Gaps and bias of the Colombian Andes flora in databases

by

Carlos Alberto Vargas Rincón

Ph.D. Dissertation. Doctorado en ciencias biomédicas y biológicas

Advisor

Prof. Dr. Adriana Sanchez Andrade. Universidad del Rosario

Other members

Prof. Dr. James Richardson. University College Cork

Prof. Dr. Tiina Sarkinen. Royal Botanic Garden Edinburgh

Escuela de medicina y ciencias de la salud

Universidad del Rosario

Bogotá 2023

INTRODUCTION

The Northern Andean Region has attracted significant scientific interest due to its exceptional biodiversity and high rates of diversification (Nürk *et al.*, 2013). It is also one of the most endangered regions on the planet and has been designated as a biodiversity hotspot (Myers *et al.*, 2000; Orme *et al.*, 2005). The region encompasses the mountainous zone from the Amotape junction, at 5° south latitude, to the Oca - Romeral system and other fault systems, in contact with the Caribbean plate, at 12° north latitude (Graham, 2009). The Colombian Andean Region, consisting of three Cordilleras, two of which are of tectonic origin (Eastern and Western), and one of volcanic origin (Central Cordillera), along with an isolated mountain system, Sierra Nevada de Santa Marta, accounts for most of the northern region of the Andes (Graham, 2009).

The Colombian plant catalog shows that the Andean Region is home to 18,491 species of vascular plants (Bernal *et al.* , 2016; accessed in December 2018; http://catalogoplantasdecolombia.unal.edu.co/es/), making it the region with the highest floristic diversity in the country. This richness is attributed to Colombia's location at the crossroads of the Isthmus of Panama and the South American continent, where the floristic elements of North and Central America blend with the those of South America. It has also been suggested that the uplift of the Andes and topographic complexity contribute to high species richness (Gentry, 1995; van der Hammen, 2000; Antonelli & Sanmartín, 2011; Bacon *et al.*, 2015).

The study of the origin of high species diversity, the link between diversity and environmental factors, and conservation and management efforts heavily relies on the

quality of digitally available data. Herbaria serve as the primary source of data for plants, followed by information from permanent plots and datasets derived from inventories. These datasets are fundamental to supporting multiple studies at different scales and disciplines. They provide a foundation for understanding evolutionary processes and mechanisms, current species distribution and patterns of diversity, and serve as the basis for decision-making on conservation and sustainable management plans.

The amount of data available for the Colombian Andes is unclear. Parra & Díaz (2016) report that 47 herbaria in Colombia contain approximately 1,685,224 specimens in their collections, with 37 of these located in the Andean Region. Given that the region has the largest herbaria, universities, botanical gardens, and research centers, it is expected to have the largest proportion of collections. Despite the existence of multiple herbaria, few of them have open access to their collections via online consultation platforms (e.g., Colombian National Herbaria [COL; http://www.biovirtual.unal.edu.co/es/colecciones/search/plants/], Herbario Forestal [UDBC; http://herbario.udistrital.edu.co/herbario/public/es]). Additionally, some of these institutions have made a portion of their collections publicly available through data repositories. SIB-Colombia has been the most widely used tool and channel for making these data available worldwide through GBIF (Global Biodiversity Information Facility).

The limited availability of plant records means that critical questions about ecology, evolution, biogeography, conservation, and resource management that rely on reliable information, are being addressed with insufficient data. Hence, increasing the digitization and curation of collections is fundamental. It could also enhance the importance of collections as sources of valuable information on biodiversity by contributing to a greater

number of records available digitally on open access platforms. Moreover, it would highlight the importance of the quality and reliability of the currently available data.

Knowledge deficit

Curiosity and the desire to interpret the world is the characteristic that has driven human beings, through observations and experimentation, to try to document the natural world. Over time we have been accumulating information and knowledge. Despite centuries of research there are still many things that we do not know about the world. The difference between the "known" and the "existing" universe has been defined as a knowledge shortfall (Hortal *et al.*, 2015). In terms of biodiversity, different aspects remain to be known, from the scale of species to that of communities and at different levels of complexity. Understanding what remains to be known and what is the magnitude of these deficiencies is the initial step to establish the routes that guide the way to reduce said deficiencies.

To understand the types and magnitudes of the knowledge shortfall, seven types have been defined (Hortal *et al.*, 2015): 1. Linnean Shortfall (species yet to be sampled and species not yet described); 2. Wallacean shortfall (species' geographic distributions); 3. Prestonian Shortfall (population dynamics); 4. Darwinian shortfall (evolutionary relationships); 5. Raunkiaeran shortfall (functional traits and ecological functions); 6. Hutchinsonian shortfall (responses and tolerances of species to abiotic conditions); 7. Eltonian shortfall (biotic interactions).

Overall, the knowledge shortfall in biodiversity research represents a significant challenge for understanding the ecological and evolutionary processes that shape biodiversity and for developing effective conservation strategies. To address these deficiencies, researchers need to prioritize sampling efforts and focus on collecting high-

quality data on species identity, distribution, population dynamics, functional traits, ecological functions, abiotic responses, and biotic interactions. Additionally, the integration of different sources of data, such as genetic, ecological, and spatial, can help to overcome some of the knowledge gaps and provide a more comprehensive understanding of biodiversity.

Information quality

The digital era has facilitated the migration of a vast amount of physical information from museums and biological collections to databases, aiming to make it accessible and easy to analyze (Meyer *et al.*, 2016). However, a significant portion of this information, ranging from 60 to 80%, is discarded due to quality issues that render it unusable (Rowe, 2005; Feeley, 2012; Veiga *et al.*, 2017; Daru *et al.*, 2018). The term "quality of information" refers to the "suitability" or "potential usefulness" of data for specific purposes. Data quality can be evaluated through two methods: 1) Quantitative, which assesses the data's suitability level, and 2) Qualitative, which determines whether the dataset is suitable for a particular use (Veiga *et al.*, 2017). The usefulness of information is influenced by multiple factors, such as accessibility, accuracy, sufficiency, relevance, detail, readability, and interpretation.

Studies based on plant records, available on online repositories, consistently indicate that the most significant quality problems are related to the geographic and taxonomic aspects of the records. A high percentage of records available in databases lack coordinates, or their precision is low. The degree of spatial grouping bias in collection locations has been incorporated as an essential element of geographic quality (García Márquez *et al.*, 2012). At the taxonomic level, a considerable number of specimens are

identified only at the family, genus, or left as indeterminate, making it challenging to use these records in studies that require higher taxonomic precision (Feeley & Silman, 2010; Feeley, 2015; Stropp *et al.*, 2016).

Gaps and biases in the biological information available in databases and their implications for the study of biodiversity

Biological collections play a crucial role in the study of biodiversity and its dynamics, as well as in providing information on diseases, pathogens, contaminants, and other social and economic aspects (Suarez & Tsutsui, 2004; Lister *et al.*, 2011). These collections store vast amounts of information on species diversity, distribution, and habitat, and are organized and systematized for future use. However, as technology advances, this information is migrating from physical formats to electronic databases, providing standardized, readily-analyzed information accessible through open-access databases such as GBIF (Veiga *et al.*, 2017).

Open access to biological data is fundamental to the study of biodiversity and has led to significant discoveries, predictions, and analyses of natural phenomena on a global scale (Goodwin *et al.*, 2015; James *et al.*, 2018). Currently there are numerous resources that allow access to biological data, the most popular being GBIF, a global coverage tool that gathers information from different taxonomic groups. There are also resources with regional information (e.g., European Natural History Specimen Information Network). and focused on particular taxonomic groups (e.g., Mammal Networked Information System; MANIS) (Graham *et al.*, 2004). This information has been used in multiple ways, from studies in biogeography that involve aspects of diversity (Cardoso *et al.*, 2017) and richness patterns (Ballesteros-Mejia *et al.*, 2013; Sousa-Baena, Garcia, & Townsend Peterson,

2014), to species distribution models (Feeley & Silman, 2011a; Ramirez-Villegas *et al.*, 2014), and biotic regionalization (Londoño-Murcia, M.C., González, I. & Bello, 2014; Serrano *et al.*, 2018), among others. From the conservation point of view, these data have been a basic tool for the design of conservation strategies (Ramirez-Villegas *et al.*, 2014), species risk analysis (Ramirez-Villegas, Jarvis, & Touval, 2012), biological invasions (Graham *et al.*, 2004), transformation analysis under climate change scenarios (Feeley, Stroud, & Perez, 2017), risk of species extinction (Panter *et al.,* 2020), etc.

However, the reliability and quality of available information in databases have raised concerns about geographic coverage, environmental representativeness, and taxonomic accuracy. Underrepresentation of environmental amplitude of records has led to an underestimation of species distribution ranges, while sampling effort has been found to be highly correlated with observed species richness (Hortal *et al.* , 2007; Feeley & Silman, 2011b). Biases in records and gaps in geographic and environmental data can distort results and decrease the predictive power of models (Boakes *et al.*, 2010; García Márquez *et al.*, 2012). Taxonomic issues such as misspelling, variations in writing, various classification systems, and unpublished names also affect diversity estimation and species conservation analysis (Meyer *et al.*, 2016; Cardoso *et al.*, 2017; Vogel *et al.*, 2017). Accurate taxonomic information is crucial for estimating species distribution and categorizing the threat of species and their risk factors.

Sources of bias and uncertainty in the biodiversity information available in databases

The lack of systematic sampling efforts, researchers' preferences, and ease of access to work areas have been associated with bias in biodiversity information. Some regions have greater sampling efforts, leading to a recurrent report of zones with more samples than

others, which suggests that many species are yet to be described (Sousa-Baena *et al.*, 2013; Meyer *et al.*, 2016; Oliveira *et al.*, 2016). This bias is linked to trends in record concentration around access roads, populated centers, and central level infrastructures (e.g., museums, herbaria), as reported in different regions of the planet (Reddy & Dávalos, 2003; Feeley & Silman, 2011a; Meyer *et al.*, 2016; Daru *et al.*, 2018). Meanwhile, gaps in the information registry are associated with difficult access areas (Daru *et al.*, 2018). Additionally, there is a preference towards "charismatic" groups such as orchids and palms, leading to a greater number of specimens in collections, as opposed to less conspicuous groups like mosses and liverworts (Troudet *et al.*, 2017; Daru *et al.*, 2018).

Uncertainty in biodiversity information stems from the lack of accuracy in measurements and predictions, as well as the lack of knowledge about a phenomenon (Hortal *et al.*, 2015). Taxonomic uncertainty is linked to misidentifications, synonyms, and spelling errors in assigned names, while geographic uncertainty relates to the imprecise georeferencing of collection locations (Goodwin *et al.*, 2015; Maldonado *et al.*, 2015; Meyer *et al.*, 2016). Due to the recent integration of geographic information systems to biodiversity studies, many collections lack geographic coordinates on their labels (Feeley & Silman, 2010). Hence, several authors emphasize the evaluation of information before constructing models and predictive maps of species distribution (Hortal *et al.*, 2008; García Márquez *et al.*, 2012; Syfert *et al.*, 2013).

Considering the challenges of biodiversity databases, it is essential to understand the limitations of the knowledge generated from the available data. This understanding can inform corrective measures in the practices and procedures of biological collections to produce reliable and sufficient information.

Diversity in Colombia: Gaps and biases in biological information in Colombia

Colombia is the second most biodiverse country on Earth. The Andean Region is home to the majority of the country's flora, with 18,491 species accounting for 75.5% of the total plant species. Additionally, the region is also where the highest number of endemic species (84.1%) are found (Bernal *et al.*, 2016; Moreno, *et al.*, 2017). The country's biological wealth is supported by almost six million biological records deposited in national and foreign collections and museums. However, only about half of them have been systematized, and 1,062,373 records are available in open access databases (Escobar *et al.*, 2016).

The Andean Region has the highest density and coverage of records, but the availability of data at the local level is still limited. The concentration of records is in the departments of Antioquia, Valle del Cauca, and Cundinamarca, which account for 37% of the biological records of the country. The middle and lower Andean orobiomes make up 42.14% of the records at the bioregional level (Londoño-Murcia *et al.*, 2014).

There is a strong geographic bias in the plant record in Colombia, which is related to the areas with the highest species richness (Sousa-Baena *et al.*, 2013; Mutke & Weigend, 2017). It is expected that the same pattern applies to the flora of the Colombian Andean Region. However, the biases and gaps in the records could also be related to the areas most affected by the armed conflict, such as Cauca, Antioquia, Valle, Caquetá, Nariño, Norte de Santander, Putumayo, Meta, and Huila.

On the other hand, the magnitude of the environmental and taxonomic biases of the digitally available information on plants from the Colombian Andes is less clear. Studies worldwide have shown preferences in the collection of records in areas with higher precipitation and temperature (Speed *et al.*, 2018), as well as towards particular biological

groups (Troudet *et al.*, 2017). In Colombia, the study of the flora has principally focused on páramos and rainforests, which may have led to information gaps in mid-altitude montane forests. There is also a taxonomic bias, with trees being the most collected and better cataloged. As a result, collections of families consisting mainly of trees have more complete data compared to herbs, shrubs, lianas, and epiphytes.

This dissertation therefore aims to establish and analyze the scope of the gaps, biases, as well as uncertainties at the taxonomic, geographic and environmental levels of the Colombian Andean flora available in databases. I also aim to analyze its implications in terms of the study of the patterns of richness of species in a floristically complex and diverse region.

In order to study the gaps and biases in the information available in databases for the Andean flora of Colombia, this research includes four chapters. Chapter one describes the information sources used, the data problems, and the data cleaning and standardization process. Chapter two uses the standardized data from chapter one to describe and analyze gaps and spatial and environmental biases in the record of the Colombian Andean flora and its implications in the estimation of wealth for the region. Chapter three studies the taxonomic and life forms representativeness of the region's flora. Finally, chapter four studies the importance of data curation in information retrieval and its potential usefulness to fill gaps using the Flora de Bogota project as a model.

References

Antonelli A, Sanmartín I. 2011. Why are there so many plant species in the Neotropics? *Taxon* 60: 403–414.

Bacon CD, Silvestro D, Jaramillo C, Smith BT, Chakrabarty P, Antonelli A. 2015. Biological evidence supports an early and complex emergence of the Isthmus of Panama. *Proceedings of the National Academy of Sciences* 112: 6110–6115.

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

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

Boakes EH, McGowan PJK, Fuller RA, Chang-Qing D, Clark NE, O'Connor K, Mace GM. 2010. Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology* 8.

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* 114: 10695–10700.

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.

Escobar, D., Gonzales, I., Amariles, D., Benítez, J.M., Londoño MC. 2016. Inventario de la biodiversidad de Colombia a nivel de especies. In: Gomez, M.F., L.A. , Andrade, G.I. y Rueda C, ed. *Biodiversidad 2015. Estado y tendencias de la biodiversidad continental de Colombia.* Bogotá D.C: Instituto Alexander von Humboldt. Bogotá, D.C., Colombia, 103.

Feeley KJ. 2012. Distributional migrations, expansions, and contractions of tropical plant species as revealed in dated herbarium records. *Global Change Biology* 18: 1335–1341.

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. 2010. Modelling the responses of Andean and Amazonian plant species to climate change: The effects of georeferencing errors and the importance of data filtering. *Journal of Biogeography* 37: 733–740.

Feeley KJ, Silman MR. 2011a. The data void in modelling current and future distributions of tropical species. *Global Change Biology* 17: 626–630.

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

Feeley KJ, Stroud JT, Perez TM. 2017. Most ‘ global ’ reviews of species ’ responses to climate change are not truly global. *Diversity and Distributions* 23: 231–234.

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.

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

Goodwin ZA, Harris DJ, Filer D, Wood JRII, Scotland RW. 2015. Widespread mistaken identity in tropical plant collections. *Current Biology* 25: R1066–R1067.

Graham A. 2009. the Andes: a Geological Overview From aGraham, A., 2009. the Andes: a Geological Overview From a Biological Perspective. Ann. Missouri Bot. Gard. 96, 371–385. https://doi.org/10.3417/2007146 Biological Perspective. *Annals of the Missouri Botanical Garden* 96: 371–385.

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT. 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution* 19: 497–503.

van der Hammen T. 2000. Aspectos de historia y ecología de la biodiversidad norandina y amazónica. *Rev Acad Colomb Cienc* 24: 231–245.

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, Jiménez-Valverde A, Gómez JF, Lobo JM, Baselga A. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* 117: 847–858.

Hortal J, Lobo JM, Jiménez-Valverde A. 2007. Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology* 21: 853–

863.

James SA, Soltis PS, Belbin L, Chapman AD, Nelson G, Paul DL, Collins M. 2018. Herbarium data: Global biodiversity and societal botanical needs for novel research: Global. *Applications in Plant Sciences* 6: 1–8.

Lister AM, Brooks SJ, Fenberg PB, Glover AG, James KE, Johnson KG, Michel E, Okamura B, Spencer M, Stewart JR, Todd JA, Valsami-Jones E, Young J. 2011. Natural history collections as sources of long-term datasets. *Trends in Ecology and Evolution* 26: 153–154.

Londoño-Murcia, M.C., González, I. & Bello LC. 2014. Registros biológicos en línea y vacíos de información. In: Bello, J.C., Báez, M., Gómez, M.F., Orrego, O. y Nagele L, ed. *Biodiversidad 2014: Reporte de estado de la biodiversidad en Colombia*. Bogotá D.C., Colombia.: Instituto Alexander von Humboldt, .

Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N, Antonelli A. 2015. Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases? *Global Ecology and Biogeography* 24: 973–984.

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

Moreno L.A., Rueda, C. y Andrade GI (Eds. . (Ed.). 2017. *Biodiversidad 2017. Estado y tendencias de la biodiversdiad continental de Colombia*. Instituto de Investigación de Recursos biológicos Alexander von Humboldt. Bogotá.D.C., Colombia.

Mutke J, Weigend M. 2017. Mesoscale patterns of plant diversity in Andean South America based on combined checklist and GBIF data. *Berichte 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.

Orme CDL, Davies RG, Burgess M, Eigenbrod F, Pickup N, Olson VA, Webster AJ, Ding TS, Rasmussen PC, Ridgely RS, Stattersfield AJ, Bennett PM, Blackburn TM, Gaston KJ, Owens IPF. 2005. Global hotspots of species richness are not congruent with endemism or threat. *Nature* 436: 1016–1019.

Panter CT, Clegg RL, Moat J, Bachman SP, Klitgård BB, White RL. 2020. To clean or not to clean: Cleaning open-source data improves extinction risk assessments for threatened plant species. *Conservation Science and Practice* 2: 1–14.

Parra C, Díaz S. 2016. *Herbarios y Jardines Botánicos : Testimonios de nuestra nuestra Biodiversidad*. Bogotá: Universidad Nacional de Colombia (sede Bogotá).

Ramirez-Villegas J, Cuesta F, Devenish C, Peralvo M, Jarvis A, Arnillas CA. 2014. Using species distributions models for designing conservation strategies of Tropical Andean biodiversity under climate change. *Journal for Nature Conservation* 22: 391–404.

Ramirez-Villegas J, Jarvis A, Touval J. 2012. Analysis of threats to South American flora and its implications for conservation. *Journal for Nature Conservation* 20: 337–348.

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

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

Serrano J, Richardson JE, Pennington TD, Cortes-B R, Cardenas D, Elliott A, Jimenez I. 2018. Biotic homogeneity of putative biogeographic units in the Neotropics: A test with Sapotaceae. *Diversity and Distributions* 24: 1121–1135.

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.

Sousa-Baena MS, Garcia LC, Townsend Peterson a. 2014. Knowledge behind conservation status decisions: Data basis for 'Data Deficient' Brazilian plant species. *Biological Conservation* 173: 80–89.

Speed JDM, Bendiksby M, Finstad AG, Hassel K, Kolstad AL, Prestø T. 2018. Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. *PLoS ONE* 13: 1–17.

Stropp J, Ladle RJ, Ana AC, Hortal J, Gaffuri J, H. Temperley W, Olav Skøien J, Mayaux P. 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa. *Global Ecology and Biogeography* 25: 1085–1096.

Suarez AV., Tsutsui ND. 2004. The Value of Museum Collections for Research and Society. *BioScience* 54: 66.

Syfert MM, Smith MJ, Coomes DA. 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLoS ONE* 8.

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.

Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D, Robertson TJ. 2017. A conceptual framework for quality assessment and management of biodiversity data. *PLoS ONE* 12: 1–20.

Vogel C, Bordignon SA de L, Trevisan R, Boldrini II. 2017. Implications of poor taxonomy in conservation. *Journal for Nature Conservation* 36: 10–13.

Chapter 1: Flo_RA: a Colombian Andean flora database

Introduction

The Andean region of Colombia is the most diverse in terms of flora, with 18,491 registered species (Bernal *et al.*, 2016). However, this region has the country's highest population density. It has been one of the most subjected to human transformation (Etter & van Wyngaarden, 2000) and many species and ecosystems are threatened (Etter *et al.*, 2015). The high floristic diversity in this region was initially investigated in 1783 with the botanical expedition of the *Nuevo Reino de Granada* led by Jose Celestino Mutis. Since then, floristic information in this region has been kept in national and international herbaria. Physical collections are not always easy to access, especially when many herbaria are involved. Only recently, some collections have been digitized, and biodiversity information is becoming more accessible. This provides excellent opportunities to study biodiversity at large scales and understand the challenges around our biodiversity knowledge. Even though more information is available through databases, we have yet to have a common framework to describe, evaluate and manage the biodiversity data quality. Therefore, it is difficult for researchers to compare and adjust the different sources of information and verify their utility (Veiga *et al.*, 2017). Among the most common problems associated with database biodiversity information are: 1. Multiplicity of formats derived from the different sources of consultation (e.g., several date and coordinates formats); 2. Digitization errors in the specimens; 3. Heterogeneous classification systems; 4. Different coordinate systems; 5. The multiplicity of forms to report dates, etc. (Stribling *et al.*, 2008).

The development of reliable and standardized consultation tools allows researchers and other users to invest less time in data processing and cleaning before the analyses. This also becomes the basic input to guide research on biodiversity and be used reliably for decision-making. These tools can also improve the power and scope of investigations due to the standardization of information, data correction and reduction of errors.

The database Flo_RA has therefore been designed to curate and standardize taxonomic, coordinate and date information of plant collections and records available in other databases. This tool is the first to integrate and systematize (so far I know) the record information of the Colombian flora, especially in the Andean Region.

Methods

Data

The information that is gathered in the Flo_RA database is composed of vascular and non-vascular plants of Colombia, available in online and public databases. I also included some smaller local collections.

Flo_RA is the most complete database in Colombia, with 2,279,519 collections. The most significant number of records was obtained from GBIF (Gbif.org, 2017; *Global Biodiversity Information Facility; data downloaded: March* 2016), which has worldwide species documentation and more than one million records. The second largest data source was the Colombian National Herbarium (COL, March 2017), with ca. 660,000 specimens, one of the country's largest and most complete botanical collections. It currently has about 80% of the collections systematized. A third source was the information obtained from

TROPICOS (Missouri Botanical Garden, April 2016), which contains ca. three million records and is the most extensive tropical flora in the World. Other sources included the Botanical Garden of Bogotá (JBB), with an up to date database of the region in and around Bogotá. This information was also complemented by collections done by the author (Table 1).

Table 1. List of sources consulted and origin of the data for the construction of the Flo_RA database

Source	**Region or area consulted**	**No. collections**	**Percent (%)**
GBIF	Colombia	1,697,974	74.5
COL	Andean Region Colombia	339,510	14.9
TROPICOS	Colombia	190,022	8.3
JBB	Bogotá - Colombia	46,053	2.0
Others	Northern Santander Colombia	5.960	0.3
TOTAL		**2,279,519**	**100**

Other sources of data

To complement the biological records, the Flo_RA database also has information such as:

- Taxonomy: Plant and lichen catalogue of Colombia (Bernal *et al.*, 2016), which was included for curation and standardization of the names of the record obtained from different sources.

