena, Couto & Townsend, 2013; Oliveira *et al.*, 2016). Environmental bias could include parts of climatic gradients being poorly represented by collections (Loiselle *et al.*, 2008). Another possible source of bias is under-collection of small and/or unattractive plants (Schmidt-Lebuhn, Knerr & Kessler, 2013). These gaps

*Corresponding author. E-mail: carlosalbe.vargas@urosario.edu.co

and biases have implications for our understanding of species richness patterns (Rowe, 2005), identification of conservation priority areas (Reddy & Dávalos, 2003) and the accuracy of species distribution models (Feeley & Silman, 2011).

Species richness is a primary biodiversity metric that indicates how many species are found in a particular locality. It is an essential ecological concept commonly used as a criterion for conservation and management purposes. Determination of total richness requires a complete census of species in a study area, which is often impossible due to financial and logistical restrictions. Therefore, different approaches have been developed to estimate species richness from incomplete sampling (e.g. Chao 1, Chao 2, bootstrapping, rarefaction; Hortal *et al.*, 2006; González-Oreja *et al.*, 2010; Gotelli & Chao, 2013; Engemann *et al.*, 2015). These estimators have been helpful in the study of richness patterns using data available in public repositories (e.g. GBIF). However, the richness estimators based on this kind of data are influenced by non-random sampling, different sampling efforts and data quality. Heterogeneous availability of data is a problem in highly diverse regions such as tropical mountains where the biodiversity is influenced by factors such as orographic, geological and edaphic heterogeneity that are a result of geological history, habitat fragmentation and a great variety of climatic characteristics (Richter, 2008).

The tropical Andes is one of the global hotspots due to the high levels of endemism and threats to biodiversity (Myers *et al.*, 2000). The topographic and climatic complexity of the Andes has created a mosaic of different ecosystems and complex species arrangements (Humboldt & Bonpland, 1807; Pennington *et al.*, 2010; Särkinen *et al.*, 2012). Despite the high species diversity of the tropical Andes, the distribution and completeness of digitally available specimen records for the flora have been little explored. Low scale analysis (grid cells size 100 × 100 km) indicated areas in the northern Andes with low record density, particularly in Colombia and Venezuela (Distler *et al.*, 2009; Jiménez, Distler & Jørgensen, 2009; Mutke, 2017), in contrast to the Ecuadorian Andes where the sampling is much better, although still poor (Engemann *et al.*, 2015). Poor sampling in Colombia can be partly explained by the geopolitical issues that the country has faced in the last 60 years. Internal conflict made fieldwork extremely risky during this period; when biological collections were generally on the increase, collecting in Colombia remained at comparatively low levels (Moura & Jetz, 2021). The recent signing of a peace agreement has allowed the return of field scientists to previously inaccessible areas (because of the conflict), leading to the discovery of hundreds of new species (Botero, 2020), and it is hoped that collection efforts will increase in the coming years so that the gaps that we highlight here may be addressed.

This study aims to determine spatial and environmental biases and gaps in the digitally available specimen records of plants in the Colombian Andes and evaluate the potential impact on predicting richness accuracy (e.g. reliability). Furthermore, such studies are needed to account for any potential biases in species diversity models and conservation planning and for formulating a strategic plan to fill the collection gaps. Therefore, we address the following questions to understand collection patterns in the region and their impact on richness estimates based on digitally available plant records. (1) Do the distribution of plant records represent the environmental and spatial variability of the Colombian Andes? (2) Are the plant records spatially structured in Colombian Andes? (3) Which areas require increased collection efforts?

MATERIAL AND METHODS

Study area

The study area comprises the Colombian Andes, as defined by Rodríguez *et al.* (2006), consisting of three mountain ranges (Western, Central and Eastern Cordilleras) and the valleys of the Cauca and Magdalena Rivers, with a lower limit of the area set at 445 m a.s.l. Additionally, we included the Sierra Nevada de Santa Marta, an isolated mountain range located in the north of Colombia where the study area reaches its highest elevation of 5659 m. Thus, the study area (Fig. 1) comprises 306 729 km^2, characterized by high climatic variability. Temperatures vary from 24 to 32 °C in the lowlands to –2 °C in the highlands. Furthermore, the topographical variability and local wind regimes determine areas of high humidity (85% of the area) and dry zones (15% of the study area). Combining these environmental variables results in a mosaic of almost 162 ecosystems in four biomes (Rodríguez *et al.*, 2006).

Specimen database

All records for plant species occurring in Colombia were downloaded from online and herbarium databases: (1) GBIF (accessed on the 30 May 2017, request available at http://doi.org/10.15468/dl.xqndaq, gbif.org, 2017); (2) Missouri Botanical Garden (MOBOT, accessed April 2016); (3) The Colombian National Herbarium (COL, request for Andean plant records throughout Colombia, accessed March 2017) and (4) Jardín Botánico de Bogotá (JBB, August 2017). The database constructed included 2 266 136 specimens (Supporting Information, Table S1). The plant records were imported into a SQL database, in which the data

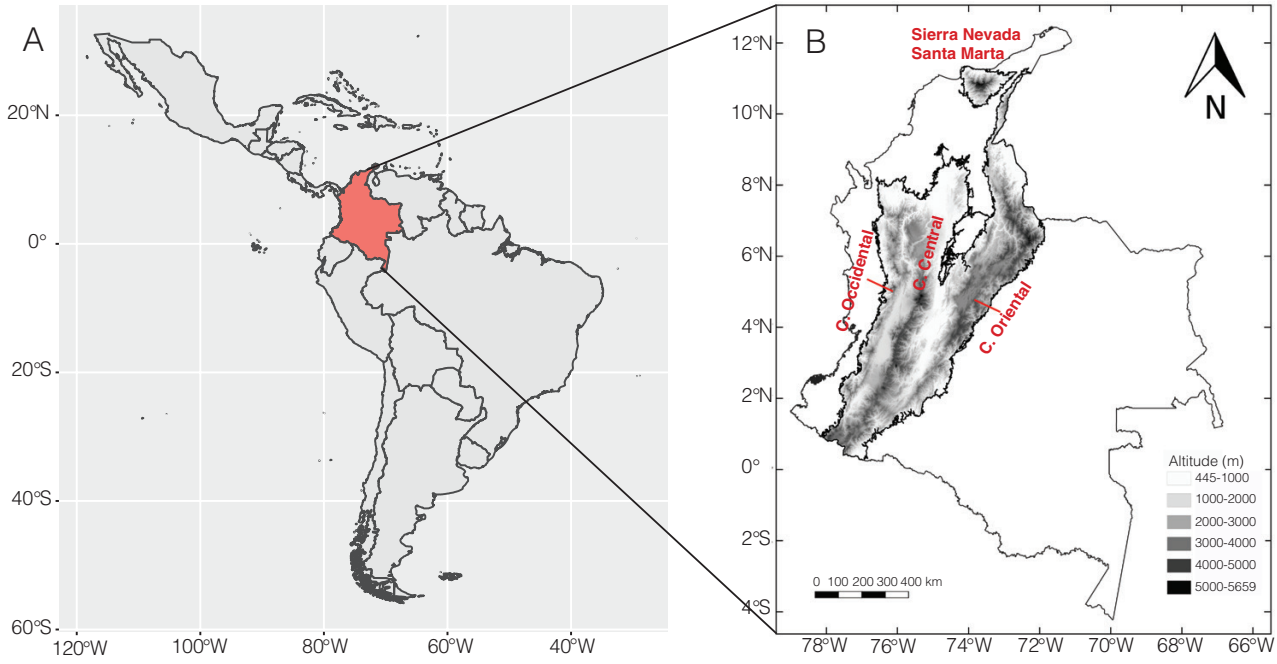

Figure 1. Map of digitally available plant specimen record data for the Colombian Andes. A, Colombia (red) in South America. B, Colombian Andean region (grey) corresponds to the montane area of the country, including the elevations between 445 and 5659 m. 'C'. is the abbreviation for Cordillera.

were indexed to reduce the computation time of the subsequent queries.

