Knowns and Unknowns: An Assessment of Knowledge Shortfalls in the Digitised Collection of Australia's Flora

---

Md. Mohasinul Haque

(MRes, MSc, BSc)

June 2018

Thesis submitted for the degree of Doctor of Philosophy

Department of Biological Sciences

Faculty of Science and Engineering

Macquarie University

MACQUARIE
University
SYDNEY·AUSTRALIA

TABLE OF CONTENTS

ABSTRACT

Massive digitisation of natural history collections (NHC), the predominant source of primary biodiversity data (i.e. species occurrence information), has provided myriad opportunities for studying biological diversity across space and time. Despite recent efforts to collate centuries of biodiversity inventories into comprehensive databases, these collections suffer inherent limitations in their spatial, temporal and taxonomic dimensions. Identifying these limitations is a priority to ensure that multiple targets specified by the Convention on Biological Diversity are met. In this thesis, which consists of four data chapters, I assess spatial, temporal and taxonomic patterns in the digitisation of data held within the Australasian Virtual Herbarium (AVH) – the largest electronic source of plant occurrence records in the country. In Chapter 2, I document spatial biases in the number of occurrence records from across Australia, with the Human Influence Index being a strong predictor of this bias. In Chapter 3, I demonstrate temporal biases, with 80% of records collected from 1970-1999. Furthermore, only 18% of the continent is represented by a relatively complete inventory consistently sampled over the last 200 years. I also found that around 25% of digitised specimens are missing key attribute information (i.e. collection date, taxonomic identification or geographic coordinates). An assessment of taxonomic bias in Chapter 4 indicates that, for one-third of Australia's plant families, the number of preserved specimens per family is not proportional to the family's known species richness. There is also a strong positive correlation between the number of collectors sampling a family and the taxonomic bias of that family. Finally, in Chapter 5, I demonstrate that digitisation effort over the last three decades varies significantly among Australia's herbaria: a time lag in digitisation means that only 30% of specimens are digitised within a year of collection. As the uses of primary biodiversity data continue to expand, my findings can direct future strategic sampling and digitisation efforts to increase our knowledge of Australia's flora.

AUTHORS DECLARATION

I hereby declare that the work presented in this thesis is my own unless otherwise stated or acknowledged. No part of this thesis has been submitted for any other degree, unless stated or acknowledged. The following lists the contributions made to the material in this thesis by the co-authors and collaborators:

Chapter 2. Published as**:** Haque M. M., Nipperess D. A., Gallagher R. V. & Beaumont L. J. (2017) How well documented is Australia's flora? Understanding spatial bias in vouchered plant specimens. *Austral Ecology* 42, 690-699.

Haque, Nipperess, Gallagher and Beaumont designed the research. Gallager helped with data retrival from the Australasian Virtual Herbarium. Data analysis was undertaken by Haque and Nipperess using custom R script. Haque prepared the manuscript with feedback and comments from co-authors. This chapter extends Haque's Master of Research Thesis, by a) focusing solely on vouchered specimens, b) assessing the relationship between sampling patterns and human influences and c) identifying regions likely to harbour species previously unrecorded within that region. Note, however, that the Introduction of Chapter 2 essentially remains very similar to the Introduction of the Master theses, which must be taken into consideration when examining the current body of work.

Chapter 3. Accepted as: Haque M. M., Nipperess D. A., Baumgartner J. B. & Beaumont L. J. (in press) A journey through time: Exploring temporal patterns among digitised plant specimens from Australia. *Systematics and Biodiversity* (accepted on 24 April 2018).

Haque, Nipperess and Beaumont designed research. Data analysis was undertaken by Haque with the assistance of Nipperess and Baumgartner, using custom R script. Haque prepared the manuscript with the feedback and comments from Beaumont, Baumgartner and Nipperess.

Chapter 4. Manuscript submitted as: Haque M. M., Beaumont L. J. & Nipperess D. A. (under review in *Biodiversity and Conservation*) Taxonomic shortfalls in digitised collections of Australia's flora.

Haque designed the research with advice from Beaumont and Nipperess. Haque also undertook data analysis with custom R script. Haque prepared the manuscript with the feedback and comments from Beaumont and Nipperess.

Chapter 5. Manuscript in prep. as: Haque M. M., Allen A. P., Klazenga N., & Beaumont L. J. Filling the gap: How quickly do Australia's herbaria digitise their vouchered specimens?

Haque designed the research, Klazenga provided guidance and expertise with data retrieval, while data analysis was undertaken by Haque and Allen with custom R script. Haque prepared the manuscript with the feedback and comments from Beaumont and Allen.

mohasin

Md. Mohasinul Haque

Date: 22nd January, 2019

ACKNOWLEDGEMENTS

First of all, I wish to thank my mentors — Dr. Linda Beaumont and Dr. David Nipperess. I am incredibly lucky to have dedicated mentors like you. My enthusiasm for work has just increased day by day during the last few years because of your care, patience, support and trust in me. Linda, I should reveal that you are the game changer in my research life. My perspectives on life and science have been refined in many ways in the last three years. I am privileged to have you as my mentor and friend. David, I have learnt from you how to dig deep into research. Thank you for your willingness to meet me at short notice every time.

I would like to thank iMQRES Scholarship Authority, Macquarie University, for giving me the opportunity to pursue my higher degree research.

Special thanks to Dr. Rachael Gallagher for your kind support in data retrieval and your external perspective on the early stage of my work has been helpful. Special thanks also goes to Dr. John Baumgartner for your support in developing R script, and I wish I could be a musician like you in R!

I am very thankful to my lab mates - Roni, Dina, Anindita, Felix, Nick, Elissa, and Hugh for their encouragement to keep me on the track. Roni thank you so much for your unconditional support in thesis formatting, and I will miss your company in playing with R coding! I am quite lucky to share my office room and spare time with such lovely mates - Lorraine, Maria, Laura, Katherine, Sajida and Robi. You people are awesome! My friend, Dr. Kinzang I will miss your company and treats at the MGSM café!

HDR Mentors, my extended family at Macquarie University, thank you very much for making this long-isolated journey colourful! Especially, I am grateful to Kim Tan, Head HDR Mentors and Dean, Prof. Nick Mansfield for giving me the opportunity to work as a peer mentor for the HDR community. I am also very thankful to Sally Purcell and Dr. Florence Chiew for your support in developing my transferable skills.

I would also like to thank the administration of the Department of Biological Sciences and the Faculty of Science and Engineering. Especially Julian, thanks much in helping me with the budget and leave approval while my mom was sick. I am also grateful to the FSE Faculty manager Dr. Jane Yang for her prompt action in approving my leave of absence while I had to fly to see my sick mother. I have thoroughly enjoyed the brief time with the 'Diversity and Inclusion' wing of the Department, and thanks Prof. Mariella for your encouragement every time we met at the corridor!

A special tribute to my GURU Dr. Swapan Kumar Sarker. I would not dare to step into ecological research if you had not allowed me to work in your forest ecology lab in the Department of Forestry at SUST, Bangladesh. I was the underdog in your lab and your endless patience, nurturing and unconditional effort ultimately given me the hope and confidence to pursue higher degree research in abroad. Thank you very much for always being there.

Father, I have lost you during this PhD journey physically, but you were always there with me, and I know no one would be proud more than you at this moment in the sky! I miss you.

Ma, I deeply regret that during your very hard time I was not there with you, such a selfish son you have! I am here because of your endless endurance and patience. I promise mom, I will be beside you until my last breath.

Lastly, this thesis is nothing but the reflections of your sacrifice, Lisa- my soulmate. I cannot imagine my life without you, and I promise from now on, we will have lunch together on the weekend! I love you.

CHAPTER ONE

INTRODUCTION

Background

Natural history collections (NHCs) are recognised as storage centres of primary information on the flora and fauna of the planet. This information encompasses three basic dimensions that characterise species' distributions – taxonomy, space and time – as it provides direct evidence that a particular species occurred at a particular location at a particular point in time. Knowledge derived from the estimated 1.5–3 billion specimens housed in museums and herbaria (Ariño 2010; Duckworth *et al.* 1993) plays a fundamental role in characterizing global patterns of biodiversity. Such information has substantial intrinsic value with respect to genetic, phylogenetic, biogeographic and ecological data, and specimens have formed the basis of many environmental and ecological studies (Pyke and Ehrlich 2010).

Digitisation of the specimen records began in the 1970s (Graham *et al.* 2004a; Thomas 2009), and involves electronically databasing information contained on the specimen label, particularly the scientific name, collector's name, date of collection, locality description, and geographic coordinates (if known) (Crovello 1972; Morris and Glen 1978). Digitisation of specimen records has now become a global enterprise, with data and images being captured by institutions around the world (Ellwood *et al.* 2018). The advent of low-cost computer processing in the late 1990s, innovations in database design and the creation of data aggregators, such as the Global Biodiversity Information Facility (GBIF, https://www.gbif.org/), the Atlas of Living Australia (ALA, https://www.ala.org.au/), Integrated Digitized Biocollections (iDigBio, https://www.idigbio.org/), the Biodiversity Heritage Library (https://www.biodiversitylibrary.org/), the European Distributed Institute of Taxonomy (http://www.ala.org.au) and Vertnet (http://www.vertnet.org), provide access to these digital primary biodiversity data. Currently > 980,000,000 digitized occurrence records have been incorporated into the largest openly accessible biodiversity distribution network, GBIF (accessed on 17 June, 2018).

Unprecedented improvement in access to primary occurrence data held within NHCs over the last two decades has opened a new window to exploring biodiversity patterns and changes from local to global scales (Andrew *et al.* 2017; Franklin *et al.* 2017; Powney and Isaac 2015), and offers great potential for providing conservation practitioners with access to relevant data (Graham *et al.* 2004b; Joshua 2011; Sullivan *et al.* 2017; Ward 2012). The integration of environmental variables and genetic data with specimen records and the introduction of state-of-the-art image-based digitization of information is also greatly expanding the scope of morphological, phenological, genetic and biogeographical studies that can be undertaken (Ke *et al.* 2013; Soltis 2017). NHC specimens are also becoming increasingly useful for studying common species to elucidate taxonomic declines (i.e. describing new species) associated with habitat destruction, climate change, non-native invasive species, and introduced pathogens (Grixti *et al.* 2009).

Despite recent efforts to collate more than two centuries of regional biodiversity inventories generated through collections into comprehensive electronic databases, these data contain biases and errors in terms of their spatial, temporal and taxonomic scope – a set of problems largely known as the biodiversity knowledge shortfalls (Hortal *et al.* 2015; Meyer *et al.* 2016). Spatial bias occurs due to opportunistic sampling effort. Collectors often focus on particular areas of interest (e.g. protected areas or hotspots of diversity) (Dennis and Thomas 2000) or in more accessible regions (close to roads, rivers, coasts, or urban areas) (Ronen *et al.* 2004; Ubirajara *et al.* 2016; Yang *et al.* 2013). Temporal bias refers to the inconsistency in temporal coverage of sampling (Meyer *et al.* 2016; Stropp *et al.* 2016), while taxonomic bias refers to the preference of collectors to sample particular taxa, which may arise due to a focus on taxonomic uniqueness (e.g. endemics), rarity or economic value (Bonnet *et al.* 2002), or from societal preferences (Wilson *et al.* 2007). Other types of bias generally connected to these three basic dimensions have also been identified. These include environmental or climate bias (Funk *et al.* 2005; Loiselle *et al.* 2008), functional biases (Schmidt-Lebuhn *et al.* 2013), and seasonal bias (ter Steege and Persaud 1991).

Errors, or uncertainties, can also arise during the digitisation of the information recorded on specimen labels. This attribute information usually includes taxonomic identity, collection date, locality or geographic coordinates, and collector. Errors may also have been made by the specialist/technician (e.g. incorrect taxonomy or not identified to species level, incorrect or missing geographic coordinates) during the curation process. Combined, spatial biases and errors/uncertainties associated with NHC records are largely responsible for creating imperfect knowledge of the spatiotemporal distribution of biodiversity (Boakes *et al.* 2010; Meyer *et al.* 2016; Nelson *et al.* 1990; Sousa-Baena *et al.* 2014), referred to as the Wallacean shortfall (Lomolino and Heaney 2004). Taxonomic bias may create artificial inflations in species numbers for certain taxa and therefore may influence decision-making regarding resource allocation and conservation actions (Farrier *et al.* 2007; Grand *et al.* 2007; Pillon and Chase Mark 2006; Walsh *et al.* 2012).

Gaps in the completeness of digitised NHCs can also arise due to delays in the digitisation of existing collections (Meyer *et al.* 2015; Vollmar *et al.* 2010). Indeed, the digitisation lag (i.e. the gap between when a specimen is collected and when it is digitised) is a major factor limiting the spatial, temporal and taxonomic coverage of digital NHCs (Meyer *et al.* 2016). For example, Meyer *et al.* (2016) calculated that only 17% of 120,000,000 terrestrial herbarium specimens collected prior to 2014 are digitally accessible via GBIF, the largest aggregator of NHC data. Around 380,000,000 plant, algae and fungi records are estimated to be held within the world's 3400 herbaria (Thiers 2017). GBIF contains > 215,000,000 records for Kingdom Plantae (as of 17 June 2018). Of these, approximately 66,500,000 represent preserved specimens while the basis of an additional 10,200,000 is unknown, indicating that around 19% of plant specimens are now digitised and accessible. However, institutions in non-western regions, especially south-east Asia, Africa and Brazil, have large numbers of specimens remaining to either be digitised (Meyer *et al.* 2015; Sousa-Baena *et al.* 2014; Stropp *et al.* 2016; Yang *et al.* 2014) or incorporated into the major data aggregators such as GBIF.

Identifying limitations in the NHC data is recognised as a priority area needed to achieve multiple targets specified by the Convention on Biological Diversity (Meyer *et al.* 2015). The credibility of the science based upon NHC data largely depends upon recognising and quantifying these shortfalls to minimize the inefficient use of limited conservation resources (Grand *et al.* 2007; Hortal *et al.* 2008).

Digitisation of Australian flora

Collection History

Australia harbours a diverse flora and has a rich history of botanical sampling. The earliest preserved specimens date to the late 17th century and were collected by European explorers such as Dirk Harthog (Webb 2003) and William Dampier (Green 1990). The first major botanical collection was undertaken by Joseph Banks and Daniel Solander in 1770 (Barker and Barker 1990) and is now preserved in the Australian National Herbarium. From 1801–1805, Robert Brown, known as "father of Australian botany", collected ~4000 specimens (see https://www.anbg.gov.au/) including those from 1700 species and 140 genera previously unknown to science. Brown's collection formed the foundation for the seven volumes of George Bentham's *Flora Australiensis* (published from 1863–1878) and the *Flora of Australia* series (ABRS and CSIRO 1981-2015, http://www.publish.csiro.au/books/series/6). However, as with material collected elsewhere around the world, many of the specimens dating to the late 18th and 19th centuries were, and in many instances remain, in overseas institutions, although some material has been returned to Australia (Webb 2003). By the mid-20th century, systematic botany in Australia had entered in a new phase whereby collections and monographic works were developed in tandem across the continent (George 1981). Today, there are 28 major Australian herbaria listed and governed by the Council of the Heads of Australasian Herbaria (CHAH) (http://www.chah.gov.au/), although more than 90% of these collections are held in the nine State or Territory herbaria in Australia (see Table 1).

The digitisation of specimens in Australian herbaria began in the mid-1970s using in-house databases (Barker 1998), following identification of the opportunities that digitisation offered for improving collection management and streamlining taxonomic processes. In 2001, the Australian Virtual Herbarium (now known as the Australasian Virtual Herbarium, AVH www.avh.chah.org.au) was established. To date, an estimated 80% of Australian plant specimens have been databased (http://avh.chah.org.au/index.php/about/), and more than 90% of these are accessible via the AVH. In sum, the AVH contains more than eight million records of plants, algae and fungi from all state and territory herbaria, as well as from several universities in Australia and New Zealand. The collation of these resources helped to inspire the development of the Atlas of Living Australia and gives anyone with an internet connection access to specimen records from around Australia and the world.

The importance of the AVH

The AVH is the main database storing information on collections of Australia's flora. Digital access to herbaria collections via the AVH has proven valuable in a range of contexts, from ecological research to citizen science projects (Cantrill, 2018. This database is increasingly used in assessments of species distributions across geographic space and environmental gradients (Crisp *et al.* 2001; Pimm *et al.* 2014), species responses to climate change (Gallagher *et al.* 2010; Mellick *et al.* 2011), hotspots of invasive species (Duursma *et al.* 2013; O'Donnell *et al.* 2012), phytogeographical analyses (Gallagher 2016; González-Orozco *et al.* 2014), prioritising regions for conservation (Baumgartner *et al.* 2018; Colloff *et al.* 2014; Lee and Mishler 2014), and patterns of endemism and evolutionary history (Bickford *et al.* 2004; Gonzalez-Orozco *et al.* 2016; Laffan and Crisp 2003; Rosauer *et al.* 2009) and biosecurity (Sultana *et al.* 2017). According to Cantrill (2018), on average every record provided by the Royal Botanic Gardens to AVH has been downloaded 220 times for ecological research since 2010.

Knowns and unknowns of digitised flora

Given the inherent multidimensional biases and gaps in NHCs (Meyer *et al.* 2016), to what extent could our assumptions of plant diversity and distributions be erroneous when based on information in the AVH? In a recent study, Schmidt-Lebuhn *et al.* (2012) found that sampling biases within the AVH may lead to erroneous perceptions of species diversity of family Asteraceae. Using AVH records, Wernberg *et al.* (2011) concluded that the distributions of numerous seaweed species have shifted southward due to climate change. However, Huisman and Millar (2013) contended that it was not species' ranges that had shifted, rather the data reflect a distinct southward skew in collection effort in more recent years.