- Environmental and Territory data: environmental information (temperature, precipitation), topography, elevation, biotic regions, ecosystems, as well as roads and cities, were also included in the database (Table 2).

Table 2. List of additional information layers incorporated into the Flo_RA database.

Information layer	Source	Name	Reference
Bioclimatic variables	WORLDCLIM	WORLDCLIM version 2, 30° resolution	http://worldclim.org/version2
Topographic and altitudinal indices	SRTM	Shuttle Radar Topography Mission	https://www.jpl.nasa.gov/
Ecosystems	IDEAM	*Instituto de hidrología, meteorología y estudios ambientales*	http://www.ideam.gov.co/geoportal
Biotic regions	IAvH	*Instituto de investigación de recursos biológicos Alexander von Humboldt*	http://datos.humboldt.org.co/
Roads and cities	DANE	National Administrative Department of Statistics	https://www.dane.gov.co/

Data processing

For managing the collected information, a database was built using free software. The PostgreSQL database engine (PostgreSQL 9.5.10 with PostGIS extension 2.4) and SQL were used as the language for database queries. This resource is highly efficient in handling large volumes of information regarding storage, processing and execution. The queries are also achieved through relatively simple computer routines. Another advantage lies in the

interoperability of the results, which can be directly analyzed with other data analysis languages such as "R" (R Development Core Team, 2019) and "QGIS" (QGIS Development Team, 2015). The database called "Flo_RA" (Flora of the Andean Region) is currently hosted in the HPC (high-performance computing) servers of the Universidad del Rosario.

The information in Table 1 required the application of different cleaning filters and data curation processes to consolidate a database with all the information on vascular (ferns and related plants, gymnosperms and angiosperms) and non-vascular plants (mosses and liverworts) from the Andean Region of Colombia. The curation and cleaning procedures were applied independently to the collected records' taxonomic, geographic and temporal components.

Geographic component

All the georeferences were cleaned and standardized using the following criteria:

- Records with coordinates: Selection of records with coordinate information. Due to the different sources used (e.g., GBIF, COL, etc.), identifying the fields containing this information was important. They are not always codified in the same manner.
- Typographic correction: Procedures to solve issues related to symbols such as commas (,), points (.), quotations (“ ”), apostrophes (‘), and degrees symbols (°), among others.
- Coordinate system standardization: Once corrected and cleaned, the record coordinate data were standardized to the "Geometry point" type format of the PostGIS system under the WGS84 geographic coordinate system.

Taxonomic component

Taxonomic information was cleaned, curated and standardized using the following procedure:

- Cleaning: In this process, I checked and excluded characters such as commas, spaces, points, etc., symbols that is not part of valid taxonomic names, mistakes that is committed in the digitalization process. I corrected those errors using regular sequences to search and correct mistakes.
- Curating and standardization: once cleaned, the data were subjected to curating by:
    - Name validation. Process in which indeterminate records and those belonging to groups other than vascular and non-vascular plants were excluded. Subsequently, misspelt names were identified and corrected using regular sequences and tools that approach the possible names given list of valid names supplied by Catálogo de plantas de Colombia (Bernal *et al.*, 2016) Finally, unpublished names were identified and reassigned to the higher taxonomic level.
    - Synonym identification and unification. This process was conducted within each taxonomic category (order, family, genera, species) following the Catalogue of plants and lichens of Colombia (Bernal *et al.*, 2016). If a name could not be validated with the Catalogue, The Name Resolution Service (TNRS) was used. This procedure was implemented at a massive scale using the R package Taxize (Scott *et al.*, 2018). Records with infra-specific classifications were treated as synonyms at the species level.

- Hierarchical correspondence of the names assigned to the records. In this step, the upper level to species were obtained from the Catálogo de plantas de Colombia (Bernal *et al.*, 2016). If a name could not be validated with the Catalogue, The Name Resolution Service (TNRS) was used. This procedure was implemented at a massive scale using the R package Taxize (Scott *et al.*, 2018).

Temporal component

The dates of the records were also curated. This was done as follows:

- Record exclusion: records with no collection date or dates before 1783 (when the Royal Botanical Expedition of the New Kingdom of Granada occurred) were excluded.
- Cleaning of symbols and characters: Characters such as dots, spaces, commas, and the field corresponding to the collection dates were curated.
- Unification of the date format: Records contained many different formats to report dates (e.g., month/day/year vs day/month/year), as well as ways to write months (e.g., May, V, mayo). Therefore, there was a unification of the date format using regular expressions. To this end, I used several tools available in SQL.
- Standardization of collection date. Once the collection dates were corrected, I used the "date_range" format in SQL as the unifying format.

Dealing with duplicated records

With a clean and curated set of records, I then identified duplicated records. This identification was conducted in three steps. Records were considered duplicates when: 1.

Records shared the same accepted name; 2. Records with collection dates differing by one day. 3. Records with the same georeference. The same set of coordinates was identified as equal when the distance between both sets was less or equal to 100 m.

Results

The most complete, curated and standardized database for the Colombian Andean Flora has been built based on database mining. Flo_RA has 2,279,519 collections equivalent to 33.950 species and 473 families of vascular and non-vascular plants. 70% of the collections available in Flo_RA are georeferenced in Colombia, and only 34% are within the Andean Region. At the taxonomic level, 97% of the collections within the Andean Region are determined to family and 75% to species (Table 3).

Table 3. Number of records obtained after data curating in the database Flo-RA.

	Number	**% georeferenced records - Andes**	**Number families**	**Number species**
Obtained records	2,219,519	N/A	473	33,950
Records with coordinates in Colombia	1,562,276	N/A	455	27,052
Records with coordinates in the Andean region	759,551	100%	422	19,464

Records at the family level within the Andean region	739,911	97%	422	19,441
Records at the species level within the Andean region	575,004	76%	390	19,463

Flo_RA database structure

In addition to the occurrence information of vascular and non-vascular plants in Colombia and the Andean Region, this database also has updated geographic and environmental data (Table 2). At the taxonomic and nomenclatural level, it incorporates updated information on the names of species registered for the country (Bernal *et al.*, 2016). It also has new tables that result from the geographic, taxonomic and temporal transformation of the information collected initially (Figure 1).

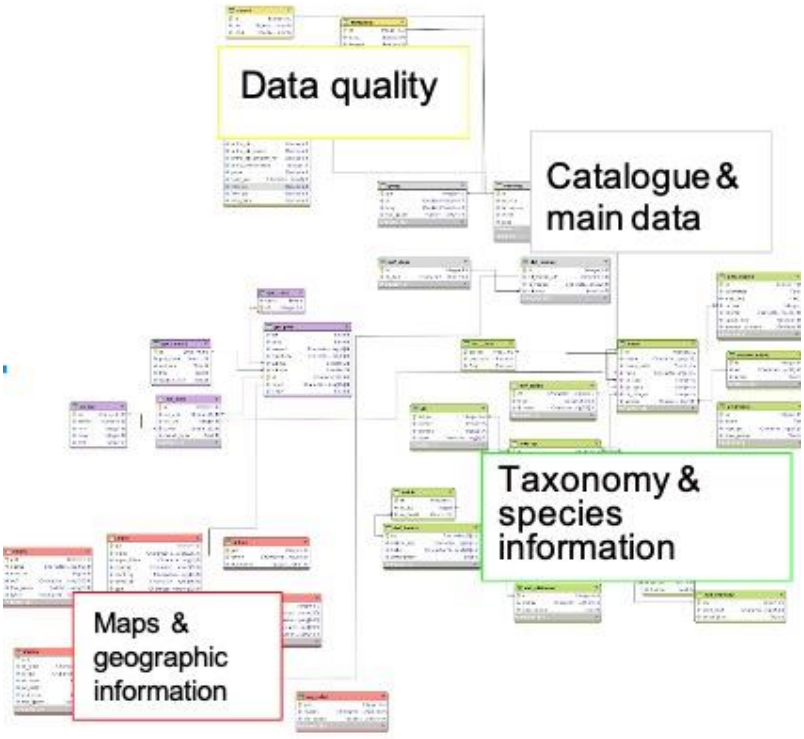


Figure 1. General scheme of the Flora Database. SQL language with Postgis GIS system.

Clean tables

Once the data was compiled and revised, new information tables were created. These tables are labelled as “cleaned data” in the Flo_RA database (Table 4).

Table 4. Description of the clean information from the Flo_RA database.

Information	Description
Geographic	Coordinates of all records corrected and standardized to the "Geometry point" type format of the PostGIS system under the WGS84 geographic coordinate system.
Taxonomic	Updated names of taxa entered into the Flo_RA database (synonyms, classification systems) according to the Colombian Plant Catalog and TNRS.
Temporal	Collection dates are standardized to the “date_range” format of the SQL system.

Discussion

The availability of a large amount of biological data in open databases has allowed the study and consolidation of information on different aspects of biodiversity at a local, regional and global scale (Soberón & Peterson, 2004; Anderson *et al.*, 2020). There are multiple platforms with biodiversity data worldwide (e.g., GBIF, BIEN), but also from

local collections (e.g., in Colombia COL, UDBC, SINCHI). Given the multiple sources of biological information available, there are also numerous formats in which the information is reported, causing difficulties in consolidating biological data for analysis, making the data cleaning and normalization process problematic and time-consuming (Dalcin, 2005; Anderson *et al.*, 2020).

Another problem with open databases, especially those that receive data from multiple sources (e.g., GBIF, BIEN), is the lack or low verification of the information received, which often depends on the source (Franklin *et al.*, 2017). Not to mention that a good part of the records come from observations, information that is difficult to verify. Among the most frequent problems are: 1. Nomenclatural problems derived from the validity of the names assigned to the records. The most common being the high proportion of synonyms (e.g., species, family) that are not constantly updated by the databases; 2. The geographic information provided from the records that in many cases does not exist given the age of the records deposited in the collections; low geographic precision (e.g., records assigned to centroids during the georeferencing process, coordinates that do not match the collection location description; coordinates located in bodies of water); coordinates in different formats? (e.g., geographic, planar coordinates) (Dalcin, 2005; Dalcin *et al.*, 2012).

The Flo_RA database is a resource that consolidates the information on plants from the Colombian Andean region available in different databases (e.g., GBIF, Tropicos, COL, JBB) and makes it open in a standardized way. The database not only gathers data from plant records but also includes environmental information (e.g., mean annual temperature and annual precipitation) (Karger *et al.*, 2017), elevation data (elevation data from the Shuttle Radar Topography Mission Global Elevation Model at 90-m resolution

(http://srtm.csi.cgiar.org) and taxonomic information (Bernal *et al.*, 2016) in separate but linked tables. The Flo-RA database structure allows for the creation of new tables with standardized information from structured database queries.

On the other hand, the processing, validation and standardization of the taxonomic, geographic and temporal data made to the records allow potential users to quickly access the standardized data of plants from the Andes of Colombia for the development of any analysis (e.g., biogeographic, floristic, species categorization, phenology, climate change). The taxonomic information of the plant records of the Andes of Colombia found in the Flo-RA database is standardized according to the Colombian plant catalogue (Bernal *et al.*, 2016) that follows the APG III classification system (Angiosperm Phylogeny Group III, 2009).

The Flo-RA database is designed to allow easy incorporation of new information from new records or the curation of existing ones with the potential of its scope to be widened to other areas of Colombia and the region. However, this database is intended to supplement, and not replace, consolidated institutional efforts at the national level (e.g., SIB - Colombia); instead, it can contribute to curated and standardized information for users who require it. As seen, a high proportion of the records included in the Flo-RA database have a limited scope in the analyses due to missing information from the records (e.g., taxonomic, geographic, temporal).

References

Anderson RP, Araújo MB, Guisan A, Lobo JM, Martínez-Meyer E, Peterson AT, Soberón JM. 2020. Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. *Frontiers of Biogeography* In press: 0–14.

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

Dalcin EC. 2005. Data Quality Concepts and Techniques Applied to Taxonomic Databases. *Life Sciences*: 266.

Dalcin EC, da Silva LAE, Cabanillas CC, Loures MGSM, Monteiro VF, da Silva GZ, de Souza JM. 2012. Data Quality Assessment at the Rio de Janeiro Botanical Garden Herbarium Database and Considerations for Data Quality Improvement. *8 th International Conference on Ecological Informatics*: 3–7.

Etter A, Andrade A, Amaya P, Arevalo P. 2015. Estado de los ecosistemas colombianos-2014: Una aplicacion de la metodologia de lista roja de ecosistemas. : 108.

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.

Franklin J, Serra-Diaz JM, Syphard AD, Regan HM. 2017. Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography* 26: 6–17.

Gbif.org. 2017. GBIF Occurrence Download.

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).

Scott A, Szoecs E, Foster Z, Boettiger C, Ram K, Baumgartner J, Donnell JO, Marchand P. 2018. Package ‘ taxize ’.

Soberón J, Peterson AT. 2004. Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B: Biological Sciences* 359: 689–698.

Stribling JB, Pavlik KL, Holdsworth SM, Leppo EW. 2008. Data quality, performance, and uncertainty in taxonomic identification for biological assessments. *Journal of the North American Benthological Society* 27: 906–919.

Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D, Robertson TJ. 2017. A conceptual framework for quality assessment and management of biodiversity data. *PLoS ONE* 12: 1–20.

Chapter 2: Environmental and geographic biases in plant specimen data from the Colombian Andes

This chapter corresponds to the accepted version of the manuscript.

Paper published: *Botanical Journal of the Linnean Society* (December 2022)

DOI: doi.org/10.1093/botlinnean/boac035

IF:2.828

Abstract

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 to understand and account 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, while mid and lowlands remain poorly sampled (i.e., montane and lowland forest). Sampling completeness (SC) median across the Colombian Andes is lower than 75% at the scales studied. We explore possible causes of sampling bias and identify critical gaps and priority areas for plant sampling, and make recommendations for strategies to increase SC and reduce biases.

Keywords: Collecting bias, Colombia herbarium specimens, Flora, 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, reaching back to the XIV century (Thiers, 2020). Today these biological data have become available through online data aggregators, such as the Global Biodiversity Information Facility (GBIF) which includes ≥ 333 million plant occurrence records (GBIF www.gbif.org, accessed 22th Feb 2021) and herbaria that have digitalized 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, several gaps and biases have been detected in datasets, particularly 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 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 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 at a particular location. 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, Borges, & Gaspar, 2006; González-Oreja *et al.*, 2010; Gotelli & Chao, 2013; Engemann *et al.*, 2015). These estimators have been helpful for 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 data availability 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 world's hotspots due to its 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 its flora's digitally available specimen records have been little explored. Low scale analysis (grid cells size 100 x100 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 over the last sixty years. An 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, conservation planning and to formulate a strategic plan to fill the area's 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 Colombian Andean region's environmental variability and spatial variability? (2) Are the plant records aggregated biased 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 (Cordillera Occidental, Cordillera Central, and the Cordillera Oriental) and the valleys of Cauca and Magdalena Rivers, with a lower limit of the area set at 445 m above sea level. Additionally, we included the Sierra Nevada de Santa Marta, an isolated mountain range located to the north of Colombia where the study area reaches its highest altitude of 5659 m. Thus, the study area (Fig. 1) comprises 306,729 km^2, characterized by high climatic variability. Annual average temperature varies from 24-32°C in the lowlands to as low as -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).

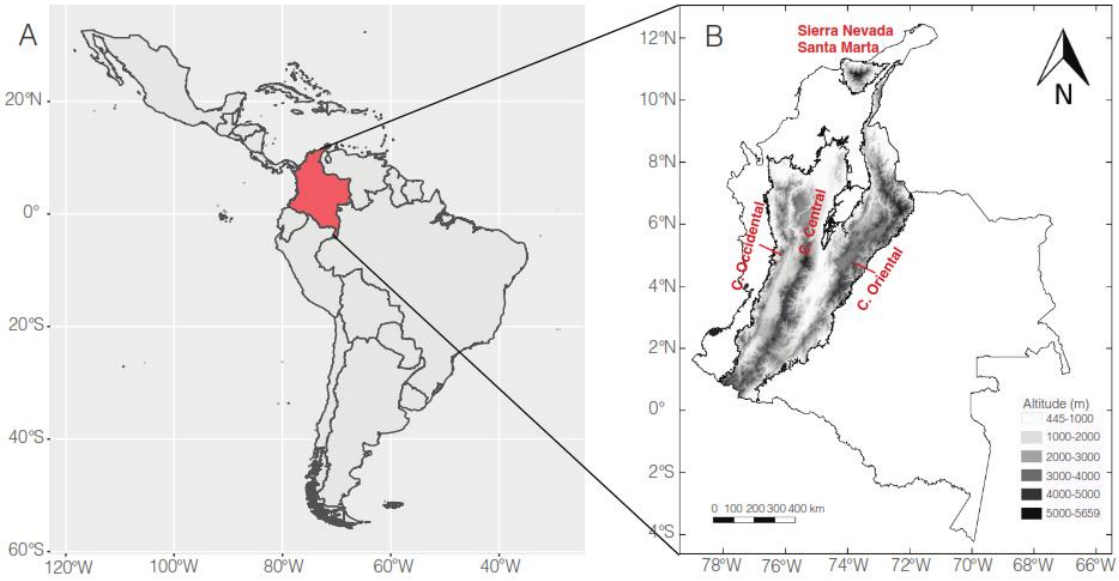


Figure 1. Ubication of the Colombian Andes. A, Colombia (red) within South America. B, Colombian Andean region (grey) corresponds to the country's montane area, including the elevations between 445 to 5659 m. “C.” is the abbreviation for Cordillera.

Specimen database

All records for plant species (kingdom Plantae) occurring in Colombia were downloaded from online databases and databases of herbaria with records of Colombian flora: (1) GBIF (accessed on the 30th 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 (Table S1).

The data were cleaned to increase the accuracy of taxonomic and geographic 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 (TNRS). After cleaning the database, we retained records with coordinates that fall in the Colombian Andean region and checked for duplicates leaving a single record for each collection in our analyses. Given different formats of 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). 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. A total of 88.2% of the records were discarded due to geographic issues such as lack of coordinates, low-precision coordinates (e.g., record with coordinated at degree resolution), or plant records outside the study area (66.5%), followed by duplicates (13.6%) and 8.1% because of incorrect species names (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), as well as 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 x 100 km, 50 x 50 km, 20 x 20 km, 10 x10 km and 5 x 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 analyzed by calculating the Moran index of spatial autocorrelation to determine if the plant collection patterns were aggregated, random or dispersed. We calculated Moran index using the "spdep" package (Bivand, Pebesma, & Gómez-Rubio, 2013) in R 3.6.1 (R 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:

$$\mathrm{Bias(d)} = \mathrm{nd} - \frac{\mathrm{pdN}}{\sqrt{\mathrm{pd}(1-\mathrm{pd})\mathrm{N}}}$$

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 the above equation is derived from the normal approximation of a binomial distribution, values are statistically significant when they are greater or less than 1.64 and -1.64, respectively (at $\alpha = 0.05$). Areas with values $\geq$ 1.64 are interpreted as over-sampled (i.e., more sampled localities than expected from a random sampling design), and areas $\leq$ -1.64 as under-sampled (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 based on 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.

Lastly, sampling completeness (SC) 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 and Jost 2012), where coverage is defined as the total relative abundance of the observed species in the sample, ranging from 0 to 1. Sample completeness was estimated using iNEXT R package (Hsieh, Ma, & Chao, 2016).

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. 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 x 100 km, 50 to 50 x 50, 20 to 20 x 20 km, 10 to 10 x 10 km and 5 to 5 x 5 km.

Scale (km)	# cells Andean region	# Cells with plant records	# 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	11047	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 grades (°), elevation in meters (m), temperature in degrees Celsius (°C), precipitation in millimeters (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 red, the only well sampled interval.

Variable	Interval	Random	% Random	Observed	% Observed	Bias	Sampling
Latitude	0-2	26445.35	9.92	25139	9.43	-8.5	undersampled

Variable	Interval	Random	% Random	Observed	% Observed	Bias	Sampling
Latitude	2-4	59628.18	22.36	27259	10.22	-150.39	undersampled
Latitude	4-6	70755.57	26.54	119303	44.75	213	oversampled
Latitude	6-8	76913.23	28.85	80007	30.01	2.34	oversampled
Latitude	8-10	19999.25	7.50	3038	1.14	-124.75	undersampled
Latitude	10-12	12883.42	4.83	5879	2.20	-63.18	undersampled
Elevation	0-1000	91753.1	34.41	54426	20.41	-152	undersampled
Elevation	1000-2000	85305.56	31.99	73337	27.51	-49.6	undersampled
Elevation	2000-3000	61565.01	23.09	75187	28.20	62.4	oversampled
Elevation	3000-4000	26274.5	9.85	59071	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	19314.94	7.24	41420	15.53	164.9	oversampled
Temperature	10-15	53449.83	20.05	77481	29.06	116.1	oversampled
Temperature	15-20	74422.5	27.91	71192	26.70	-13.9	undersampled
Temperature	20-27.1	117766.8	44.17	72701	27.27	-175.6	undersampled
Precipitation	688-1177	15612.79	5.86	34306	12.87	154.28	oversampled

Variable	Interval	Random	% Random	Observed	% Observed	Bias	Sampling
Precipitation	1177-1666	58986.19	22.12	63504	23.82	21.02	oversampled
Precipitation	1666-2155	70060.11	26.28	62721	23.52	-32.37	undersampled
Precipitation	2155-2646	55221.77	20.71	55200	20.70	0.012	well sampled
Precipitation	2646-6963	65255.58	24.47	50717	19.02	-65.49	undersampled
Precipitation	6963-11281	1352.22	0.51	177	0.07	-32.03	undersampled

Results

Geographic 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 to 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 to 4 °N, which is an east to west direction, corresponding with the foothills of Caquetá and Meta, in 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 Cordillera Central (Córdoba, Sucre and Bolívar Departments), and the third to the north of the Cordillera Oriental, 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 under-sampling 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). Altitudinally, collection efforts were concentrated above 2000 m, with the highest magnitude in the range of 3000 to 4000 m. In contrast, localities below 2000 m were under-represented, particularly lowland forests (localities between 445 to 1000 m) (Table 2; Fig. 3A). Because of the negative correlation between temperature and elevation, the lower temperature regimes (from -5 to 15 °C) were over-represented by plant records in the database, while the higher temperatures (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 more than 2646 mm/year. Specimen records were under-represented in areas with high precipitation (above 1666 mm/year), located in the foothills to the west of the Cordillera Occidental, east of the Cordillera Oriental and north of the northwest Cordillera Central (Table 2; Fig. 3C).

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 analyzed include altitude, 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 x 100 km, 50 to 50 x 50, 20 to 20 x 20 km, 10 to 10 x 10 km and 5 to 5 x 5 km.