The data were cleaned to increase the accuracy of taxonomic and geographical information. Taxonomic cleaning consisted of revising species names, checking for spelling errors and synonyms and validating species names according to the *Catálogo de las Plantas de Colombia* (Bernal, Grandstein & Celis, 2016). A few unresolved names were further checked using the Taxonomic Name Resolution Service. Coordinates were also thoroughly checked. Our standardization procedure consisted of an extensive regular expression treatment of the diverse raw formats of coordinates, to use as much information as possible. We also discarded records with no clear mention of coordinate uncertainty, and we corrected obvious errors, such as inversion of latitude and longitude. After cleaning the database, we retained records with coordinates that fall in the Colombian Andean region with a precision < 0.0016 decimal degrees in a WGS84 projection (equivalent to a distance of 100 m at the Equator). To standardize the elevation data for each record, we used the coordinates and extracted this information using the Shuttle Radar Topography Mission Global Elevation Model at a 90-m resolution (http://srtm.csi.cgiar.org), using the plant record coordinates. Finally, we checked for duplicates leaving a single record for each collection in our analyses. Given that there were different formats for number and collector name in the database, we defined duplicates as those records matching collection dates, collection numbers, species names and coordinates within 100 m of each other (Feeley, 2015). Among the groups of duplicated records, we selected the samples that had a consistent identification before and after checking for synonyms and spelling. In the case of duplicated plant records with the different valid names, priority was given to local databases (e.g. JBB followed by COL).

The final database consisted of 266 625 georeferenced plant records representing 19 638 species. Therefore, despite the initial number of plant records gathered for the study, only 11.8% of them were used in our analyses. In total, 88.2% of the records were discarded due to geographical issues such as lack of coordinates, low-precision coordinates or plant records outside the study area (66.5%), followed by duplicates (13.6%) and 8.1% because of incorrect species names (Supporting Information, Table S2).

Environmental data

To study the environmental representativeness of plant records in the Colombian Andean region, we used mean annual temperature and annual precipitation from the CHELSA database at 1-km (30 arcsec) resolution (Karger *et al.*, 2017) and elevation data from the Shuttle Radar Topography Mission Global Elevation Model at 90-m resolution (http://srtm.csi.cgiar.org).

Data analysis

The data analysis consisted of three steps: (1) description of the spatial collection pattern; (2) quantification of bias based on spatial and environmental variables and (3) description of sampling completeness (SC) of plant richness. To describe plant collection patterns, the spatial coverage, defined as the number of grid cells with plant records, and plant collection density were calculated using five scales: 100 × 100, 50 × 50, 20 × 20, 10 × 10 and 5 × 5 km (Table 1). The spatial coverage was measured at the five scales, as the proportion of grid cells with specimen records over the total possible number of grid cells (Table 1). The record density was measured based on the number of specimen records per grid cell. Density and coverage maps were created in QGIS (QGIS Development Team, 2015).

The spatial pattern of plant records was analysed by calculating the Moran index of spatial autocorrelation to determine whether the plant collection patterns were aggregated, random or dispersed. We calculated Moran index using the 'spdep' package (Bivand, Pebesma & Gómez-Rubio, 2013) in R v.3.6.1 (R Development Core Team, 2019), and *P* value was estimated through Monte Carlo simulation (Bivand & Wong, 2018).

The environmental bias of the plant records was estimated performing six intervals for elevation, mean annual temperature and annual precipitation (Table 2). The magnitude of the records bias was calculated using the Kadmon Index (Kadmon, Farber & Danin, 2004) that was originally designed to assess roadside bias, but may be equally applied to other forms of geographical bias:

$$\text{Bias}\,(d) = \text{nd} - \frac{\text{pd}n}{\sqrt{\text{pd}\,(1-\text{pd})\,n}}.$$

where nd is the number of collection localities within a specified interval (d), n is the total number of collection localities in the database and pd is the probability that a given collection locality is within an interval (d). Since this equation is derived from the normal approximation of a binomial distribution, values are statistically significant when they are ≥ 1.64 and < −1.64, (at $\alpha = 0.05$). Areas with values ≥ 1.64 are interpreted as oversampled (i.e. more sampled localities than expected from a random sampling design), and areas < −1.64 as undersampled (i.e. fewer sampled localities than expected from a random sampling design). To approximate pd for each interval (i.e. to account for differences in spatial coverage of environmental conditions), the same number of points as collection localities were generated on the basis of a spatial random sampling design. The fraction of random points within each interval was taken to be pd. The generation of random points and the bias index estimation was repeated 100 times (Kadmon *et al.*, 2004; García Márquez *et al.*, 2012).

Environmental representativeness of plant records was calculated at the five different spatial scales, to study the congruence between environmental variability of plant records and environmental variability of grid cells. Median values of the environmental variables (elevation, mean annual temperature and annual precipitation) were calculated per grid cell and for specimen records on each grid cell. Next, we calculated the difference of environmental variable median values per grid cell and specimen records. Plant records were considered representative of the grid cells when the differences between the median environmental values of grid cells and the environmental values given by plant records were close to zero.

Last, SC by rarefaction was calculated for each grid cell at the five different scales using a threshold of 20 plant records as the minimum sample size (Gotelli & Colwell, 2011).This analysis uses sample coverage as a proxy (based on Chao & Jost, 2012), where coverage is defined as the total relative abundance of the observed species in the sample, ranging from 0 to 1. SC by rarefaction have showed the best performance as a richness estimator for big data

Table 1. Information of the number of occurrence records by grid cell and spatial coverage of localities at different scales in the Colombian Andean region. The total number of grid cells (# cells) per scale on the Colombian Andean region, plant records median and mean are given by grid cell. Scale refers to cell size: 100 × 100 km, 50 to 50 × 50, 20 to 20 × 20 km, 10 to 10 × 10 km and 5 to 5 × 5 km

Scale (km)	Number of cells Andean region	Number of cells with plant records	Number of cells with > 20 plant records	Median plant records/grid cell	Mean plant records/grid cell
100	52	47 (90%)	46 (88%)	2497	5672
50	154	140 (91%)	131 (85%)	704	1904
20	804	694 (86%)	542 (67%)	122	382
10	2916	2125 (73%)	1223 (42%)	30	124
5	11 047	5606 (51%)	2089 (19%)	10	47

Table 2. Evaluation of the occurrence record bias for latitude, elevation, temperature and precipitation ranges in the Colombian Andes. Intervals for latitude are in degrees, elevation in metres (m), temperature in degrees Celsius (°C), precipitation in mm/year. Temperature corresponds to mean annual temperature and precipitation to annual precipitation. Observed corresponds to the total sample records on the Colombian Andes at every environmental interval; random corresponds to the points randomly generated in Colombian Andes at every environmental interval. In bold, the only well sampled interval

Variable	Interval	Random	Percentage random	Observed	Percentage observed	Bias	Sampling
Latitude	0-2	26 445.35	9.92	25 139	9.43	−8.5	undersampled
Latitude	2–4	59 628.18	22.36	27 259	10.22	−150.39	undersampled
Latitude	4–6	70 755.57	26.54	119 303	44.75	213	oversampled
Latitude	6–8	76 913.23	28.85	80 007	30.01	2.34	oversampled
Latitude	8–10	19 999.25	7.50	3038	1.14	−124.75	undersampled
Latitude	10–12	12 883.42	4.83	5879	2.20	−63.18	undersampled
Elevation	0–1000	91 753.1	34.41	54 426	20.41	−152	undersampled
Elevation	1000–2000	85 305.56	31.99	73 337	27.51	−49.6	undersampled
Elevation	2000–3000	61 565.01	23.09	75 187	28.20	62.4	oversampled
Elevation	3000–4000	26 274.5	9.85	59 071	22.16	213.1	oversampled
Elevation	4000–5000	1690.61	0.63	4331	1.62	63.42	oversampled
Elevation	5000–5659	36.22	0.01	247	0.09	35	oversampled
Temperature	–5–0	109.25	0.04	212	0.08	9.8	oversampled
Temperature	0–5	1376.16	0.52	3619	1.36	60.5	oversampled
Temperature	5–10	19 314.94	7.24	41 420	15.53	164.9	oversampled
Temperature	10–15	53 449.83	20.05	77 481	29.06	116.1	oversampled
Temperature	15–20	74 422.5	27.91	71 192	26.70	−13.9	undersampled
Temperature	20–27.1	117 766.8	44.17	72 701	27.27	−175.6	undersampled
Precipitation	688–1177	15 612.79	5.86	34 306	12.87	154.28	oversampled
Precipitation	1177–1666	58 986.19	22.12	63 504	23.82	21.02	oversampled
Precipitation	1666–2155	70 060.11	26.28	62 721	23.52	−32.37	undersampled
Precipitation	**2155–2646**	**55 221.77**	**20.71**	**55 200**	**20.70**	**0.012**	**well sampled**
Precipitation	2646–6963	65 255.58	24.47	50 717	19.02	−65.49	undersampled
Precipitation	6963–11281	1352.22	0.51	177	0.07	−32.03	undersampled