The recent decadal plan (2018-2027) for Australian taxonomy and biosystematics reported that since the 1990s, there has been a substantial decline in the databasing of plant specimens and the annual rate of naming new species has begun to decline (ACS 2018). The ACS report also highlighted the potential consequences of poor knowledge of taxa, including compromising the effectiveness of research into diverse areas ranging from biosecurity to the effects of climate change and other environmental stresses on biodiversity. Moreover, a sound understanding of biodiversity is critical for mega-diverse countries such as Australia, which harbour numerous unique and endemic species many of which are globally important in understanding the evolutionary history of the planet (Mittermeier *et al.* 2011).

To date, there has been no comprehensive study that quantifies knowledge shortfalls across spatial, temporal and taxonomic space among the specimens included in the AVH. Given that this is the main database for describing the flora of the Australian continent, identifying its limitations is paramount of importance for the validity of conservation and environmental studies, and to improve our knowledge on Australian flora.

Table 1. *List of herbaria in Australia with their estimated number of specimens and estimated number of specimens digitised (data from www.ala.org.au). N.B., in addition to these herbaria there are also a number of plant pathology herbaria across the country.*

Herbaria (Code)	**Date established**	**Estimated number of specimens**	**Estimated number of specimens digitised (%)**	**Notes**	**Predominant geographic range of specimens**
Australian National Herbarium (CANB)	1994	1,142,785	810,768 (70.9)		Australia
Australian Tropical Herbarium (CNS)	1971	180,000	180,000 (100)		Australia's tropics, especially north-eastern Australian rainforests
Charles Sturt University (CSU)	1999	4,264	4,100 (96.2)		Upper Murray and Murrumbidgee regions
Downing Herbarium (Macquarie University)	1972	13,000	9,750 (75.0)		Mostly New South Wales
The Robert Brown Herbarium (ECU)	1950	10,000	10,000 (100)		Western Australia, especially south-west region, Pilbara, some Eastern Goldfields
Eurobodalla Regional Botanic Gardens Wallace Herbarium	1987	12548	7780 (62.2)		Catchments of the Clyde, Deuda and Tuross Rivers
Gauba Herbarium (GAUBA)	1961	26900	NA	Not accessible through ALA	Mostly Australian Capital Territory and New South Wales
James Cook University Herbarium (JCT)	1960	24,046	25,671*		Tropical Northern Queensland
John Ray Herbarium (SYD)	1916	62503	NA	Not accessible through ALA	Mostly south-eastern Australia
John T Waterhouse Herbarium (UNSW)	1960	53,150	6,380 (12.0)		Mainly New South Wales and northern Australia
Kings Park and Botanic Garden Herbarium (KPBG)	1963	18,307	5,355 (29.3)		Western Australia

Herbaria (Code)	**Date established**	**Estimated number of specimens**	**Estimated number of specimens digitised (%)**	**Notes**	**Predominant geographic range of specimens**
La Trobe University (LTB)	1970	25,000	5,393 (21)		Tropical and temperate eastern Australia
Murdoch University (MURU)	1975	7,000	NA	Not accessible through ALA	Mainly Western Australia
National Herbarium of New South Wales (NSW)	1896	1,425,000	720,000 (50.5)		Australia, especially New South Wales
National Herbarium of Victoria (MEL)	1853	1,386,403	868,232 (62.6)		Australia
North Coast Regional Botanic Garden Herbarium (CFSHB)	1940	28,776	28,776 (100)		North-eastern New South Wales
Northern Territory Herbarium (NT)	1954	45,000	45,000 (100)		Arid zone of Northern Territory
Northern Territory Herbarium (DNA)	1966	243,000	242,000 (99.6)		NT monsoonal tropics and arid zone; tropical Western Australia and Queensland
Queensland Herbarium (BRI)	1855	849,023	839,023 (98.8)		Mainly Queensland
State Forests of New South Wales Herbarium	1936	4,500	1,500 (33.0)	Not accessible through ALA	Mostly fungi, but some plants; New South Wales
State Herbarium of South Australia (AD)	1954	1,030,000	723,000 (70.2)		Mainly South Australia
Tasmanian Herbarium (HO)	1930s	285,700	177,000 (62.0)		Tasmania
The Janet Cosh Herbarium (WOLL)	1989	10,905	10,605 (97.2)		Illawarra, Southern Highlands, Sydney Basin
The University of Melbourne Herbarium (MELU)	1926	93,290	9,000 (9.6)		Mainly Victoria

Herbaria (Code)	Date established	Estimated number of specimens	Estimated number of specimens digitised (%)	Notes	Predominant geographic range of specimens
University of New England N.C.W. Beadle Herbarium	1938	82,752	68,000 (82.2)		Northern New South Wales; south-eastern Queensland
University of Western Australia Herbarium (UWA)	1914	10,000	NA	Not accessible through ALA	Southwestern Australia and Pilbara
Western Australian Herbarium (PERTH)	1929	751,803	751,803 (100)		Australian states

*Note a discrepancy exists between the estimated number of specimens within this collection and the number digitised.

Aims of the thesis

This thesis explores the pattern in knowledge shortfalls among the Australian flora based on digitised preserved specimens in the Australasian Virtual Herbarium (AVH). In addition, this thesis also assesses the digitisation effort of the key State and Territory herbaria. As such, the thesis has four Aims, each of which constitutes a separate chapter.

Aim 1. To explore spatial patterns among digitised plant specimens by assessing sampling effort and inventory completeness, to ascertain the extent to which human influences characterise patterns of sampling effort and to identify areas likely to harbour species previously unrecorded for that area.

Aim 2. To explore the temporal patterns among digitised plant specimens from Australia by *assessing* the temporal consistency in collection effort, data quality and inventory completeness across spatio-temporal space.

Aim 3. To explore the extent to which taxonomic shortfalls exist among the digitised plant specimens from Australia.

Aim 4. To assess lags in the digitisation effort of key Australian herbaria.

Structure and format of thesis

I have organized this thesis into six chapters, including the Introduction, and have written and structured it to comply with the format of "thesis by publication". As a result, Chapters 2 – 5 are written as standalone chapters, with each structured to conform to the format of the journal to which it is published/submitted/will be submitted. The titles of the chapters are:

Chapter One: Introduction

Chapter Two: How well documented is Australia's flora? Understanding spatial bias in vouchered plant specimens. (**Published** in Austral Ecology, DOI: 10.1111/aec.12487)

Chapter Three: A journey through time: Exploring temporal patterns among digitised plant specimens from Australia. (**In press** in Systematics and Biodiversity)

Chapter Four: Taxonomic shortfalls in digitised collections of Australia's flora. (under review in Biodiversity and Conservation)

Chapter Five: Filling the gap: How quickly do Australia's herbaria digitise their vouchered specimens? (manuscript in preparation)

Chapter Six: Discussion and Conclusion

REFERENCES

ABRS & CSIRO. (1981-2015) *Flora of Australia*. Australian Government Publishing Service (AGPS)

ACS. (2018) *Discovering Biodiversity: A decadal plan for taxonomy and biosystematics in Australia and New Zealand 2018–2027*. Australian Academy of Science (ACS), Australia.

Andrew C., Heegaard E., Kirk P. M., Bässler C., Heilmann-Clausen J., Krisai-Greilhuber I., Kuyper T. W., Senn-Irlet B., Büntgen U. & Diez J. (2017) Big data integration: Pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. *Fungal Biology Reviews* **31**, 88-98.

Ariño A. H. (2010) Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics* **7**, 81-92.

Barker R. M. & Barker W. R. (1990) Botanical contributions overlooked: the role and recognition of collectors, horticulturists, explorers and others in the early documentation of the Australian flora. In: *In History of Systematic Botany in Australasia* (ed P. Short) pp. 37-85. Australian Systematic Botany Society: Melbourne.

Barker W. R. (1998) *The Virtual Australian Herbarium*. Australian Herbarium Information Systems Committee (HISCOM). .

Baumgartner J. B., Esperón-Rodríguez M. & Beaumont L. J. (2018) Identifying in situ climate refugia for plant species. *Ecography* **41**, 1-14.

Bickford S. A., Laffan S. W., Kok R. P. & Orthia L. A. (2004) Spatial analysis of taxonomic and genetic patterns and their potential for understanding evolutionary histories. *Journal of Biogeography* **31**, 1715-1733.

Boakes E. H., Mcgowan P. J., Fuller R. A., Chang-Qing D., Clark N. E., O'connor K. & Mace G. M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology* **8**, e1000385.

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

Cantrill D. J. (2018) The Australasian Virtual Herbarium: Tracking data usage and benefits for biological collections. *Applications in Plant Sciences*, e1026.

Colloff M. J., Ward K. A. & Roberts J. (2014) Ecology and conservation of grassy wetlands dominated by spiny mud grass Pseudoraphis spinescens in the southern Murray–Darling Basin, Australia. *Aquatic Conservation: Marine and Freshwater Ecosystems* **24**, 238-255.

Crisp M. D., Laffan S., Linder H. P. & Monro A. (2001) Endemism in the Australian flora. *Journal of Biogeography* **28**, 183-198.

Crovello T. J. (1972) Computerization of specimen data from the Edward Lee Greene Herbarium (ND-G) at Notre Dame. *Brittonia* **24**, 131-141.

Dennis R. & Thomas C. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation* **4**, 73-77.

Duckworth W. D., Genoways H. H. & Rose C. L. (1993) *Preserving natural science collections: chronicle of our environmental heritage*. National Institute for the Conservation of Cultural Property.

Duursma D. E., Gallagher R. V., Roger E., Hughes L., Downey P. O. & Leishman M. R. (2013) Next-generation invaders? Hotspots for naturalised sleeper weeds in Australia under future climates. *PLoS One* **8**, e84222.

Ellwood E. R., Kimberly P., Guralnick R., Flemons P., Love K., Ellis S., Allen J. M., Best J. H., Carter R., Chagnoux S., Costello R., Denslow M. W., Dunckel B. A., Ferriter M. M., Gilbert E. E., Goforth C., Groom Q., Krimmel E. R., Lafrance R., Martinec J. L., Miller A. N., Minnaert-Grote J., Nash T., Oboyski P., Paul D. L., Pearson K. D., Pentcheff N. D., Roberts M. A., Seltzer C. E., Soltis P. S., Stephens R., Sweeney P. W., Von Konrat M., Wall A., Wetzer R., Zimmerman C. & Mast A. R. (2018) Worldwide Engagement for Digitizing Biocollections (WeDigBio): The Biocollections Community's Citizen-Science Space on the Calendar. *Bioscience* **68**, 112-124.

Farrier D., Whelan R. & Mooney C. (2007) Threatened species listing as a trigger for conservation action. *Environmental Science & Policy* **10**, 219-229.

Franklin J., Serra-Diaz J. M., Syphard A. D. & Regan H. M. (2017) Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography* **26**, 6-17.

Funk V. A., Richardson K. S. & Ferrier S. (2005) Survey-gap analysis in expeditionary research: where do we go from here? *Biological Journal of the Linnean Society* **85**, 549-567.

Gallagher R. V. (2016) Correlates of range size variation in the Australian seed-plant flora. *Journal of Biogeography* **43**, 1287-1298.

Gallagher R. V., Hughes L., Leishman M. R. & Wilson P. D. (2010) Predicted impact of exotic vines on an endangered ecological community under future climate change. *Biological Invasions* **12**, 4049-4063.

George A. (1981) The background to the flora of Australia. In: *In Flora of Australia* Vol. 1 pp. 3-24. Bureau of Flora and Fauna: Canberra, ACT, Australia.

González-Orozco C. E., Ebach M. C., Laffan S., Thornhill A. H., Knerr N. J., Schmidt-Lebuhn A. N., Cargill C. C., Clements M., Nagalingum N. S. & Mishler B. D. (2014) Quantifying phytogeographical regions of Australia using geospatial turnover in species composition. *PLoS One* **9**, e92558.

Gonzalez-Orozco C. E., Pollock L. J., Thornhill A. H., Mishler B. D., Knerr N., Laffan S. W., Miller J. T., Rosauer D. F., Faith D. P., Nipperess D. A., Kujala H., Linke S., Butt N., Kulheim C., Crisp M. D. & Gruber B. (2016) Phylogenetic approaches reveal biodiversity threats under climate change. *Nature Climate Change* **10.1038/nclimate3126**.

Graham C., Ferrier S., Huettman F., Moritz C. & Peterson A. (2004a) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution* **19**, 497-503.

Graham C. H., Ferrier S., Huettman F., Moritz C. & Peterson A. T. (2004b) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution* **19**, 497-503.

Grand J., Cummings M. P., Rebelo T. G., Ricketts T. H. & Neel M. C. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecology Letters* **10**, 364-374.

Green J. (1990) History of early Western Australian herbaria. In: *History of systematic botany in Australasia* (ed P. Short) pp. 23-27.

Grixti J. C., Wong L. T., Cameron S. A. & Favret C. (2009) Decline of bumble bees (Bombus) in the North American Midwest. *Biological Conservation* **142**, 75-84.

Hortal J., De Bello F., Diniz-Filho J. a. F., Lewinsohn T. M., Lobo J. M. & Ladle R. J. (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 J. F., Lobo J. M. & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847-858.

Huisman J. M. & Millar A. J. K. (2013) Australian seaweed collections: use and misuse. *Phycologia* **52**, 2-5.

Joshua D. (2011) The Role of Natural History Institutions and Bioinformatics in Conservation Biology. *Conservation Biology* **25**, 1250-1252.

Ke B., Tyler L., Dan V., M. G. J., Rasmus N. & Craig M. (2013) Unlocking the vault: next-generation museum population genomics. *Molecular Ecology* **22**, 6018-6032.

Laffan S. W. & Crisp M. D. (2003) Assessing endemism at multiple spatial scales, with an example from the Australian vascular flora. *Journal of Biogeography* **30**, 511-520.

Lee A. C. & Mishler B. (2014) Phylogenetic diversity and endemism: metrics for identifying critical regions of conifer conservation in Australia. *Berkeley Scientific Journal* **18**, 48-58.

Loiselle B. A., Jørgensen P. M., Consiglio T., Jiménez I., Blake J. G., Lohmann L. G. & Montiel O. M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography* **35**, 105-116.

Lomolino M. V. & Heaney L. R. (2004) *Frontiers of Biogeography. New Directions in the Geography of Nature*. Sunderland, Mass.. Sinauer Associates.

Mellick R., Lowe A. & Rossetto M. (2011) Consequences of long-and short-term fragmentation on the genetic diversity and differentiation of a late successional rainforest conifer. *Australian Journal of Botany* **59**, 351-362.

Meyer C., Kreft H., Guralnick R. & Jetz W. (2015) Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6**, 8221-8229.

Meyer C., Weigelt P. & Kreft H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**, 992-1006.

Mittermeier R. A., Turner W. R., Larsen F. W., Brooks T. M. & Gascon C. (2011) Global Biodiversity Conservation: The Critical Role of Hotspots. In: *Biodiversity Hotspots: Distribution and Protection of Conservation Priority Areas* (eds F. E. Zachos and J. C. Habel) pp. 3-22. Springer Berlin Heidelberg, Berlin, Heidelberg.

Morris J. & Glen H. (1978) PRECIS, the National Herbarium of South Africa (PRE) computerized information system. *Taxon* **27**, 449-462.

Nelson B. W., Ferreira C. a. C., Da Silva M. F. & Kawasaki M. L. (1990) Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature* **345**, 714.

O'Donnell J., Gallagher R. V., Wilson P. D., Downey P. O., Hughes L. & Leishman M. R. (2012) Invasion hotspots for non-native plants in Australia under current and future climates. *Global Change Biology* **18**, 617-629.

Pillon Y. & Chase Mark W. (2006) Taxonomic Exaggeration and Its Effects on Orchid Conservation. *Conservation Biology* **21**, 263-265.

Pimm S. L., Jenkins C. N., Abell R., Brooks T. M., Gittleman J. L., Joppa L. N., Raven P. H., Roberts C. M. & Sexton J. O. (2014) The biodiversity of species and their rates of extinction, distribution, and protection. *Science* **344**, 1246752.

Powney G. D. & Isaac N. J. (2015) Beyond maps: a review of the applications of biological records. *Biological Journal of the Linnean Society* **115**, 532-542.

Pyke G. H. & Ehrlich P. R. (2010) Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biological Reviews (Cambridge)* **85**, 247-266.

Ronen K., Oren F. & Avinoam D. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* **14**, 401-413.

Rosauer D., Laffan S. W., Crisp M. D., Donnellan S. C. & Cook L. G. (2009) Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. *Molecular Ecology* **18**, 4061-4072.

Schmidt-Lebuhn A. N., Knerr N. J. & González-Orozco C. E. (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. *Journal of Biogeography* **39**, 2072-2080.

Schmidt-Lebuhn A. N., Knerr N. J. & Kessler M. (2013) Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and Conservation* **22**, 905-919.

Soltis P. S. (2017) Digitization of herbaria enables novel research. *American Journal of Botany* **104**, 1281-1284.

Sousa-Baena M. S., Garcia L. C., Peterson A. T. & Brotons L. (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions* **20**, 369-381.

Stropp J., Ladle R. J., Malhado M., Ana C., 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.

Sullivan B. L., Phillips T., Dayer A. A., Wood C. L., Farnsworth A., Iliff M. J., Davies I. J., Wiggins A., Fink D., Hochachka W. M., Rodewald A. D., Rosenberg K. V., Bonney R. & Kelling S. (2017) Using open access observational data for conservation action: A case study for birds. *Biological Conservation* **208**, 5-14.