Scale (km)	Variable	min	1Q	median	mean	3Q	max
100	altitude	-3656.00	-919.00	-384.00	-605.12	-83.25	394.50
50	altitude	-2980.50	-518.62	-158.50	-306.12	0.75	859.00
20	altitude	-2666.50	-229.75	-19.00	-78.29	117.75	1235.00
10	altitude	-1419.00	-124.00	3.00	-1.60	144.63	1175.50
5	altitude	-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

Scale (km)	Variable	min	1Q	median	mean	3Q	max
5	temperature	-9.65	-0.60	0.00	-0.03	0.45	6.50

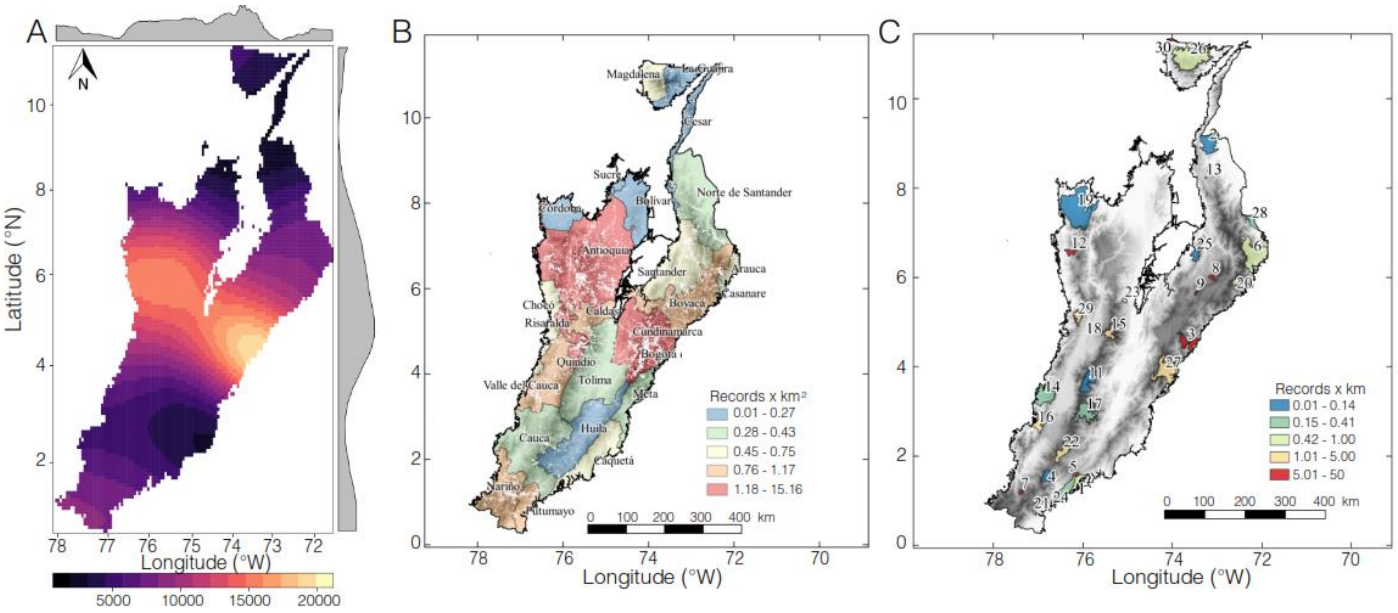


Figure 2. A, Plant record density across the Colombian Andean region. The grey area above the map indicates the longitudinal concentration of records, while the grey to the right, the latitudinal concentration. B, Variation in record number per km^2 for each of the Andean Departments of Colombia. The white areas correspond to the actual records. C, Collection pattern per km^2 in Protected Areas of the Colombian Andes. In grey the altitudinal variation is show, with darker colors 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.

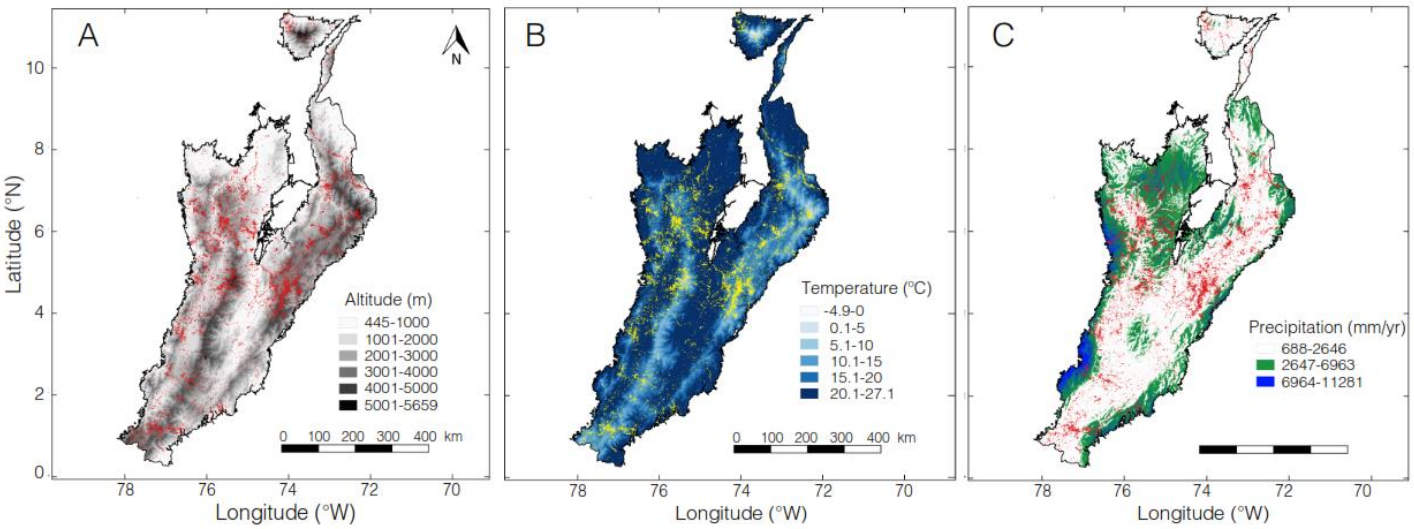


Figure 3. Spatial distribution of plant records (red dots) across the Colombian Andean Region in relation to: A, elevation (m). B, mean annual temperature (°C). C, annual precipitation (mm/year). The yellow and red dots correspond to the plant records collected in the area.

Coverage, density plant records representativity and scale effect

Decreasing resolution inflated SC and increased the number of plant records by grid cell. For example, while at low resolutions (e.g., cell size 100 x 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 x 5 km) dropped to 51%, and the median record number was 57 (Fig. S1; Table S1).

Despite this, plant records at a high resolution were better able to represent grid cells environmental variability than those at a low resolution. Meanwhile, the difference between grid cells and plant records was close to zero in grid cells of 5 x 5 km for the environmental variables considered (annual precipitation, mean annual temperature, altitude); the difference was maximum in grid cells of 100 x 100 km (Fig. 4; Table S3).

Completeness of Colombian Andean flora

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

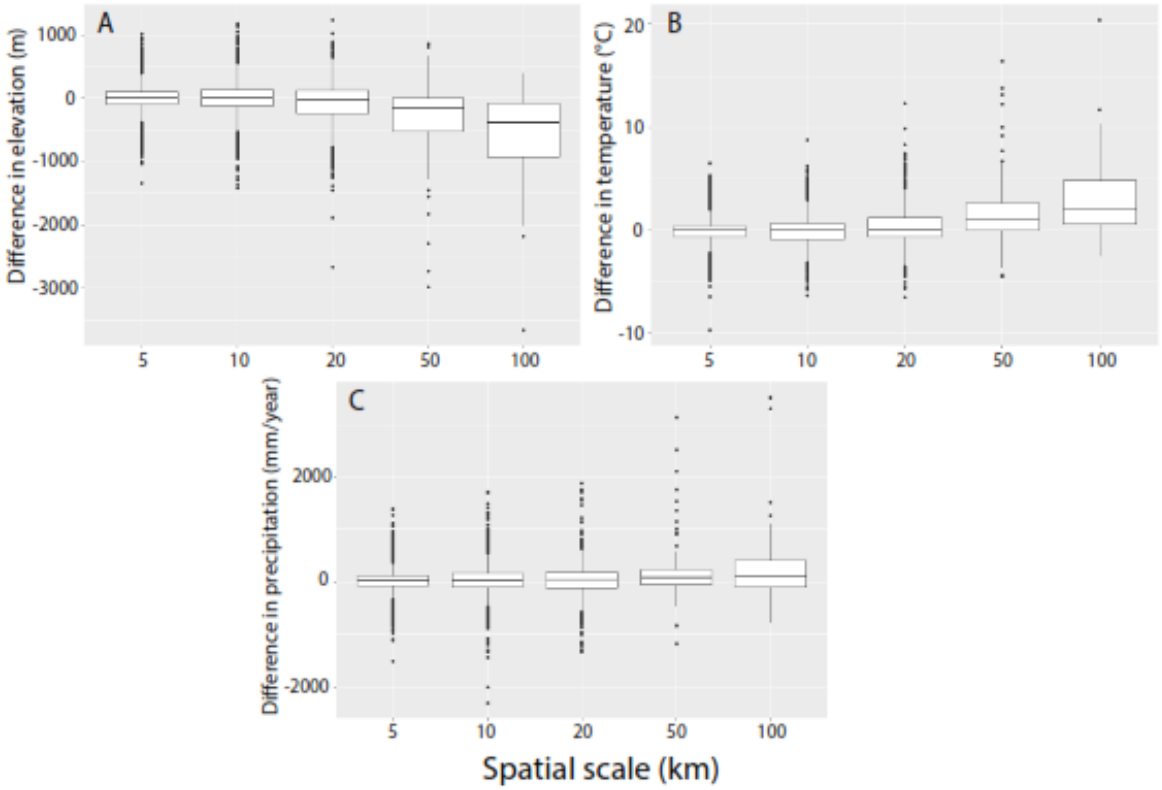


Figure 4. Scale effect (5 x 5, 10 x 10, 20 x 20, 50 x 50 and 100 x 100 km) on the environmental difference between specimen records and grid cells across Colombian Andes. Boxplots show the differences for A, Elevation (m). B, Mean annual temperature (°C). C, Mean annual precipitation (mm/year). The difference 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 x 100 km, 50 to 50 x 50, 20 to 20 x 20 km, 10 to 10 x 10 km and 5 to 5 x 5 km.

Discussion

Geographic bias and gaps

Our results showed strong geographic 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 to 43% of specimens to our database (Table S3). Three gaps (i.e., low specimen record density or no records; Figs. 2, 3) were located: the first in the northern part of the Cordillera Central; the second in the northern part of the Cordillera Oriental, including Serranía del Perijá and the Sierra Nevada de Santa Marta, and the third in Tolima, Huila and Cauca Departments, as well as the Cordillera Oriental's eastern foothills in the 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; Fig. 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; Yang, Ma, & Kreft, 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 physically access and thus less likely to have been collected. Other areas 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. Other 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; 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 (Table S4).

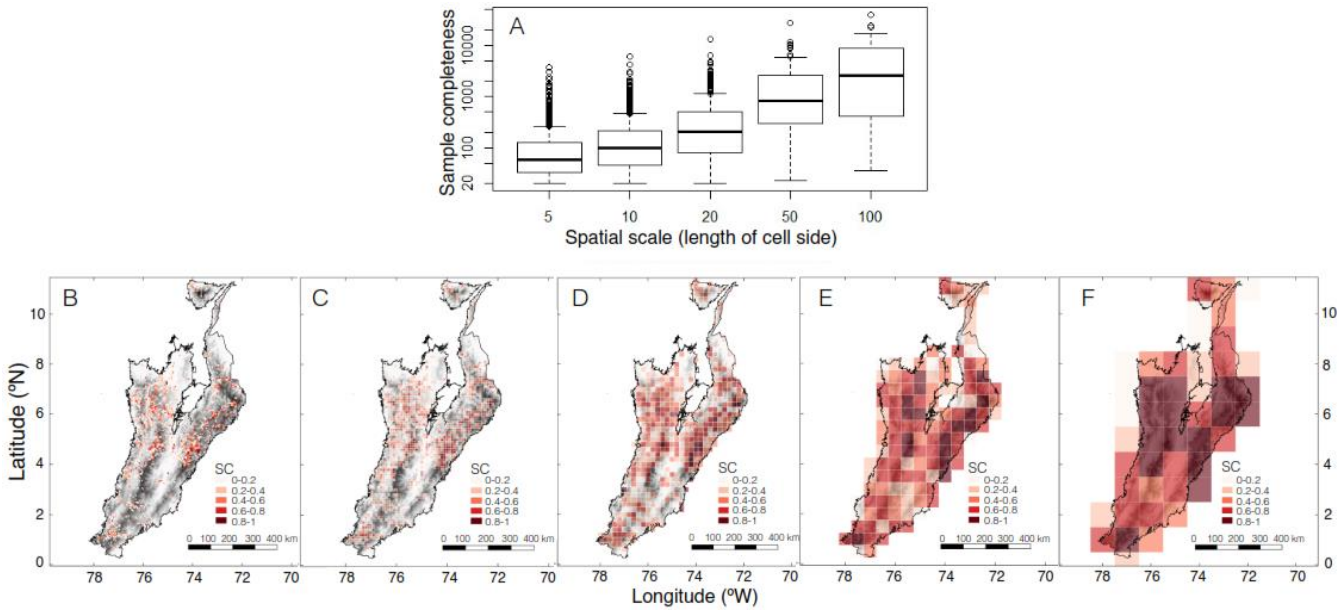


Figure 5. Variation in sampling completeness (SC) of plants at different spatial scales (5 x 5, 10 x 10, 20 x 20, 50 x 50 and 100 x 100 km). A, SC boxplot in the Colombian Andes at different spatial scales. The bottom and top part of the boxplot indicates the 25th and 75th percentile (respectively),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

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, while in montane areas the sampling effort has been focused on the highlands (Yang *et al.*, 2014; Engemann *et al.*, 2015). In the

Colombian Andes more samples were collected in areas above 2000 m, especially between 3000 to 4000 m elevation. 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 páramo's relevance 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 explain partly 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 altitude were under-sampled, even though the forest at around 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 Cordillera Occidental lowlands, the foothills of the Cordillera Oriental and the Magdalena and Cauca river valleys. Other biomes under-represented with restricted distribution in the Colombian Andes were the subxerofitic tropical biomes 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 species inventory's completeness 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 x 100 km) is about 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 x 100 km, the altitudinal 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 more than 50% of the Colombian Andean region lacking information.

The SC of cells with specimen records show that more than 60% of the plant diversity remains unregistered at scales of 20 x 20 km, 10 x 10 km and 5 x 5 km. In fact, less than 10% of the grid cells at 100 x 100 km and less than 1% of grid cells at 5 x 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. 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 were sharing collections or the same record was in multiple databases.

Data availability of Colombian Andean region

Plant occurrences from Colombia are scattered across different national and international herbaria, and only a part of them 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 where the Colombian flora is well represented (Table S1). As well as GBIF, the central repository of biodiversity records includes records from herbaria that we cannot directly access (Tabla S3). However, despite the number of plant occurrences gathered, important quality issues related to the data's geographic dimension made 88.2% of the Colombian data unusable. Together with institutional sharing policies, these issues make data access for biodiversity research difficult.

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, that 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 altitude, 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 digitalized come from three herbaria (COL, HUA, FMB; http://rnc.humboldt.org.co/admin/index.php/registros/colecciones, consulted October 2021) that are focused on Colombia's Flora. The small herbaria that are focused on regional floras

(e.g., Universidad de Pamplona [HECASA], Norte de Santander; Instituto Tecnológico 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 *et al.* (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 need to be focused on as, although the biodiversity is protected, the median sampling is 1 record/km^2 (Fig. 2C; 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.

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. *Biodivesidad 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* 114: 10695–10700.

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.

Distler T, Jørgensen PM, Graham A, Davidse G, Jiménez I. 2009. Determinats 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 Ltd., 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*. Nueva 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.

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: Second Edition*. New York: Elsevier Ltd., 195–211.

Gotelli NJ, Colwell RK. 2011. Estimating species richness. In: Magurran AE,, In: 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 of the United States of America* 103: 10334–10339.

Humboldt A, Bonpland A. 1807. *Essai sur la Géografie des Plantes*. Chez Levrault, Schoell.

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 I V, 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*. Brooklyn: 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 : Testimonios de nuestra 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* 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 Develpment 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.

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, Oregon: 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.

Chapter 3: Taxonomic collection biases in plant occurrence data in the Colombian Andes

Paper will be submitted to: *Plant Ecology and Diversity* or journal alike.

Abstract

The accelerated change in the ecosystems by human actions has driven a worldwide crisis of high biodiversity loss. Many species are lost before being discovered, and many others need better knowledge. Although the tropical flora is one of the most diverse, it is also little known. Through the implementation of representativeness analysis, we study the taxonomic gaps and biases at the phylum, family and species level for the Colombian Andean flora. We also analyze the representativeness of the Colombian Andean flora at the level of life forms and discuss implications for management and conservation. We found that 33% of species recognized for the Colombian Andes are not represented in the currently available occurrence databases and that 77% of species are poorly represented with less than 20 records. The best represented phyla were Magnoliosida, Pteridophyta and Bryophyta in contrast to Marchantiophyta, Anthocerophyta and Cycadopsida, Asteraceae, Melastomataceae were the best represented families while Orchidaceae, and Bromeliaceae were underrepresented. In addition, we found a high correlation between plant family representativeness and sampling effort. An analysis of life forms showed that epiphytes are more poorly represented than expected, based on their species diversity. Trees, shrubs and herbs were overrepresented. We suggest that future collection efforts should acknowledge these biases and aim to focus on underrepresented plant families and growth forms. This would provide more comprehensive and less biased information on plant species diversity and distribution, which is crucial for conservation efforts.

Introduction

The accelerated change in the ecosystems by human actions has driven a worldwide crisis of high biodiversity loss. Many species are lost before being discovered, and many others need better knowledge. Hortal *et al.* (2015) described various knowledge gaps including the Linnean shortfall for species taxonomy; the Wallacean shortfall for species distributions; the Prestonian shortfall for species abundance; and the Darwinian shortfall for evolutionary patterns. While many knowledge gaps exist, the current biodiversity data therefore, has several known biases. One of the strongest biases is spatial. Our taxonomic knowledge is concentrated towards temperate regions, where biodiversity has a long history of being recorded, in contrast to tropical regions where documentation is still at a relatively early stage (Meyer, Weigelt, & Kreft, 2016a; Lagomarsino & Frost, 2020; Vargas *et al.*, 2022).

Taxonomic knowledge has also been argued to be biased by societal preferences, given that research and investment can be focused in particular taxonomic groups (Troudet *et al.*, 2017). These societal preferences vary across temperate and tropical areas, and have distinct causes. Taxonomic biases in temperate regions seem to be associated with a long history of explorations and differential rates of societal and cultural development (Brummitt, Araújo, & Harris, 2021). In contrast, taxonomic biases in tropical regions seem to be associated with individual collectors. This is likely because our current knowledge on the vast tropical floras is mostly based on collections that were made in the last 50 to 100 years by a relatively few collectors. Their individual interests and collecting behavior still shapes our understanding of diversity in many cases (Lagomarsino and Frost 2020; Meyer *et al.*, 2016a). Collector behavior resulting in an overrepresentation of some taxa in

specimen collections has been demonstrated for Australia's flora (Haque *et al.*, 2017; Haque, Beaumont, & Nipperess, 2020).

Floristic data published in open databases in recent years has enabled a new wave of studies of plant diversity patterns at continental and regional scales (Pérez-Escobar *et al.*, 2022; Ulloa *et al.*, 2017). These studies estimate that of the 124,993 plant species known to exist in the Americas, occurrence data exist for 23% (28,691) Andean species in open databases (Pérez-Escobar *et al.*, 2022). Within Colombia, the Andes includes 18,494 species, which represents 14% of the plants known from the Americas and 33% of them are represented in open databases.

The Andes is a topographically complex 7,000 km long mountain chain that runs along the western side of South America. There are a high variety of habitats, from lowland rainforest through to high elevation páramo, that have been configured by a geological history of non-continuous rises across space and time. Given the differences in the geological genesis of the Andes, the cordillera has been divided into three sub regions with exceptional biodiversity endemic to each (Graham, 2009; Josse, Cuesta, & Navarro, 2011; Pérez-Escobar *et al.*, 2022): (1) the Southern Andes, including Argentina and Chile; (2) the Central Andes ,including Peru and Bolivia; and (3) the Northern Andes, including Venezuela, Colombia, and Ecuador.

The northern Andes is one of the 25 world biodiversity hotspots. A region highly threatened by heavy land use change (Myers *et al.*, 2000) and extremely vulnerable to climatic change, particularly the highland areas (Peyre *et al.*, 2020; Trew & Maclean, 2021). Around 63% of the northern Andes is placed in Colombia (Josse *et al.*, 2011; Vargas

et al., 2022) and includes the main cities of the country and hosts a populations of c. 31 million people (65% of Colombia's population) (https://www.dane.gov.co/).

Given the intense pressure the Andean flora is under, it is important to know the taxonomic bias for a highly biodiverse region in order to speed up the advance in knowledge in the light of accelerated land use transformation and to drive research and conservation policies. However, sampling in Colombia is challenging given the lack of economic support for this activity, and the lack of adequate infrastructure for herbaria collections.

Here we analyse the taxonomic representativeness of currently available data in open access databases of plant occurrence data for the Colombian Andes and ask: (1) How well is the Andean flora taxonomically represented in databases; and (2) What are the best and worst represented phyla, plant families, and life forms in the Colombian Andes? Our results show clear biases and we discuss our findings in light of their effect on our understanding of current diversity patterns in the Andean flora of Colombia.

Materials and Methods

Study area

The Colombian Andes includes lowland to highland biomes in an elevational range from 445 to 6,000 m. Rodríguez *et al.* (2006). We include the Sierra Nevada de Santa Marta, an isolated mountain in northern Colombia, in our area definition, as well as the Inter-Andean valleys of Magdalena and Cauca rivers, following Rodríguez *et al.* (2006) (Figure 1).

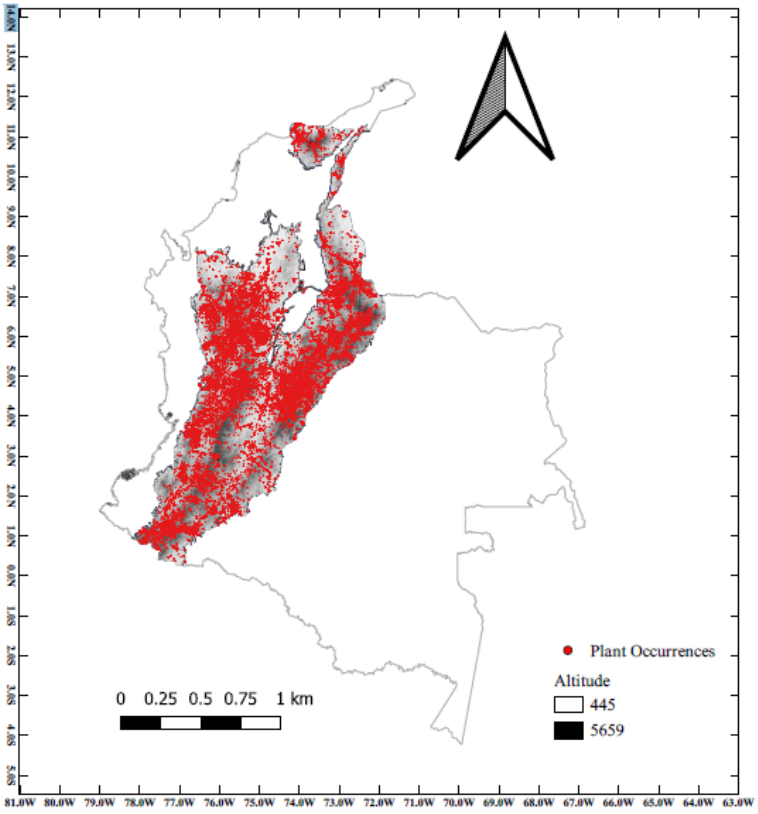


Figure 1. Map of the Colombian Andes and the plant occurrence data used. The map shows the elevational variation (grey scale) of the study area from 445 m to c. 6,000 m elevation, and the native plant occurrence records used (red dots).

Plant data

In order to characterize the taxonomic representativeness of currently available plant records for the Colombian Andes, we compared the plant species occurrences available in the Flo_RA dataset (Vargas *et al.*, 2022) with the official list of native plant species for the Colombian Andes available in Bernal *et al.* (2016) where we selected the list of plant species recorded for Andes region. The catalogue includes a total of 24,530 species of vascular and non-vascular plants. The Flo_RA dataset is the most reliable dataset for the Colombian Andes and was built with records available from open databases. Flo_RA has

266,625 plant records that represent 19,638 species. After a cleaning and standardization process (details in Vargas *et al.*, 2022), the dataset contains 218,399 plant occurrences that represent 11,581 native species in 345 families. Flo_RA also has information about life form, species origin and taxonomic hierarchy (phylum and family) based on the Catalogue of plants and lichens of Colombia (Bernal *et al.*, 2016).