and is less susceptible to sample effort (Engemann *et al.*, 2015), however, it requires sufficient and random sampling (Gotelli & Colwell, 2011). Sample completeness was estimated using iNEXT R package (Hsieh *et al.*, 2016).

RESULTS

Geographical bias

The distribution of plant records in the Colombian Andes was highly uneven with a strong spatial autocorrelation (I = 0.118; P = 0.001), which showed an aggregated distribution. The highest collection densities are located in two hotspots between 4 and 8 °N (Fig. 1B), around the largest cities, Bogotá (Cundinamarca Department) and Medellín (Antioquia Department) (Fig. 2). In contrast, three zones had the lowest record density: the first one was located between 2 and 4 °N, which is an east to west direction, corresponding with the foothills of Caquetá and Meta, on the border with the Amazonian and Orinoquia regions and the mountains of Huila, Tolima, Cauca and Valle Departments. The second zone was located to the north of the Central Cordillera (Córdoba, Sucre and Bolívar Departments), and the third to the north of the Eastern Cordillera, including Serranía del Perijá and the Sierra Nevada de Santa Marta (Norte de Santander, Cesar, Guajira and Magdalena Departments) (Fig. 3).

Environmental bias

Significant bias and gaps were found on environmental, topographical and spatial variables. Spatially, localities from latitudes 0° to 4° N and 8° to 12°N had fewer records than expected at random, with the highest undersampling between 2 to 4°N followed by 8 to 10°N. In contrast, 4 to 8°N had more records than expected at random with the highest bias from 4 to 6°N latitude (Table 2). For elevation, collection efforts were concentrated above 2000 m, with the highest

Table 3. Descriptive values of the difference between the environmental median values obtained for specimen records and the median values of the grid cells at different scales. The environmental variables analysed include elevation, annual precipitation and mean annual temperature. Minimum values (min), quartile 25 (1Q), quartile 75 (3Q) and maximum values (max) are given. Scale of 100 refers to 100 × 100 km, 50 to 50 × 50, 20 to 20 × 20 km, 10 to 10 × 10 km and 5 to 5 × 5 km

Scale (km)	Variable	Minimum	1Q	Median	Mean	3Q	Maximum
100	elevation	−3656.00	−919.00	−384.00	−605.12	−83.25	394.50
50	elevation	−2980.50	−518.62	−158.50	−306.12	0.75	859.00
20	elevation	−2666.50	−229.75	−19.00	−78.29	117.75	1235.00
10	elevation	−1419.00	−124.00	3.00	−1.60	144.63	1175.50
5	elevation	−1334.50	−90.75	3.00	2.33	104.50	1016.00
100	precipitation	−799.00	−106.00	109.00	444.20	402.80	3517.00
50	precipitation	−1180.00	−54.25	71.25	164.18	231.25	3130.00
20	precipitation	−1346.00	−121.00	35.50	32.11	177.00	1873.00
10	precipitation	−2317.00	−102.25	14.00	29.77	152.75	1719.50
5	precipitation	−1526.00	−72.88	8.50	12.42	103.50	1396.00
100	temperature	−2.50	0.60	2.00	3.20	4.85	20.40
50	temperature	−4.50	−0.03	0.95	1.66	2.65	16.40
20	temperature	−6.50	−0.65	0.10	0.42	1.20	12.30
10	temperature	−6.30	−0.80	0.00	−0.01	0.70	8.75
5	temperature	−9.65	−0.60	0.00	−0.03	0.45	6.50

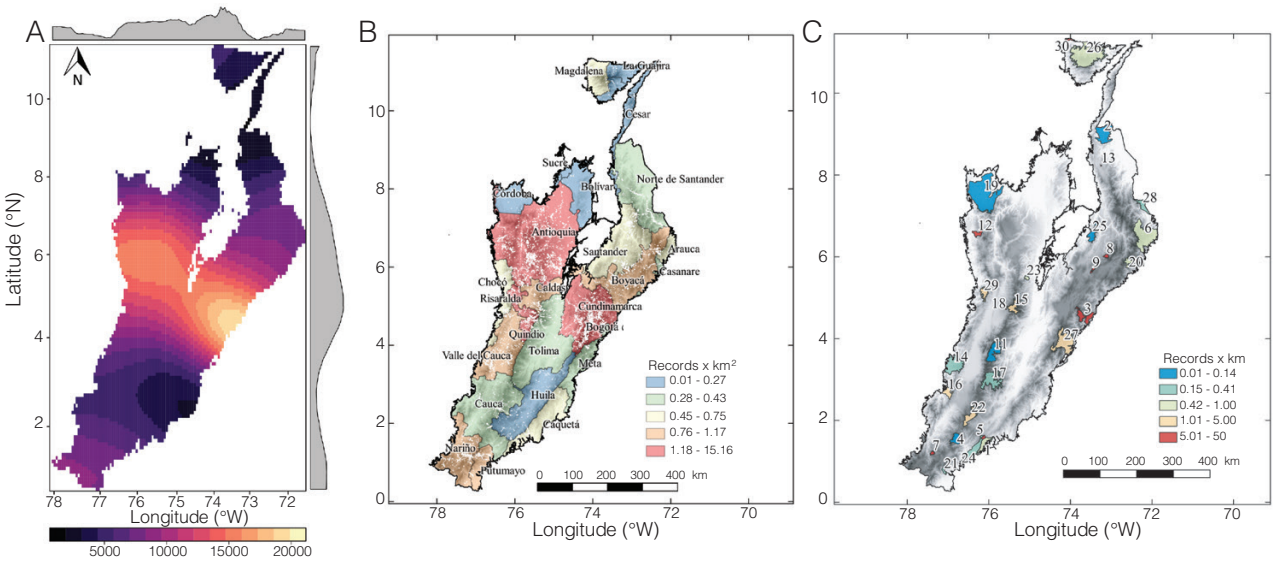

Figure 2. A, Plant record density across the Colombian Andean region. The grey area above the map indicates the longitudinal concentration of records, and the grey to the right, the latitudinal concentration. B, Variation in record number per km² for each of the Andean Departments of Colombia. The white areas correspond to the actual records. C, Collection pattern per km² in Protected Areas of the Colombian Andes. The elevation variation is shown in grey, with darker colours indicating higher elevations. Park names: 1 Alto Fragua Indiwasi; 2 Catatumbo Bari; 3 Chingaza; 4 Complejo Volcánico Doña Juana Cascabel; 5- Cueva de los Guácharos; 6 El Cocuy; 7 Galeras; 8 Guanentá Alto Río Fonce; 9 Iguaque; 10 Isla de la Corota; 11 Las Hermosas; 12 Las Orquídeas; 13 Estoraques; 14 Farallones de Cali; 15 Los Nevados; 16 Munchique; 17 Nevado del Huila; 18 Otún Quimbaya; 19 Paramillo; 20 Pisba; 21 Plantas Medicinales Orito Ingi Ande; 22 Puracé; 23 Selva de Florencia; 24 Serranía de los Churumbelos; 25 Serranía de los Yariguíes; 26 Sierra Nevada de Santa Marta; 27 Sumapaz; 28 Tamá; 29 Tatamá; 30 Tayrona.