Sultana S., Baumgartner J. B., Dominiak B. C., Royer J. E. & Beaumont L. J. (2017) Potential impacts of climate change on habitat suitability for the Queensland fruit fly. *Scientific Reports* **7**, 13025.

Ter Steege H. & Persaud C. A. (1991) The phenology of Guyanese timber species: a compilation of a century of observations. *Vegetatio* **95**, 177-198.

Thiers B. (2017) Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/science/ih/.

Thomas C. (2009) Biodiversity databases spread, prompting unification call. *Science* **324**, 1632-1633.

Ubirajara O., Pereira P. A., Brescovit A. D., de Carvalho C. J. B., Silva D. P., Rezende D. T., Sa Fortes Leite F., Batista J. A. N., Barbosa J. P. P. P., Stehmann J. R., Ascher J. S., de Vasconcelos M. F., De Marco Jr P., Lowenberg-Neto P., Dias P. G., Ferro V. G. & Santos A. J. (2016) The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Diversity and Distributions* **22**, 1232-1244.

Vollmar A., Macklin J. A. & Ford L. (2010) Natural History Specimen Digitization: Challenges and Concerns. *2010* **7**, 93-112.

Walsh J. C., Watson J. E. M., Bottrill M. C., Joseph L. N. & Possingham H. P. (2012) Trends and biases in the listing and recovery planning for threatened species: an Australian case study. *Oryx* **47**, 134-143.

Ward D. F. (2012) More than just records: analysing natural history collections for biodiversity planning. *PLoS One* **7**, e50346.

Webb J. B. (2003) *The Botanical Endeavour: Journey Towards a Flora of Australia*. Surrey Beatty & Sons, New South Wales, Australia.

Wernberg T., Russell B. D., Thomsen M. S., Gurgel C. F., Bradshaw C. J., Poloczanska E. S. & Connell S. D. (2011) Seaweed communities in retreat from ocean warming. *Current Biology* **21**, 1828-1832.

Wilson J. R., Procheş Ş., Braschler B., Dixon E. S. & Richardson D. M. (2007) The (bio)diversity of science reflects the interests of society. *Frontiers in Ecology and the Environment* **5**, 409-414.

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

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

CHAPTER TWO

HOW WELL DOCUMENTED IS AUSTRALIA'S FLORA? UNDERSTANDING SPATIAL BIAS IN VOUCHERED PLANT SPECIMENS

(This Chapter has been published in Austral Ecology)

ABSTRACT

Massive digitisation of natural history collections (NHC) has opened the door for researchers to conduct inferential studies on the collection of biological diversity across space and time. The widespread use of NHCs in scientific research makes it essential to characterise potential sources of spatial bias. Here, we assessed spatial patterns in records from the Australian Virtual Herbarium (AVH), based on > 3,000,000 vouchered specimens of around 21,000 native plant species. The AVH is the main database for describing Australia's flora, and identifying its limitations is of paramount interest for the validity of conservation and environmental studies. We characterised how sampling effort is distributed across each Interim Bioregion of Australia (IBRA), then asked: 1) How complete are species inventories for each bioregion? We define completeness (*C*) as the ratio of observed to estimated species richness, using the Chao 1 estimator, 2) How is sampling effort related to a commonly used Human Influence Index (HII)? and 3) What is the probability that additional collections would result in the identification of previously unrecorded species in each bioregion? Sampling effort across bioregions is unequal, which partially reflects the collecting behaviour of naturalists in relation to species richness patterns. The density of records in bioregions ranges from 0.02-8.37/km^2. At the bioregional scale, completeness is generally high with 79% of bioregions estimated to have records for at least 80% of their species. Completeness is partly explained by sampling effort (r = 0.43, p = 0.01), although some bioregions (e.g. Northern Kimberley and Burt Plain) have high completeness yet relatively low sampling effort. The inventory of Hampton, however, is

substantially less complete than other bioregions (*C* = 0.66). Bioregions with high HII consistently have high completeness, while regions with low HII span the full range of completeness values. We calculated that an additional specimen collected from a bioregion has a 0.33% (Wet Tropics) to 11.7% (Arnhem Coast) probability of representing a new species for that region. Our assessment can assist with directing future systematic survey efforts by identifying bioregions where additional surveying may result in the greatest return, in terms of increasing knowledge of species richness and diversity.

INTRODUCTION

Natural history collections (NHCs), mostly housed in museums and herbaria, are regarded as a cornerstone resource for understanding biological diversity across space and time. Such information has substantial intrinsic value with respect to genetic, phylogenetic, biogeographic and ecological data, and specimens have formed the basis of many environmental and ecological studies (Pyke and Ehrlich 2010).

The advent of online databases has substantially increased access to information stored in NHCs. Digital databases recording details on specimen labels were initiated in the 1970s (Graham *et al.* 2004a; Thomas 2009). To date > 624,000,000 digitised records have been incorporated into the largest publicly accessible database, the Global Biodiversity Information Facility (GBIF, see http://www.gbif.org/).

Ready access to NHC data has enabled researchers to conduct inferential studies on the spatial distribution of biological diversity from global to continental and regional scales (Ballesteros-Mejia *et al.* 2013; Barthlott *et al.* 2007; Lavoie 2013; Ter Steege *et al.* 2006). But, while there has been a dramatic rise over the last 20 years in the number of studies using information from NHCs to explore ecological and environmental research questions (Pyke and Ehrlich 2010), a major concern that remains is how comprehensive are these data across space and time, in terms of the extent to which the species inventory is captured?

The validity of studies utilising records from NHCs is strongly dependent upon the abundance and representativeness of records (Hijmans *et al.* 2000; Santos *et al.* 2010; Yesson *et al.* 2007b) and data quality (Soberón & Peterson, 2004). These factors, in turn, are influenced by the haphazard collection of most specimens, and historical and taxonomic biases in sampling effort (Beck *et al.* 2014; Hortal *et al.* 2008b; Isaac and Pocock 2015). For instance, sampling effort has often focused on areas of particular interest (protected areas or hotspots of diversity) or in more accessible regions (close to roads, rivers, coasts, or urban areas). This may lead to some areas being inadequately sampled (Nelson *et al.* 1990) and frequently no information on collection effort or method is recorded. As a consequence incorrect conclusions may be drawn with regards to the spatial distribution of biodiversity (Boakes *et al.* 2010; Soria-Auza and Kessler 2008).

Australia's Virtual Herbarium (AVH) (www.chah.gov.au/avh) contains digitised specimen records from the country's major herbaria. This database is increasingly used in assessments of species distributions across geographic space and environmental gradients (Crisp *et al.* 2001), species responses to climate change (Gallagher *et al.* 2010; Mellick *et al.* 2011), hotspots of invasive species (Duursma *et al.* 2013; O'Donnell *et al.* 2012), phytogeographical analyses (Gallagher 2016; González-Orozco *et al.* 2014), prioritising regions for conservation (Colloff *et al.* 2014; Lee and Mishler 2014), and patterns of endemism and evolutionary history (Bickford *et al.* 2004; Gonzalez-Orozco *et al.* 2016; Laffan and Crisp 2003; Rosauer *et al.* 2009). Yet, as with other NHC's, biases and errors will be present in this collection, and have already been shown to lead to erroneous perceptions of species diversity (e.g. Schmidt-Lebuhn *et al.* (2012).

To date, there has been no comprehensive study of spatial patterns among the specimens included in the AVH. Given that this is the main database for describing the flora of the Australian continent, identifying its limitations is necessary for the validity of conservation and environmental studies (Beck *et al.* 2014; Hortal *et al.* 2007), and for prioritising bioregions for future sampling effort. Here, we explore spatial patterns among vouchered specimens in the AVH, at the bioregional scale. First, we characterise how sampling effort is distributed across bioregions. We then use the AVH data to ask: 1) How complete are species inventories for each bioregion? 2) How is sampling effort across Australia related to the Human Influence Index

(HII)? and 3) What is the probability that additional collections would result in the identification of previously unrecorded species in each bioregion?

METHODS

Dataset

Australia's Virtual Herbarium (AVH) provides digitised records from vouchered specimens held within Australia's nine major herbaria and is the largest source of occurrence data for this continent's flora (CHAH, 2009). We downloaded these records from the Atlas of Living Australia (ALA) (http://www.ala.org.au/; accessed January 2015) by accessing data for each plant family identified by the Australian Plant Census (APC) (n = 361 families). This resulted in a preliminary dataset of 10,102,447 occurrence records. A multi-step procedure was used to clean these raw data prior to analysis by removing observations that: (1) were not identified to species level (i.e. consisted of a genus name and the epithet "sp."); (2) lacked georeferencing information; (3) represented cultivated specimens (e.g. in a garden or agricultural trial) or non-native species; (4) were hybrids; and (5) were outside the geographic boundary of the Australian coastline (i.e. coastal waterways or on offshore islands). The final dataset incorporated 3,046,617 records belonging to 20,618 native species and 300 families.

Data Analysis

We conducted our analysis at the bioregional level using the Interim Biogeographic Regions of Australia (IBRA, Thackway and Cresswell, 1995). These bioregions are management units frequently used for describing species diversity patterns and for developing national conservation strategies (Mackey *et al.* 2008; Polak *et al.* 2015; Williamson *et al.* 2011). Using the spatial package 'sp' (Roger *et al.* 2013) for R version 3.0.1 (R Development Core Team, 2013), we overlaid occurrence records with a shapefile of the bioregions (downloaded from http://www.environment.gov.au/fed/catalog/main/home.page), and calculated the number of occurrence records, species and families within each bioregion (Appendix S2). Four bioregions (Coral Sea, Indian Tropical Islands, Pacific Subtropical Islands, and Sub-Antarctic Islands) were excluded from analysis as they were beyond the terrestrial extent of the Australian continent.

Sampling completeness in bioregions

For each bioregion we calculated sampling effort (simply the number of records/km^2). This a useful first step to describing collecting patterns, but lacks information on the likely completeness of sampling across species (Soberón *et al.* 2007). Hence, we calculated inventory completeness (C), which is defined as the ratio of observed to estimated species richness in a given region (Soberón *et al.* 2007). We estimated species richness using the non-parametric Chao 1 estimator (Colwell and Coddington 1994). This is one of the most accurate non-parametric estimators across landscapes with varying biophysical conditions and is a widely used technique for presence-only records (Ballesteros-Mejia *et al.* 2013; Brose *et al.* 2003; Hortal *et al.* 2006; Schmidt-Lebuhn *et al.* 2012; Soria-Auza and Kessler 2008). The Chao 1 estimator calculates the total number of species likely to be present, including species that were not sampled, by extrapolating the asymptote of a rarefaction curve. For a given region i, estimated species richness ($S_{\mathrm{est(i)}}$) can be calculated as:

$$S_{\mathrm{est(i)}} = S_{\mathrm{obs(i)}} + (f_1^2 / 2f_2)*(n-1/n) \qquad \text{(equation 1)}$$

where $S_{\mathrm{obs}(i)}$ = observed species richness in region i, f_1 and f_2 are the number of singletons and doubletons (species represented by one or two occurrence records), respectively, and n is the number of records in region i. The completeness index (C) was then calculated as:

$$C = S_{\mathrm{obs(i)}}/S_{\mathrm{est(i)}} \qquad \text{(equation 2)}$$

This analysis was conducted using the ‘vegan’ package (Oksanen *et al.* 2016) for R version 3.0.1 (R Development Core Team, 2013).

Sampling effort and Human Influence Index

We evaluated whether sampling effort was correlated with human influence, defined here by the Human Influence Index (HII) (accessed at http://sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-footprint-geographic at a 0.01 arc-second (~1km) resolution). HII is based on population density, human land use and infrastructure, and accessibility (roads, railroads, coastlines, navigable rivers) (Sanderson *et al.* 2002). Using the Spatial Analyst Extension in ArcGIS v 10.4 (ESRI, 2015), we calculated the median HII of each bioregion as a measure of proximity to inhabited areas. We performed a Pearson correlation test between the median HII, sampling effort and C of each bioregion.

A boundary test procedure was used to assess whether there was a significant upper or lower bound to the distribution of data points in a bivariate plot. The boundary was defined as an upper or lower triangle delineated by the median x and y values on the plot, and the numbers of points placed beyond the boundary were counted. A randomisation test (1000 random draws of 85 x, y coordinates from a random uniform distribution bounded by the minima and maxima of the observed data) was then used to estimate a p-value (the proportion of times the point count beyond the boundary was equal to or less than the observed count). We used EcoSim v 7 (Gotelli and Entsminger 2004) to implement this test.

Probability of collecting a previously unrecorded species

To prioritise IBRA bioregions for future sampling effort, we calculated the probability that one additional record collected from a bioregion would represent a species previously unrecorded in that bioregion, using the following formula (see detail in Appendix S1):

$P = f_0 * [f_1 / (n*f_0 + f_1)]$ (equation 3)

where P = probability that an additional occurrence record would represent a new species for that bioregion; f_0 = estimated number of unseen species; n=number of total records and f_1=number of singletons.

RESULTS

As would be expected, sampling effort varied across bioregions. The bioregions with the lowest density of records were Gibson Desert (0.02 records/km^2), Nullarbor (0.03 records/km^2), Great Sandy Desert (0.04 records/km^2), Little Sandy Desert (0.04 records/km^2) and Gascoyne (0.05 records/km^2). In contrast, bioregions with the highest density of records were the Wet Tropics (8.37 records/km^2) followed by Kanmantoo (5.87 records/km^2), Australian Alps (4.22 records/km^2), Sydney Basins (4.05 records/km^2) and Warren (3.9 records/km^2) (Fig. 1a, b) (see details in Appendix S2).

How complete are inventories for each bioregion?

Our analysis of sampling completeness revealed that the inventory of most bioregions (79%) was relatively complete ($C \geq 0.8$, i.e. ≥ 80% of the species are estimated to have been sampled). Completeness was highest ($C \geq 0.87$) for Esperance Plains, Kanmantoo, and Burt Plain, followed by Northern Kimberley, Avon Wheat-belt and Wet Tropics. The inventory of Hampton (south-eastern Western Australia) was the least complete ($C = 0.66$) (Fig. 1c, d). Completeness was partly explained by sampling effort ($r = 0.43$, $p = 0.01$), although some bioregions (e.g. Northern Kimberley and Burt Plain) have high completeness yet relatively low sampling effort.

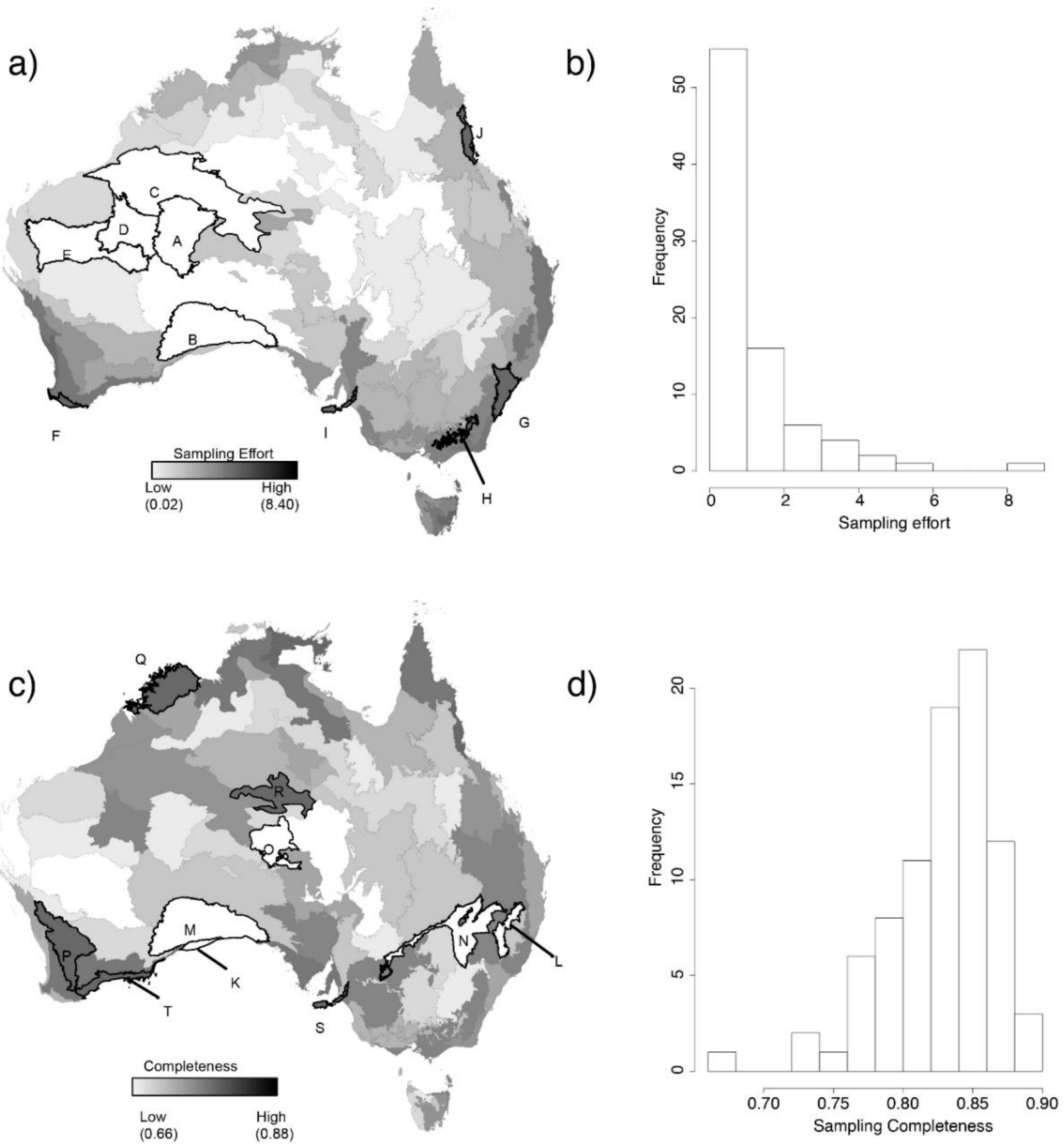


Figure 1. *Spatial distribution and frequency histograms of (a, b) sampling effort (number of vouchered specimens in Australia's Virtual Herbarium, per km^2) and (c, d) inventory completeness (C; the ratio of observed to estimated species richness, using the Chao 1 estimator) for 85 IBRA (Interim Biogeographic Regionalisation of Australia) bioregions across Australia (http://www.environment.gov.au/land/nrs/science/ibra). a) The five bioregions with the lowest sampling effort are (A-E) Great Sandy Desert, Gascoyne, Little Sandy Desert, Gibson Desert, and Nullarbor, while the five bioregions with the highest sampling effort are (F-J) Wet Tropics, Kanmantoo, Australian Alps, Sydney Basin, and Warren. c) The five bioregions with the lowest and highest completeness are (K-O) Nullarbor, Hampton, Finke, Darling Riverine Plains, Nandewar and (P-T) Northern Kimberley, Avon Wheatbelt, Esperance Plains, Burt Plain, Kanmantoo, respectively.*

Are sampling effort and completeness related to HII?