Representativeness

We evaluated the representativeness of plants in the Colombian Andes based on taxonomic hierarchy (phylum and family), life form, and spatial distribution. For phyla, we considered four phyla within vascular plants (Magnoliopsida, Pteridophyta, Coniferopsida, Cycadopsida), and three for non-vascular plants (Bryophyta, Anthocerophyta, Marchantiophyta). All angiosperms families in APGIII were recognized, with a total of 420 native plant families found in the Colombian Andes (Bernal *et al.*, 2016; Vargas *et al.*, 2022).

Life form representativeness was analysed based on the categories used in the Flo_RA database (Vargas *et al.*, 2022; Bernal *et al.*, 2016). Due to the wide variety of terms in Bernal *et al.* (2016) we simplified the categorization into eight life forms, including aquatic herbs, climber (e.g., include herbaceous and woody), epiphytes, parasitic plants, saprophytic plants, terrestrial herbs, shrubs, and trees.

The representativeness analysis followed the proposal of Troudet *et al.* (2017) that estimates the ideal sampling of established groups based on the relationship of the richness (e.g., taxonomic, life form) and the record number as:

$$I = NBocc * (\frac{N}{Ntot})$$

where I is the record number expected given the number of species in a given group (e.g., taxonomic, life form), N is the number of native species sampled for each group, Ntot is the total number of native species, and NBocc is the number of occurrence across all groups. We then measured the taxonomic representativeness as a difference (O-I) and the ratio (O/I) between the observed (O) and expected specimens (I) for each group.

The representativeness analysis was complemented by the calculation of the proportion of species per family with more than one and 20 occurrences in the Flo_RA database. This threshold was established to measure the sampling state of species at the family level, to know the data sufficiency to perform models to estimate species distribution (Feeley & Silman, 2011a) and categorization of endangered species (Mace *et al.*, 2008) .

We analyzed the spatial distribution of taxonomic groups by counting the number of grid cells where phylum, families and species were recorded by plant occurrences. The spatial coverage of taxonomic hierarchy was studied at 20 km x 20 km grid cell resolution across the study area. We chose this scale because this size best represents the environmental variability of the geographic area and environmental information from plants records (Vargas *et al.*, 2022). We calculated the proportion of species present in more than one grid cell and 20 or more grid cells. We chose a threshold of 20 spatially distinct occurrences because it is a common threshold in niche modelling analyses (Feeley & Silman, 2011b).

All analyses were done in R (R Develpment Core Team, 2019) using the packages dplyr (Yarberry & Yarberry, 2021), ggplot2 (Wickham, 2011) and forcats (Wickham, 2021). Maps were generated using QGIS (QGIS Development Team, 2015).

Results

Taxonomic representativeness

82% of the families and 67% of species recognized in the Catalogue of Colombian plants and lichens are represented in the currently available plant occurrence dataset with coordinates in Flo_RA. The currently available occurrence data is biased towards vascular plants at the phylum level (with the exception of Cycadopsida), with less records available for non-vascular plants (with the exception of Bryophyta) (Table 1). For instance, Magnoliopsida had 60,309 records more than expected, while Marchantiophyta had 413 records less than expected.

Table 1. Number of occurrences observed (O) and expected (I) at phylum level, as well as the differences between the observed occurrences and expected occurrences (O-I). Positive numbers represent over sampled groups, while negative, under sampled.

Phylum	Occurrences (O)	Species Number	I	O-I
Magnoliopsida	180,418	9,496	120,109	60,309
Bryophyta	18,190	684	8,652	9,539
Pteridophyta	12,723	950	12,016	707
Coniferopsida	169	4	51	118
Cycadopsida	11	4	51	-40
Anthocerophyta	5	4	51	-46
Marchantiophyta	4,178	363	4,591	-413

At family level, the top ten that had the highest numbers of occurrences represented 40% of the data. These families are Asteraceae, Melastomataceae, Rubiaceae, Ericaceae, Solanaceae, Euphorbiaceae, Dicraniaceae, Fabaceae, Lauraceae, and Primulaceae (Fig. 2). In general, the most species rich families are those that have the highest number of occurrences, with some exceptions (Table S1). For instance, Poaceae, Piperaceae, and Bromeliaceae all have fewer occurrences than expected (Fig. 2; Table S2). In contrast, Ericaceae, Solanaceae and Euphorbiaceae are not in the top 10 most species rich families in the Colombian Andes but have some of the highest numbers of occurrences, even more than expected (Fig. 2; Table S2). In total, 60% of the families showed more records than expected in relation to their species richness, while 40% showed less than expected (Table S2). Asteraceae are the best represented family with the highest number of occurrence records and has an excess of 8,187 plant records. In contrast to Orchidaceae has 7,990 plant records less than expected (Fig. 2).

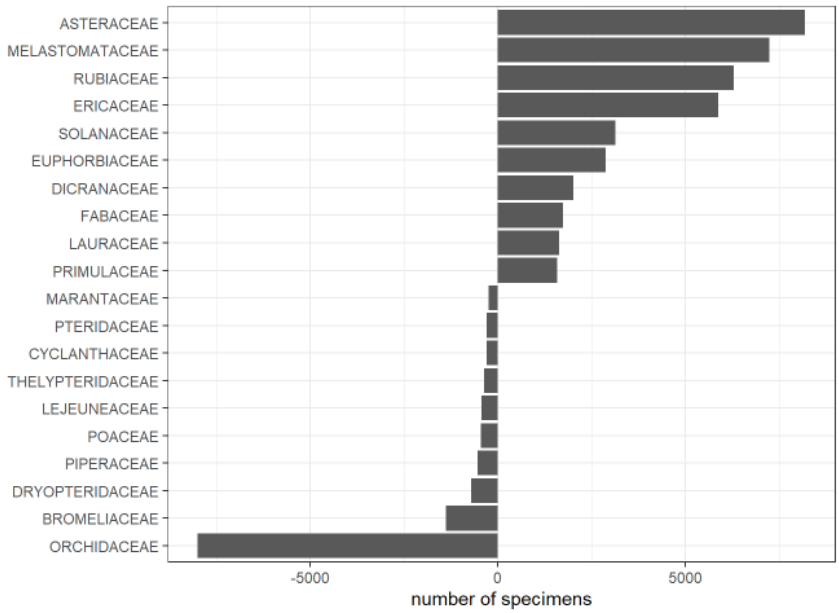


Figure 2. The top 10 overrepresented and underrepresented Andean plant families. Number of specimens refers to the number of occurrence records in excess or deficit per family, compared to the ideal number calculated based on Troudet *et al.* (2017). Table S2 has the complete dataset for all families.

A total of 72% (249) of the Andean plant families had species with at least 20 plant records, and 97% had species represented by more than one record. A total of 157 families had all of their species represented by more than one record and 31 families had all their species represented by at least 20 records (Fig. 3). However, although, the majority of families had overrepresentation of plant records, less than 50% of its species were well represented (> 20 plant records). For example, Asteraceae the best represented family (Fig. 2), has only 28% of its species with more than 20 occurrences; 84% with more than one; and 16% of its species were represented by only one plant record. In contrast, 6% of the Orchidaceae

species were represented by more than 20 plant records; 58% by more than one; and 42% by just a single plant record (Fig. 4).

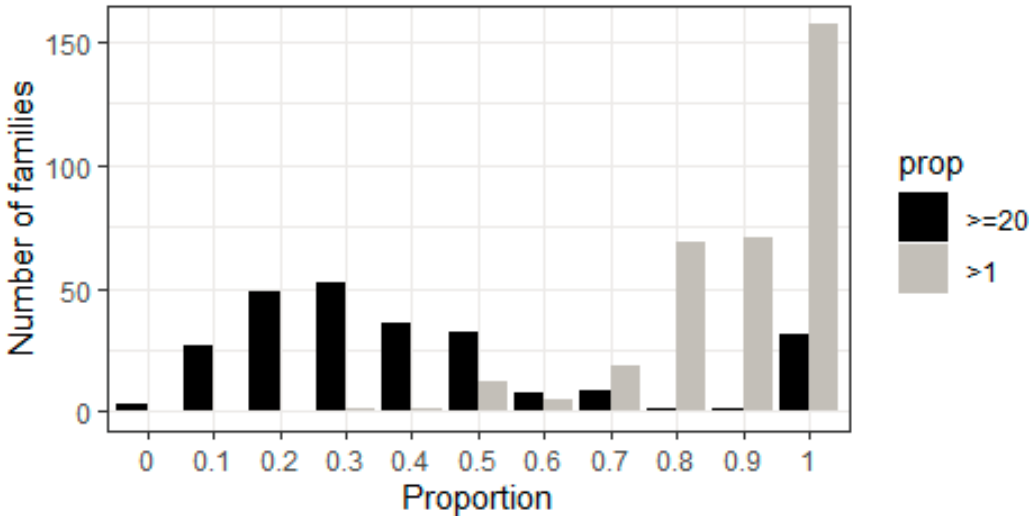


Figure 3. Proportion of species per family with > 1 and ≥ 20 occurrence across the 345 native plant families found in the Colombian Andes.

The number of occurrences per species ranged from one to 742, with a median of six occurrences per species. The top 10 occurrences per species represent 2.6% of the total plant occurrences for the Colombian Andes (Table 2).

Table 2. Top 10 plant species of Colombian Andes based on the number of records.

Family	**Species**	**Record Number**
Ericaceae	*Cavendishia bracteata*	742
Ericaceae	*Pernettya prostrata*	689
Euphorbiaceae	*Acalypha macrostachya*	611
Loranthaceae	*Gaiadendron punctatum*	562
Primulaceae	*Myrsine coriacea*	523
Ericaceae	*Gaultheria anastomosans*	505
Rubiaceae	*Palicourea angustifolia*	500
Rubiaceae	*Arcytophyllum nitidum*	496
Ericaceae	*Gaultheria erecta*	495
Polytrichaceae	*Polytrichum juniperinum*	489
Total top 10 species		**5,612 (3%)**
Remaining species		**212,787 (97%)**

Regarding the representatives in terms of occurrences, nearly 23% of the species (n = 2,720) had at least 20, and 18% (n = 2,065) were represented only by one. The remaining 59% of species were represented by between two and 19 occurrences (Table S3).

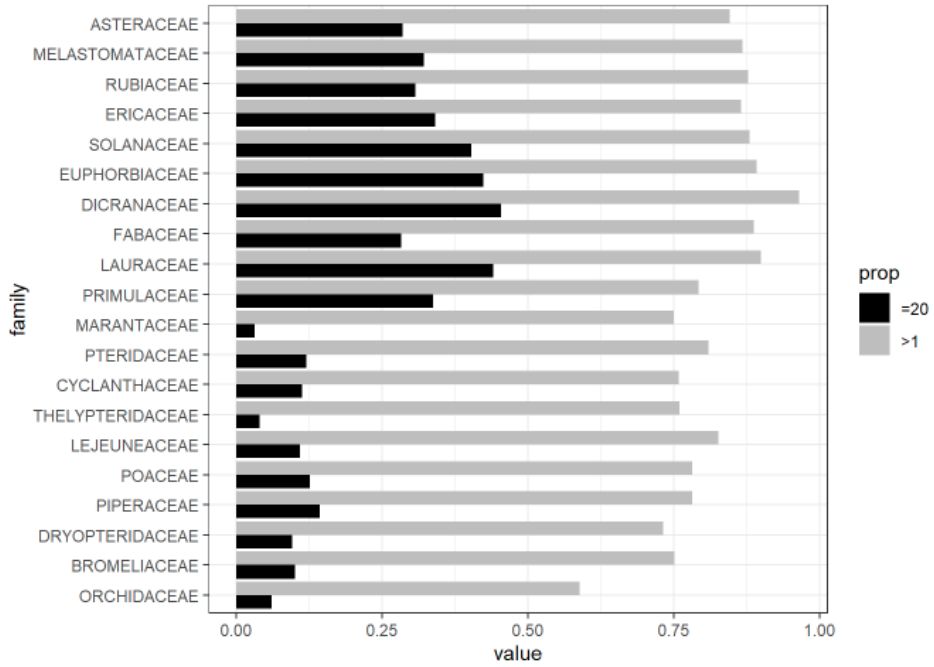


Figure 4. Top 10 Andean families that are over- and underrepresented. The figure shows the proportion of species represented by > 1 (grey bar) and 20 or more occurrence records (black bar).

Life forms

From the total native plant species reported for the Colombian Andes, just 60% (10,361) are represented in Flo_RA and have life form information. These all correspond to vascular plant species. Most species in Flo_RA are herbs, followed by shrubs, trees and epiphytes. These four life forms represent 87% of species (9,064 species) (Fig. 5). However, the main proportion of life forms in the plant occurrences of Flo_RA were herbs, trees and shrubs (81.8%; 158,360 occurrences; Fig. 6). Epiphytes were only marginally represented (6.38%)

almost as low as the climber's (10.6%) group (Fig. 6), which have half of the species represented compared with epiphytes.

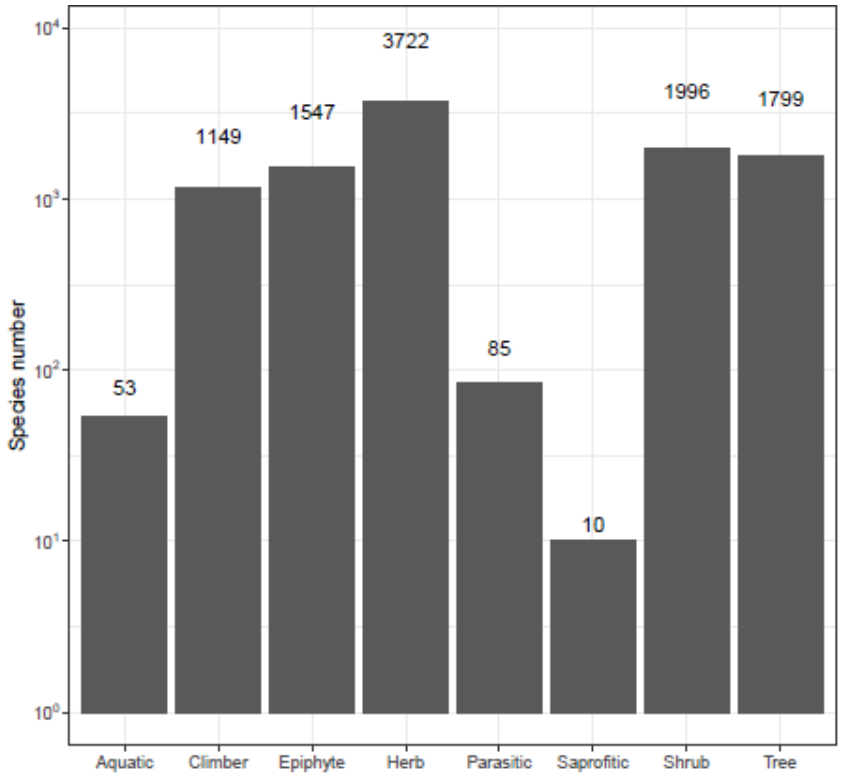


Figure 5. Number of species by life form represented in Flo_RA database for the Colombian Andes (note that the y-axis is in log scale).

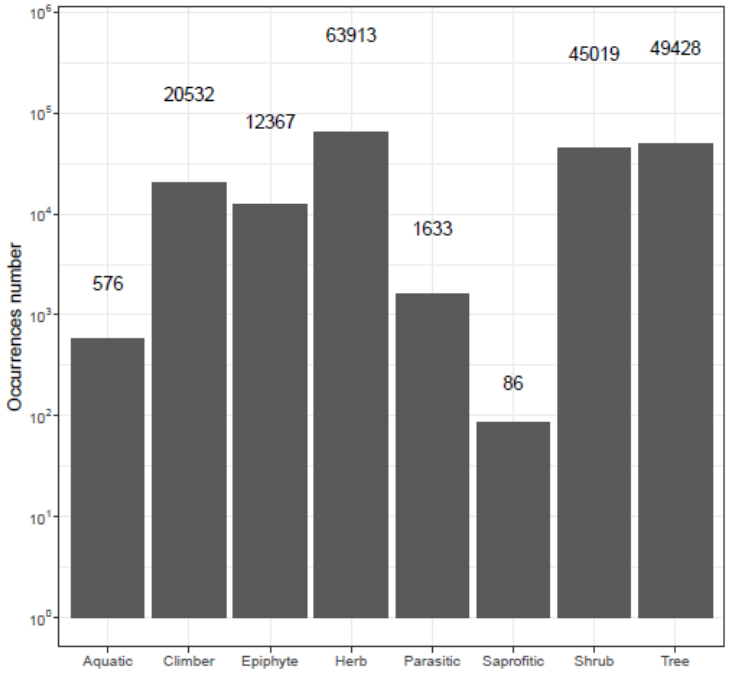


Figure 6. Number of plant occurrences available in Flo_RA database grouped by life form for Colombian Andes (note that the y-axis is in log scale).

The representativity analysis by plant life form showed that there are more occurrences than expected for trees and shrubs (Table 3). There were fewer occurrences than expected for saprophytic and aquatic life forms, but perhaps more dramatic is the low representativeness of epiphytic plants (21,839 plant occurrences less than expected; Table 3).

Table 3. Representation of life forms of the Colombian Andes based on their richness and occurrences. I = occurrences number expected; O-I = difference between occurrences expected and observed in the Flo_RA database.

Life form	Occurrences (O)	Species number	I	O-I
Herb	63913	3722	49969	13944
Tree	49428	1799	24152	25276
Shrub	45019	1996	26797	18222
Climber	20532	1149	15426	5106
Epiphyte	12367	1547	20769	-8402
Parasitic	1633	85	1141	492
Aquatic	576	53	712	-136
Saprophytic	86	10	134	-48

Spatial patterns

The spatial analysis showed that 86.3% (e.g., 694) of the 20 x 20 km grid cells across the Colombian Andes have records of native plants (Fig. 7).

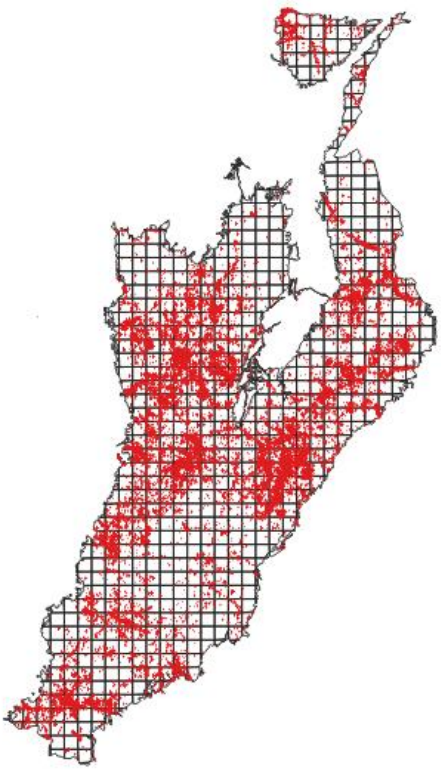


Figure 7. Native plant records (red dots) in the Colombian Andes on grid cell sizes of 20 x 20 km.

At the phylum level we found that Angiosperms are the best represented group with several hotspots of density occurrences (e.g., mainly Cundinamarca and Antioquia departments); Pteridophyta is widely represented in the Colombian Andes but with much fewer occurrences (e.g., mainly Cundinamarca department). Coniferophyta and Cycadophyta are poorly represented with very few occurrences (Fig. S1a). Bryophytes are the best represented non-vascular plant group followed by Marchantiophyta and Anthocerotophyta, with the latter represented by very few occurrences (Fig. S1b).

The number of cells occupied by families ranged between one and 526 (e.g., ca. 65% of all possible grid cells available for the study area). The range at the species level was between one to 206 (e.g., 25% of total cells available for the Colombian Andes). The

spatial representativity of species by family showed that 22% of the species were reported in more than 20 grid cells and 89% in more than one grid cell. For example, the best represented family, Asteraceae, had 79% of its species in more than one grid cell, and only 14% species were in more than 20 (Fig. S2a). In contrast, 54% of the species in Orchidaceae (the less represented family when taking into account species richness), were reported in more than one grid cell, and only 3% were in more than 20 (Fig. 8, Fig. S2b).

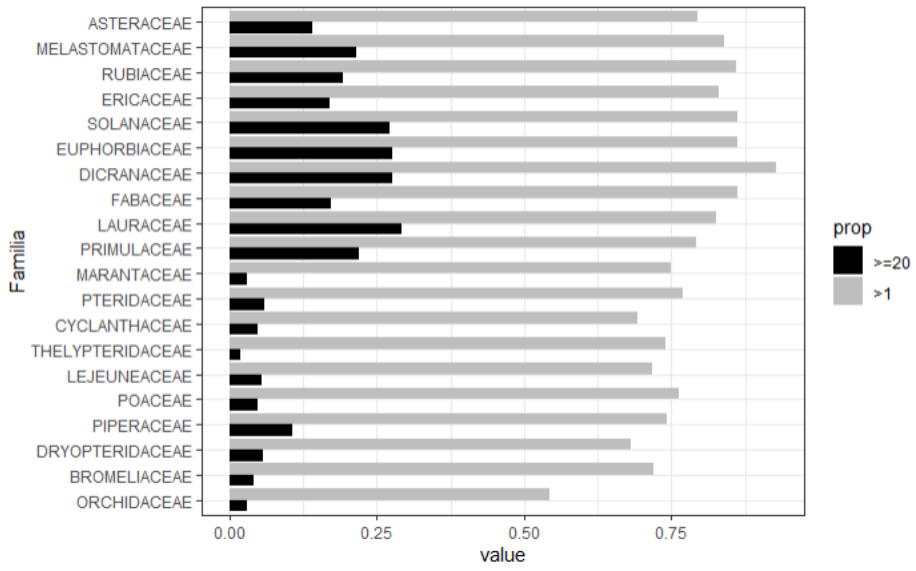


Figure 8. Species proportion by family present in > 1 grid cells (grey bar) and 20 or more grid cells (black bar). The families shown correspond to the top ten over- and under-represented by plant records (Fig. 2).

Discussion

There are few studies of taxonomic biases for the tropical regions despite them being the most diverse and threatened. In the case of the Colombian Andean flora, we found only 67% of the registered native plant species represented in the database (Flo_RA). Because our study only included records with coordinates, it is possible that some of the species are represented in botanical collections by specimens that have not yet been georeferenced, cannot be georeferenced, or are in collections not yet digitized and accessible by the public domain. In other cases, specimens may exist (digitized or undigitized) but they have not yet been identified to the species level.

At a taxonomic level, we found that flowering plants (Magnoliopsida) are by far the best-represented plant group in the Colombian Andes, unlike bryophytes (Bryophyta), ferns (Pteridophyta) and conifers (Coniferopsida) which, although they also showed more records than expected given their richness, their magnitude was much lower. In contrast, cycads (Cycadopsida), anthocerotes (Anthocerophyta) and Marchantiophyta were the least represented groups, reporting fewer records than expected. This behavior can be in part explained due to the origin of the information collected in the Flo_RA database. This information comes from records of plants deposited in herbaria. In Colombia, these collections result from explorations carried out by botanists, who have contributed to these collections since the botanical expedition led and initiated by Mutis in 1789 (Lagomarsino & Frost, 2020). As a consequence, the information in the Colombian Andean flora is the result of preferences, mainly of scientific origin, which contrasts with the pattern that has been reported in temperate countries, where plant collection preferences have to do with

scientific interests and with the establishment of botanical societies (these do not necessarily include scientists or trained botanists) (Troudet *et al.*, 2017).