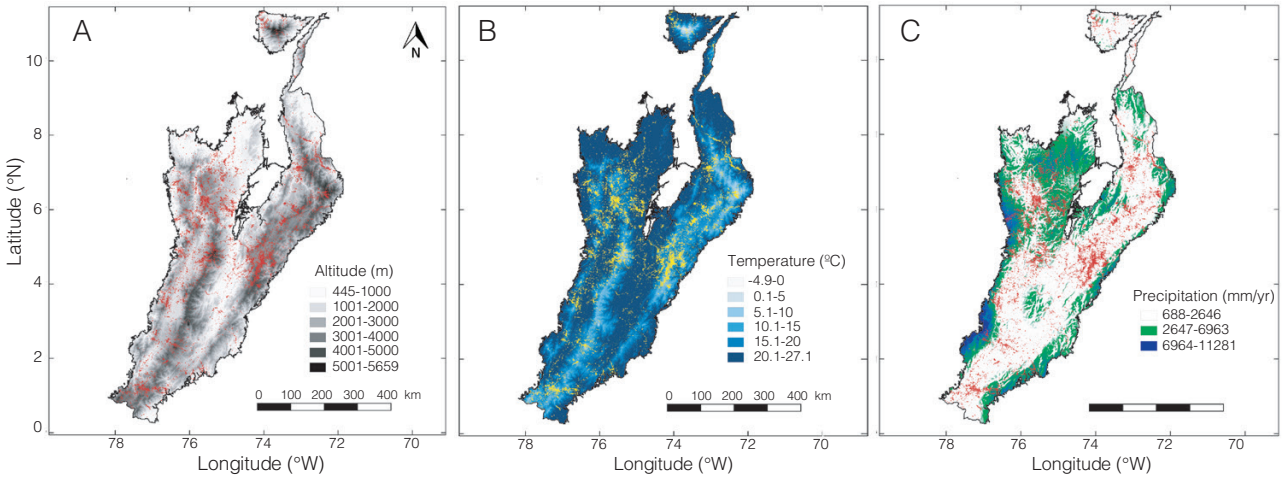

Figure 3. Spatial distribution of plant records (red dots) across the Colombian Andean Region in relation to: A, elevation (m), B, mean annual temperature (°C) and C, annual precipitation (mm/year). The yellow and red dots correspond to the plant records collected in the area.

magnitude between 3000 and 4000 m. In contrast, localities below 2000 m were under-represented, particularly lowland forests (localities between 445 and 1000 m) (Table 2; Fig. 3A). Because of the negative correlation between temperature and elevation, lower temperature regimes (from −5 to 15 ° C) were over-represented by plant records in the database, whereas higher temperature regimes (from 15 to 27 ° C) were under-represented (Table 2; Fig. 3B).

Seventy-five percent of the Colombian Andean region receives 688–2646 mm of rain per year, with few areas (25%) receiving > 2646 mm/year. Specimen records were under-represented in areas with high precipitation (> 1666 mm/year), located in the foothills to the west of the Western Cordillera, east of the Eastern Cordillera and north of the north-western Central Cordillera (Table 2; Fig. 3C).

Coverage, density plant records representativity and scale effect

Decreasing resolution inflated SC and increased the number of plant records by grid cell. For example, whereas at low resolutions (e.g. cell size 100 × 100 km), 90% of the area of the Colombian Andean region was covered, and the median number of records by grid cell was 2540, the coverage at high resolution (e.g. cell size 5 × 5) dropped to 51% and the median record number was 57 (Supplementary Information, Fig. S1; Table S1).

Despite this, plant records at a high resolution were better able to represent environmental variability of grid cells than those at a low resolution. Meanwhile, the difference between grid cells and plant records was close to zero in grid cells of 5 × 5 km for the environmental variables considered (annual precipitation, mean annual temperature, elevation); the difference was maximum in grid cells of 100 × 100 km (Fig. 4; Table 3).

Completeness of Colombian Andean flora

Sample completeness decreased from low to high resolution (e.g. SC median at 100 × 100 km cell size was 0.68, whereas the SC decreased progressively to 0.22 at 5 × 5 km). A low proportion of grid cells were well sampled at all scales studied; for example, whereas the quartile 75 (Q75) of grid cells of 100 × 100 km had SC > 0.8, the Q75 decreased significantly in cells of 10 × 10 km and 5 × 5 km where the Q75 were 0.45 and 0.39, respectively. Grid cells with SC > 0.8 were atypical (Fig. 5). Spatially low resolutions (e.g. grid cells of 100 × 100, 50 × 50 and 20 × 20 km) with SC > 0.8 were concentrated between 3 ° and 7 °N. This section corresponds to the central and northern area of the Western and Central cordilleras (e.g. Antioquia, Caldas Risaralda, Quindio y Valle del Cauca Departments); the northern part of the Eastern Cordillera (e.g. Cundinamarca, Boyacá and Santander Departments) and the southern area of the Colombian Andean region (e.g. Nariño and western Putumayo Departments). The areas with SC < 0.8 were in the northern part of the Eastern Cordillera (e.g. Serranía del Perijá), Sierra Nevada de Santa Marta and the central area of Colombian Andean region (e.g. Cauca, Huila, Tolima and western Caquetá Departments). Higher resolutions (e.g. grid cell sizes 10 × 10 and 5 × 5) showed the grid cells with SC > 0.8 on areas > 2000 m elevation (Fig. 5).