We found a significant correlation between sampling effort (log-transformed) and the Human Influence Index (HII) ($r = 0.68$, $p = 0.001$) at the bioregional level. That is, bioregions with high human activities tend to have a high density of records (Fig. 2a) (see spatial distribution of HII values in Appendix S4). In contrast, there was no significant correlation between HII and inventory completeness ($p = 0.11$), although we identified a strong boundary effect for the lower right triangle of the plot (observed and expected points beyond the boundary were 1 and 10.28, respectively, $p \leq 0.001$). That is, there were far fewer regions with a combination of high HII and low *C* than expected by chance (Fig. 2b).

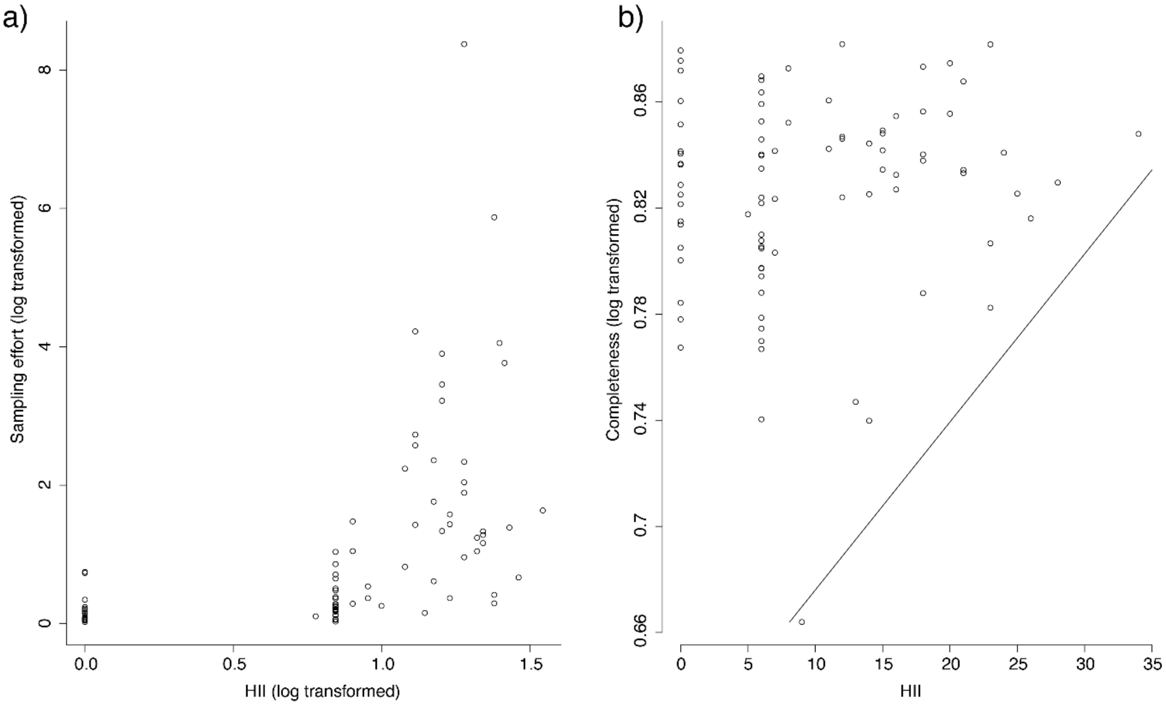


***Figure 2**. Relationship between Human Influence Index (HII) and a) sampling effort (Pearson correlation: r = 0.68, p = 0.001), b) inventory completeness (C) (Pearson correlation: r = 0.11, p = 0.101) across 85 IBRA (Interim Biogeographic Regionalisation of Australia) bioregions of Australia. The diagonal line marks a boundary connecting the median values of HII and C. A randomisation test indicated that IBRA bioregions were highly unlikely to have high HII values with low C values (p = 0.001).*

Which bioregions are most likely to contain previously unrecorded species?

The probability that an additional specimen collected from a given bioregion would belong to a species previously unrecorded *from that bioregion* was highest for Central Arnhem (11.69%), followed by Gibson Desert (4.82%), Hampton (4.77%), Tasmanian Northern Slopes (4.6%) and Gulf Coastal (4.55%). The lowest probabilities occurred in the Wet Tropics (0.33%), Eyre Yorke Block (0.38%), Sydney Basin (0.38%), South Eastern Queensland (0.41%) and Flinders Lofty Block (0.42%) (Fig. 3). The probability that an additional record would represent a new species was negatively correlated with sampling intensity and *C* (*r* = -0.53 and -0.55, respectively), indicating that bioregions with higher sampling intensity and *C*-index are less likely to yield new species (Fig. 4).

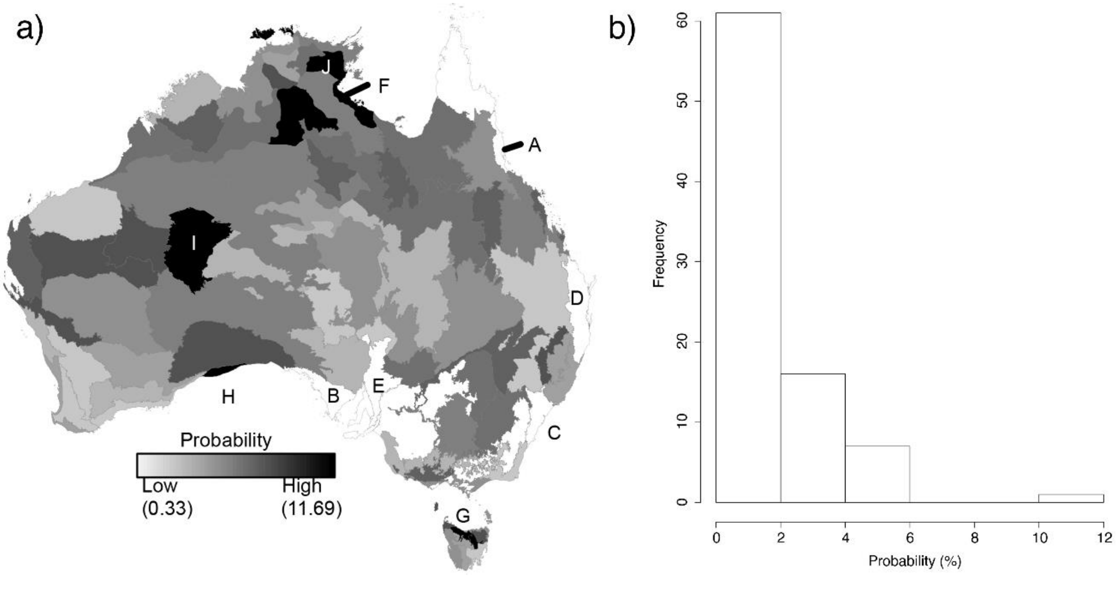


***Figure 3**. a) Spatial distribution and b) frequency histogram of the probability that an additional sample would represent a species previously unrecorded in that bioregion, across the 85 IBRA (Interim Biogeographic Regionalisation of Australia) bioregions of Australia. The five bioregions with the lowest ([A-E] Wet Tropics, Eyre Block, Sydney Basin, South Eastern Queensland, Flinders Lofty Block) and highest ([F-J] Central Arnhem, Gibson Desert, Hampton, Tasmanian Northern Slopes, Gulf Coastal) probabilities are shown.*

DISCUSSION

In this study, we explored the spatial patterns in sampling effort and inventory completeness across the 85 bioregions of continental Australia, as represented by vouchered specimens digitised in Australia's Virtual Herbarium. We demonstrated that: a) there is considerable variation in the density of records across bioregions, with sampling effort strongly influenced by HII; b) the inventories of most bioregions are relatively complete; and c) the bioregions most likely to yield a species previously unreported in that region are patchily distributed across the continent, and not necessarily the least sampled or the most incomplete (Fig. 3a).

Exploring spatial patterns in natural history collections is necessary to understand biases in sampling effort and identify geographic regions that have been under-sampled (Wieringa *et al.* 2004) with respect to their true richness. Certainly, the density of specimen records varies across Australia's bioregions, ranging from 0.02 records/km^2 in Gibson Desert to 8.37 records/km^2 in Wet Tropics. This, in itself, is unsurprising: historically, spatial patterns in specimen collections have been driven by human settlement and accessibility (Aikio *et al.* 2010; Reddy and Dávalos 2003) as well as established patterns of plant diversity.

Spatial characteristics of *ad hoc* specimen collection often reflect the "roadmap effect" (Küper *et al.* 2006; Nelson *et al.* 1990), whereby the accessibility of areas close to, or at junctions of, major roads means these areas are likely to be better sampled than regions further away. Indeed, we found a strong positive correlation between HII and sampling effort, i.e. the most intensively sampled bioregions tended to be those most densely populated and urbanised, such as the Sydney Basin. In contrast, bioregions in the arid interior are generally inaccessible and sparsely populated, thereby having a lower density of records - a pattern that is consistent with other arid regions of the world (Newbold 2010). The exception in central Australia is

MacDonnell Ranges: this bioregion is home to the remote township of Alice Springs and stands out from surrounding bioregions in terms of its higher sampling effort.

Bias in collection effort can also occur in hotspots of biodiversity or protected areas (Dennis and Thomas 2000). For example, the Warren bioregion in the south-west corner of Australia and the Wet Tropics bioregion in the north-east have among the highest density of records. These regions are recognised global hotspots of biodiversity (Sloan *et al.* 2014), making them attractive areas to conduct scientific research and for land managers to prioritise for data collection (Ens *et al.* 2014).

We analysed inventory completeness (C) for each bioregion by calculating species richness based on the Chao 1 estimator. An area with $C \geq 0.80$ is often regarded as representative of a well-collected sample (Mora *et al.* 2008; Schmidt-Lebuhn *et al.* 2012; Soberón *et al.* 2007). According to this threshold, inventory completeness for most bioregions across Australia (79%) is high, indicating that the Australian flora is well known at a *bioregional* level.

Sampling effort, however, only partly explains inventory completeness ($r = 0.43$, $p = 0.01$) (Appendix S3). For example, Northern Kimberley and Burt Plain have among the highest levels of completeness, but both have low levels of sampling effort. Incompleteness is partly due to sampling strategy. More coherent and systematic surveys will capture greater species richness than opportunistic surveys (Lister and Group 2011). Where sampling is spatially random, regions with lower species richness or where species tend to have larger ranges, will have more complete inventories, given the same number of samples.

The value of C may be artificially inflated due to two key factors. Firstly, the Chao 1 estimator was developed for abundance data (Chao 1984). However, NHCs usually consist of presence only data, and the number of collected specimens for a species may not be proportional to local or regional abundance. Secondly, the Chao 1 estimator is scale dependent, and extrapolating species richness to coarser scales may lead to some inflation of C (Soberón *et al.* 2007; Sousa-Baena *et al.* 2014a). We point out, however, that we explored C at the bioregional level as this is the spatial unit frequently used for describing species diversity patterns and developing national conservation strategies (Polak *et al.* 2015; Williamson *et al.* 2011). Further, earlier sensitivity analyses at a variety of spatial scales (25 km^2, 50 km^2, 100 km^2) showed consistent patterns, although at finer spatial scales reliable completeness scores could not be calculated for an increasingly larger number of cells (Haque 2015).

Despite the limitations of using the Chao 1 estimator with incidence data, non-parametric estimators are widely used with presence-only data (Ballesteros-Mejia *et al.* 2013; Soria-Auza and Kessler 2008; Sousa-Baena *et al.* 2014a).

Extending our knowledge of Australia's flora

NHCs are prone to inherent spatial biases in sampling effort, which have the potential to distort study outcomes (Newbold 2010). Given budget and logistic support limitations for surveys, this makes it all the more important to optimise future sampling strategies to increase the coverage and representativeness of NHCs (Gioia 2010; Hardisty and Roberts 2013; Vos *et al.* 2014). As such, we explored the probability that the next sample taken from a bioregion would represent a species previously unreported for that bioregion. At this spatial scale, we found Central Arnhem, a region known to be rich in endemic plants (Whiting *et al.* 2000), to have the highest probability of yielding a new species (11.69%).

Interestingly, the least densely sampled bioregions are not necessarily the regions of greatest opportunity. Similarly, the HII is weakly correlated with the probability that the next sample will represent a new species (r = -0.35) (see Appendix S5). We have also found that the probability of sampling an unrecorded species in a bioregion is not entirely explained by inventory completeness (r = -0.55) (fig. 4b). This is because the probability depends on the slope of the sampling curve at that point (see Appendix S1), which varies even among bioregions with similar inventory completeness. What this means is that completeness and past sampling effort are only partial indicators of where future sampling effort should be prioritised. The more informative indicator is the slope of the sampling curve, which directly measures our expected gain in knowledge for a given gain in sampling effort.

To conclude, although the amount of data available from online repositories of NHCs has increased, there are concerns about their completeness, quality and biases, not only in the spatial domain but also in taxonomic, environmental and temporal space (Beck *et al.* 2014; Meyer *et al.* 2016; Sousa-Baena *et al.* 2014b; Yang *et al.* 2013). These biases may limit our ability to anticipate biological responses to climate, and other environmental, changes. We found a general pattern of the Australian flora whereby, at the bioregional level, taxonomic sampling is relatively complete. We point out, however, that our study focuses on vouchered specimens. Observational records within the AVH and elsewhere may, of course, increase completeness of bioregional inventories, although unlike vouchered specimens the taxonomic accuracy of observations cannot be reassessed. We also emphasise that at spatial scales finer than bioregional levels there remain vast swathes of the Australian continent for which we do not have specimen records. However, prioritisation of future survey efforts, whether at bioregional or finer spatial scales, should not be guided purely by lack of occurrence records but should focus on locations where effort is most likely to yield new information. Our method for determining the probability of encountering a novel species provides a means for a more strategic approach to future sampling effort.

ACKNOWLEDGEMENTS

I thank Dr. John Baumgartner for his assitance in calculating HII index.

REFERENCES

Aikio S., Duncan R. P. & Hulme P. E. (2010) Herbarium records identify the role of long-distance spread in the spatial distribution of alien plants in new zealand. *Journal of Biogeography.* **37**, 1740-1751.

Andrew C., Heegaard E., Kirk P. M., Bässler C., Heilmann-Clausen J., Krisai-Greilhuber I., Kuyper T. W., Senn-Irlet B., Büntgen U. & Diez J. (2017) Big data integration: Pan-european fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. *Fungal Biology Reviews* **31**, 88-98.

Ballesteros-Mejia L., Kitching I. J., 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.

Barker R. M. & Barker W. R. (1990) Botanical contributions overlooked: The role and recognition of collectors, horticulturists, explorers and others in the early documentation of the australian flora. In: *In history of systematic botany in australasia* (ed P. Short) pp. 37-85. Australian Systematic Botany Society:Melbourne.

Barker W. R. (1998) The virtual australian herbarium. Australian Herbarium Information Systems Committee (HISCOM).

Barthlott W., Hostert A., Kier G., Küper W., Kreft H., Mutke J., Rafiqpoor M. D. & Sommer J. H. (2007) Geographic patterns of vascular plant diversity at continental to global scales. *Erdkunde* **61**, 305-315.

Baumgartner J. B., Esperón-Rodríguez M. & Beaumont L. J. (2018) Identifying in situ climate refugia for plant species. *Ecography* **41**, 1-14.

Bebber D. P., Carine M. A., Wood J. R., Wortley A. H., Harris D. J., Prance G. T., Davidse G., Paige J., Pennington T. D. & Robson N. K. (2010) Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences* **107**, 22169-22171.

Beck J., Böller M., Erhardt A. & Schwanghart W. (2014) Spatial bias in the gbif database and its effect on modeling species' geographic distributions. *Ecological Informatics* **19**, 10-15.

Bickford S. A., Laffan S. W., Kok R. P. & Orthia L. A. (2004) Spatial analysis of taxonomic and genetic patterns and their potential for understanding evolutionary histories. *Journal of Biogeography.* **31**, 1715-1733.