In Colombia, the number of botanists that are specialized in vascular plants, particularly flowering plants, is more significant compared to other groups. Gymnosperms are not very diverse and have a restricted distribution, which explains their low representation in collections. On the other hand, the preference for Magnoliophyta may be related to their high diversity in the tropics, with new species being described almost daily. At the family level, we found that more than half of those registered in the Andes of Colombia report more records than expected, with respect to their species richness. The best-represented families are those with the greatest richness, but they also count with specialists (e.g., Asteraceae, Melastomataceae, Rubiaceae, Ericaceae). These families also correspond to highly diverse families above 1,500 meters of elevation (Gentry, 1995; Vargas Rincón, 2011; Bernal *et al.*, 2016)**,** which is consistent with Vargas *et al.* (2022), who report collection biases above this altitude.

Even though the highest proportion of families is overrepresented for the Colombian Andes at the species level, the average number of records is deficient (e.g., six average occurrences per species). Only a very low proportion of species per family are well represented (i.e., >20 occurrences). When analyzing which are the best collected species, they correspond to common and widely distributed species, typical of the sub-páramo and páramo ecosystems (Table 2). This could reflecting botanists' preference for of the páramo biome, which is attractive due to its high rates of endemism, diversity and also strategic for ecosystem services, such as water provisioning (Madriñán, Cortés, & Richardson, 2013; Nürk, Scheriau, & Madriñán, 2013; Valencia-Leguizamón & Tobón, 2017). The proximity

of research centers to particular ecosystems has also been associated with biases in the collection patterns (Engemann *et al.*, 2015; Daru *et al.*, 2018; Vargas *et al.*, 2022), and might explain why the highest proportion of plant records are from the páramo.

The collector's expertise and the recurrence in the exploration of some sites is related to the state of knowledge of the flora of the Colombian Andes. The expertise of the collectors is associated with reporting new species, species with a restricted distribution, and life forms that are more difficult to collect, such as epiphytes (Scott & Caroline J. Hallam, 2003). While inexperienced collectors prefer to collect common species, more experienced collectors report in their collections a higher proportion of rare species as well as difficult life forms. Recurrence of collection in the tropics is also low which, as a consequence, results in areas collected once or with few collection events whose periods between collections exceed 80 years (Meyer *et al.*, 2016b).

The preference of collections in the Andes also shows groups with fewer records than expected given their richness; these include valuable groups for their ornamental nature, such as Orchidaceae and Bromeliaceae in which, although in Colombia they have specialists and cultivator societies, species are represented by only a few records. The social preference towards these groups is not manifested in the availability of records in collections (e.g., herbaria), perhaps due to some difficulty for these sectors of society to include these records in collections. In contrast, taxonomic complexity and inconspicuous character are likely explanations why some highly diverse families are underrepresented in collections and databases (e.g., Poaceae, Piperaceae; Daru *et al.*, 2018; Schmidt-Lebuhn *et al.*, 2013). In general, unstructured sampling schemes (e.g., free collections) direct the collection to specimens in a reproductive state. This type of collection scheme produces a

"collection blindness" towards families with inconspicuous flowers or fruits, resulting in families with small/inconspicuous flowers being less frequently collected (such as Poaceae and Piperaceae).

Regarding life forms, trees, shrubs, and herbs far exceed the number of records expected given their richness. This is in contrast to aquatic, saprophytic and epiphytic, the latter group being the most poorly represented in Flo_RA. In the case of trees, shrubs and grasses, these are conspicuous and dominant groups in ecosystems from lowlands to high elevations in the Colombian Andes. Therefore, it is not uncommon to find them well represented in collections and in databases. Trees are also commonly used to characterize plant communities. In contrast, the sampling of aquatic plants is scarce given the restricted nature of these systems in the Colombian Andes, being isolated, small, or yet to be discovered. In the case of saprophytes and epiphytes, their sampling has been complementary to the sampling of plots and transects. Added to this is the very biology of the species, with species that only grow in the canopy, with restricted population size, or short and irregular flowering periods, among others. This may explain the low number of records available by species of these groups of plants, especially if it is considered that the material in the vegetative state makes their identification almost impossible (e.g., Orchidaceae).

Spatially, a low proportion of well-represented species per family was also found (e.g., 22% of the species reported for the Andes of Colombia were reported in 20 or more cells of 20 x 20 km). The high rate of endemism and restricted species distributions is a characteristic of the flora of the tropical Andes (Hughes & Eastwood, 2006; Antonelli & Sanmartín, 2011; Madriñán *et al.*, 2013). However, our analysis supports the hypothesis of

a lack of sampling for most of the plant families reported for the Colombian Andes, given the high rate of species with less than 20 records together with an increased number of species represented by only one record. and those that are only known from the type collections. Additionally, for the sampling bias that concentrates the highest density of plant collections in the Colombian Andes, mainly in Cundinamarca and Antioquia, Vargas *et al.* (2022) seems to support the hypothesis.

The state of documentation of the Colombian Andean flora presents multiple challenges from the point of view of species conservation in the face of local threats and global climate change driven transformation. In the case of the Andean flora, a high proportion of the registered species have very few records. The scarcity of records of Andean species has implications at more complex levels of the study of biodiversity than those that are eminently taxonomic. For example, the lack of information at the species level affects the knowledge of their distribution (Wallacean shortfall), the structure of their populations (Prestonian shortfall), the understanding of their phylogenetic relationships (Darwinian shortfall), the knowledge of their ecological traits and functions (Raunkerian shortfall), their physiological tolerance (Hutchinsonian shortfall) and the interactions with other species (Eltonian shortfall) (Hortal *et al.*, 2015).

The lack of information regarding species records for the case of the Colombian Andes limits the application of distribution models, a powerful tool used in conservation biology, ecology and biogeography (Feeley and Silman, 2011a). Modelling tools such as Maxent have shown accuracy in generating species distribution maps with a minimum of 20-30 (Elith *et al.*, 2006; Hernandez *et al.*, 2006; Wisz *et al.*, 2008) unique localities per species. A condition that is not met for a large proportion of the species reported here

(77%). There needs to be more information at the species level to permit adequate categorization of their risk of extinction.

The mobilization of collections, from analogue to digital formats, has been proposed as an alternative to the lack of data (Peterson, Soberón, & Krishtalka, 2015; Davis, 2023). However, understanding what types of taxonomic groups need more work in collection numbers through fieldwork allows for directing and orienting biodiversity research resources (Vargas *et al.*, 2023). The training of new botanists is also of the utmost importance, not only due to the decrease in taxonomists dedicated to the description of biodiversity, but also because they could be oriented towards poorly collected groups. This would allow to fill gaps in many important plant groups and provide a more comprehensive picture of biodiversity, especially in highly complex and diverse systems such as the tropical Andes. This would, in turn, enable better management, use and conservation of the plant biodiversity.

References

Antonelli A, Sanmartín I. 2011. Why are there so many plant species in the Neotropics? *Taxon* 60: 403–414.

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

Brummitt N, Araújo AC, Harris T. 2021. Areas of plant diversity—What do we know? *Plants People Planet* 3: 33–44.

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.

Davis CC. 2023. The herbarium of the future. *Trends in Ecology and Evolution* 38: 412–423.

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.

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

Feeley KJ, Silman MR. 2011b. The data void in modelling current and future distributions of tropical species. *Global Change Biology* 17: 626–630.

Gentry AH. 1995. Patterns of diversity and floristic composition in Neotropical montane

forest. In: Churchill SP,, In: Balslev H,, In: Forero E,, In: Luteyn JL, eds. *Biodiversity and conservation of neotropical montane forests*. Nueva York: The New York Botanical Garden, 103–126.

Graham A. 2009. The Andes: a Geological Overview From a Biological Perspective. Ann. Missouri Bot. Gard. 96, 371–385. https://doi.org/10.3417/2007146 Biological Perspective. *Annals of the Missouri Botanical Garden* 96: 371–385.

Haque M, Beaumont LJ, Nipperess DA. 2020. Taxonomic shortfalls in digitised collections of Australia ' s flora. *Biodiversity and Conservation* 29: 333–343.

Haque MM, Nipperess DA, Gallagher R V., Beaumont LJ. 2017. How well documented is Australia's flora? Understanding spatial bias in vouchered plant specimens. *Austral Ecology* 42: 690–699.

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 of the United States of America* 103: 10334–10339.

Josse C, Cuesta F, Navarro G. 2011. Physical geography and ecosystems in the tropical Andes. *... in the tropical Andes ...*: 152–169.

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.

Mace GM, Collar NJ, Gaston KJ, Hilton-Taylor C, Akçakaya HR, Leader-Williams N, Milner-Gulland EJ, Stuart SN. 2008. Quantification of extinction risk: IUCN's system for classifying threatened species. *Conservation Biology* 22: 1424–1442.

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. 2016a. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology letters* 19: 992–1006.

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

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.

Pérez-Escobar OA, Zizka A, Bermúdez MA, Meseguer AS, Condamine FL, Hoorn C, Hooghiemstra H, Pu Y, Bogarín D, Boschman LM, Pennington RT, Antonelli A, Chomicki G. 2022. The Andes through time: evolution and distribution of Andean floras. *Trends in Plant Science* 27: 364–378.

Peterson AT, Soberón J, Krishtalka L. 2015. A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC Ecology* 15: 1–9.

Peyre G, Lenoir J, Karger DN, Gomez M, Gonzalez A, Broennimann O, Guisan A. 2020. The fate of páramo plant assemblages in the sky islands of the northern Andes. *Journal of Vegetation Science*: 1–14.

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

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

Rodríguez N, Armenteras D, Morales M, Romero M. 2006. *Ecosistemas de los Andes colombianos*. Bogotá: Instituto de investigación de recursos biológicos Alexander von Humboldt.

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.

Scott WA, Caroline J. Hallam. 2003. Assessing species misidentification rates through quality assurance of vegetation monitoring. *Plant Ecology* 165: 101–115.

Trew BT, Maclean IMD. 2021. Vulnerability of global biodiversity hotspots to climate change. *Global Ecology and Biogeography* 30: 768–783.

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.

Valencia-Leguizamón JM, Tobón C. 2017. Influencia de la vegetación en el funcionamiento hidrológico de cuencas de humedales de alta montaña tropical. *Ecosistemas, revista científica de ecología y medio ambiente* 26: 10–17.

Vargas CA, Bottin M, Särkinen T, Richardson JE, Raz L, Garzon-Lopez CX, Sanchez A. 2022. Environmental and geographical biases in plant specimen data from the Colombian Andes. *Botanical Journal of the Linnean Society*: 1–14.

Vargas Rincón CA. 2011. Caracterización Florística Y Fitogeográfica Del Sector Sur De La

Serranía De Perijá Y Áreas Adyacentes De La Cordillera Oriental Colombiana.

Vargas CA, Bottin M, Sarkinen T, Richardson JE, Celis M, Villanueva B, Sanchez A. 2023. How to fill the biodiversity data gap : Is it better to invest in fieldwork or curation ? *Plant Diversity*. *In press.*

Wickham H. 2011. ggplot2. *Wiley interdisciplinary reviews: computational statistics* 3.

Wickham H. 2021. forcats: Tools for Working with Categorical Variables (Factors). *R package version 0.5. 1.*

Yarberry, W., & Yarberry W. 2021. DPLYR, Stringr, Lubridate, and RegEx in R. *CRAN*: 1–58.

Chapter 4: How to fill the biodiversity data gap: Is it better to invest in fieldwork or curation?

This chapter corresponds to the accepted version of the manuscript.

Paper published: *Plant diversity* (2023)

DOI: doi.org/10.1016/j.pld.2023.06.003

IF: 3.359

Abstract

Data gaps and biases are two important issues that affect the quality of biodiversity information and downstream results. Understanding how best to fill existing gaps and account for biases is necessary to improve our current information most effectively. Two current main approaches for obtaining and improving data include (1) curation of biological collections, and (2) fieldwork. However, the comparative effectiveness of these approaches in improving biodiversity data remains little explored. We used the Flora de Bogotá project to study the magnitude of change in species richness, spatial coverage, and sample coverage of plant records based on curation versus fieldwork. The process of curation resulted in a decrease in species richness (synonym and error removal), but it significantly increased the number of records per species. Fieldwork contributed to a slight increase in species richness, via accumulation of new records. Additionally, curation led to increases in spatial coverage, species observed by locality, the number of plant records by species, and localities by species compared to fieldwork. Overall, curation was more efficient in producing new information compared to fieldwork, mainly because of the large number of records available in herbaria. We recommend intensive curatorial work as the first step in increasing biodiversity data quality and quantity, to identify bias and gaps at the regional scale that can then be targeted with fieldwork. The stepwise strategy would enable fieldwork to be planned more cost-effectively given the limited resources for biodiversity exploration and characterization.

Keywords: Colombia, Flora de Bogotá, sample coverage, species richness, Tropical Andes.

Introduction

Identifying spatial patterns of biodiversity distribution is fundamental to understanding how they were established. They are also essential for designing effective conservation and management strategies. However, it is well known that biodiversity information is incomplete or biased, limiting the generalization of results and the predictive power of models (Feeley and Silman, 2011a; García Márquez *et al.*, 2012; Sousa-Baena *et al.*, 2013; Vargas *et al.*, 2022).

Although researchers are generally aware of the deficiencies of biological data, and different strategies have been developed to reduce their impact (Elith *et al.*, 2006; Syfert, Smith, & Coomes, 2013; Engemann *et al.*, 2015), many agree on the need to improve data availability (Graham *et al.*, 2004; Feeley & Silman, 2010; Feeley, 2015; Ball-Damerow *et al.*, 2019) and to generate more data through field explorations (Hopkins, 2007; Feeley and Silman, 2011a, 2011b). Museums and biological collections centralise a significant amount of biodiversity data that is currently available through global public online repositories (e.g., GBIF (https://www.gbif.org/), BIEN (https://bien.nceas.ucsb.edu/bien/)). The online information is used for different purposes that include the study of evolutionary process, causes and limitations of species distributions, the response of species to climate change, and designing protected areas (Soberón & Peterson, 2004; Bebber *et al.*, 2010; Feeley & Silman, 2010; Gaira, Dhar, & Belwal, 2011). This availability is possible due to the mass digitalisation of biological collections through standard formats, which can be handled with different information analysis programs. Digitalisation opens up centuries of data and the possibility of studying biodiversity dynamics, evaluating changes through time, and modelling of future scenarios for management, use and conservation purposes (Feeley,

2012; Morueta-Holme *et al.*, 2015; Nualart *et al.*, 2017). However, around 40% to 80% of biological records available online or databased are discarded because of taxonomic (e.g., poor determination, undetermined material and nomenclatural issues), geographical (e.g., poor georeference precision or samples assigned to a centroid), or temporal (i.e., without date) data deficiencies (Gueta & Carmel, 2016; Meyer *et al.*, 2016; Daru *et al.*, 2018). This data can be rescued by curatorial work (Feeley & Silman, 2011b), although it is a time-consuming effort. Additionally, not all museum collections are digitised, particularly small and local collections, and much work remains to be done in curating these collections to a high taxonomic standard.

Fieldwork is an alternative way to increase the amount and the accuracy of data. However, this requires investment of financial resources by institutions whose role is to describe biodiversity (e.g., governmental agencies, botanic gardens). Given the limited financial resources for research, it is therefore necessary to analyze where funds could best be invested (O'Connell, Gilbert, & Hatfield, 2004; Franco, Palmeirim, & Sutherland, 2007; Gardner *et al.*, 2008; Targetti *et al.*, 2014), in order to increase resources with the best return.

Biological collections contain large amounts of data that could be retrieved through curatorial work. Curatorial work has the potential to increase the richness by the discovery of new species (Goodwin *et al.*, 2015), expanding species distributions, increasing the environmental envelop of species (Feeley & Silman, 2011b), categorizing threatened species (Nualart *et al.*, 2017), and/or filling geographical gaps (Daru *et al.*, 2018). At the regional scale, curatorial work has the potential to increase the geographical coverage of collections by filling gaps on the collection information. However, it is not well known how curatorial work impacts biodiversity knowledge compared with fieldwork.

In this paper, we explore how curatorial work of herbarium collections increases biodiversity data quality and quantity in contrast to fieldwork. Using the Flora de Bogotá project as a model, we analyze the change in species richness, spatial coverage, and sample coverage of plant records in Bogotá (capital of Colombia). We evaluate the impact of both curation and fieldwork on increasing the taxonomic and geographical robustness of biodiversity information and highlight their unique contributions.

Materials and Methods

Study area

Bogotá, the capital of Colombia and the most populated city in the country (ca. 7.5 million people over ca. 1,630 km^2; http://www.sdp.gov.co/, accessed on 9th Sep 2022), is located in the Colombian Cordillera Oriental between 2,510-3,780 m elevation (Fig. 1). The climate is typical of tropical mountains with daily temperatures varying between 6 and 22°C and low annual seasonality (Secretaría Distrital de Ambiente, 2007). The rainfall regime is bimodal, with two peaks occurring in April - May and October - November. Bogotá is situated on two physiographic units: a flat area north of the city's urban area, where an enormous lake disappeared 30,000 years ago (Van der Hammen, 1986), and a mountainous area surrounding the metropolitan area to the east and south which includes the most extensive area of páramo vegetation on Earth (Sumapaz Páramo). Seventy-five per cent of Bogotá's territory is rural, while the remaining 25% is occupied by the urban area, where 80% of the population lives. Urban ecosystems are represented mainly through metropolitan parks and the wetland system associated with the Bogotá River. Meanwhile, natural ecosystems are

concentrated in the city's rural areas where the páramo ecosystem predominates, and relicts of Andean and high Andean forests are also found.

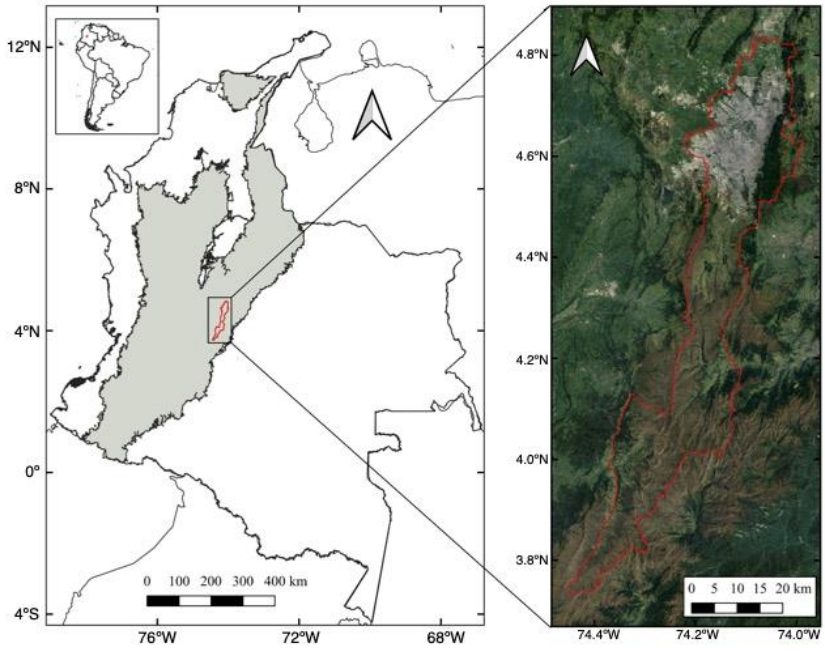


Figure 1. Map of Bogotá (satellite image, red line left panel) in the context of South America (top left) and Colombia (left). Grey area corresponds to the Colombian Andes.

Data

The Flora de Bogotá database is an initiative of the *Jardín Botánico de Bogotá* to study the plant diversity of the city. The database was established in 2013 to gather information of Bogotá's plant records deposited in herbaria and contains 37,468 plant records obtained from local and worldwide herbaria. Additionally, 5,401 new plant records from fieldwork were made by *Jardín Botánico de Bogotá* between 2011 and 2016 and are included in the database. The total database consists of 42,869 plant records (Table 1).

Table 1. Data sources of vascular plant records gathered in the Flora de Bogotá database.

Source	Number of records
Herbario Nacional Colombiano (COL)	12,869
Bibliographic review	10,966
Herbario Jardín Botánico de Bogotá (JBB)	8,671
Herbario Forestal Gilberto Emilio Mahecha (UDBC)	1,533
Missouri Botanical Garden (MO)	1,186
Pontificia Universidad Javeriana (HPUJ)	1,077
Instituto de investigación Alexander von Humboldt (FMB)	544
Smithsonian Institution (US)	452
New York Botanical Garden (NY)	170
Jardín Botánico de Bogotá Fieldwork	5,401
Total	42,869

For this study, records with identical collection numbers were screened, leaving only one of each in the database. Plant records with coordinates outside of Bogotá were excluded. This reduced the dataset to 21,926 plant records that, after curatorial and fieldwork represent 2384 species, 903 genera and 187 families. All specimens of this research are available in open databases (see: Vargas *et al.*, 2022; https://doi.org/10.1093/botlinnean/boac035). The specimens at the Herbario Nacional Colombiano (COL, http://www.biovirtual.unal.edu.co/es/colecciones/search/plants/) and Jardín Botanico de Bogotá (JBB, https://herbario.jbb.gov.co/especimen/simple) have

images in high resolution for many of the specimens that support the current manuscript. The list of flora de Bogotá is online (https://florabog.jbb.gov.co/) and the names are supported by specimens that can be checked by users. The flora of Bogotá is also continuously revised and improved. Therefore, the two main datasets used in this manuscript are digitized and available for checking by specialists.

Data treatment

To study the effect of curatorial vs. fieldwork in the biodiversity patterns of Bogotá, we analyzed the change in species richness, spatial coverage of plant records and completeness at four different stages of data collection, during the first stage of the Flora de Bogotá project (2012 -2016):

1. Raw dataset: Plant records of the Flora of Bogotá database obtained from herbaria without nomenclatural and coordinate corrections.
2. Curated dataset: Plant records on the Flora database obtained from herbaria that were corrected for nomenclatural and coordinate metadata. The taxonomic work consisted of revising herbarium specimens and correcting obvious orthographic and spelling errors on the names assigned to plant records, as well as screening for synonymy. The taxon names were standardized using the *Catálogo de plantas y líquenes de Colombia* (Bernal, Grandstein, & Celis, 2016). Geographical work was advanced on every plant record by correcting and standardizing coordinates. Plant records without coordinates were georeferenced using the locality information from the specimen label, following the point-radius method (Wieczorek, Guo, & Hijmans, 2004). The specimens with insufficient locality data were excluded from the analysis, such as the specimens older than the 1900's, mainly collected by Jose

Jeronimo Triana, who reports specimens from "Bogotá province", which was a wider region than the current area of Bogotá city.

3. Fieldwork dataset: The fieldwork conducted between 2012-2016 by *Jardín Botánico* in different parts of Bogotá to characterize areas without data. The characterization of those areas was conducted by the plant collection in reproductive condition. The plant records were deposited in the *Jardín Botánico de Bogotá* herbarium and identified by the botanical team to species level. The geographical and taxonomic data was checked to correct for any mistakes.
4. Total (Curated – fieldwork): Curated dataset in addition to plant records obtained from fieldwork done between 2012-2016 by *Jardín Botánico* in Bogotá city.

Data analysis

For the analysis, the taxonomic and geographical changes in plant records were used to evaluate the difference in richness species, spatial coverage and sample coverage of the Flora of Bogotá, made through curatorial work compared with fieldwork. To analyze the effect of curatorial work (made in the database and herbarium specimens) and fieldwork to the Flora of Bogotá data quality, we evaluated the changes between 2012 to 2016. We conducted the taxonomic and spatial analysis for each dataset and compared differences between datasets to evaluate the improvement of data through both curatorial and fieldwork.