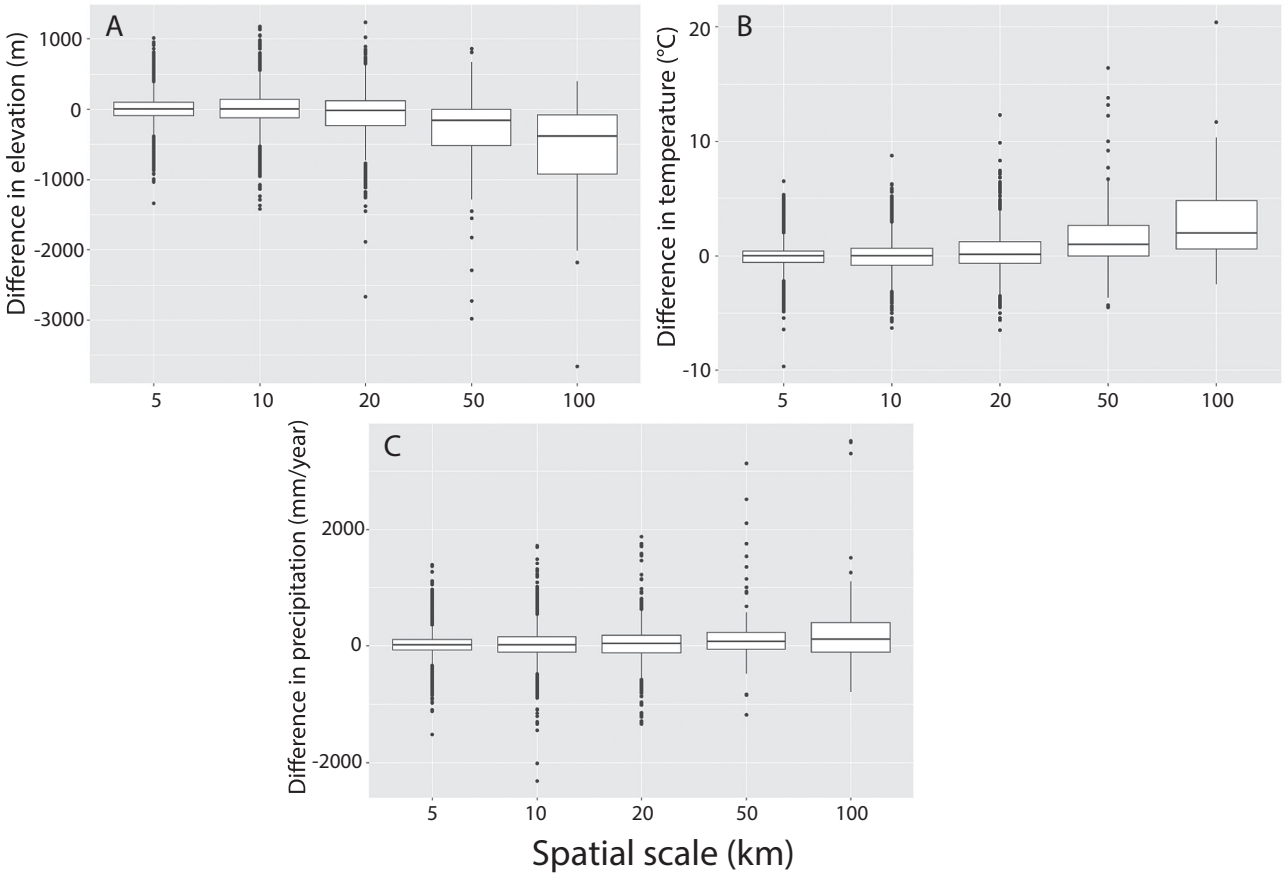

Figure 4. Scale effect (5 × 5, 10 × 10, 20 × 20, 50 × 50 and 100 × 100 km) on the environmental difference between specimen records and grid cells across the Colombian Andes. Boxplots show the differences for: A, elevation (m), B, mean annual temperature (°C) and C, mean annual precipitation (mm/year). The difference was calculated as the environmental median for plant records per grid cell minus the environmental median per grid cell. The bottom and top part of the boxplot indicates the 25th and 75th percentile, respectively, the horizontal line within the box, the median value and the dots, the outliers. Scale of 100 refers to 100 × 100 km, 50 to 50 × 50, 20 to 20 × 20 km, 10 to 10 × 10 km and 5 to 5 × 5 km.

DISCUSSION

Geographical bias and gaps

Our results showed strong geographical and environmental bias in the digitally available plant data for the Colombian Andes. In our study, the highest plant collection density was around Bogotá and Medellín where the largest and oldest herbaria are located (Parra & Díaz, 2016). These herbaria contribute 43% of specimens in our database (Supplementary Information, Table S3). Three gaps (low specimen record density or no records; Figs 2, 3) were located: the first in the northern part of the Central Cordillera; the second in the northern part of the Eastern Cordillera, including Serranía del Perijá and the Sierra Nevada de Santa Marta, and the third in Tolima, Huila and Cauca Departments and the eastern foothills of the Eastern Cordillera in Caquetá and Meta Departments (Fig. 2B). Thus, these areas may potentially host higher plant biodiversity than current estimates suggest but are poorly known, or collections from these regions may exist in smaller herbaria that have not been databased, digitized or are not publicly available.

Sampling bias has been associated with several factors such as proximity to roads (Kadmon *et al.*, 2004; Oliveira *et al.*, 2016), accessibility to research facilities (e.g. herbaria), seasonality (Daru *et al.*, 2018), research (Bonnet, Shine & Lourdais, 2002) or societal preferences (Troudet *et al.*, 2017). In the Colombian Andes, plant records are concentrated around the major cities (Bogotá and Medellín; Figs 2, 3), where research infrastructure (e.g. herbaria such as COL, UDBC, HUA, JAUM and COA; Parra & Díaz, 2016) and plant specialists are concentrated. The strong sampling bias discovered here reflects the worldwide tendency in which botanists tend to collect near their homes and research facilities (Moerman & Estabrook, 2006;

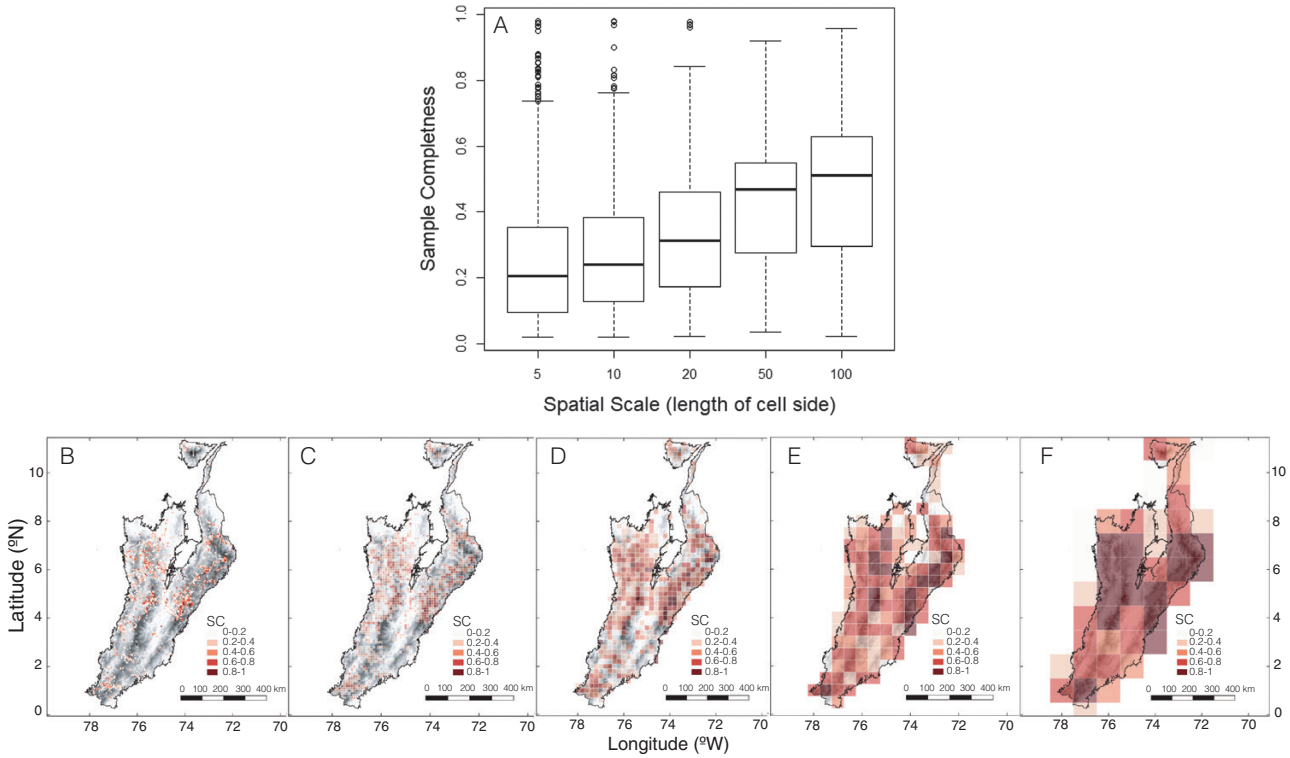

Figure 5. Variation in sampling completeness (SC) of plants at different spatial scales (5 × 5, 10 × 10, 20 × 20, 50 × 50 and 100 × 100 km). A, SC boxplot in the Colombian Andes at different spatial scales. The bottom and top part of the boxplot indicates the 25th and 75th percentile, respectively, the horizontal line within the box, the median value and the dots, the outliers. B, SC maps at 5 × 5, where dark red represents high completeness and areas with no collections are shown in grey. C, SC at 10 × 10. D, SC at 20 × 20. E, SC at 50 × 50. F, SC at 100 × 100. Areas considered as well sampled (SC values > 90%) shown at different spatial scales. Scale of 100 refers to 100 × 100 km, 50 to 50 × 50, 20 to 20 × 20 km, 10 to 10 × 10 km and 5 to 5 × 5 km.

Yang *et al.*, 2014; Engemann *et al.*, 2015; Lagomarsino & Frost, 2020). In addition, due to the topographic complexity of the study region, many areas are extremely remote and difficult to access physically, and they are thus less likely to have been collected. Others have been occupied by armed groups that coincide with areas with low plant collections, such as Cauca, Tolima and Meta Departments where FARC guerrillas have been present during the last 60 years. Some areas may have had their native vegetation removed and replaced by other crops, pastures or illicit crop plantations (Etter & van Wyngaarden, 2000).