Bloom T. D. S., Flower A. & Dechaine E. G. (2018) Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis. *Ecology and Evolution* **8**, 765-777.

Boakes E. H., Mcgowan P. J., Fuller R. A., Chang-Qing D., Clark N. E., O'connor K. & Mace G. M. (2010) Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology.* **8**, e1000385.

Brose U., Martinez N. D. & Williams R. J. (2003) Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology* **84**, 2364-2377.

Bystriakova N., Peregrym M., Erkens R. H. J., Bezsmertna O. & Schneider H. (2012) Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. *Systematics and Biodiversity.* **10**, 305-315.

Ceballos G., Ehrlich P. R., Barnosky A. D., García A., Pringle R. M. & Palmer T. M. (2015) Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances* **1,** e1400253.

Chao A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**, 265-270.

Chao A., Chiu C. H., Hsieh T., Davis T., Nipperess D. A. & Faith D. P. (2015) Rarefaction and extrapolation of phylogenetic diversity. *Methods in Ecology and Evolution* **6**, 380-388.

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

Colloff M. J., Ward K. A. & Roberts J. (2014) Ecology and conservation of grassy wetlands dominated by spiny mud grass pseudoraphis spinescens in the southern murray–darling basin, australia. *Aquat. Conservation: Marine Freshwater Ecosystems.* **24**, 238-255.

Colwell R. K. & Coddington J. A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **345**, 101-118.

Crisp M. D., Laffan S., Linder H. P. & Monro A. (2001) Endemism in the australian flora. *Journal of Biogeography.* **28**, 183-198.

Dennis R. & Thomas C. (2000) Bias in butterfly distribution maps: The influence of hot spots and recorder's home range. *Journal of Insect Conservation.* **4**, 73-77.

Drew J. (2011) The role of natural history institutions and bioinformatics in conservation biology. *Conservation Biology.* **25**, 1250-1252.

Duursma D. E., Gallagher R. V., Roger E., Hughes L., Downey P. O. & Leishman M. R. (2013) Next-generation invaders? Hotspots for naturalised sleeper weeds in australia under future climates. *PloS One* **8**, e84222.

Ens E. J., Pert P., Budden M., Clarke P. A., Clubb L., Doran B., Douras C., Gaikwad J., Gott B. & Leonard S. (2014) Indigenous biocultural knowledge in ecosystem science and management: Review and insight from australia. *Biological Conservation.* **181**, 133-149.

Ferrier S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? *Systematic Biology* **51**, 331-363.

Franklin J., Serra-Diaz J. M., Syphard A. D. & Regan H. M. (2017) Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeogr.* **26**, 6-17.

Gallagher R. V. (2016) Correlates of range size variation in the australian seed-plant flora. *Journal of Biogeography.* **43**, 1287-1298.

Gallagher R. V., Hughes L., Leishman M. R. & Wilson P. D. (2010) Predicted impact of exotic vines on an endangered ecological community under future climate change. *Biological Invasions* **12**, 4049-4063.

Gardner J. L., Amano T., Sutherland W. J., Joseph L. & Peters A. (2014) Are natural history collections coming to an end as time-series? *Frontiers in Ecology and the Environment.* **12**, 436-438.

George A. (1981) The background to the flora of australia. In: *In flora of australia* Vol. 1 pp. 3-24. Bureau of Flora and Fauna: Canberra, ACT, Australia.

Gioia P. (2010) Managing biodiversity data within the context of climate change: Towards best practice. *Austral Ecology* **35**, 392-405.

González-Orozco C. E., Ebach M. C., Laffan S., Thornhill A. H., Knerr N. J., Schmidt-Lebuhn A. N., Cargill C. C., Clements M., Nagalingum N. S. & Mishler B. D. (2014) Quantifying phytogeographical regions of australia using geospatial turnover in species composition. *PloS One* **9**, e92558.

Gonzalez-Orozco C. E., Pollock L. J., Thornhill A. H., Mishler B. D., Knerr N., Laffan S. W., Miller J. T., Rosauer D. F., Faith D. P., Nipperess D. A., Kujala H., Linke S., Butt N., Kulheim C., Crisp M. D. & Gruber B. (2016) Phylogenetic approaches reveal biodiversity threats under climate change. *Nature Climate Change* **6**, 1110-1114.

González-Orozco C. E., Pollock Laura j., Thornhill Andrew h., Mishler Brent d., Knerr N., Laffan Shawn w., Miller Joseph t., Rosauer Dan f., Faith Daniel p., Nipperess David a., Kujala H., Linke S., Butt N., Külheim C., Crisp Michael d. & Gruber B. (2016) Phylogenetic approaches reveal biodiversity threats under climate change. *Nature Climate Change* **6**, 1110-1115.

Gotelli N. & Entsminger G. (2004) Ecosim: Null models software for ecology. Version 7. Acquired intelligence inc. & kesey-bear. Jericho, vt 05465. *Computer software. http://garyentsminger. com/ecosim/index. htm.(accessed on 19-03-2016).*

Gotelli N. J. & Colwell R. K. (2011) *Estimating species richness,* Oxford University Press, Oxford, UK. .

Graham C., Ferrier S., Huettman F., Moritz C. & Peterson A. (2004a) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution.* **19**, 497-503.

Graham C. H., Ferrier S., Huettman F., Moritz C. & Peterson A. T. (2004b) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution.* **19**, 497-503.

Grand J., Cummings M. P., Rebelo T. G., Ricketts T. H. & Neel M. C. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecological Letters.* **10**, 364-374.

Green J. (1990) History of early western australian herbaria. In: *History of systematic botany in australasia* (ed P. Short) pp. 23-27.

Haque M. M. (2015) A legacy of sampling: Exploring spatial patterns among occurrence records in australia's virtual herbarium (Master of Research dissertation). Department of Biological Sciences, Macquarie University NSW Australia, http://hdl.handle.net/1959.14/1050711.

Haque M. M., Nipperess D. A., Gallagher R. V. & Beaumont L. J. (2017) How well documented is australia's flora? Understanding spatial bias in vouchered plant specimens. *Austral Ecol.* **42**, 690-699.

Hardisty A. & Roberts D. (2013) A decadal view of biodiversity informatics: Challenges and priorities. *BMC Ecology.* **13**, 16.

Hewitt J. E., Thrush S. F. & Ellingsen K. E. (2016) The role of time and species identities in spatial patterns of species richness and conservation. *Conservation Biology.* **30**, 1080-1088.

Hijmans R., Garrett K., Huaman Z., Zhang D., Schreuder M. & Bonierbale M. (2000) Assessing the geographic representativeness of genebank collections: The case of bolivian wild potatoes. *Conservation Biology.* **14**, 1755-1765.

Hortal J., Borges P. A. & Gaspar C. (2006) Evaluating the performance of species richness estimators: Sensitivity to sample grain size. *Journal of Animal Ecology.* **75**, 274-287.

Hortal J., De Bello F., Diniz-Filho J. a. F., Lewinsohn T. M., Lobo J. M. & Ladle R. J. (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46**, 523-549.

Hortal J., Jimenez-Valverde A., Gomez J. F., Lobo J. M. & Baselga A. (2008a) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847-858.

Hortal J., Jiménez-Valverde A., Gómez J. F., Lobo J. M. & Baselga A. (2008b) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847-858.

Hortal J., Lobo J. M. & Jiménez-Valverde A. (2007) Limitations of biodiversity databases: Case study on seed-plant diversity in tenerife, canary islands. *Conservation Biology.* **21**, 853-863.

Isaac N. J. & Pocock M. J. (2015) Bias and information in biological records. *Biological Journal of the Linnean Sociesty.* **115**, 522-531.

Jansen F. & Dengler J. (2010) Plant names in vegetation databases – a neglected source of bias. *Journal of Vegetation Science* **21**, 1179-1186.

Jetz W., Mcpherson J. M. & Guralnick R. P. (2012) Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology & Evolution.* **27**, 151-159.

Küper W., Sommer J., Lovett J. & Barthlott W. (2006) Deficiency in african plant distribution data–missing pieces of the puzzle. *Botanical Journal of the Linnean Society.* **150**, 355-368.

Laffan S. W. & Crisp M. D. (2003) Assessing endemism at multiple spatial scales, with an example from the australian vascular flora. *Journal of Biogeography.* **30**, 511-520.

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.

Lee A. C. & Mishler B. (2014) Phylogenetic diversity and endemism: Metrics for identifying critical regions of conifer conservation in australia. *Berkeley Scientific Journal* **18**, 48-58.

Lister A. M. & Group C. C. R. (2011) Natural history collections as sources of long-term datasets. *Trends in Ecology and Evolution.* **26**, 153-154.

Mackey B. G., Berry S. L. & Brown T. (2008) Reconciling approaches to biogeographical regionalization: A systematic and generic framework examined with a case study of the australian continent. *Journal of Biogeography.* **35**, 213-229.

Mellick R., Lowe A. & Rossetto M. (2011) Consequences of long-and short-term fragmentation on the genetic diversity and differentiation of a late successional rainforest conifer. *Australian Journal of Botany*. **59**, 351-362.

Meyer C., Kreft H., Guralnick R. & Jetz W. (2015) Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6**, 8221-8229.

Meyer C., Weigelt P. & Kreft H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecological Letters*. **19**, 992-1006.

Mihoub J. B., Henle K., Titeux N., Brotons L., Brummitt N. A. & Schmeller D. S. (2017) Setting temporal baselines for biodiversity: The limits of available monitoring data for capturing the full impact of anthropogenic pressures. *Scientific Reports* **7**, 1-10.

Moerman D. E. & Estabrook G. F. (2006) The botanist effect: Counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography*. **33**, 1969-1974.

Mora C., Tittensor D. P. & Myers R. A. (2008) The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society B: Biological Sciences* **275**, 149-155.

Nelson B. W., Ferreira C. A., Da Silva M. F. & Kawasaki M. L. (1990) Endemism centres, refugia and botanical collection density in brazilian amazonia. *Nature* **345**, 714-716.

Newbold T. (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography* **34**, 3-22.

O'donnell J., Gallagher R. V., Wilson P. D., Downey P. O., Hughes L. & Leishman M. R. (2012) Invasion hotspots for non-native plants in australia under current and future climates. *Global Change Biology*. **18**, 617-629.

Oksanen J., Blanchet F. G., Kindt R., Legendre P., Minchin P. R., O'hara R., Simpson G. L., Solymos P., Stevens M. H. H. & Wagner H. (2016) Vegan: Community ecology package. R package version 2.3-3. Https://cran.R-project.Org/package=vegan.

Olszewski T. D. (2004) A unified mathematical framework for the measurement of richness and evenness within and among multiple communities. *Oikos* **104**, 377-387.

Page L. M., Macfadden B. J., Fortes J. A., Soltis P. S. & Riccardi G. (2015) Digitization of biodiversity collections reveals biggest data on biodiversity. *Bioscience* **65**, 841-842.

Pebesma E. J. & Bivand R. S. (2005) Classes and methods for spatial data in r. *R news* **5**, 9-13.

Peterson A. T., Navarro-Siguenza A. & Scachetti Pereira R. (2004) Detecting errors in biodiversity data based on collectors' itineraries. *Bulletin of the British ornithologists' club* **124**, 143-151.

Polak T., Watson J. E., Fuller R. A., Joseph L. N., Martin T. G., Possingham H. P., Venter O. & Carwardine J. (2015) Efficient expansion of global protected areas requires simultaneous planning for species and ecosystems. *Royal Society open science* **2**, 150107.

Powney G. D. & Isaac N. J. (2015) Beyond maps: A review of the applications of biological records. *Biol. J. Linn. Soc.* **115**, 532-542.

Pyke G. H. & Ehrlich P. R. (2010) Biological collections and ecological/environmental research: A review, some observations and a look to the future. *Biological Reviews (Cambridge)* **85**, 247-266.

R Development Core Team. (2010) R development core team. In: *A language and environment for statistical computing* pp. 275-286. R Founder for Statistical Computing, Vienna, Austria.

Reddy S. & Dávalos L. M. (2003) Geographical sampling bias and its implications for conservation priorities in africa. *Journal of Biogeography*. **30**, 1719-1727.

Robbirt K. M., Davy A. J., Hutchings M. J. & Roberts D. L. (2011) Validation of biological collections as a source of phenological data for use in climate change studies: A case study with the orchid ophrys sphegodes. *Journal of Ecology*. **99**, 235-241.

Rocchini D., Hortal J., Lengyel S., Lobo J. M., Jimenez-Valverde A., Ricotta C., Bacaro G. & Chiarucci A. (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography* **35**, 211-226.

Roger S., Bivand R. & Pebesma E. (2013) Applied spatial data analysis with r. New York: Springer.

Rondinini C., Wilson K. A., Boitani L., Grantham H. & Possingham H. P. (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecological Letters.* **9**, 1136-1145.

Rosauer D., Laffan S. W., Crisp M. D., Donnellan S. C. & Cook L. G. (2009) Phylogenetic endemism: A new approach for identifying geographical concentrations of evolutionary history. *Molecular Ecology.* **18**, 4061-4072.

Sanderson E. W., Jaiteh M., Levy M. A., Redford K. H., Wannebo A. V. & Woolmer G. (2002) The human footprint and the last of the wild: The human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. *Bioscience* **52**, 891-904.

Santos A., Jones O. R., Quicke D. L. & Hortal J. (2010) Assessing the reliability of biodiversity databases: Identifying evenly inventoried island parasitoid faunas (hymenoptera: Ichneumonoidea) worldwide. *Insect Conservation and Diversity* **3**, 72-82.

Schmidt-Lebuhn A. N., Knerr N. J. & González-Orozco C. E. (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: The example of asteraceae in australia. *Journal of Biogeography.* **39**, 2072-2080.

Schmidt-Lebuhn A. N., Knerr N. J. & González-Orozco C. E. (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: The example of asteraceae in australia. *Journal of Biogeography.* **39**, 2072-2080.

Sloan S., Jenkins C. N., Joppa L. N., Gaveau D. L. & Laurance W. F. (2014) Remaining natural vegetation in the global biodiversity hotspots. *Biological Conservation.* **177**, 12-24.

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 T. (2004) Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 689-698.

Soria-Auza R. W. & Kessler M. (2008) The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: A case study of ferns from bolivia. *Diversity & Distributions.* **14**, 123-130.

Sousa-Baena M. S., Garcia L. C. & Townsend Peterson A. (2014b) Knowledge behind conservation status decisions: Data basis for “data deficient” brazilian plant species. *Biological Conservation.* **173**, 80-89.

Sousa-Baena M. S., Garcia L. C. & Peterson A. T. (2014a) Completeness of digital accessible knowledge of the plants of brazil and priorities for survey and inventory. *Diversity & Distribution.* **20**, 369-381.

Stropp J., Ladle R. J., Malhado M., Ana C., 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.

Sultana S., Baumgartner J. B., Dominiak B. C., Royer J. E. & Beaumont L. J. (2017) Potential impacts of climate change on habitat suitability for the queensland fruit fly. *Scientific Reports* **7**, 13025.

Ter Steege H., Pitman N. C., Phillips O. L., Chave J., Sabatier D., Duque A., Molino J.-F., Prévost M.-F., Spichiger R. & Castellanos H. (2006) Continental-scale patterns of canopy tree composition and function across amazonia. *Nature* **443**, 444-447.

Thackway, R., & Cresswell, I. D. (1995). An interim biogeographic regionalisation for Australia: a framework for setting priorities in the national reserves system cooperative program, version 4.0. Canberra: Australian Nature Conservation Agency.

Thomas C. (2009) Biodiversity databases spread, prompting unification call. *Science* **324**, 1632-1633.

Tingley M. W. & Beissinger S. R. (2009) Detecting range shifts from historical species occurrences: New perspectives on old data. *Trends in Ecology and Evolution.* **24**, 625-633.

Troia M. J. & Mcmanamay R. A. (2016) Filling in the gaps: Evaluating completeness and coverage of open-access biodiversity databases in the united states. *Ecology and Evolution* **6**, 4654-4669.

Troia M. J. & Mcmanamay R. A. (2017) Completeness and coverage of open-access freshwater fish distribution data in the united states. *Diversity & Distribution.* **23**, 1482-1498.

Tulig M., Tarnowsky N., Bevans M., Kirchgessner A. & Thiers B. M. (2012) Increasing the efficiency of digitization workflows for herbarium specimens. *Zookeys* **209**, 103-113.

Vollmar A., Macklin J. A. & Ford L. (2010) Natural history specimen digitization: Challenges and concerns. *Biodiversity informatics* **7**, 93-112.

Vos R. A., Biserkov M. J. V., Balech B., Beard N., Blissett M., Brenninkmeijer C., Van Dooren T., Eades D., Gosline G. & Groom Q. J. (2014) Enriched biodiversity data as a resource and service. *Biodiversity data journal* **2**, e1125.

Ward D. F. (2012) More than just records: Analysing natural history collections for biodiversity planning. *PloS one* **7**, e50346.

Webb J. B. (2003) *The botanical endeavour: Journey towards a flora of australia*. Surrey Beatty & Sons, New South Wales, Australia.

Whiting A. S., Lawler S. H., Horwitz P. & Crandall K. A. (2000) Biogeographic regionalization of australia: Assigning conservation priorities based on endemic freshwater crayfish phylogenetics. *Animal Conservation.* **3**, 155-163.

Wieringa J., Poorter L., Bongers F., Kouamé F. & Hawthorne W. (2004) Biodiversity hotspots in west africa; patterns and causes. *Biodiversity of West African forests: an ecological atlas of woody plant species*, 61-72.