The analysis was conducted at two levels:

1. Taxonomic

To understand the change in taxonomic quality through curatorial and fieldwork, we calculated the number of taxon names and plant records by species at every data stage. We compared between stages using non-parametric tests (Kruskal-Wallis and the Wilcoxon test) in R 3.6.1 (R Development Core Team, 2019).

2. Spatial and sample coverage

In order to understand the geographical contribution of curatorial and fieldwork, we created a grid of cell size 1 km by 1 km (1 x 1) over the city. We analyzed the change in spatial coverage (number of grid cells with plant records), density records (number of plant records by grid cell of 1 x1 km), richness observed (number of species observed by grid cell of 1 x1 km), and sample coverage in the raw, curated and total datasets. The same analysis was conducted at the ecosystem level using the Colombian ecosystem map (Etter, 1998) to delimit Bogotá's ecosystems.

We conducted a spatial coverage analysis to observe the representativity of plant records in Bogotá, calculating it as the proportion of grid cells of 1 x 1 km with plant records over the total grid cells in Bogotá (1,842 grid cells of 1 km x 1 km) at the four stages of the data:

$$\boldsymbol{Cob}\% = \frac{\boldsymbol{Nr}}{\boldsymbol{N}} * \mathbf{100}$$

Where N is total number of grid cells in Bogotá and Nr the number of grid cells with plant records at every stage of data treatment.

We also analyzed the effect of curatorial and fieldwork on the record density by km^2, as a first step to describing the collection patterns on the territory (Soberón *et al.*, 2007); observed richness and sample coverage for every grid cell. For the sample coverage, rarefaction was calculated in cells with more than 20 plant records. The sample coverage is a measure of sample completeness, giving the proportion of the total number of individuals in a community that belongs to the species represented in the sample (Chao and Jost, 2012). Sample coverage is defined as the total relative abundance of the observed species in the sample, ranging from 0 to 1. Sample completeness was estimated using the iNEXT R package (Hsieh, Ma, & Chao, 2016).

Finally, we tested for differences in richness and sample coverage between raw, curated and total datasets using non-parametric tests (Kruskal-Wallis and the Wilcoxon test) using R 3.6.1 (R Develpment Core Team, 2019)(R Develpment Core Team, 2019) and illustrated the results in maps created in QGIS (QGIS Development Team, 2015).

Results

Taxonomic changes following cleaning and fieldwork

Taxonomic cleaning decreased the number of taxon names, while fieldwork added new ones to the Flora de Bogotá database. As a result, taxa decreased by 24% for family and species levels, and 7% for genera (Table 2). On the other hand, fieldwork added 83 (3.5% of total species diversity) new names at the species level, most of which were herbs and epiphytes (Table S1).

Table 2. Number of species, genera and family in the Flora de Bogotá database at four data stages: raw data, clean data, fieldwork and total (combination of curated and fieldwork datasets), evidencing the change in the data quality resulting from curation and fieldwork.

	Raw	Curated	Fieldwork	Total
Family	225	187	110	187
Genera	967	904	312	904
Species	2878	2301	749	2384

The curatorial process and fieldwork significantly improved the number of records by species ($p > 0.05$) (Fig. S1), where the probability of species with less than five plant records decreased from 0.35 in the raw dataset to 0.21 in the curated dataset and 1.9 in the total dataset (curated – fieldwork), respectively (Fig. S1). It is important to note that for 744 species the number of records increased by fieldwork, for 690 species through the curatorial process, for 1157 by either of the two approaches (curatorial or fieldwork), and 277 species by the combined effect of both curatorial and fieldwork. *Vaccinium floribundum* (Ericaceae) showed the maximum number of plant records (143) and species with more than 100 plant records represented 0.6% (15 species) of the species recorded for Bogotá (Table S2).

Spatial Representation

The spatial distribution of species showed significant differences between datasets ($p > 0.05$) (Fig. S2), where the probability of species being in one cell decreased from raw to curated and total datasets (curated – fieldwork) (Fig. S2). At the same time, fieldwork

added grid cells to 28% of species, while the curatorial process added grid cells to 26% of species. The combination of cleaning and fieldwork added grid cells to 46% of species reported in the Flora database (Table S3). *Gaultheria anastomosans* (Ericaceae) was the most widely distributed species, recorded in 78 (4.2%) grid cells in Bogotá.

Spatial and ecosystem changes by curatorial and fieldwork

Grid cells and density records. The curatorial work on georeferences of plant records and fieldwork increased the number of plant records with coordinates in the Flora de Bogotá database. The number of plant records with coordinates increased by 77% from raw to total datasets, but the main contribution was due to the curatorial work that added coordinates to 59% of records, while fieldwork only added 18% (Fig. 2). Additionally, curatorial and fieldwork increased the number of cells with plant records from 364 in the raw dataset (19.8% of total grid cells) to 753 (41%) in the total dataset (curated – fieldwork). However, fieldwork only added three new grid cells (0.1%) (Fig. 3).

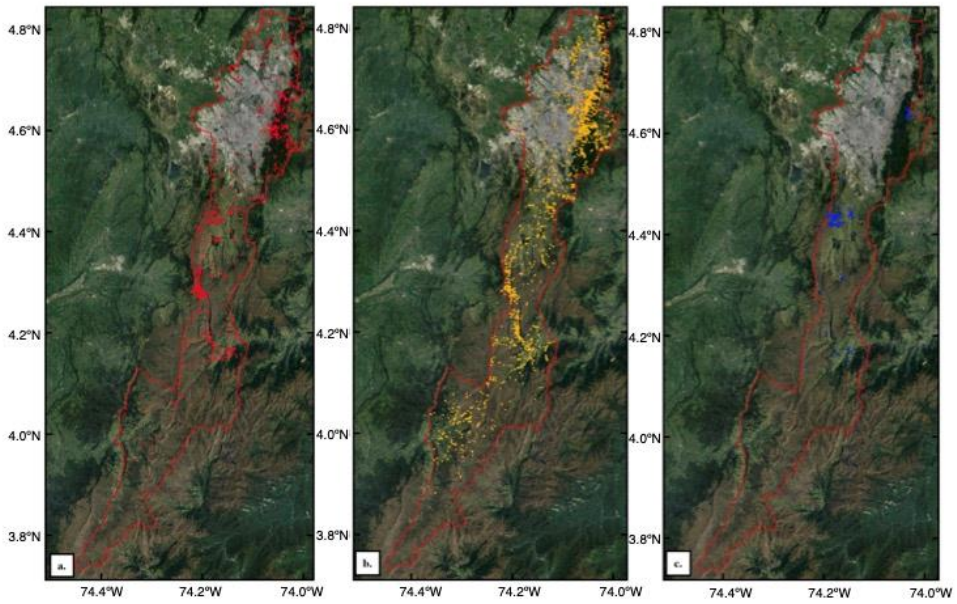


Figure 2. Plant records with species names and coordinates available in the different datasets of the vascular plants of the Flora of Bogotá: (a) raw dataset (red dots), (b) curated dataset (yellow dots), and (c) fieldwork dataset (blue dots). The red line delimits the area of Bogotá.

On the other hand, cells with low-density records predominate at the three stages of data, although georeferenced and fieldwork slightly increased density. The number of grid cells with very low (1 to 10 plant records by grid cells), low (11-100) and medium (100-500) density increased by 11, 8 and 2%, respectively, from the raw dataset to the total dataset. The increase in density of grid cells with high (501-1000 plant records) and very high (1001-1500 plant records) density records increased, although it was 0.1 and 0.2, respectively (Fig. 4).

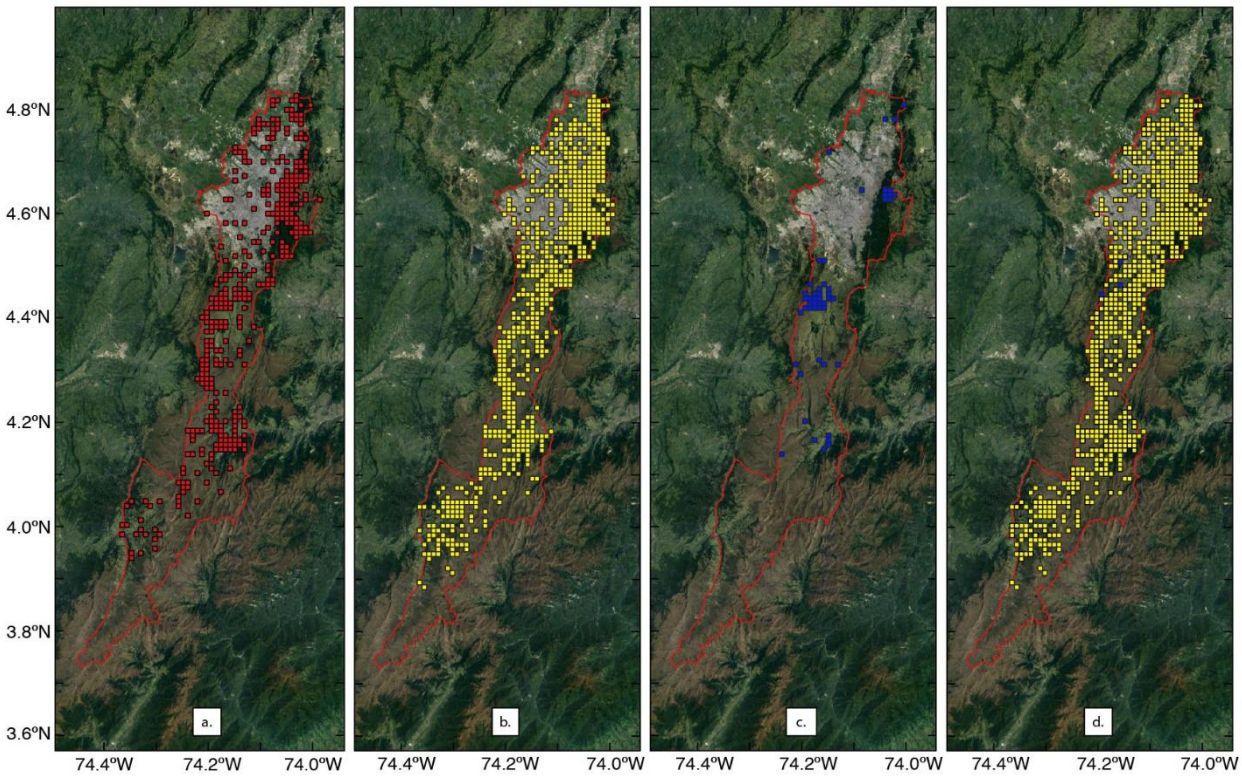


Figure 3. Number of 1km^2 grid cells covered by the different datasets of the vascular plants of the Flora of Bogotá: (a) raw dataset (red dots), (b) curated dataset (yellow dots), (c) fieldwork dataset (blue dots), and (d) total dataset (combination of curated and fieldwork datasets). The red line delimits the area of Bogotá.

At ecosystem level, plant records increased in Bogotá as a result of curatorial and fieldwork. Furthermore, ecosystems that were not represented in the raw dataset were represented by plant records after curatorial and fieldwork. Ecosystem representation was higher after curatorial work (ecosystem mean 57.1%) than through fieldwork (mean 10%) and raw data (mean 32.9%). For instance, plant density records increased in all ecosystems and a lake ecosystem appeared in the floristic record (i.e., La Regadera lake). In contrast,

fieldwork increased the density of records in just five ecosystems with only one (i.e., western dry páramo of the city), representing 73% of those plant records (Table 3).

Table 3. Changes in ecosystem assignment when analysing plant records at the three stages of data (raw data, clean data, fieldwork and total [combination of curated and fieldwork datasets]). Every stage shows the input in plant records and its contribution to the total number of records by ecosystem (in %). Although several ecosystems are repeated (i.e., Dry Andean forest, Mixed agroecosystems, Humid Andean forest and Lake), they are considered as independent in the ecosystem classification of Bogotá (they occur at different locations). For a spatial representation of the ecosystems, please refer to Fig. S4.

Ecosystem	Ecosystem code	Area (km2)	Raw (%)	Clean (%)	Fieldwork (%)	Total (%)
Rural areas trasnformed by human activities	II	266,202	1439 (20.4)	2987 (42,4)	2614 (37.1)	7040 (100)
Humid páramos	19	582,543	1034 (22.4)	3351 (72,7)	222 (4.8)	4607(100)
Dry Andean forest	18b	42,964	748 (21.1)	2795 (78.8)	0(0)	3543(100)
Urban Area	U	316,03	423 (17.9)	1912 (80.9)	26 (1.1)	2361(100)
Mix agrosystems	C3	133,219	766 (39.1)	538 (27.4)	655 (33.4)	1959 (100)
Dry páramos	20	40,307	248 (24.8)	750 (75.1)	0 (0)	998 (100)
Dry Andean forest	18b	16,81	87 (16.5)	382 (72.4)	58 (11)	527 (100)
Dry páramos	20	15,373	102 (24.9)	9 (2.2)	298 (72.8)	409 (100)

Ecosystem	Ecosystem code	Area (km2)	Raw (%)	Clean (%)	Fieldwork (%)	Total (%)
Milky agrosystems	C4	62,366	47 (28.8)	116 (71.1)	0 (0)	163 (100)
Humid Andean forest	18a	29,658	14 (12.5)	98 (87.5)	0 (0)	112 (100)
Oak Andean forest	18c	31,903	37 (41.1)	53 (58.8)	0 (0)	90 (100)
Mix agrosystems	C3	10,93	54 (98.1)	1 (1.8)	0 (0)	55 (100)
Humid Andean forest	18a	2,492	20 (90.9)	2 (9.1)	0 (0)	22 (100)
Humid Andean forest	18a	36,445	3 (17.6)	14 (82.3)	0 (0)	17 (100)
Lake	La	1,859	6 (50)	6 (50)	0 (0)	12 (100)
Lake	La	1,859	0 (0)	11 (100)	0 (0)	11 (100)
Humid Andean forest	18a	1,531	0 (0)	0 (0)	0 (0)	0 (0)

Overall, 80% of the Flora of Bogotá species are found in just four ecosystems. Two of those (humid páramo and dry high Andean Forest) are natural ecosystems, while the others consist of ecosystems with human intervention around the urban area. The most conserved ecosystems (e.g., cloud and high Andean humid forests; wet páramos) are located in the rural area, some distance from the urban area of Bogotá.

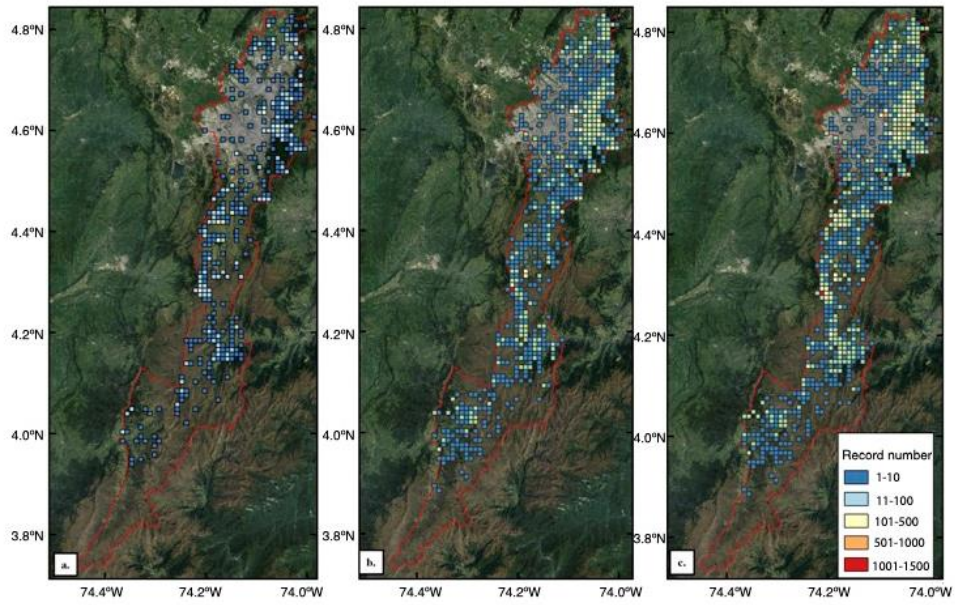


Figure 4. Number of plant records per 1 km^2 grid cell size with panels indicating change of density through curatorial and fieldwork in the vascular plant dataset of Flora of Bogotá: (a) raw dataset, (b) curated dataset, and (c) total dataset (combination of curated and fieldwork datasets). The blue – red scale indicates the spatial variation of record density in every stage of data: low density is indicated in blue while high density is indicated in red. The red line delimits the area of Bogotá.

Richness and completeness changes through curatorial and fieldwork

Richness. Curatorial processes and fieldwork increased the median number of species observed in grid cells; however, there were no significant differences between datasets. While the median of observed richness at the raw dataset was four, the median at the curated dataset and total dataset (curated - fieldwork) was five and six, respectively. 75% of grid cells recorded 11 species at the raw dataset, 15 at curated and 18 at the total dataset. On the other hand, very few grid cells showed a high number of species observed. For example, 4% of grid cells contained more than 50 species in the raw stage, meanwhile, 7%

of grid cells in the curated dataset and 9% of grid cells in the total dataset recorded more than 50 species observed (Fig. S3).

Sample coverage. This analysis discarded many grid cells with plant records because of the low sample size, even if the curatorial process and fieldwork added new ones. For example, while the proportion of grid cells with plant records in the raw dataset was 19.8%, the grid cells that reached the threshold (e.g., 20 plant records by cell) for the sample coverage analysis were only 2.8%. The curated dataset had 40.9% of the grid cells with plant records, but in the curated dataset only 8.8% were valid for this analysis. The total dataset had 41% of grid cells with plant records, but only 10.3% were valid for sample coverage. However, the sample coverage values between data stages did not show significant differences ($p > 0.05$). Median values on the grid cells were 0.22 in raw dataset, 0.21 in curated dataset and 0.24 in the total dataset. The 75% of grid cells showed sample coverage values below 0.5 with a maximum of 0.6 in the raw dataset, while for the curated and total, the maximum values were 0.91 (Fig. 5). Significant differences were found in the grid cells where fieldwork was done ($p > 0.05$). During the time of the study, only ten grid cells had fieldwork. Nevertheless, those grid cells showed a significant sample coverage increase by fieldwork. In those same grid cells, the sample coverage in the raw and curated datasets did not show significant differences (0.25 each), but in the total dataset (clean – fieldwork) the sample coverage increased to 0.68, showing the input of fieldwork.

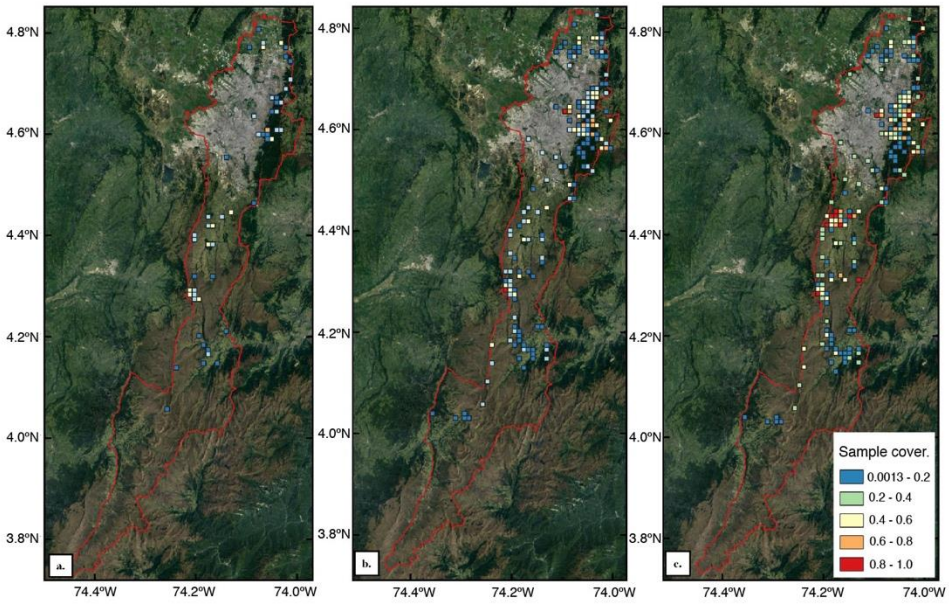


Figure 5. Sample coverage variation by curatorial and fieldwork in the vascular plant dataset of Flora of Bogotá: (a) raw dataset, (b) curated dataset, and (c) total dataset (combination of curated and fieldwork datasets). The blue – red scale indicates the spatial variation of record density in every stage of data: low density is indicated in blue while high density is indicated in red. The red line delimits the area of Bogotá.

At the ecosystem level, the sample coverage mainly increased in the curated dataset with a slight increase in fieldwork dataset. In only one ecosystem (C3), fieldwork's sample coverage increase was higher than in the curated dataset. Although the overall sample coverage of ecosystems was improved in the curated dataset compared to fieldwork, non-significant differences were observed between datasets ($p < 0.05$) (Fig. S4).

Discussion

In this study, we analyzed the change in magnitude of floristic knowledge and information quality of Bogotá city through both curatorial (nomenclatural cleaning and georeferencing process of plant records) and fieldwork in a time window between 2012 – 2016. These two activities were done simultaneously in the first stage of the flora de Bogotá project. We also evaluate their impact on the taxonomic, geographical, richness and sample coverage aspects of biodiversity information. We found the highest change in the Bogotá's floristic data was due to curatorial processes rather than fieldwork. There was a decrease in alpha diversity through taxonomic curatorial work, because of synonyms or orthographical mistakes in the species names. At the same time, there was an increase in spatial coverage because of geographical curatorial work that increased the number of plant records with georeferences.

Taxonomic changes resulting from curatorial and fieldwork

Our work significantly improved the taxonomic quality of the Flora of Bogotá database, which decreased the number of species names by 24%, by removing synonyms and orthographical errors. The loss of names is not surprising on account of different data sources with distinct curatorial levels, which is also evident in open databases such as GBIF, where much inaccurate and incorrect information is published (Maldonado *et al.*, 2015). Different classification systems were found in the Flora of Bogotá database that refer to the same taxon with a different name (e.g., Compositae = Asteraceae, Palmae =Arecaceae, and Gramineae = Poaceae) that together with the orthographical mistakes, had inflated the alpha diversity of Bogotá. Although checking for nomenclatural issues is the first step in obtaining an accurate list of species for a region, this is not obvious, since local and regional species inventories are full of nomenclatural mistakes, artificially increasing the diversity. For example Cardoso *et al.* (2017) indicate that for the Amazon basin, almost

7% of species reported were mistakes, because individual species were listed more than once as synonyms and spelling variants. The problem is worsened because much information reported in open databases is not reviewed by experts (Goodwin *et al.*, 2015) and does not utilize up to date nomenclature.

On the other hand, the contribution to Bogotá plant species richness by fieldwork was small (only 3% of new species for the Bogotá species list), compared with the data obtained from collection databases. Several factors contribute to the low rate of new species records obtained by fieldwork. For instance, all the analyzed fieldwork collections were based on biased sampling towards grid cells that were already intensively sampled. Additionally factors related with the low rate of detection of new species in the Bogotá area are also related with collector expertise (Ahrends *et al.*, 2011), the sensitivity of sampling methods that exclude some groups of plants, preference (e.g., preference for angiosperms against ferns or non-vascular plants) (Daru *et al.*, 2018), detectability of plants (e.g., that depend on phenology, life form) (Chen *et al.*, 2009), species density (McCarthy *et al.*, 2013), and sampling bias.

This study found that fieldwork had limited reach compared to the data from collections, which resulted from the combined efforts of multiple explorers and researchers over a long period of time. The data registered in collections reflects contributions from many collectors and the exploration of diverse locations. As a result, improving the data associated with existing herbarium specimens through curatorial work may be more effective in producing a more accurate representation of the flora. This can be achieved by adding new collection sites and enhancing the representation of species' climatic niches (Feeley and Silman, 2011a), as well as expanding geographical coverage. Alternatively, the

curated information could assist in planning efficient fieldwork strategies for areas and plant groups with limited data.