Vast land areas in the Colombian Andean region are still in their natural state or little transformed. Many of those are in national parks where the sampling is low (less than one record per km^2 was found in 15 of the 30 national parks; Fig. 2C; Supplementary Information, Table S4) or biased (e.g. PNN Chingaza, PNN Sumapaz and PNN El Cocuy, where sampling is biased in páramo areas). Only nine protected areas (out of 30) exceeded six records per km^2 (Supplementary Information, Table S4).

Environmental bias and gaps

Several studies have shown that there is an environmental sampling bias reflected in that particular biomes or ecosystems are better sampled than others. For example, Sousa-Baena *et al.* (2013) found more sampling effort in the Amazonian region, whereas Caatinga and Cerrado lacked biodiversity information. In montane areas, the sampling effort has been focused on the highlands (Yang *et al.*, 2014; Engemann *et al.*, 2015). For instance, in the Colombian Andes more samples were collected in areas above 2000 m, especially between 3000 and 4000 m. These latter elevations correspond to the páramo ecosystem, recognized for its high speciation rate, diversity and endemism (e.g. Luteyn, 1999; Hughes & Eastwood, 2006; Madriñán, Cortés & Richardson, 2013; Nürk, Scheriau & Madriñán, 2013) (Fig. 4). Beyond the relevance of páramo in terms of climate change studies (e.g. Peyre *et al.*, 2015; Lasso *et al.*, 2021) and evolutionary processes (e.g. Madriñán *et al.*, 2013; Flantua & Hooghiemstra, 2018), scientific preferences related to proximity to research facilities

and concentration of botanists may partly explain its prominence in floristic data. There may also have been a greater focus on the páramo due to its importance as a water source for many Colombian Andean cities. The provision of this fundamental resource by this ecosystem has perhaps resulted in more studies and hence more collections of species found in páramo.

Surprisingly, areas below 2000 m were undersampled, even though the forest at *c.* 1500 m has been recognized as having the greatest species richness in the Andes (Gentry, 1995; Särkinen *et al.*, 2012; Engemann *et al.*, 2015). The same tendency was observed in the high rainfall areas, where high plant diversity is also expected (e.g. Pennington, Hughes & Moonlight, 2015; Cardoso *et al.*, 2017), but the sampling was again comparatively poor (Fig. 3). These areas correspond to the humid tropical forest of the Western Cordillera lowlands, the foothills of the Eastern Cordillera and the Magdalena and Cauca river valleys, areas where the roads are scarce and armed groups have been active. Other biomes under-represented with restricted distributions in the Colombian Andes were the subxerofitic tropical biomes (scattered and highly transformed) of the inter-Andean valleys and the sub-Andean biome or 'selva subandina'. This biome is located between 1000 and 2400 m and includes a heterogeneous mix of 25 different biogeographic districts as recognized by Rodríguez *et al.* (2006).

Sampling bias and biological implications

The usefulness of database information depends on completeness of species inventories and even distribution of sampling in time, space and environmental dimensions (Troia & McManamay, 2016). The bias and gaps in the digitally available records of the Colombian Andes have resulted in a different level of SC at different scales. Our analysis indicates that the flora registered at broader scales (e.g. 100 × 100 km) is *c.* 60 to 68% of the total richness expected (Fig. 5). However, the level of knowledge of the Colombian Andes floristic richness could be underestimated, because of the high topographic and environmental variability in these grid cells and the sampling bias we demonstrate here. For example, in an area of 100 × 100 km, the elevational range may exceed 4000 m and include many ecosystems (from lowlands to highlands), many of them not represented by specimen records.

In contrast, environmental variability at higher spatial resolutions is low (Fig. 4). Therefore, specimen records are more likely to represent the environmental conditions within the cells, having more even sampling and increased accuracy of the SC estimation. However, the total area covered by plant collections is small at these resolutions, with > 50% of the Colombian Andean region lacking information.

The SC of cells with specimen records show that > 60% of the plant diversity remains unregistered at scales of 20 × 20, 10 × 10 and 5 × 5 km. In fact, < 10% of the grid cells at 100 × 100 km and < 1% of grid cells at 5 × 5 km could be considered floristically well studied (SC > 90%). These results agree with Engemann *et al.* (2015), who reported severe undersampling for Ecuador, indicating that much more sampling or different methods are needed to provide reliable richness estimation for countries with poor data collection. According to their study, large cell sizes can contain many different habitats, resulting in higher number of species (Engemann *et al.*, 2015) as is the case in Colombia (Fig. 4). This topographic complexity is the most important driver of species richness in the Andes (Distler *et al.*, 2009).

Some of the information that could help fill the gaps could be recovered from small collections not yet databased or digitized or from information (already available in databases) that was discarded due to quality issues. Only 12% of the dataset downloaded from sources used for this study proved useful. The main reason for discarding records was issues with georeferencing, such as records without coordinates or coordinates in the ocean. Another source of loss was duplicate records, as different herbaria shared collections or the same record was in multiple databases.

Data availability in the Colombian Andean Region

Plant occurrences from Colombia are scattered across national and international herbaria, and only some have been digitalized and made publicly available. Although this study did not have access to all digital plant data from the Colombia Andes, we created a comprehensive database compiled from national and international herbaria databases in which the Colombian flora is well represented (Supplementary Information, Table S1). As well as GBIF, the central repository of biodiversity records includes records from herbaria that we cannot directly access (Supplementary Information, Table S3). However, despite the number of plant occurrences gathered, important quality issues related to the geographical dimension of the data made 88.2% of the Colombian data unusable. Together with institutional sharing policies, these issues make data access for biodiversity research difficult.

In this study we found a low SC and a high sampling bias for the Colombian Andes, with > 260 000 records (close to 20 000 species) for this region alone. Some recent studies have attempted to propose a biogeographic regionalization for the whole country (*c.* 270 000 records; González-Orozco, 2021). This study

shows that there will be areas with low sampling and low SC, and therefore the areas proposed may not reflect true biogeographic regions. A good practice when calculating species richness or species occurrences would be to explicitly take into account the uncertainty by creating Distributional Uncertainty Maps (or maps of ignorance) (Rocchini *et al.*, 2011). These maps provide a spatially explicit quantification of uncertainty and would reflect the areas that need more fieldwork to attain reliable knowledge on species distributions.

RECOMMENDATIONS

The analysis of completeness for the digitally available records of the Colombian Andean flora indicated that vast areas of the region are yet to be explored and sampled, and this is even more important given the accelerated rate of land use transformation. It is therefore necessary to increase the sampling effort and improve floristic knowledge of undersampled regions to fill gaps on distributions (Wallacean shortfall) (Hortal *et al.*, 2015) and the environmental tolerance of species (Hutchinsonian shortfall), both criteria important for conservation. In this study we used elevation, temperature and precipitation at 1 km (30 arcsec) to study the environmental representation of plant records in the Colombian Andean region. We found sampling bias in areas around main cities of Colombia and in the high and cold Andean forest and páramos, whereas the lowlands and humid areas are poorly collected. This could also have consequences in terms of representing unique conditions (such as refugia) and vegetation limited to small areas that are more likely to be encountered in regions of high topographic complexity such as in our study area.