Williamson G. J., Christidis L., Norman J., Brook B. W., Mackey B. & Bowman D. M. (2011) The use of australian bioregions as spatial units of analysis to explore relationships between climate and songbird diversity. *Pacific Conservation Biology.* **17**, 354-360.

Willis C. G., Ellwood E. R., Primack R. B., Davis C. C., Pearson K. D., Gallinat A. S., Yost J. M., Nelson G., Mazer S. J., Rossington N. L., Sparks T. H. & Soltis P. S. (2017) Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology and Evolution* **32**, 531-546.

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

Yesson C., Brewer P. W., Sutton T., Caithness N., Pahwa J. S., Burgess M., Gray W. A., White R. J., Jones A. C. & Bisby F. A. (2007a) How global is the global biodiversity information facility? *PLoS One* **2**, e1124.

Yesson C., Brewer P. W., Sutton T., Caithness N., Pahwa J. S., Burgess M., Gray W. A., White R. J., Jones A. C., Bisby F. A. & Culham A. (2007b) How global is the global biodiversity information facility? *PLoS One* **2**, e1124.

SUPPLEMENTARY INFORMATION

Appendix S1. *Probability of sampling a species previously unrecorded in that bioregion*

Following Olszewski (2004), the probability (*P*) that the next specimens collected in a given bioregion will represent a species previously unrecorded in that bioregion is equivalent to the expected gain in species richness from adding one additional individual (observation) to an individuals-based rarefaction curve. Thus, the value can be expressed as:

$$P = \hat{S}_{n+1} - S_n \qquad \text{(equation 1)}$$

where S_n is the observed species richness and $\hat{S}_{n+1}$ is the expected species richness after adding one additional observation.

To determine the value of $\hat{S}_{n+1}$, we must extrapolate beyond the known rarefaction curve (i.e. beyond the recorded number of individuals). Chao and Jost (2012) and Chao *et al.* (2015) derived a formula for the smooth extrapolation of the curve for any number of additional observations as follows:

$$\hat{S}_{n+m} = S_n + \hat{f}_0 \left[1 - \left(1 - \frac{f_1}{n\hat{f}_0 + f_1}\right)^m\right] \qquad \text{(equation 2)}$$

where *m* is the number of additional observations, f_1 is the number of singletons (species observed only once) and $\hat{f}_0$ is the number of species that are present but unobserved (calculated using the Chao 1 species richness estimator).

In the case where *m*=1, equation 2 simplifies to:

$$\hat{S}_{n+1} = S_n + \hat{f}_0 \left(\frac{f_1}{n\hat{f}_0 + f_1}\right) \qquad \text{(equation 3)}$$

Substituting into equation 1:

$$P = \hat{f}_0 \left(\frac{f_1}{n\hat{f}_0 + f_1}\right) \qquad \text{(equation 4)}$$

Reference

Sanderson E. W., Jaiteh M., Levy M. A., Redford K. H., Wannebo A. V. & Woolmer G. (2002) The Human Footprint and the Last of the Wild: the human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. *BioScience* **52**, 891-904.

Appendix S2. *Voucher specimens of native plant species digitised in Australia's Virtual Herbarium (AVH) and incorporated into our study, summarised for 85 bioregions in the Interim Biogeographic Regionalisation of Australia (IBRA). Information for each bioregion includes area, number of occurrence records, species richness, number of families, and four measures: Sampling intensity and C (index of inventory completeness), HII (Human Influence Index, based on (Sanderson et al. 2002)), probability that one additional record collected from a bioregion would represent a species previously unrecorded in that bioregion.*

Bioregion	Area (km^2)	No. records	Richness	No. Families	Sampling effort	C	HII-index	Probability (in %)
Arnhem Coast	33,356	17,935	1,662	163	0.54	0.88	8	1.51
Arnhem Plateau	23,060	17,245	1,588	157	0.75	0.87	0	1.69
Australian Alps	12,329	52,067	1,620	138	4.22	0.82	12	0.63
Avon Wheatbelt	95,171	99,544	4,095	124	1.05	0.88	20	0.70
Ben Lomond	6,575	9,387	1,115	132	1.43	0.85	12	2.62
Brigalow Belt North	136,745	39,219	2,889	197	0.29	0.84	7	1.52
Brigalow Belt South	272,197	99,833	3,833	203	0.37	0.85	8	0.72
Broken Hill Complex	56,354	11,267	841	83	0.20	0.79	6	1.85
Burt Plain	7,797	14,934	1,009	93	0.20	0.88	0	1.06
Cape York Peninsula	122,564	86,979	3,108	209	0.71	0.87	6	0.50
Carnarvon	84,301	14,689	1,312	102	0.17	0.77	6	2.28
Central Arnhem	34,624	2,897	921	129	0.08	0.77	0	11.70
Central Kimberley	76,755	13,562	1,386	126	0.18	0.84	0	2.24
Central Mackay Coast	14,642	21,012	2,238	203	1.44	0.83	16	2.40
Central Ranges	101,640	24,469	1,092	88	0.24	0.81	0	0.87
Channel Country	304,094	27,370	1,278	100	0.09	0.83	0	0.99
Cobar Peneplain	73,853	17,980	1,283	116	0.24	0.80	6	1.76
Coolgardie	129,122	46,707	2,473	96	0.36	0.81	6	1.18
Daly Basin	20,922	10,032	1,298	136	0.48	0.84	6	3.00
Dampierland	83,608	15,109	1,323	124	0.18	0.85	6	1.79
Darling Riverine Plains	106,997	16,479	1,453	114	0.15	0.74	13	2.54
Darwin Coastal	28,431	24,442	1,872	163	0.86	0.86	6	1.28
Davenport Murchison Ranges	58,051	8,864	927	87	0.15	0.84	0	2.40
Desert Uplands	69,410	16,845	1,519	129	0.24	0.80	6	2.08
Einasleigh Uplands	116,257	58,838	3,318	216	0.51	0.82	6	1.23
Esperance Plains	29,213	75,234	3,258	129	2.58	0.89	12	0.64
Eyre Yorke Block	61,204	71,205	1,625	108	1.16	0.87	21	0.38
Finke	72,674	14,177	933	90	0.20	0.76	6	1.40
Flinders Lofty Block	66,157	97,705	2,293	141	1.48	0.82	7	0.43
Furneaux	5,375	18,581	1,315	152	3.46	0.85	15	1.24
Gascoyne	180,752	10,021	1,266	91	0.06	0.79	6	3.14
Gawler	120,028	31,888	1,306	102	0.27	0.85	6	0.76
Geraldton Sandplains	31,421	64,153	2,862	120	2.04	0.84	18	0.79
Gibson Desert	156,289	3,503	581	63	0.02	0.78	0	4.82
Great Sandy Desert	394,861	17,580	1,351	96	0.04	0.84	0	1.46
Great Victoria Desert	422,465	23,555	1,415	88	0.06	0.82	0	1.38
Gulf Coastal	27,117	5,924	1,088	124	0.22	0.84	0	4.56
Gulf Fall and Uplands	118,479	21,587	1,778	133	0.18	0.86	0	1.54
Gulf Plains	220,418	26,055	2,053	162	0.12	0.84	6	1.65
Hampton	10,881	2,784	368	66	0.26	0.66	9	4.78
Jarrah Forest	45,090	105,400	3,725	147	2.34	0.86	18	0.57
Kanmantoo	8,124	47,711	1,553	124	5.87	0.89	23	0.46

Bioregion	Area (km^2)	No. records	Richness	No. Families	Sampling effort	C	HII-index	Probability (in %)
King	4,255	6,717	838	133	1.58	0.83	16	2.95
Little Sandy Desert	110,898	5,137	828	74	0.05	0.85	0	3.81
MacDonnell Ranges	39,294	28,742	1,228	113	0.73	0.80	0	0.77
Mallee	73,975	60,706	3,095	103	0.82	0.86	11	0.88
Mitchell Grass Downs	334,687	21,780	1,654	117	0.07	0.80	6	1.93
Mount Isa Inlier	67,782	12,490	1,099	104	0.18	0.80	6	2.15
Mulga Lands	251,883	26,455	1,534	119	0.11	0.82	5	1.30
Murchison	281,205	36,261	2,001	100	0.13	0.77	6	1.37
Murray Darling Depression	199,583	73,246	2,238	137	0.37	0.86	16	0.50
Nandewar	27,019	16,531	1,766	150	0.61	0.74	14	3.07
Naracoorte Coastal Plain	24,582	32,754	1,564	129	1.33	0.83	21	0.83
New England Tablelands	30,022	52,893	2,517	176	1.76	0.83	14	1.00
Northern Kimberley	84,201	29,010	1,831	153	0.34	0.88	0	0.97
NSW North Coast	39,965	49,562	3,105	209	1.24	0.86	20	1.10
NSW South Western Slopes	86,811	36,027	2,309	154	0.42	0.78	23	1.61
Nullarbor	197,227	5,964	677	70	0.03	0.74	6	3.76
Ord Victoria Plain	125,407	17,675	1,498	125	0.14	0.80	0	1.93
Pilbara	178,231	37,101	1,496	103	0.21	0.81	6	0.74
Pine Creek	28,517	29,557	1,972	160	1.04	0.87	6	1.05
Riverina	97,044	28,354	1,822	126	0.29	0.81	23	1.54
Simpson Strzelecki Dunefields	279,842	17,340	1,002	90	0.06	0.78	0	1.33
South East Coastal Plain	17,492	28,609	1,943	158	1.64	0.85	34	1.30
South East Corner	25,320	59,760	2,599	184	2.36	0.85	14	0.74
South Eastern Highlands	83,759	112,031	3,412	194	1.34	0.85	15	0.55
South Eastern Queensland	78,049	147,546	4,077	233	1.89	0.84	18	0.42
Southern Volcanic Plain	24,403	16,256	1,655	142	0.67	0.83	28	2.17
Stony Plains	131,663	30,870	1,093	93	0.23	0.84	6	0.70
Sturt Plateau	98,575	6,470	971	104	0.07	0.81	0	4.08
Swan Coastal Plain	15,257	57,475	3,025	135	3.77	0.83	25	1.07
Sydney Basin	36,295	147,197	3,557	209	4.06	0.84	24	0.38
Tanami	259,972	15,796	1,188	95	0.06	0.83	0	1.60
Tasmanian Central Highlands	7,678	17,190	1,050	126	2.24	0.84	11	1.17
Tasmanian Northern Midlands	4,154	5,766	898	113	1.39	0.82	26	4.02
Tasmanian Northern Slopes	6,231	5,972	978	135	0.96	0.79	18	4.67
Tasmanian South East	11,318	36,451	1,638	153	3.22	0.84	15	0.76
Tasmanian Southern Ranges	7,572	20,681	1,317	141	2.73	0.85	12	1.19
Tasmanian West	15,650	16,392	993	138	1.05	0.80	7	1.37
Tiwi Cobourg	10,105	6,585	1,127	142	0.65	0.82	6	4.18
Victoria Bonaparte	73,012	27,997	1,965	152	0.38	0.87	6	1.20
Victorian Midlands	34,697	44,455	2,117	151	1.28	0.83	21	0.86

Bioregion	Area (km^2)	No. records	Richness	No. Families	Sampling effort	*C*	HII-index	Probability (in %)
Warren	8,447	32,951	1,811	131	3.90	0.83	15	1.03
Wet Tropics	19,891	166,598	4,020	250	8.38	0.88	18	0.34
Yalgoo	50,875	14,499	1,544	97	0.28	0.78	6	2.64

CHAPTER THREE

A JOURNEY THROUGH TIME: EXPLORING TEMPORAL PATTERNS AMONG DIGITISED PLANT SPECIMENS FROM AUSTRALIA.

(This chapter has been accepted in Systematics and Biodiversity, in Press)

ABSTRACT

Online access to species occurrence records has opened new windows to investigating biodiversity patterns across multiple scales. The value of these records for research depends on their spatial, temporal and taxonomic quality. We assessed temporal patterns in records from the Australasian Virtual Herbarium, asking: 1) how temporally consistent has collecting been across Australia?; 2) which areas of Australia have the most reliable records, in terms of temporal consistency and inventory completeness?; 3) are there temporal trends in the completeness of attribute information associated with records? We undertook a multi-step filtering procedure, then estimated temporal consistency and inventory completeness for sampling units (SUs) of 50km × 50km. We found temporal bias in collecting, with 80% of records collected over the period 1970-1999. South-eastern Australia, the Wet Tropics in north-east Queensland, and parts of Western Australia have received the most consistent sampling effort over time, whereas much of central Australia has had low temporal consistency. Of the SUs, 18% have relatively complete inventories with high temporal consistency in sampling. We also determined that 25% of digitised records had missing attribute information. By identifying areas with low reliability, we can limit erroneous inferences about distribution patterns and identify priority areas for future sampling.

INTRODUCTION

Over the last few decades, there has been considerable effort to digitise specimen records held in natural history collections (NHCs) (Page *et al.* 2015). Unprecedented improvement in access to NHCs via global databases, such as the Global Biodiversity Information Facility (GBIF) (http://www.gbif.org/), has opened a new window to exploring biodiversity patterns and changes from local to global scales (Andrew *et al.* 2017; Franklin *et al.* 2017; Powney and Isaac 2015).

NHC records constitute an important source of information on where a member of a species was at a given point in time. However, the value of NHCs for biodiversity research depends on the accuracy, adequacy and continuity of these collections (Isaac and Pocock 2015). In the context of NHCs, accuracy refers to the similarity of the record to the true details of the original observation or specimen (i.e. correct taxonomy, geographic location, date of observation or collection), adequacy refers to the completeness of the inventory of a collection at a particular spatial or temporal scale, and continuity refers to the temporal evenness (i.e. consistency of sampling effort over time) of the collection.

An important bias in NHCs arises due to uneven sampling effort, which leads to imperfect knowledge of the spatiotemporal distribution of biodiversity, referred to as the Wallacean shortfall (Hortal *et al.* 2015; Meyer *et al.* 2016). Deficiencies in the attribute information associated with species occurrence records – mainly due to outdated taxonomy (Bebber *et al.* 2010), errors in species identification, and imprecise georeferencing or dating format (Stropp *et al.* 2016) – can also preclude the use of some records. As a result, the usefulness of these databases may be diminished, compromising our capacity to describe existing biodiversity or to make accurate predictions of future patterns (Hortal *et al.* 2008; Ward 2012). Such biases and gaps in accessible digital information can also lead to misidentification of ecological and evolutionary processes and inefficient use of limited conservation resources (Grand *et al.* 2007; Sousa-Baena *et al.* 2014).

Identifying limitations in the primary biodiversity data (e.g. preserved specimens) that may arise from poor accuracy, adequacy or continuity of these records, is recognised as a priority area to achieve multiple targets specified by the Convention on Biological Diversity (Meyer *et al.* 2015). By identifying limitations and assessing their consequences, researchers will be better positioned to use and interpret NHC data appropriately. In addition, identification of information gaps will guide efforts to bridge them and thereby improve the quality of these collections (Stropp *et al.* 2016; Troia and McManamay 2016).

While spatial biases in NHCs have been scrutinised elsewhere (Beck *et al.* 2014; Bystriakova *et al.* 2012; Haque *et al.* 2017; Schmidt-Lebuhn *et al.* 2012; Sousa-Baena *et al.* 2014), assessments of temporal patterns and inconsistencies or incompleteness with respect to attribute information (i.e. taxonomic identity, collection date and geographic coordinates) have received little attention. Previous studies have found that NHC data often lack quality in terms of the detail of attribute information associated with specimen records, and that records from a given area have frequently been collected within a short period rather than consistently over longer time-frames (Boakes *et al.* 2010; Stropp *et al.* 2016). Limited attribute information and temporal gaps in collections of records may result in an inaccurate representation of a species' distribution or biodiversity patterns, and prevent the detection of long-term changes (Gardner *et al.* 2014; Tingley and Beissinger 2009).

Australia harbours a diverse flora and has a rich history of botanical sampling. The earliest preserved specimens date to the late 17th century and were collected by European explorers such as Dirk Harthog (Webb 2003) and William Dampier (Green 1990), while the first major botanical collection was undertaken by Joseph Banks and Daniel Solander in 1770

(Barker and Barker 1990) . As with material collected elsewhere around the world, many of the specimens dating to the late 18th and 19th centuries were, and in many instances remain, in overseas institutions, although some material has since been returned to Australia (Webb 2003). By the mid-20th century, systematic botany in Australia had entered in a new phase whereby collections and monographic works were developed in tandem across the continent (George 1981).

The digitisation of specimens in Australian herbaria began in the mid-1970s using in-house databases (Barker 1998). To date, currently an estimated 80% of specimens have been databased (http://avh.chah.org.au/index.php/about/). In 2001, the Australian Virtual Herbarium (now known as the Australasian Virtual Herbarium, AVH, www.avh.chah.org.au) was established, and now contains more than eight million records from all of the state and territory herbaria, as well as from several universities. Digital access to herbaria collections via the AVH has proven valuable in a range of contexts, from conservation planning (Gallagher 2016) and biosecurity (Sultana *et al.* 2017) to evolutionary studies (Gallagher 2016; González-Orozco *et al.* 2016).

Like other NHCs, AVH suffers from spatial biases (Haque *et al.* 2017). However, to date, there has been no analysis of temporal biases in this collection, or patterns in the quality of the attribute information associated with records. Hence, here we specifically address the following questions: 1) How temporally consistent has the collection of specimens been across Australia?; 2) For which areas of Australia do we have the most reliable history of species' occurrences, where reliability is defined as high inventory completeness and high temporal sampling consistency? and 3) Are there temporal trends in the completeness of the attribute information associated with records?