Geographical changes resulting from curatorial and fieldwork

Spatial and environmental changes. Our study showed an important increase in the spatial coverage of the Flora through curatorial work. The geographical dimension of biological records has been an important issue since most biological records, especially old ones (e.g., collections before 1990 were GPS was not popular) (Feeley & Silman, 2010) are deposited in collections without coordinates. As a result, many records are discarded from ecological analyses, and these could represent new areas and environmental combinations. Curatorial work allowed the recovery of important floristic information such as ecosystems not previously represented in Bogotá, that were revealed through georeferencing (e.g., La Regadera lake). More of the environmental and climatic spectrum of the Flora of Bogotá not represented by the non-curated raw data, were elucidated through georeferencing (clean dataset). In contrast, fieldwork was carried out in ecosystems and grid cells that had already been sampled before, resulting in a low number of new samples. This finding suggests that there were flaws in the sampling design at the regional scale, as supported by previous information.

Taxonomic perspective changes. The recovery of data from collections could improve species niche information by adding new environmental variables not previously recorded, information that would help to improve species distribution models (Feeley & Silman, 2011a). As expected, our work improved species distribution data. However, we found that the species that increased the number of distinct localities (new species records in new grid

cells) by curatorial and fieldwork, are the most common ones. In contrast, for many species, particularly rare ones, new localities were not added after cleaning and fieldwork. It is possible that rare species are not represented in collections due to low detectability of species or collection preferences. However curatorial work could help identify rare species and with this information, focus on targeted fieldwork and add new environment information.

Richness perspective. Although the number of species names in the database decreased as a result of curatorial work, those corrections have improved the taxonomic quality, resulting in a more reliable species list for the city. Taxonomic corrections increased the number of records for some species, increasing their range of geographic distribution. Fieldwork added new species to the Flora checklist, but the increase was low (e.g., 83 species that correspond to 3.5% of the total species diversity). The low rate of new species recorded by fieldwork could be due to an already exhaustive sampling of Bogotá. However, our analysis showed low sampling rates in the grid cells with plant records and a high proportion of grid cells without records (e.g., 59% of grid cells of Bogotá). Spatial sampling bias, collection preferences of some taxonomic groups (e.g., collections made by experts in certain groups that prefer angiosperms to ferns), and collector expertise could explain the low rate of recorded novel species (e.g., new species (undescribed) or new species records (new species for the area)). In our case, we found biased sampling around the urban area of Bogotá city, especially “Cerros Orientales” and some places such as the páramo of Sumapaz (e.g., Laguna de Chisacá and Nazareth). Only in one place, “Páramos de Pasquilla”, where the Flora of Bogotá project undertook fieldwork intensively, did the

sampling increase significantly, and the observed richness and the completeness increased above 40% with grid cells over 80%.

Locally (e.g., grid cells), curatorial work significantly improved the number of species observed in the grid cells with plant records. On the other hand, 50% of new grid cells were represented by plant records increasing the number of species observed in areas without plant information. However, many grid cells did not suffer changes in density and observed richness, especially those far away from the urban area where the lack of access, and social conflicts (e.g., guerrillas presence) can make it difficult for them to be reached (Negret *et al.*, 2017).

Georeferencing increased the coverage of plant records, the number of species observed and sample coverage at local and regional scales in Bogotá. On the other hand, fieldwork (Fig. 5c) made significant changes at the local scale. For example, in the grid cells where fieldwork was performed, few plant records were recorded at the raw data stage, and few were recovered by curatorial work. After fieldwork, completeness in those grid cells increased significantly, reaching values above 0.8. Sample coverage analysis and richness estimators depend on sample size (Gotelli & Colwell, 2011). Although the number of plant records increased at a regional scale by curatorial work, at the local scale (e.g., grid cells), 80% of grid cells had less than 10 plant records (Fig. 4) and only 20% of grid cells had more than 20 plant records (threshold used to calculate sample coverage). The main changes were observed on the grid cells where fieldwork was undertaken.

Our study fixed taxonomic and geographical issues in the data recovery information that increased richness observed and completeness locally. Given the limited resources to explore the territory, especially in low or middle-income countries, it is essential to carefully invest those scarce resources in order to obtain as much information as possible.

Biological collections contain information that recompiles the efforts of several researchers and projects through the years. As many researchers have pointed out physical collections have vast amounts of data that is not usable because of issues in three basic dimensions (taxonomic, geography and time) (Lavoie, 2013; Feeley, 2015; Hortal *et al.*, 2015; Meyer *et al.*, 2016). As we showed in this study, investing in curatorial work (e.g., physical and digital) as the first step of describing biodiversity could unveil those aspects that are necessary to make the use of few resources for biodiversity studies most efficiently. Fieldwork is a crucial activity to study biodiversity but should be targeted to under-collected areas that can be defined by improving collections data through curatorial work.

However, despite the great efforts that have been made with digitalization further improvements could be made that would enhance biodiversity studies. Small herbaria such as that of the Jardín Botánico de Bogotá would also benefit from citizen science contributions. Label digitization, for example, could allow data to be read from home by retirees. At the Jardín Botánico de Bogotá herbarium volunteer work contributes to scanning and photographing the specimen collection, as well as with mounting. We believe that by being involved in these activities, it is possible to inspire new taxonomists who will contribute to enhancing herbarium research. Research could also be assisted by new technologies associated with Artificial Intelligence approaches that can check the consistency of identifications and indicate which specimens are problematic requiring expert review (see for example, Hussein *et al.*, 2022). All of these approaches can contribute to the concept of the 'global meta-herbarium', linking digitized specimens with other digital data (Davis, 2023).

Conclusions

Curatorial work of biodiversity collections and fieldwork are not distinct processes, but rather exist in continuous feedback that generates and improves biodiversity knowledge. Therefore, to maximize the scarce resources invested by research organizations and institutions in biodiversity, it is crucial that this circular process continually informs the subsequent steps. From the point of view of research institutes in charge of biodiversity characterization, curatorial work, facilitated by digitalization, is an investment that would offer a large amount of improved data that could be retrieved from biological collections at relatively low cost and requiring little time (Suarez & Tsutsui, 2004; Lavoie, 2013). However, many herbaria have decreased the investment in curators and care for collections decreases every day (Vogel *et al.*, 2017), especially in local and small ones.

In contrast fieldwork requires high investment in personnel, logistics, preparation and time (Suarez & Tsutsui, 2004) with limited capacity to capture new data. Although fieldwork remains essential for acquiring biodiversity information, its significant investment underscores the importance of maintaining constant feedback between curatorial work and fieldwork. This enables better identification of collection areas or taxonomic groups that require further investigation. In order to improve baseline biodiversity information, we advocate for increased investment in curation and maintenance of herbaria, both in terms of trained personnel and infrastructure. This investment is particularly necessary in smaller, local herbaria that have been shown to contain important collections that contribute to a better understanding of the overall distribution of biodiversity (e.g., Marsico *et al.*, 2020; Monfils *et al.*, 2020).

References

Ahrends A, Rahbek C, Bulling MT, Burgess ND, Platts PJ, Lovett JC, Kindemba VW, Owen N, Sallu AN, Marshall AR, Mhoro BE, Fanning E, Marchant R. 2011. Conservation and the botanist effect. *Biological Conservation* 144: 131–140.

Ball-Damerow JE, Brenskelle L, Barve N, Soltis PS, Sierwald P, Bieler R, LaFrance R, Ariño AH, Guralnick R. 2019. Research applications of primary biodiversity databases in the digital age. *bioRxiv*: 1–26.

Bebber DP, Carine M a., Wood JRI, Wortley a. H, Harris DJ, Prance GT, Davidse G, Paige J, Pennington TD, Robson NKB, Scotland RW. 2010. Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences* 107: 22169–22171.

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

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* 114: 10695–10700.

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

Chen G, Kéry M, Zhang J, Ma K. 2009. Factors affecting detection probability in plant distribution studies. *Journal of Ecology* 97: 1383–1389.

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.

Elith J, H. Graham C, P. Anderson R, Dudík M, Ferrier S, Guisan A, J. Hijmans R, Huettmann F, R. Leathwick J, Lehmann A, Li J, G. Lohmann L, A. Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, McC. M. Overton J, Townsend Peterson A, J. Phillips S, Richardson K, Scachetti-Pereira R, E. Schapire R, Soberón J, Williams S, S. Wisz M, E. Zimmermann N. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.

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.

Feeley KJ. 2012. Distributional migrations, expansions, and contractions of tropical plant species as revealed in dated herbarium records. *Global Change Biology* 18: 1335–1341.

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. 2010. Modelling the responses of Andean and Amazonian plant species to climate change: The effects of georeferencing errors and the importance of data filtering. *Journal of Biogeography* 37: 733–740.

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

1140.

Feeley KJ, Silman MR. 2011b. The data void in modelling current and future distributions of tropical species. *Global Change Biology* 17: 626–630.

Franco AMA, Palmeirim JM, Sutherland WJ. 2007. A method for comparing effectiveness of research techniques in conservation and applied ecology. *Biological Conservation* 134: 96–105.

Gaira KS, Dhar U, Belwal OK. 2011. Potential of herbarium records to sequence phenological pattern: A case study of Aconitum heterophyllum in the Himalaya. *Biodiversity and Conservation* 20: 2201–2210.

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.

Gardner TA, Barlow J, Araujo IS, Ávila-Pires TC, Bonaldo AB, Costa JE, Esposito MC, Ferreira L V., Hawes J, Hernandez MIM, Hoogmoed MS, Leite RN, Lo-Man-Hung NF, Malcolm JR, Martins MB, Mestre LAM, Miranda-Santos R, Overal WL, Parry L, Peters SL, Ribeiro MA, Da Silva MNF, Da Silva Motta C, Peres CA. 2008. The cost-effectiveness of biodiversity surveys in tropical forests. *Ecology Letters* 11: 139–150.

Goodwin ZA, Harris DJ, Filer D, Wood JRII, Scotland RW. 2015. Widespread mistaken identity in tropical plant collections. *Current Biology* 25: R1066–R1067.

Gotelli NJ, Colwell RK. 2011. Estimating Species Richness. *Biological diversity. frontiers in measurement and assessment*. New York: Oxford University press, .

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT. 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution* 19: 497–503.

Gueta T, Carmel Y. 2016. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological Informatics* 34: 139–145.

van der Hammen T. 1986. La Sabana de Bogotá y su lago en el Pleniglacial Medio. *Caldasia* 15: 249–262.

Hopkins MJG. 2007. Modelling the known and unknown plant biodiversity of the Amazon Basin. *Journal of Biogeography* 34: 1400–1411.

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.

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.

Lavoie C. 2013. Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics* 15: 68–76.

Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N, Antonelli A. 2015. Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases? *Global Ecology and Biogeography* 24: 973–984.

Marsico TD, Krimmel ER, Carter JR, Gillespie EL, Lowe PD, McCauley R, Morris AB, Nelson G, Smith M, Soteropoulos DL, Monfils AK. 2020. Small herbaria contribute unique biogeographic records to county, locality, and temporal scales. *American Journal of Botany* 107: 1577–1587.

McCarthy MA, Moore JL, Morris WK, Parris KM, Garrard GE, Vesk PA, Rumpff L, Giljohann KM, Camac JS, Bau SS, Friend T, Harrison B, Yue B. 2013. The influence of

abundance on detectability. *Oikos* 122: 717–726.

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

Monfils AK, Krimmel ER, Bates JM, Bauer JE, Belitz MW, Cahill BC, Caywood AM, Cobb NS, Colby JB, Ellis SA, Krejsa DM, Levine TD, Marsico TD, Mayfield-Meyer TJ, Miller-Camp JA, Gil Nelson RM, Phillips MA, Revelez MA, Roberts DR, Singer RA, Zaspel JM. 2020. Regional collections are an essential component of biodiversity research infrastructure. *BioScience* 70: 1045–1047.

Morueta-Holme N, Engemann K, Sandoval-Acuña P, Jonas JD, Segnitz RM, Svenning JC. 2015. Strong upslope shifts in Chimborazo's vegetation over two centuries since Humboldt. *Proceedings of the National Academy of Sciences* 112: 12741–12745.

Negret PJ, Allan J, Braczkowski A, Maron M, Watson JEM. 2017. Need for conservation planning in postconflict Colombia. *Conservation Biology* 31: 499–500.

Nualart N, Ibáñez N, Soriano I, López-Pujol J. 2017. Assessing the Relevance of Herbarium Collections as Tools for Conservation Biology. *Botanical Review* 83: 303–325.

O'Connell AF, Gilbert AT, Hatfield JS. 2004. Contribution of natural history collection data to biodiversity assessment in national parks. *Conservation Biology* 18: 1254–1261.

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

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

Secretaría Distrital de Ambiente. 2007. *Atlas ambiental de Bogotá DC. Imprenta Nacional de Colombia.* Bogota (Colombia).

Soberón J, Jiménez R, Golubov J, Koleff P. 2007. Assessing completeness of biodiversity

databases at different spatial scales. *Ecography* 30: 152–160.

Soberón J, Peterson AT. 2004. Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B: Biological Sciences* 359: 689–698.

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.

Suarez A V., Tsutsui ND. 2004. The Value of Museum Collections for Research and Society. *BioScience* 54: 66.

Syfert MM, Smith MJ, Coomes DA. 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLoS ONE* 8.

Targetti S, Herzog F, Geijzendorffer IR, Wolfrum S, Arndorfer M, Balàzs K, Choisis JP, Dennis P, Eiter S, Fjellstad W, Friedel JK, Jeanneret P, Jongman RHG, Kainz M, Luescher G, Moreno G, Zanetti T, Sarthou JP, Stoyanova S, Wiley D, Paoletti MG, Viaggi D. 2014. Estimating the cost of different strategies for measuring farmland biodiversity: Evidence from a Europe-wide field evaluation. *Ecological Indicators* 45: 434–443.

Vogel C, Bordignon SA de L, Trevisan R, Boldrini II. 2017. Implications of poor taxonomy in conservation. *Journal for Nature Conservation* 36: 10–13.

Wieczorek J, Guo Q, Hijmans RJ. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18: 745–767.

FINAL CONCLUSIONS

The study of biodiversity can be based on biological records, represented as specimens in biological collections. Each of these records attest to the presence of a species in a particular place and time. The set of records makes it possible to establish the state of biodiversity at a local, regional and global level and study its variation over time. With the development and implementation of computer tools, the biological information that remained physical on the specimen tags has been digitized and mobilized to databases of the collections and repositories of biodiversity information (e.g., GBIF). This advance has allowed it to consolidate large volumes of data with which biodiversity-related processes are described and analyzed at different scales.

Despite the large volume of information available, only a fraction is usable, given the multiple deficiencies in the source data of the records (temporal, geographic, taxonomic). In the case of the information on the Colombian Andean flora available in databases, the data from the recovered records showed multiple deficiencies that resulted in a low proportion of usable data. In my dissertation, the geographic variable was the one that contributed the most to discarding of records, in contrast with the taxonomic component.

The analysis of the recovered and available information showed that the floristic record of the Andes does not adequately represent its environmental and spatial variability, with the high Andean and páramo fringes close to the main cities where the highest proportion of records is concentrated. However, despite the records concentration, the region's richness estimation models show low completeness values at different spatial scales, even in the best-represented areas.

Although the concentration of records in the Colombian Andean region is clear, there is no formal analysis of its causes, an aspect that is relevant to reduce the factors that affect sampling bias and guiding strategies aimed at completing the missing information, emphasizing zones and regions poorly known or poorly represented in their physical-environmental combinations (e.g., dry forests). However, whatever the factors causing bias, this study shows the administrative units (e.g., departments, protected areas) and environmental combinations (e.g., strips below 2,000 meters of altitude, high humidity) where it is necessary to increase the information available. Some of these administrative units are protected areas where information is marginal and where a high proportion of undescribed or unregistered species is presumed.

On the other hand, this study also revealed taxonomic deficiencies, with only a fraction of species with sufficient records to implement predictive distribution or threat status categorization models with any confidence. For most Andean plant species, the information needs to be more comprehensive; for several species (ca. 33%), no records are available in databases.

Based on the above, it is essential to consider that the analyses presented here are based on a fraction of the records publicly available. As already noted, a high proportion of records were useless due to deficiencies in geographic and/or taxonomic information. In this sense, I have shown that the work of taxonomic and geographic curation of the records has a high potential to fill gaps (Chapter 4), increasing the number of available records of the species, but also increasing the spatial coverage of the floristic record.

Considering the accelerated rate of transformation and disappearance of Andean ecosystems, it is essential to promptly and systematically increase the information available

on the flora of this region, for the management and conservation of Andean ecosystems. This is why the next step, after this dissertation, would be to consolidate and recover records available in physical and local collections, while promoting explorations to little-known areas. These poorly represented areas and taxonomic groups are described in Chapters 2 and 3 (e.g., humid forests of the foothills of the western and eastern cordilleras; Orchididacae, epiphytes).

This study opens perspectives for the study of multiple aspects of the biology of the species (e.g., phenology) and macroecological processes in the region based on information from collections that must be made available through digital means that allow access to multiple researchers.

SUPPLEMENTARY MATERIAL – ALL CHAPTERS

Note. Given the size of several documents, additional supplementary material can be found in the following link: https://uredu-my.sharepoint.com/:f:/g/personal/carlosalbe_vargas_urosario_edu_co/EnU3W0asgoNIgJKDHYKwjk4BsjZalGU8obsxNBvWUS9-6w?e=oRgZJB

Chapter 2: Supplementary Material

Table S1. Plant record sources used to analyze bias in the Colombian Andean flora.

Data source	Record number
GBIF	1,697,974
Institute of Natural Sciences Herbarium	345,470
Tropicos	190,022
Jardín Botánico de Bogotá	32,670
Total plant records	**2,266,136**

Table S2. Plant records used to analyze gaps in the Colombia's Andean flora.

Description of plant records	Record number	%
Plant records in Flo_RA database	2,266,136	100.0
Plant records with coordinates	1,618,079	71.4
Plant records with coordinates in Colombia	1,594,912	70.4
Plant records with coordinates in the Colombian Andean region	759,551	33.5
Plant records with species name	1,824,241	80.5

Plant records with species name's synonyms	187,978	8.3
Plant records with coordinates in the Colombian Andean region and with species name	575,040	25.4
Duplicated records	308,415	13.6
Total plant records analyzed	**266,625**	**11.8**

Table S3. Plant records from Colombia gathered from GBIF**. Collection (Coll).**

Herbarium	**Country**	**Coll focus**	**Coll code**	**Total**
Herbario Nacional Colombiano	COL	Colombia	COL	384,920
Instituto de Investigación de Recursos Biológicos Alexander von Humboldt	COL	Colombia	FMB	127,181
Instituto amazónico de investigaciones cientificas_SINCHI	COL	Región amazónica de Colombia	COAH	91,099
Universidad de Antioquia	COL	Andes; Colombia	HUA	84,155
Smithsonian Institution	US	Worldwide	US	63,790
Field Museum of Natural History	US	Worldwide	Botany	37,428
Pontificia Universidad Javeriana	COL	Colombia Orchids	HPUJ	23,805
Universidad de Nariño	COL	Nariño, Putumayo, and southern Cauca	PSO	18,029
Universidad Industrial de Santander	COL	Santander	HUIS	14,407
The New York Botanical Garden	US	Worldwide with greatest strength in tropical America and North America	NY	14,324
Real Jardín Botánico Madrid_España	ES	Worldwide, with greatest strength in Western Mediterranean, Central and South America, Africa, Australia and New Zealand	MA	11,344
Muséum National d'Histoire Naturelle	FR	Worldwide	P	10,174
Universidad de La Salle	COL	Colombia	BOG	8,383

Herbarium	Country	Coll focus	Coll code	Total
Universidad Distrital "Francisco José de Caldas"	COL	Colombia	UDBC	7,000
Royal Botanic Gardens	UK	Worldwide	K	6,648
Universidad del Valle	COL	Neotropics, especially western Colombia.	CUVC	5,437
Universidad Catolica de Oriente	COL	Antioquia oriente	HUCO	4,500
Universidad de Antioquia	COL	Colombia Antioquia	HUA	4,175
Jardín Botánico de Cartagena "Guillermo Piñeres"	COL	Colombia Caribbean Coast	JBGP	3,752
Swedish Museum of Natural History	SWEDEN	Worldwide	SPA	3,604
Jardín Botánico Juan Maria Cespedes	COL	Valle de Cauca	TULV	2,091
Universidad de Córdoba	COL	Colombia, especially Córdoba, Sucre, and northern Antioquia	HUC	2,021
Universidad del Tolima	COL	Colombia; eastern llanos; Caquetá; Tolima; Nevado del Ruíz; Nevado del Tolima; Amazon	TOLI	2,009
Universidad ICESI	COL	Southwestern Colombia, especially the western and central cordilleras in the Departments of Caldas, Risaralda, Quindío and Valle del Cauca. Dry forest, montane forests of the tropical Andes and páramo	ICESI	2,007

Herbarium	Country	Coll focus	Coll code	Total
Museu Botânico Municipal Brazil-Paraná	BR	Worlwide	MBM	1,942
Botanischer Garten und Botanisches Museum Berlin, Zentraleinrichtung der Freien Universität Berlin	GE	Worldwide	B	1,888
Royal Botanic Garden Edinburgh	UK	Southwestern and southeastern Asia, Arabia, Turkey, Bhutan, Brazil, Britain, China, Himalayas, Mediterranean, Chile, Argentina, and southern Africa	E	1,339
Universidad de los Andes	COL	Colombia; neotropics; Páramos; Amazon	ANDES-H	1,137
Université de Montpellier	FR	Worldwide	MPU	1,121
Universidad de la Amazonia	COL	Amazonas	HUAZ	988
Fundación Jardín Botánico Joaquín Antonio Uribe de Medellín	COL	Antioquia, especially regions of Urabá, central Magdalena Valley, National Natural Park Las Orquídeas, and Río Claro municipality San Luís; Chocó ecoregion; Amazonía, Caribbean region in Colombia; Department of Nariño.	JAUM	628
Naturhistorisches Museum Wien	AUSTRIA	Worldwide	W	508
Universität Wien	AUSTRIA	Worldwide	WU	121

Herbarium	Country	Coll focus	Coll code	Total
Missouri Botanical Garden	US	Worldwide	MO	64
Karl-Franzens-Universität Graz	AUSTRIA	Worldwide	GZU	57
Universidad del Quindío	COL	Colombia, especially the Andean region	HUQ	42
University of Kansas	US	Kansas; Great Plains; western United States	KANU	32
Jardim Botânico do Rio de Janeiro	BR	Worldwide, especially Brazil	RB	4
Total number of records of Colombian plants in herbaria registered in GBIF				**942,154**
Total number of records of plants in Colombia registered in GBIF				**1,697,974**

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.