It is crucial to promote strategies to obtain new data to improve the accuracy of richness inferences and ensure that conservation policies are based on sufficient information. In the future, encouraging the mobilization of data and strategically increasing sampling efforts will result in better information and diminished biodiversity shortfalls (Hortal *et al.*, 2015). In Colombia, 50% of plant collections and 40% of those digitized come from three herbaria (COL, HUA, FMB; http://rnc.humboldt.org.co/wp/colecciones/, consulted October 2021) that are focused on the flora of Colombia. The small herbaria that are focused on regional floras, e.g. Universidad de Pamplona (HECASA), Norte de Santander; Instituto Tecnologico del Putumayo (HEAA), Putumayo; Universidad Nacional de Colombia (VALLE), Valle and Universidad Surcolombiana (SURCO), Huila, are not adequately databased. The importance of local herbaria has been outlined by Delves (2021), who pointed out that promoting the mobilization of specimen data from physical to digital formats could uncover new localities or new species while also reducing spatial and environmental bias and increasing sampling completeness. Therefore, we recommend that more funding be directed toward smaller regional herbaria to allow them to curate, digitize and database their collections. We also consider it essential to encourage the formation of new botanists at regional levels to strengthen local collections and to incorporate the indigenous knowledge base. National Parks also require more focus as, although the biodiversity is protected, the median sampling is 1 record/km^2 (Fig. 2C; Supplementary Information, Table S4). It is reasonable to assume that protected areas contain many species that remain to be described, but bureaucratic issues prevent researchers from exploring those areas. It is important to strengthen ways to work together with communities and institutions to improve floristic knowledge in those areas. Finally, and perhaps most obviously, more investment in fieldwork is needed in under-collected areas.

ACKNOWLEDGEMENTS

We are grateful to all the collaborators and contributors of the Global Biological Information Facility (GBIF), the Biodiversity Informatics Program of the Instituto de Ciencias Naturales, which administers the database of the Herbario Nacional Colombiano (COL), Jardín Botánico de Bogotá (Proyecto flora de Bogotá) and the Missouri Botanical Garden (Tropicos) for making their herbarium data available. We would like to thank the group 'Genética evolutiva, filogeografía y ecología de biodiversidad Neotropical' and the High Performance Computing service of the Universidad del Rosario for hosting our PostgreSQL database on their servers. This study would not have been possible without the support of Colciencias Doctoral funds and the support of Universidad del Rosario. We would also like to thank Iván Jiménez (curator at the Missouri Botanical Garden) and Orlando Rivera-Díaz (Instituto de Ciencias Naturales, Universidad Nacional de Colombia) for their valuable comments on the document.

FUNDING

This project was supported by Colciencias Doctoral funding (727-2015) and Universidad del Rosario, through a teaching assistantship and a doctoral grant.

AUTHOR CONTRIBUTIONS

CV, MB, TS and AS conceived and designed the research. CV, MB and LR obtained and processed the plant records. CV, MB and CG analysed the data. CV, TS, JR and AS wrote and edited the manuscript.

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

The datasets we used were deposited in ZENODO (http://doi.org/10.5281/zenodo.4726190).

REFERENCES

Bernal R, Grandstein R, Celis M, eds. 2016. *Catálogo de plantas y líquenes de Colombia*. Bogotá: Editorial Universidad Nacional de Colombia.

Bivand RS, **Pebesma E**, **Gómez-Rubio V. 2013.** *Applied spatial data analysis with R*. New York: Springer US.

Bivand RS, **Wong DWS. 2018.** Comparing implementations of global and local indicators of spatial association. *TEST* **27:** 716–748.

Bonnet X, **Shine R**, **Lourdais O. 2002.** Taxonomic chauvinism. *Trends in Ecology & Evolution* **17:** 1–3.

Botero CA. 2020. La paz produce ciencia. Expediciones biológicas en reemplazo de la guerra. *Biodiversidad en la práctica. Documentos de trabajo del Instituto Humboldt* **5:** 1–14.

Cardoso D, **Särkinen T**, **Alexander S**, **Amorim AM**, **Bittrich V**, **Celis M**, **Daly DC**, **Fiaschi P**, **Funk VA**, **Giacomin LL**, **Goldenberg R**, **Heiden G**, **Iganci J**, **Kelloff CL**, **Knapp S**, **Cavalcante de Lima H**, **Machado AFP**, **dos Santos RM**, **Mello-Silva R**, **Michelangeli FA**, **Mitchell J**, **Moonlight P**, **de Moraes PLR**, **Mori SA**, **Nunes TS**, **Pennington TD**, **Pirani JR**, **Prance GT**, **de Queiroz LP**, **Rapini A**, **Riina R**, **Vargas-Rincon CA**, **Roque N**, **Shimizu G**, **Sobral M**, **Stehmann JR**, **Stevens WD**, **Taylor CM**, **Trovó M**, **van den Berg C**, **van der Werff H**, **Viana PL**, **Zartman CE**, **Forzza RC. 2017.** Amazon plant diversity revealed by a taxonomically verified species list. *Proceedings of the National Academy of Sciences, USA* **114:** 10695–10700.

Chao A, Jost L. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* **93:** 2533–2547.

Daru BH, **Park DS**, **Primack RB**, **Willis CG**, **Barrington DS**, **Whitfeld TJS**, **Seidler TG**, **Sweeney PW**, **Foster DR**, **Ellison AM**, **Davis CC. 2018.** Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* **217:** 939–955.

Delves J. 2021. *Decolonise herbaria and specimen data: quantifying the contribution of local herbaria to biodiversity patterns.* Unpublished MSc thesis, The University of Edinburgh and Royal Botanic Garden Edinburgh.

Distler T, **Jørgensen PM**, **Graham A**, **Davidse G**, **Jiménez I. 2009.** Determinants and prediction of broad-scale plant richness across the western Neotropics. *Annals of the Missouri Botanical Garden* **96:** 470–491.

Engemann K, **Enquist BJ**, **Sandel B**, **Boyle B**, **Jørgensen PM**, **Morueta-Holme N**, **Peet RK**, **Violle C**, **Svenning JC. 2015.** Limited sampling hampers 'big data' estimation of species richness in a tropical biodiversity hotspot. *Ecology and Evolution* **5:** 807–820.

Etter A, **van Wyngaarden W. 2000.** Patterns of landscape transformation in Colombia, with emphasis in the Andean Region. *AMBIO: A Journal of the Human Environment* **29:** 432–439.

Feeley KJ. 2015. Are we filling the data void? An assessment of the amount and extent of plant collection records and census data available for tropical South America. *PLoS One* **10:** 1–17.

Feeley KJ, **Silman MR. 2011.** Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions* **17:** 1132–1140.

Flantua S, **Hooghiemstra H. 2018.** Historical connectivity and mountain biodiversity. In: Hoorn C, Perrigo A, Antonelli A, eds. *Mountains, climate and biodiversity*. Oxford: John Wiley & Sons, 171–185.

García Márquez J, **Dormann C**, **Sommer JH**, **Schmidt M**, **Thiombiano A**, **Sylvestre Da S**, **Chatelain C**, **Dressler S**, **Barthlott W. 2012.** A methodological framework to quantify the spatial quality of biological databases. *Biodiversity & Ecology* **4:** 25–39.

gbif.org. 2017. GBIF occurrence download.

Gentry AH. 1995. Patterns of diversity and floristic composition in Neotropical montane forest. In: Churchill SP, Balslev H, Forero E, Luteyn JL, eds. *Biodiversity and conservation of Neotropical montane forests*. New York: The New York Botanical Garden, 103–126.

González-Oreja JA, **de la Fuente-Díaz-Ordaz AA**, **Hernández-Santín L**, **Buzo-Franco D**, **Bonache-Regidor C. 2010.** Evaluación de estimadores no paramétricos de la riqueza de especies. Un ejemplo con aves en áreas verdes de la Ciudad de Puebla, México. *Animal Biodiversity and Conservation* **33:** 31–45.

González-Orozco CE. 2021. Biogeographical regionalisation of Colombia: a revised area taxonomy. *Phytotaxa* **484:** 247–260.

Gotelli NJ, **Chao A. 2013.** Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: Levin SA, ed. *Encyclopedia of biodiversity*, *2nd edn*. New York: Elsevier, 195–211.

Gotelli NJ, **Colwell RK. 2011.** Estimating species richness. In: Magurran AE, McGill BJ, eds. *Biological diversity: frontiers in measurement and assessment*. Oxford: Oxford University Press, 359.

Hortal J, **de Bello F**, **Diniz-Filho JAF**, **Lewinsohn TM**, **Lobo JM**, **Ladle RJ. 2015.** Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46:** 523–549.