METHODS

Data filtering process

Initially, we retrieved 4,528,541 digitised plant occurrence records available in the Australasian Virtual Herbarium (AVH) via the Atlas of Living of Australia (ALA, www.ala.org.au), by accessing data for each native plant family identified by the Australian Plant Census (APC) (*n* = 299 families) (data extracted on 15 November 2016). Species names were matched to those classified as native in the Australian Plant Census list (www.anbg.gov.au/chah/apc/about-APC.html). APC covers all published scientific names used in an Australian context in the taxonomic literature and is endorsed by the Council of Heads of Australasian Herbaria (CHAH).

We excluded records (a) for which there was not a preserved specimen, (b) were hybrids, or (c) were located outside the geographic boundary of the Australian coastline. Of the original set of records, we retained 83% (3,756,475) representing 25,487 taxa within 299 families.

We then applied several data filters to subset records that did not meet all of the following criteria: (a) included geographic coordinates, (b) included the complete collection date (i.e. day, month and year), and (c) identified to species level or lower. From this subset we excluded duplicate specimens, which we defined following Stropp *et al.* (2016) as: two or more records with the same species name and date of collection, and which were collected from within 1 km of each other (to account for records with rounded or truncated geographic coordinates). We note that duplicate records may arise when a collector submits material taken from the same specimen on the same date to two or more institutions, each of which then gives a unique identification code to their record.

Temporal consistency in sampling effort

We overlayed collection records with a 50km × 50km grid across Australia using the spatial package ‘sp’ (Pebesma and Bivand 2005) for R version 3.0.1 (R Development Core Team 2010). At this spatial resolution, there are 3414 grid cells (or sampling units; SUs) across Australia, of which 30 have no specimen records. For each SU we calculated the median year of collection across all records. Temporal consistency of each SU was then defined as the coefficient of variation (CV) of record counts per decade (1800–2009). Because CV is unitless, it can be used to compare temporal variation in collecting effort between SUs that have very different mean numbers of records.

Inventory completeness

We estimated inventory completeness (the number of species recorded in an area as a proportion of all native, vascular plant species that exist in that area) for each SU, based on the cumulative number of specimens and species collected from 1800–2009. To do so, we calculated the number of sampling events for each SU, where a sampling event represents a unique combination of location (i.e. latitude and longitude) and date of collection. We then calculated the number of species recorded for each sampling event. Subsequently, we used the non-parametric Chao 1 estimator (Colwell and Coddington 1994) to calculate the expected number of species. This technique has been demonstrated to produce accurate non-parametric estimates across landscapes with varying biophysical conditions and is most frequently used for presence-only records (Ballesteros-Mejia *et al.* 2013; Hortal *et al.* 2006; Schmidt-Lebuhn *et al.* 2012; Soria-Auza and Kessler 2008). The Chao 1 estimator calculates the total number of species likely to be present based on the number of rare species, identified as species sampled only once (i.e. singletons) or twice (i.e. doubletons), and an estimate of the number of unsampled species (based on extrapolating the asymptote of a rarefaction curve) in a sample. The conceptual basis of this estimator is the 'stopping rule' used in biodiversity sampling, i.e. additional species are unlikely to be found when all species in a sample are represented by at least two individuals (or samples) (Gotelli and Colwell 2011).

We calculated the expected number of species ($S_{\exp(i)}$) for SU$_i$ as:

$$S_{exp(i)} = S_{obs(i)} + \frac{a_i^2}{2b_i}$$

where $S_{obs(i)}$ is the number of species observed in SU$_i$, a_i and b_i are the number of species observed in only one or two sampling events in SU$_i$, respectively. The inventory completeness (C_i) for SU$_i$ was then estimated as:

$$C_i = \frac{S_{obs(i)}}{S_{exp(i)}}$$

The value of C_i can range from zero to one, with one indicating a complete inventory (all species present are recorded). To define the range at which values of *C* are stable, and thus more reliable, we assessed the relationship between *C* and the number of unique records (i.e. a unique combination of date, location and species name (Sousa-Baena *et al.* 2014)). As with Stropp *et al.* (2016), we found that a monotonic relationship exists above ~200 unique records (Appendix S1). Therefore, we present estimates of inventory completeness only for SUs that have more than 200 unique records. Finally, to identify areas of Australia with the most reliable records, we grouped SUs into four categories based whether their values for temporal consistency of sampling (CV) and *C* were above or below the median of all SUs.

RESULTS

Temporal consistency

The temporal consistency (i.e. CV of decadal record counts) of the collection of specimens within SUs ranged from 0.90–4.70, with a median of 2.31 (Fig. 1). In general, areas in central Australia (particularly in Western Australia) had higher variation in the number of specimens per decade, indicating that collecting has been undertaken inconsistently over time. In contrast, south eastern Australia, the Wet Tropics in north-east Queensland, and parts of south western Western Australia had a lower CV (Fig. 1).

Of the 3384 SUs across Australia for which there were specimen records, approximately 30% (spanning much of New South Wales, Victoria and Tasmania) had a median year of specimen collection between 1980–1989 (Fig. 2). Approximately 11% (381) had a median collection date before 1970 – these SUs, representing much older collections, were mostly in central Australia. Interestingly, substantial areas of northern central Australia, Western Australia and Queensland (regions that are very sparsely settled) had median collection dates in the 21st century. It is also important to note that between 1990–2009, no records were collected for 161 SUs, which are mostly located in south-western Australia.

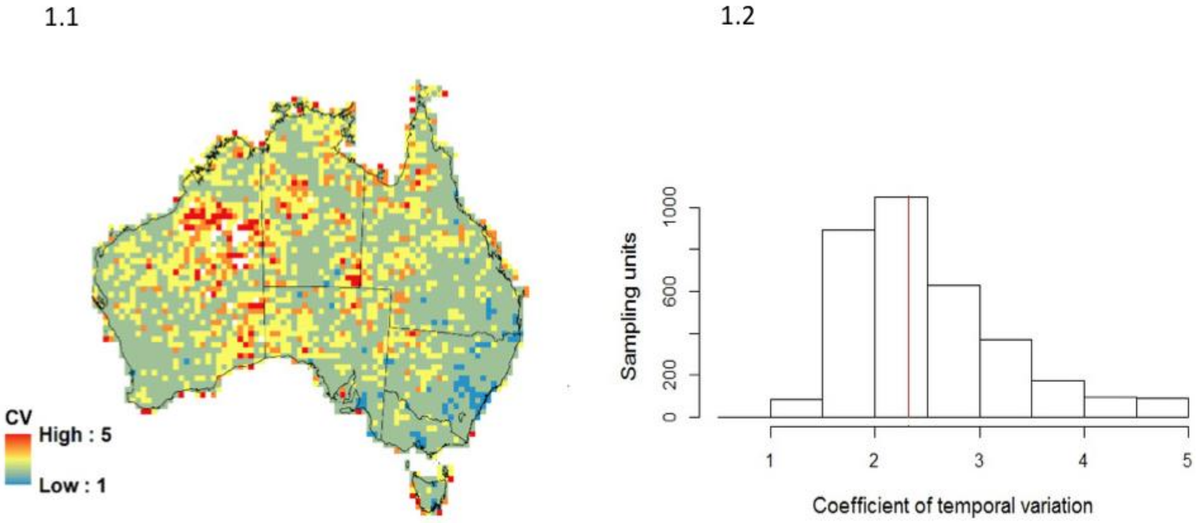


Figure 1. *(1) Spatial distribution and (2) frequency histogram of the coefficient of variation (CV) of plant specimens collected per decade (1800-2009) in each 50 km × 50 km sampling unit across Australia. Data are based on plant records digitised and included in the Australasian Virtual Herbarium. Lower values of CV indicate more consistent sampling effort over time. The red line in Figure 1.2 indicates the median value of CV (2.31).*

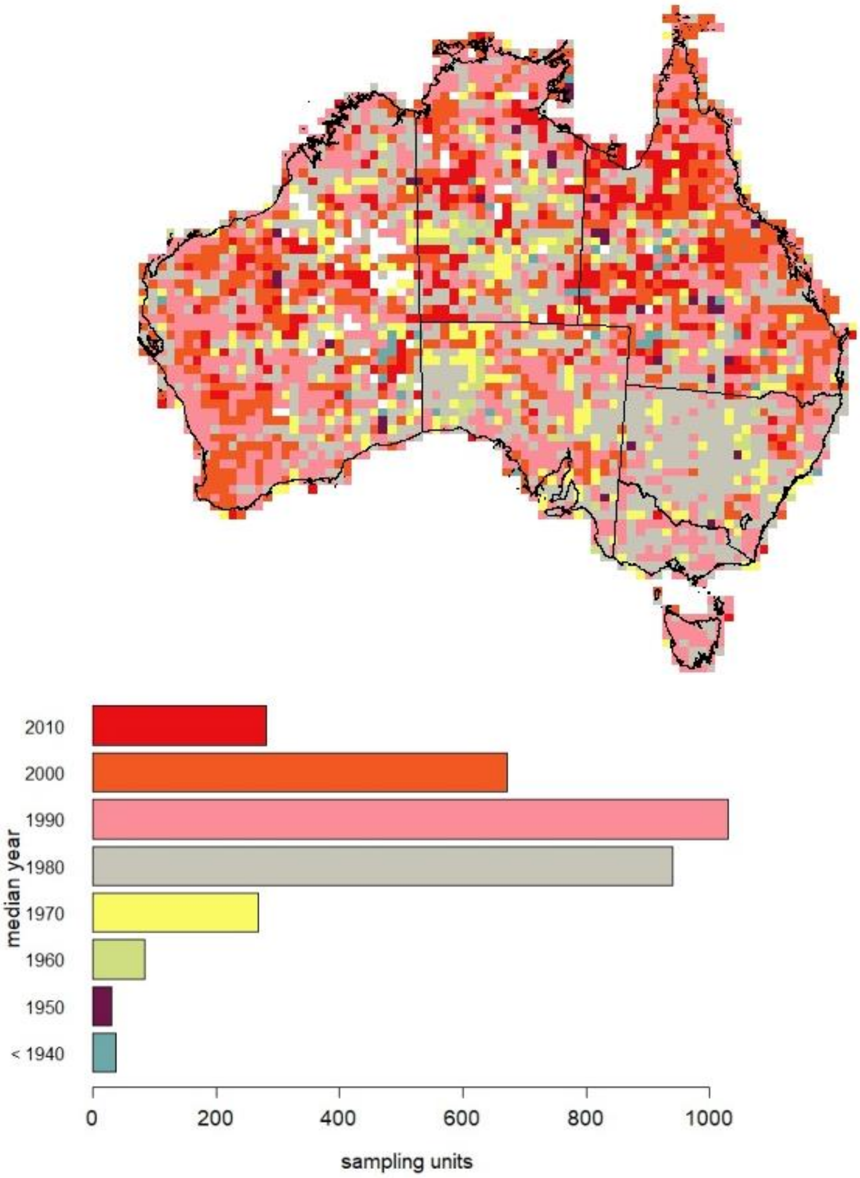


Figure 2. *The median year in which native plant specimens were collected in each 50 km × 50 km sampling unit across Australia, based on specimens in the Australasian Virtual Herbarium. The colours in the map refer to the appropriate category in the bar chart.*

Reliability of collection history

Of the 3384 SUs, 65% (1882) had sufficient records (i.e. at least 200 unique records) to estimate inventory completeness (*C*), which ranged from 0.29–0.75 with a median of 0.64. We grouped SUs into four categories based on their values for *C* and CV: a) $C > 0.64$ and $CV \leq 2.31$; b) $C > 0.64$ and $CV > 2.31$; c) $C \leq 0.64$ and $CV > 2.31$; d) $C \leq 0.64$ and $CV \leq 2.31$ (Fig. 3). The first of these categories represents SUs with relatively complete inventories that have been consistently sampled over time (i.e. values of *C* and CV are above the median). Only 18% (626) of SUs fell into this category, and these primarily occur around human settlements, particularly in the south-east and south-west of the continent, as well as the centre (around the town of Alice Springs), north-east Queensland (the Wet Tropics), eastern Tasmania, and the tip of the Northern Territory. The inventories of 9.5% (323) of SUs had high completeness despite also having high temporal variability (i.e. $C > 0.64$ and $\mathrm{CV} > 2.31$). Approximately 17% (607) of SUs fell into the third category, with low completeness and high temporal inconsistency ($C \leq 0.64$ and $\mathrm{CV} > 2.31$). Finally, the fourth category ($C \leq 0.64$ and $\mathrm{CV} \leq 2.31$, i.e. sampled consistently through time but have low completeness) was represented by 10% (334) of SUs, which were located sparsely across the continent.

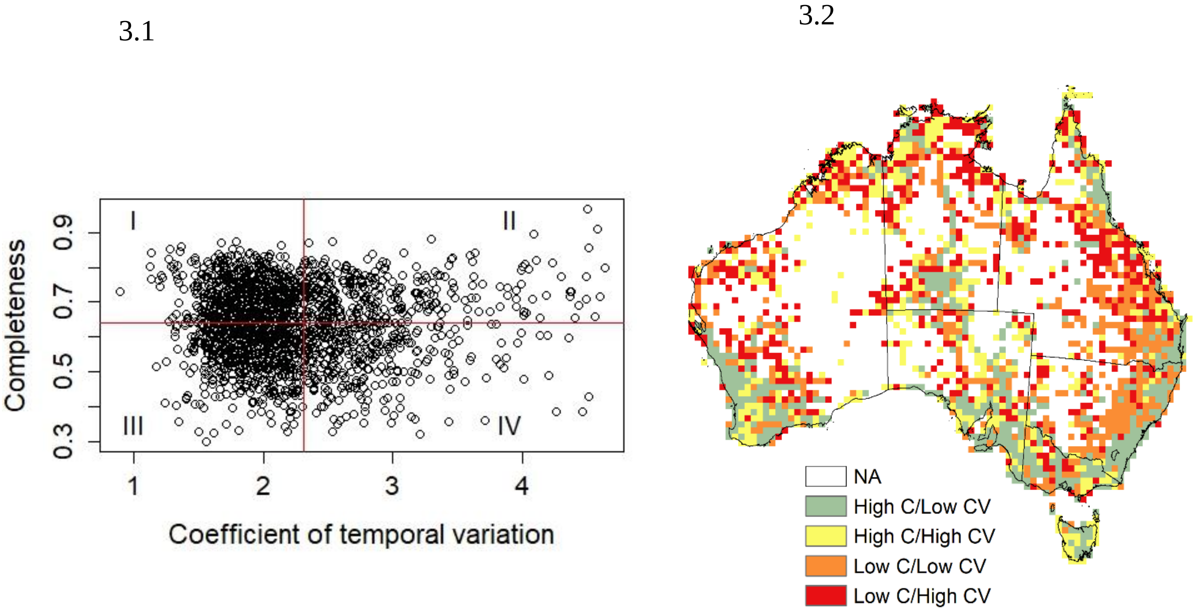


Figure 3. *(1) Scatter plot of inventory completeness (C, estimated for SUs with ≥ 200 records) and the coefficient of temporal variation (CV) in native Australian plant specimen records included in the Australasian Virtual Herbarium. The red lines indicate the median values of CV (2.31) and C (0.64), dividing the plot into four quadrats: (I) high C/low CV; (II) high C/high CV; (III) low C/low CV; (IV) low C/high CV. (2) The map displays the spatial distribution of the four quadrats. White areas represent cells with < 200 records, for which C could not be estimated.*

Completeness of attribute information

We assessed data quality in terms of the completeness of the attribute information included with AVH's records of preserved specimens. From the set of 3,756,475 native species records, we excluded 24.7% (929,765) as they lacked the event date (641,882), geographical coordinates (174,730), or were recorded as taxa above the species level (e.g., genus) (113,153). The exclusion of duplicates eliminated 12.8% of the remaining specimens (360,866). Our final dataset, therefore, contained 2,465,844 specimens belonging to 19,731 vascular plant species, within 296 families. We also note that the attribute information was incomplete for all physical specimens for three families (Rhacocarpaceae, Actinidiaceae and Dipteridaceae).

The coverage of attribute information associated with specimen records increased substantially from the 1960s, i.e. after data filtering 73% of post-1960 records were retained compared to only 20% of pre-1960 records (Fig. 4.1). This period also coincides with a substantial increase in the number of records collected, with the three decades spanning 1970–1999 contributing more records than had been accumulated over the period 1800–1969 (Fig. 4.2).