Protect Area in Colombian Andes	Category	Area (km^2)	# records	Rec/km^2
Isla de la Corota	SFF	0.16	190	1187.50
Otún Quimbaya	SFF	4.51	225	49.89
Iguaque	SFF	68.88	1155	16.77
Los Estoraques	ANU	6.35	95	14.96
Las Orquídeas	PNN	290.76	3647	12.54
Cueva de los Guácharos	PNN	71.33	855	11.99
Galeras	SFF	83.29	707	8.49
Chingaza	PNN	771.70	5198	6.74
Tayrona	PNN	97.90	623	6.36
Guanentá Alto Río Fonce	SFF	102.65	540	5.26
Los Nevados	PNN	613.88	1731	2.82
Sumapaz	PNN	2145.40	5182	2.42
Tatamá	PNN	434.98	483	1.11
Munchique	PNN	481.08	526	1.09

Protect Area in Colombian Andes	Category	Area (km^2)	# records	Rec/km^2
Puracé	PNN	899.23	914	1.02
El Cocuy	PNN	3017.53	2943	0.98
Selva de Florencia	PNN	100.13	54	0.54
Pisba	PNN	355.08	168	0.47
Sierra Nevada de Santa Marta	PNN	4013.98	1895	0.47
Alto Fragua Indiwasi	PNN	593.39	273	0.46
Los Farallones de Cali	PNN	1739.14	536	0.31
Tamá	PNN	510.21	147	0.29
Nevado del Huila	PNN	1649.52	438	0.27
Serranía de los Churumbelos	PNN	971.30	175	0.18
Plantas Medicinales Orito Ingi Ande	SF	104.01	15	0.14
Paramillo	PNN	5322.87	613	0.12
Serranía de los Yariguies	PNN	596.92	54	0.09
Complejo Volcánico Dona Juana Cascabel	PNN	658.57	56	0.09
Catatumbo Barí	PNN	1607.85	58	0.04
Las Hermosas	PNN	1248.83	12	0.01

6

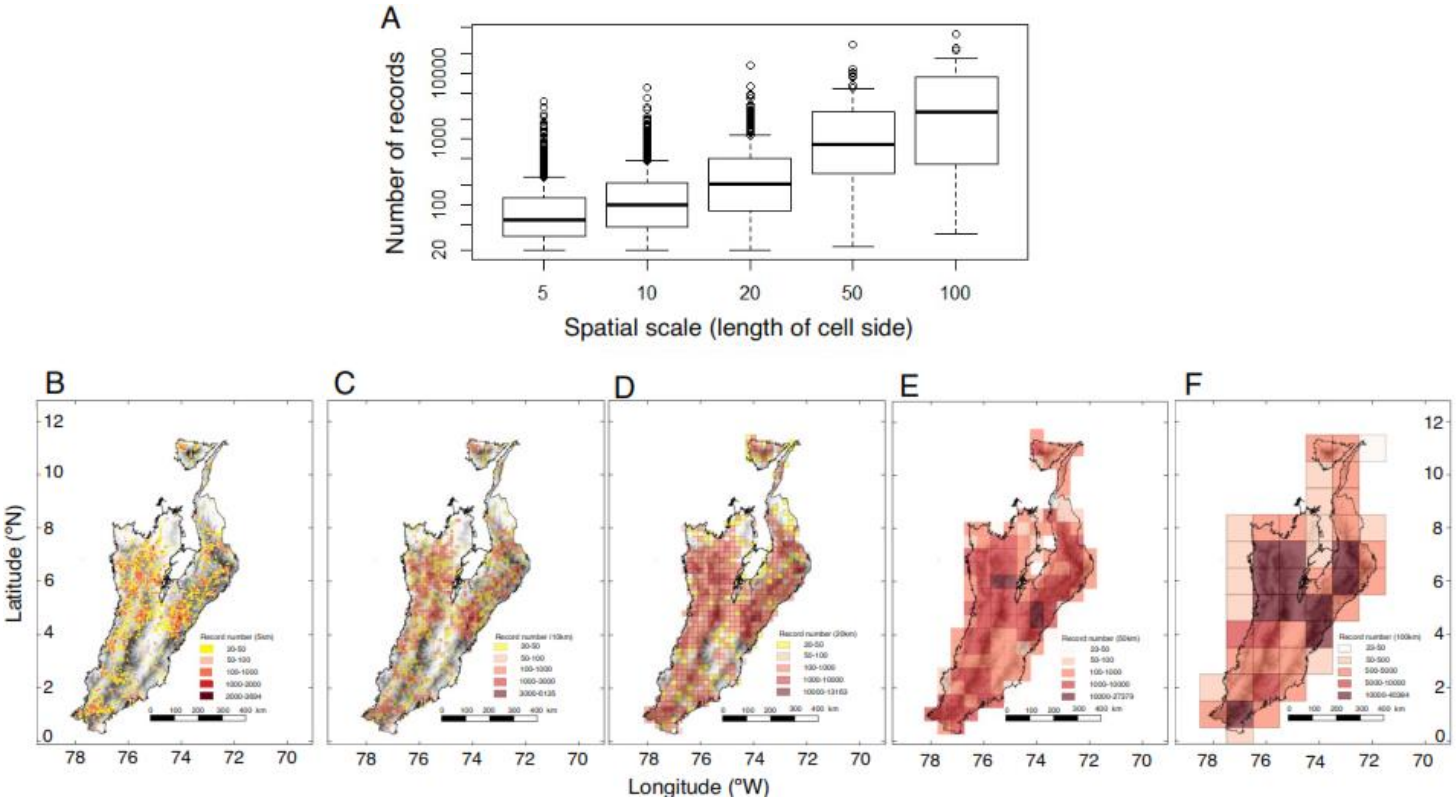


Figure S1. Collection density (i.e., number) of digitally available plant specimen records across the Colombian Andes at different grid cell sizes. A, Boxplots of the number of plant specimen records by grid cell across different scales (5 x 5, 10 x 10, 20 x 20, 50 x 50 and 100 x 100 km). The bottom and top part of the boxplot indicates the 25th and 75th percentile (respectively), the horizontal line within the box, the median value and the circles, 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 x 100 km (B), 50 to 50 x 50 (C), 20 to 20 x 20 km, 10 to 10 x 10 km and 5 to 5 x 5 km.

Chapter 3: Supplementary Material

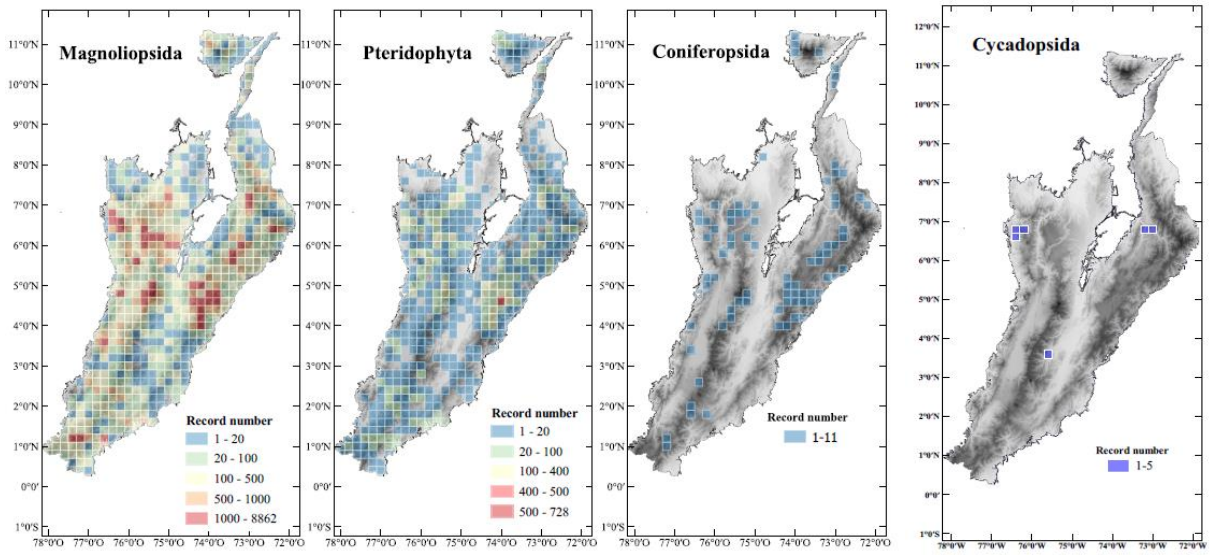


Figure S1a. Distribution of the occurrence records of the main groups of vascular plants across the Colombian Andes at 20 x 20 km cell size.

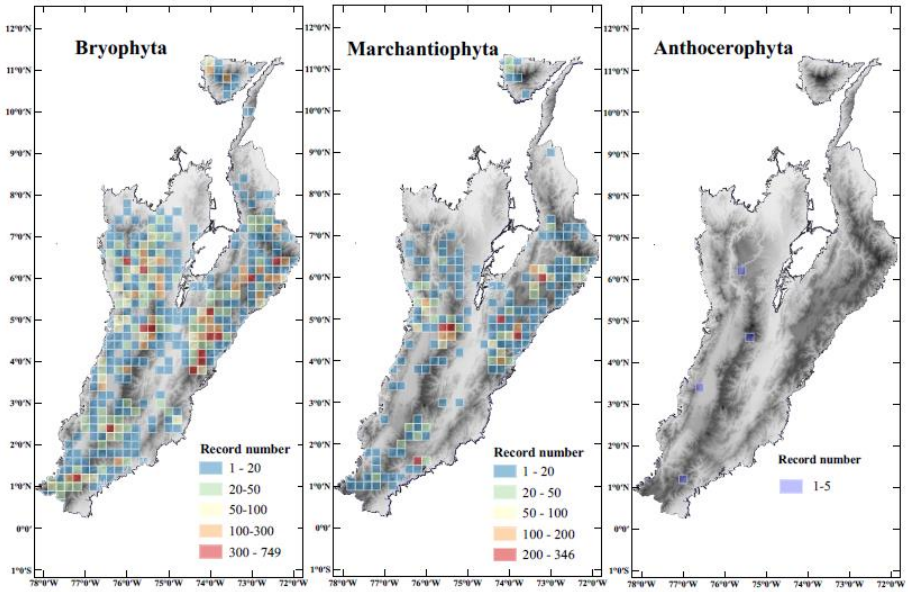
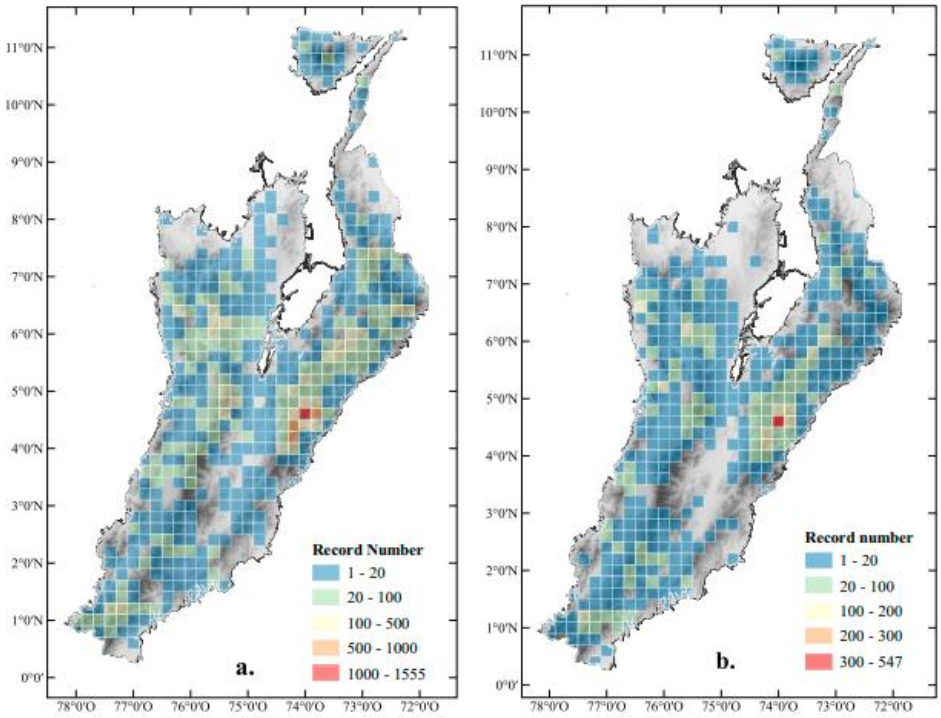


Figure S1b. Distribution of the occurrence records of the main groups of non-vascular plants at phylum level across the Colombian Andes at 20 x 20 km cell size.



24 **Figure S2.** Spatial distribution of Asteraceae (a) and Orchidiaceae (b) in the Colombian

25 Andes.

26

27

Chapter 4. Supplementary Material

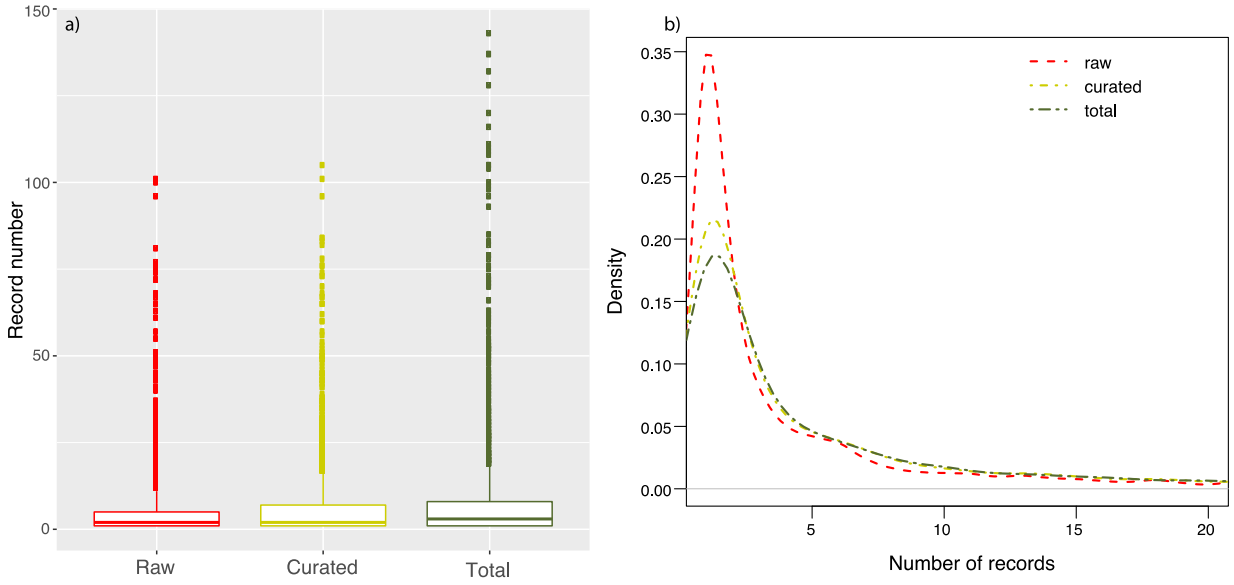


Figure S1. Boxplot (a) and plot of density function (b) comparing the number of species records at the three stages of data cleaning: 1. Raw dataset (red line); 2. Curated dataset, 3. Total (curated – fieldwork) dataset (dark green line).

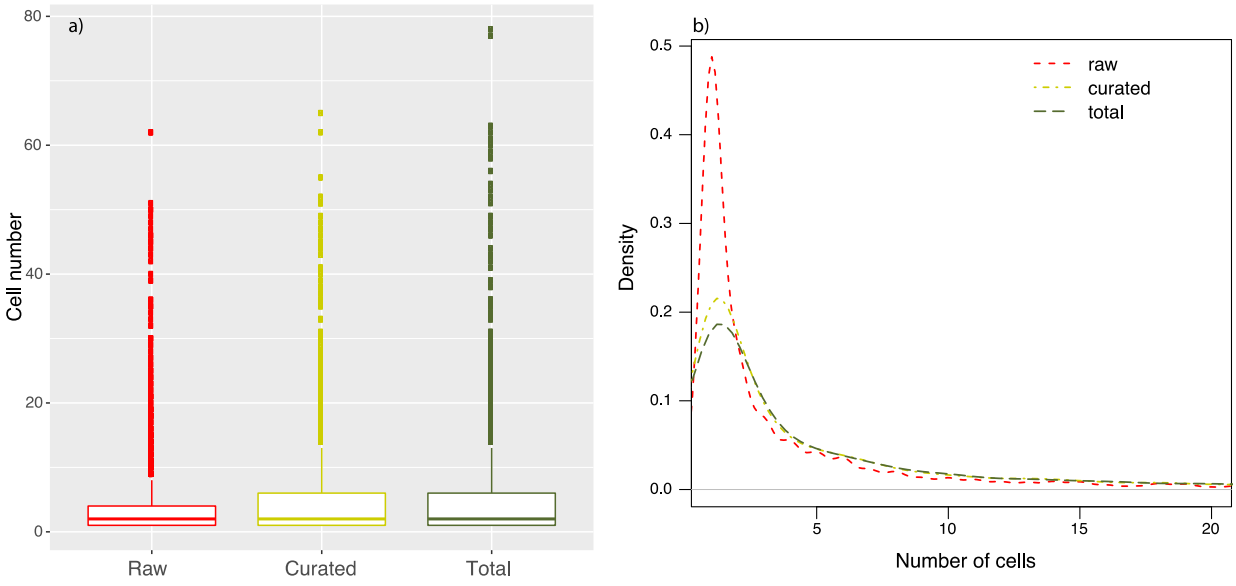


Figure S2. Boxplots (a) showing the numbers of grid cells of 1 x 1 km where species has been recorded in raw dataset (red), curated dataset (yellow) and total dataset (green). The probability function (b) shows the decrease in the probability of finding species in one cells from raw (red line) to curated (yellow line) and total dataset (dark green line).

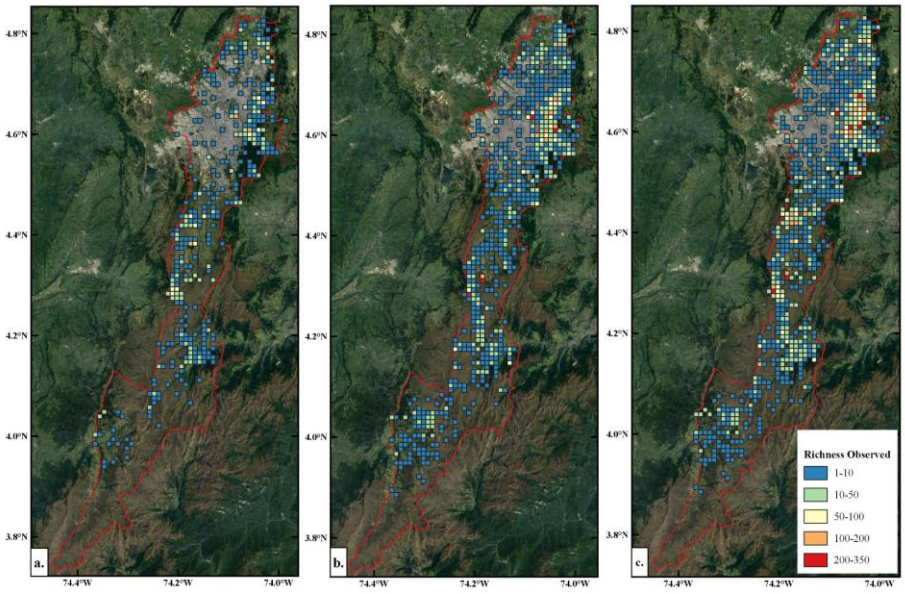


Figure S3. Variation in richness observed by curatorial and fieldwork: (a) raw dataset, (b) curated dataset, (c) total dataset (combination of curated and fieldwork datasets). The blue – red scale indicates the spatial variation of observed richness in every stage of data: low richness is indicated in blue while high richness is indicated in red. The red line delimits the area of Bogotá.

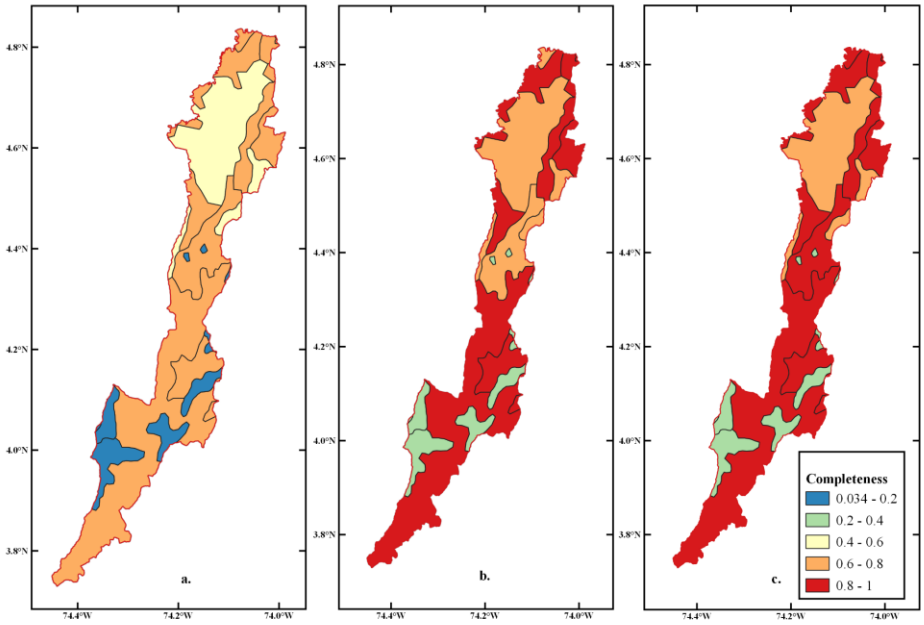


Figure S4. Ecosystem completeness variation by curatorial and fieldwork in the three stages of data: (a) raw data, (b) clean data, (c) total dataset (combination of curated and fieldwork datasets). The blue – red scale indicates the completeness ecosystem variation in every stage of data: low completeness is indicated in blue while high completeness is indicated in red. The red line delimits the area of Bogotá.

ACKNOWLEDGEMENTS

The study of the floristic diversity of the Andes of Colombia has been an important part of my scientific development. Understanding the patterns and causes of the diversity of this complex area have led me to implement new approaches for its study.

The doctorate at the Universidad del Rosario allowed this approximation through complex processes of reflection, analysis, trials, errors, questions and implementation of tools to approximate answers to the expression of diversity of the Colombian Andean flora. Although the research process seems solitary, it is not possible without the valuable contribution of people and entities that got involved and facilitated this reflection.

First of all, I would like to thank my supervisors, Dr. Adriana Sanchez and Dr. James Richardson, who believed in the project and its scope and allowed me to carry out the research independently but contributing with their valuable observations.

In the same way to Dr. Marius Bottin who during his postdoctoral stay at the Universidad del Rosario contributed notably to the consolidation of the Flo_RA database, assured and instructed in the use of bioinformatics tools for the analysis of information and that through the exchange of ideas contributed to the development of the research. To Dr. Carol X.Garzon-Lopez for her support in developing some of the analyzes in chapter 2 of the thesis.

A very special thanks to Dr. Tiina Sarkinen from the Royal Botanic Garden Edinburgh (Scotland) who was present at fundamental moments of the research contributing with her

valuable ideas and contributions to the documents and facilitated my doctoral stay in Edinburgh (Scotland).

I also thank the entities that shared her information for the development of the investigation. To the Jardín Botánico de Bogotá Jose Celestino Mutis - Flora de Bogotá research line its coordinators and professionals. Especially to Dra Marcela Celis and Diego Moreno. To the Institute of Natural Sciences of the University of Colombia, especially to Dr. Lauren Raz for sharing the data from the Herbario Nacional Colombiano collections and to Professor Orlando Rivera for his valuable observations during our pleasant talks. To the Missouri Botanical Garden for sharing data from their collections.

To the "evolutionary genetics, phylogeography and ecology of neotropical biodiversity group of the Universidad del Rosario for allowing the Flo-RA database to be hosted on their servers.

To my parents and siblings who have always been the fundamental support of my career. To Catalina who always believed in me and encouraged me even in the moments of greatest discouragement and to Ana Victoria whose presence illuminated this last year and a half of research.

A special mention to the usual friends "los Regios", "GRUFIT", Andres Orejuela and to my office colleagues, adventures and doctoral misfortunes Nicol Rueda, Carlos Pedraza, Osvaldo Gil and Carolina Alvarez.

Finally, this doctoral work was possible thanks to funds from MinCiencias (Colciencias call 727-2015) and the scholarship for doctoral students from the Universidad del Rosario.