Hortal J, **Borges PA V**, **Gaspar C. 2006.** Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology* **75:** 274–287.

Hsieh TC, **Ma KH**, **Chao A. 2016.** iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution* **7:** 1451–1456.

Hughes C, **Eastwood R. 2006.** Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. *Proceedings of the National Academy of Sciences, USA* **103:** 10334–10339.

Humboldt A, **Bonpland A. 1807.** *Essai sur la géografie des plantes*. Schoell: Levrault.

Jiménez I, **Distler T**, **Jørgensen PM. 2009.** Estimated plant richness pattern across northwest South America provides similar support for the species-energy and spatial heterogeneity hypotheses. *Ecography* **32:** 433–448.

Kadmon R, **Farber O**, **Danin A. 2004.** Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* **14:** 401–413.

Karger DN, **Conrad O**, **Böhner J**, **Kawohl T**, **Kreft H**, **Soria-Auza RW**, **Zimmermann NE**, **Linder HP**, **Kessler M. 2017.** Climatologies at high resolution for the earth's land surface areas. *Scientific Data* **4:** 1–20.

Lagomarsino LP, **Frost LA. 2020.** The central role of taxonomy in the study of Neotropical biodiversity. *Annals of the Missouri Botanical Garden* **105:** 405–421.

Lasso E, **Matheus-Arbeláez P**, **Gallery RE**, **Garzón-López C**, **Cruz M**, **Leon-Garcia IV**, **Aragón L**, **Ayarza-Páez A**, **Llambi LD. 2021.** Homeostatic response to three years of experimental warming suggests high intrinsic natural resistance in the páramos to warming in the short term. *Frontiers in Ecology and Evolution* **9:** 1–22.

Loiselle BA, **Jørgensen PM**, **Consiglio T**, **Jiménez I**, **Blake JG**, **Lohmann LG**, **Montiel OM. 2008.** Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography* **35:** 105–116.

Luteyn JL. 1999. *Páramos: a checklist of plant diversity, geographical distribution, and botanical literature*. The Bronx: New York Botanical Garden Press.

Madriñán S, **Cortés AJ**, **Richardson JE. 2013.** Páramo is the world's fastest evolving and coolest biodiversity hotspot. *Frontiers in Genetics* **4:** 1–7.

Meyer C, **Weigelt P**, **Kreft H**, **Lambers JHR. 2016.** Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19:** 992–1006.

Moerman DE, **Estabrook GF. 2006.** The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography* **33:** 1969–1974.

Moura MR, **Jetz W. 2021.** Shortfalls and opportunities in terrestrial vertebrate species discovery. *Nature Ecology and Evolution* **5:** 631–639.

Mutke J. 2017. Mesoscale patterns of plant diversity in Andean South America based on combined checklist and GBIF data. *Berichten der Reinhold-Tüxen-Gesellschaft* **29:** 83–97.

Myers N, **Mittermeier R**, **Mittermeier C**, **da Fonseca G**, **Kent J. 2000.** Biodiversity hotspots for conservation priorities. *Nature* **403:** 853–858.

Nürk NM, **Scheriau C**, **Madriñán S. 2013.** Explosive radiation in high Andean *Hypericum*-rates of diversification among New World lineages. *Frontiers in Genetics* **4:** 1–14.

Oliveira U, **Pereira Paglia A**, **Brescovit AD**, **de Carvalho CJB**, **Paiva Silva D**, **Rezende DT**, **Leite FSF**, **Nogueira Batista JA**, **Pena Barbosa JPP**, **Stehmann JR**, **Ascher JS**, **Ferreira de Vasconcelos M**, **De Marco P**, **Lowenberg-Neto P**, **Guimaraes Dias P**, **Gianluppi Ferro V**, **Santos AJ. 2016.** The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Diversity and Distributions* **22:** 1232–1244.

Parra C, **Díaz S. 2016.** *Herbarios y jardines botánicos:esttimonios de nuestra Biodiversidad*. Bogotá: Universidad Nacional de Colombia (sede Bogotá).

Pennington RT, **Hughes M**, **Moonlight PW. 2015.** The origins of tropical rainforest hyperdiversity. *Trends in Plant Science* **20:** 693–695.

Pennington RT, **Lavin M**, **Sarkinen T**, **Lewis GP**, **Klitgaard BB**, **Hughes CE. 2010.** Contrasting plant diversification histories within the Andean biodiversity hotspot. *Proceedings of the National Academy of Sciences, USA* **107:** 13783–13787.

Peyre G, **Balslev H**, **Martí D**, **Sklenář P**, **Ramsay P**, **Lozano P**, **Cuello N**, **Bussmann R**, **Cabrera O**, **Font X. 2015.** VegPáramo, a flora and vegetation database for the Andean páramo. *Phytocoenologia* **45:** 195–201.

QGIS Development Team. 2015. QGIS geographic information system, open source Geospatial Foundation project, version 3.8.0.

R Development Core Team. 2019. R: a language and environment for statistical computing (Version 3.6.1).

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

Richter M. 2008. Tropical mountain forest - distribution and general features. In: Gradstein SR, Homeier J, Gansert D, eds. *The tropical mountain forest. Patterns and processes in a biodiversity hotspot*. Göttingen: Universitätsverlag Göttingen, 1–224.

Rocchini D, **Hortal J**, **Lengyel S**, **Lobo JM**, **Jiménez-Valverde A**, **Ricotta C**, **Bacaro G**, **Chiarucci A. 2011.** Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography* **35:** 211–226.

Rodríguez N, **Armenteras D**, **Morales M**, **Romero M. 2006.** *Ecosistemas de los Andes colombianos*. Bogotá: Instituto de Investigación de Recursos Biológicos Alexander von Humboldt.

Rowe R. 2005. Elevational gradient analyses and the use of historical museum specimens: a cautionary tale. *Journal of Biogeography* **32:** 1883–1897.

Särkinen T, **Pennington RT**, **Lavin M**, **Simon MF**, **Hughes CE. 2012.** Evolutionary islands in the Andes: persistence and isolation explain high endemism in Andean dry tropical forests. *Journal of Biogeography* **39:** 884–900.

Schmidt-Lebuhn AN, **Knerr NJ**, **Kessler M. 2013.** Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and Conservation* **22:** 905–919.

Sousa-Baena MS, **Couto L**, **Townsend A. 2013.** Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions* **20:** 1–13.

Thiers BM. 2020. *Herbarium: the quest to preserve and classify the world's plants*. Portland: Timber Press.

Troia MJ, **McManamay RA. 2016.** Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. *Ecology and Evolution* **6:** 4654–4669.

Troudet J, **Grandcolas P**, **Blin A**, **Vignes-Lebbe R**, **Legendre F. 2017.** Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7:** 1–14.

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.

SUPPORTING INFORMATION

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

Table S1. Plant record sources used to analyse bias in the Colombian Andean flora.
Table S2. Plant records used to analyse gaps in the Colombian Andean flora.
Table S3. Plant records from Colombia gathered from GBIF**. Collection (Coll).**
Table S4. Protected areas in the Colombian Andes region indicating category, area, number of plant records (# records) and the density of plant records by km^2 (Rec/km^2). PNN = Parque Nacional Natural SFF = Santuario de Flora y Fauna; ANU = Area Natural Unica; SF = Santuario de Flora.
Figure S1. Collection density (i.e. number) of digitally available plant specimen records across the Colombian Andes at different grid cell sizes. A, Boxplots of the number of plant specimen records by grid cell across different scales (5 × 5, 10 × 10, 20 × 20, 50 × 50 and 100 × 100 km). The bottom and top part of the boxplot indicates the 25th and 75th percentile, respectively, the horizontal line within the box represents the median value and the circles represent the outliers. B, Map of collection density at different spatial scales, where dark red denotes areas with high density, yellow areas with < 20 records, and white areas without records. Scale of 100 refers to 100 × 100 km, 50 to 50 × 50, 20 to 20 × 20 km, 10 to 10 × 10 km and 5 to 5 × 5 km.