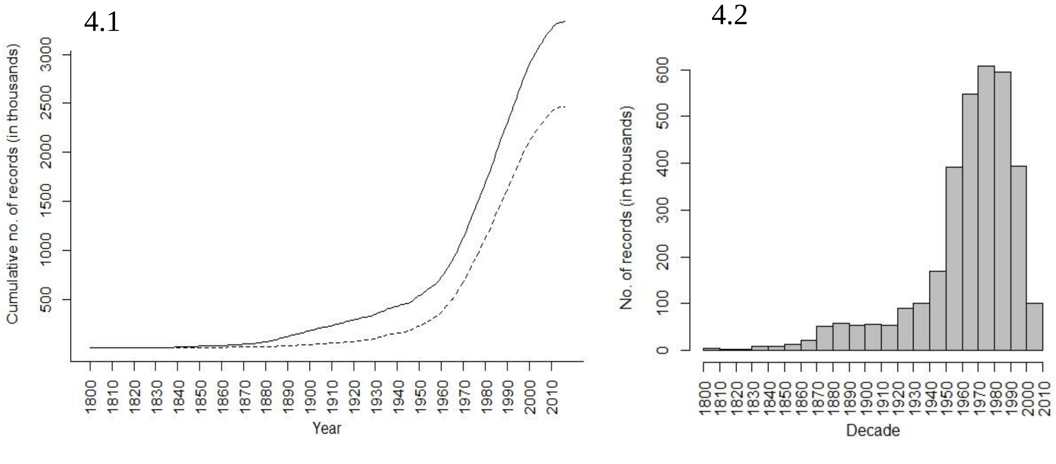


Figure 4. *(1) Records from the Australasian Virtual Herbarium were filtered to exclude those with incomplete information on specimen labels. The solid line indicates the number of records (in thousands) collected in a given year, prior to data filtering (i.e. removal of records with incomplete/uncertain information coverage on taxonomy, geographical coordinates and collection date). The dashed line indicates the number of records retained after filtering. (2) The number (in thousands) of native plant specimens (prior to filtering) indexed in the Australasian Virtual Herbarium collected in each decade from 1800 to 2009.*

DISCUSSION

Rapid improvements in the accessibility of the information that has accumulated in NHCs over the past few centuries is now enabling researchers to explore the temporal consistency of these collections (Hortal *et al.* 2008; Troia and McManamay 2016; 2017). Identifying and surveying areas that have not been sampled for a considerable time may uncover new spatio-temporal patterns and priorities for biodiversity (Meyer *et al.* 2016; Mihoub *et al.* 2017). Across Australia, the regions that have been sampled most consistently over the last ~200 years are mainly located in the north-east and south-east (see Fig. 2). This is likely a result of these areas being among the earliest sites of European settlement of the continent, and due to the ongoing growth of urban centres that followed. Indeed, Haque *et al.* (2017) found that areas with either high accessibility (i.e. Sydney Basin) or that are biodiversity hotspots (i.e. Wet Tropics) tend to have higher inventory completeness and sampling intensity, which is a common phenomenon of NHCs collections (Nelson *et al.* 1990). In contrast, a substantial proportion of the continent has experienced temporally inconsistent sampling (CV > 2.31; Fig. 1 & 2), likely because many of these areas represent remote regions far from human settlement.

Inventory completeness of species occurrence records has often been evaluated from a spatial perspective (Schmidt-Lebuhn *et al.* 2012; Sousa-Baena *et al.* 2014). However, the period over which occurrence records have accumulated is one of the key factors affecting the completeness of species inventories (Troia and McManamay 2016). Moreover, temporal bias in inventory completeness can yield an incomplete picture of biodiversity patterns over time (Rondinini *et al.* 2006), and have flow-on effects for tools that use these data, such as species distribution models (Mihoub *et al.* 2017). Therefore, the validity of scientific outcomes derived from studies based on these collections largely depends on the degree of the collections' completeness over time.

Our study suggests that the SUs with the most reliable samples (i.e. with both high completeness and high temporal consistency), span less than a quarter of the continent, and are mostly located in south-eastern Australia, north-eastern Queensland (Wet Tropics), and south-western Australia (see Fig. 3). These locations likely have the most rigorous baseline for measuring changes in biodiversity over time.

We also found that there is high inventory completeness for ~10% of SUs, despite these regions having high temporal variability. One explanation for this is that data may have been collected over a short period of time from these areas, a common phenomenon in natural history collections (Ward 2012). For example, it is entirely plausible that a recently surveyed area could be almost completely documented for its current biodiversity without any record of previous patterns. Therefore, high inventory completeness does not necessarily indicate that an area is well known: there may be considerable knowledge gaps in the historical pattern of species turnover.

Estimation of inventory completeness is also sensitive to spatial scale (Hortal *et al.* 2006; Soberón *et al.* 2007), and determining the spatial scale over which temporal patterns in biodiversity are consistent has direct implications for conservation planning (Hewitt *et al.* 2016). The spatial scale of any analysis will depend on the availability and quality of data, and the objectives of the study. Given the spatial biases and gaps in NHCs (Haque *et al.* 2017), estimating completeness at finer scales may not always be possible (Ferrier 2002) while adopting coarser scales may lead to its overestimation (Sousa-Baena *et al.* 2014).

Digitised NHCs possess enormous potential for directing biodiversity assessments and monitoring schemes (DREW 2011; Graham *et al.* 2004). However, failing to understand the temporal biases and gaps in these collections may hinder implementation of effective conservation strategies, particularly when habitat destruction and climate change have led to conspicuous reductions in biological diversity (Ceballos *et al.* 2015). Despite having more than 200 years of flora collection history, less than one-fifth of Australia is represented by a consistently-sampled set of digitised preserved specimens, with most of these locations being either easily accessible or in biodiversity hotspots (Haque *et al.* 2017). As such, we have poor knowledge of the magnitude of temporal changes in native flora for vast regions of the continent (see Fig. 3).

Our analysis identifies areas of lost opportunities as well as areas presenting future opportunities. In addition, we identify areas that have a history of sustained collection effort and for which future monitoring will enable long-term patterns in turnover to be recorded. Areas where we have lost the opportunity to capture previous patterns in flora are those with poor historical sampling (i.e. high CV and recent median year). These areas are mainly located in the northern region of Western Australia and upper Queensland (see Fig. 1 & 2). There are, however, two important sources of data that we have not accounted for: specimens yet to be digitised and specimens stored in institutions overseas. Repeating our study with data from GBIF may help to elucidate the extent to which specimens sent overseas, particularly those by early collectors, may distort estimates of inventory completeness and patterns in species' distributions.

Areas of future opportunity are those where the median year of specimen collection is now dated. This includes the western region of South Australia and central parts of the Northern Territory where the median year of collection occurs in the 1970s-80s. In contrast, areas of sustained collection effort are those with a long history of consistent and intensive sampling (high completeness and low CV). These areas are mainly located in the south-east and south-west of the continent (see Fig. 3) and are important contributors to our understanding of historical trends in Australia's flora. However, these areas may come to represent lost opportunities if collections do not take place in the future. As such, our findings may assist decision makers in establishing temporal baselines of Australia's native flora and implementing strategies for future sampling and digitisation effort.

Completeness of attribute information

The detail and accuracy of information recorded on specimen labels is rarely quantified but constitutes an important step in assessing the quality of data held in natural history collections (Boakes *et al.* 2010; Meyer *et al.* 2016; Stropp *et al.* 2016). We found that incomplete attribute information on specimen labels (i.e. lacking geographic coordinates or a complete date) were most common among older records, although more than half of the specimens collected in the last three decades (1980-2009) also had incomplete information. In addition, the duplicate records we omitted from our study (n = 360,866) may have also suffered from incorrect or incomplete labelling.

Missing attribute information from older records likely occurred for a range of reasons such as the lack of a standard taxonomic congurence in species identification due to difference sources of the collections (Jansen and Dengler 2010; Soberón and Peterson 2004), the unavailability of appropriate maps (leading to exclusion of geographic coordinates from records), and the difficulty in foreseeing how these data may be used in future (e.g. for species distribution modeling (Bloom *et al.* 2018; Willis *et al.* 2017) or phenological studies (Robbirt *et al.* 2011).

In addition, attribute errors can occur during the sampling event, during label transcription if the labels were not prepared at the time of collection (Peterson *et al.* 2004), or during the digitisation process (Vollmar *et al.* 2010; Willis *et al.* 2017). Further, errors may occur from a lack of precision or accuracy of the geographic coordinates assigned to the specimen’s location. Note that given the size of our database it was beyond the scope of this study to make judgments on the geographical accuracy of records, except for obvious errors such as when records of terrestrial species were assigned coordinates that placed them in the ocean.

The improvement of data quality involves error detection, correction and prevention (Yesson *et al.* 2007). While some errors may be corrected by reviewing the original specimen, this option may not always be viable. As such, it is crucial to scrutinize records before they are used and explicitly communicate their related spatial and temporal errors to avoid erroneous inferences (Hortal *et al.* 2015). To prevent future errors, it is important to promote best practices by adhering to standard protocols and methods for collecting and vouchering botanical data, and to implement error identification techniques to reduce errors that arise during the digitisation process.

CONCLUSION

Species occurrence data held within electronic databases are increasingly being applied to broad scale biogeographical and biodiversity research (Baumgartner *et al.* 2018; Jetz *et al.* 2012). However, the usefulness of these data for discerning biodiversity patterns largely depends on the quality, consistency and completeness of these collections along spatial and temporal dimensions. Given the magnitude of historical biases and limitations in the completeness of information within these massive collections, rapid improvement may not be possible due to limited resources and the lack of institutional investments (Gioia 2010; Moerman and Estabrook 2006; Tulig *et al.* 2012). Under such circumstances, one approach is to explore maps of opportunities and identify areas for which we can make reliable inferences about biodiversity patterns (Rocchini *et al.* 2011). The findings of our study may help to characterise species’ distributions and biodiversity patterns and identify where best to allocate limited resources to improve the quality and coverage of species occurrence data.

ACKNOWLEDGEMENTS

We thank Dr Rachael Gallagher and Stuart Allen for their assistance in data retrieval from Australasian Virtual Herbarium (AVH) via Atlas of Living Australia (ALA).

REFERENCES

Andrew C., Heegaard E., Kirk P. M., Bässler C., Heilmann-Clausen J., Krisai-Greilhuber I., Kuyper T. W., Senn-Irlet B., Büntgen U. & Diez J. (2017) Big data integration: Pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. *Fungal Biology Reviews* **31**, 88-98.

Ballesteros-Mejia L., Kitching I. J., 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.

Barker R. M. & Barker W. R. (1990) Botanical contributions overlooked: the role and recognition of collectors, horticulturists, explorers and others in the early documentation of the Australian flora. In: *In History of systematic botany in Australasia* (ed P. Short) pp. 37-85. Australian Systematic Botany Society:Melbourne.

Barker W. R. (1998) The Virtual Australian Herbarium. Australian Herbarium Information Systems Committee (HISCOM). .

Baumgartner J. B., Esperón-Rodríguez M. & Beaumont L. J. (2018) Identifying in situ climate refugia for plant species. *Ecography* **41**, 1-14.

Bebber D. P., Carine M. A., Wood J. R., Wortley A. H., Harris D. J., Prance G. T., Davidse G., Paige J., Pennington T. D. & Robson N. K. (2010) Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences* **107**, 22169-22171.

Beck J., Böller M., Erhardt A. & Schwanghart W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* **19**, 10-15.

Bloom T. D. S., Flower A. & Dechaine E. G. (2018) Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis. *Ecology and Evolution* **8**, 765-777.

Boakes E. H., Mcgowan P. J., Fuller R. A., Chang-Qing D., Clark N. E., O'connor K. & Mace G. M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology* **8**, e1000385.

Bystriakova N., Peregrym M., Erkens R. H. J., Bezsmertna O. & Schneider H. (2012) Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. *Systematics and biodiversity* **10**, 305-315.

Ceballos G., Ehrlich P. R., Barnosky A. D., García A., Pringle R. M. & Palmer T. M. (2015) Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances* **1**.

Colwell R. K. & Coddington J. A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **345**, 101-118.

Drew J. (2011) The Role of Natural History Institutions and Bioinformatics in Conservation Biology. *Conservation Biology* **25**, 1250-1252.

Ferrier S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology* **51**, 331-363.

Franklin J., Serra-Diaz J. M., Syphard A. D. & Regan H. M. (2017) Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography* **26**, 6-17.

Gallagher R. V. (2016) Correlates of range size variation in the Australian seed-plant flora. *Journal of Biogeography* **43**, 1287-1298.

Gardner J. L., Amano T., Sutherland W. J., Joseph L. & Peters A. (2014) Are natural history collections coming to an end as time-series? *Frontiers in Ecology and the Environment* **12**, 436-438.

George A. (1981) The background to the flora of Australia. In: *In Flora of Australia* Vol. 1 pp. 3-24. Bureau of Flora and Fauna: Canberra, ACT, Australia.

Gioia P. (2010) Managing biodiversity data within the context of climate change: towards best practice. *Austral Ecology* **35**, 392-405.

González-Orozco C. E., Pollock Laura j., Thornhill Andrew h., Mishler Brent d., Knerr N., Laffan Shawn w., Miller Joseph t., Rosauer Dan f., Faith Daniel p., Nipperess David a., Kujala H., Linke S., Butt N., Külheim C., Crisp Michael d. & Gruber B. (2016) Phylogenetic approaches reveal biodiversity threats under climate change. *Nature Climate Change* **6**, 1110-1115.

Gotelli N. J. & Colwell R. K. (2011) *Estimating species richness*, Oxford University Press, Oxford, UK. .

Graham C. H., Ferrier S., Huettman F., Moritz C. & Peterson A. T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution* **19**, 497-503.

Grand J., Cummings M. P., Rebelo T. G., Ricketts T. H. & Neel M. C. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecology Letters* **10**, 364-374.

Green J. (1990) History of early Western Australian herbaria. In: *History of systematic botany in Australasia* (ed P. Short) pp. 23-27.

Haque M. M., Nipperess D. A., Gallagher R. V. & Beaumont L. J. (2017) How well documented is Australia's flora? Understanding spatial bias in vouchered plant specimens. *Austral Ecology* **42**, 690-699.

Hewitt J. E., Thrush S. F. & Ellingsen K. E. (2016) The role of time and species identities in spatial patterns of species richness and conservation. *Conservation Biology* **30**, 1080-1088.

Hortal J., Borges P. A. & Gaspar C. (2006) Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology* **75**, 274-287.

Hortal J., De Bello F., Diniz-Filho J. a. F., Lewinsohn T. M., Lobo J. M. & Ladle R. J. (2015) Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46**, 523-549.

Hortal J., Jimenez-Valverde A., Gomez J. F., Lobo J. M. & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847-858.

Isaac N. J. & Pocock M. J. (2015) Bias and information in biological records. *Biological Journal of the Linnean Society* **115**, 522-531.

Jansen F. & Dengler J. (2010) Plant names in vegetation databases – a neglected source of bias. *Journal of Vegetation Science* **21**, 1179-1186.

Jetz W., Mcpherson J. M. & Guralnick R. P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology & Evolution* **27**, 151-159.

Meyer C., Kreft H., Guralnick R. & Jetz W. (2015) Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6**, 8221-8229.

Meyer C., Weigelt P. & Kreft H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**, 992-1006.

Mihoub J. B., Henle K., Titeux N., Brotons L., Brummitt N. A. & Schmeller D. S. (2017) Setting temporal baselines for biodiversity: the limits of available monitoring data for capturing the full impact of anthropogenic pressures. *Scientific Reports* **7**, 1-10.

Moerman D. E. & Estabrook G. F. (2006) The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography* **33**, 1969-1974.

Nelson B. W., Ferreira C. a. C., Da Silva M. F. & Kawasaki M. L. (1990) Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature* **345**, 714.

Page L. M., Macfadden B. J., Fortes J. A., Soltis P. S. & Riccardi G. (2015) Digitization of biodiversity collections reveals biggest data on biodiversity. *Bioscience* **65**, 841-842.

Pebesma E. J. & Bivand R. S. (2005) Classes and methods for spatial data in R. *R news* **5**, 9-13.

Peterson A. T., Navarro-Siguenza A. & Scachetti Pereira R. (2004) Detecting errors in biodiversity data based on collectors' itineraries. *Bulletin of the British ornithologists' club* **124**, 143-151.

Powney G. D. & Isaac N. J. (2015) Beyond maps: a review of the applications of biological records. *Biological Journal of the Linnean Society* **115**, 532-542.

R Development Core Team. (2010) R development core team. In: *A language and environment for statistical computing* pp. 275-286. R Founder for Statistical Computing, Vienna, Austria.

Robbirt K. M., Davy A. J., Hutchings M. J. & Roberts D. L. (2011) Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid Ophrys sphegodes. *Journal of Ecology* **99**, 235-241.

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

Rondinini C., Wilson K. A., Boitani L., Grantham H. & Possingham H. P. (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters* **9**, 1136-1145.

Schmidt-Lebuhn A. N., Knerr N. J. & González-Orozco C. E. (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. *Journal of Biogeography* **39**, 2072-2080.

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 T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 689-698.

Soria-Auza R. W. & Kessler M. (2008) The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: a case study of ferns from Bolivia. *Diversity and Distributions* **14**, 123-130.

Sousa-Baena M. S., Garcia L. C., Peterson A. T. & Brotons L. (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions* **20**, 369-381.

Stropp J., Ladle R. J., Malhado M., Ana C., 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.

Sultana S., Baumgartner J. B., Dominiak B. C., Royer J. E. & Beaumont L. J. (2017) Potential impacts of climate change on habitat suitability for the Queensland fruit fly. *Scientific Reports* **7**, 13025.

Tingley M. W. & Beissinger S. R. (2009) Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in Ecology & Evolution* **24**, 625-633.

Troia M. J. & Mcmanamay R. A. (2016) Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. *Ecology and Evolution* **6**, 4654-4669.

Troia M. J. & Mcmanamay R. A. (2017) Completeness and coverage of open-access freshwater fish distribution data in the United States. *Diversity and Distributions* **23**, 1482-1498.

Tulig M., Tarnowsky N., Bevans M., Kirchgessner A. & Thiers B. M. (2012) Increasing the efficiency of digitization workflows for herbarium specimens. *ZooKeys* **209**, 103-113.

Vollmar A., Macklin J. A. & Ford L. (2010) Natural history specimen digitization: challenges and concerns. *Biodiversity Informatics* **7**, 93-112.

Ward D. F. (2012) More than just records: analysing natural history collections for biodiversity planning. *PLoS One* **7**, e50346.

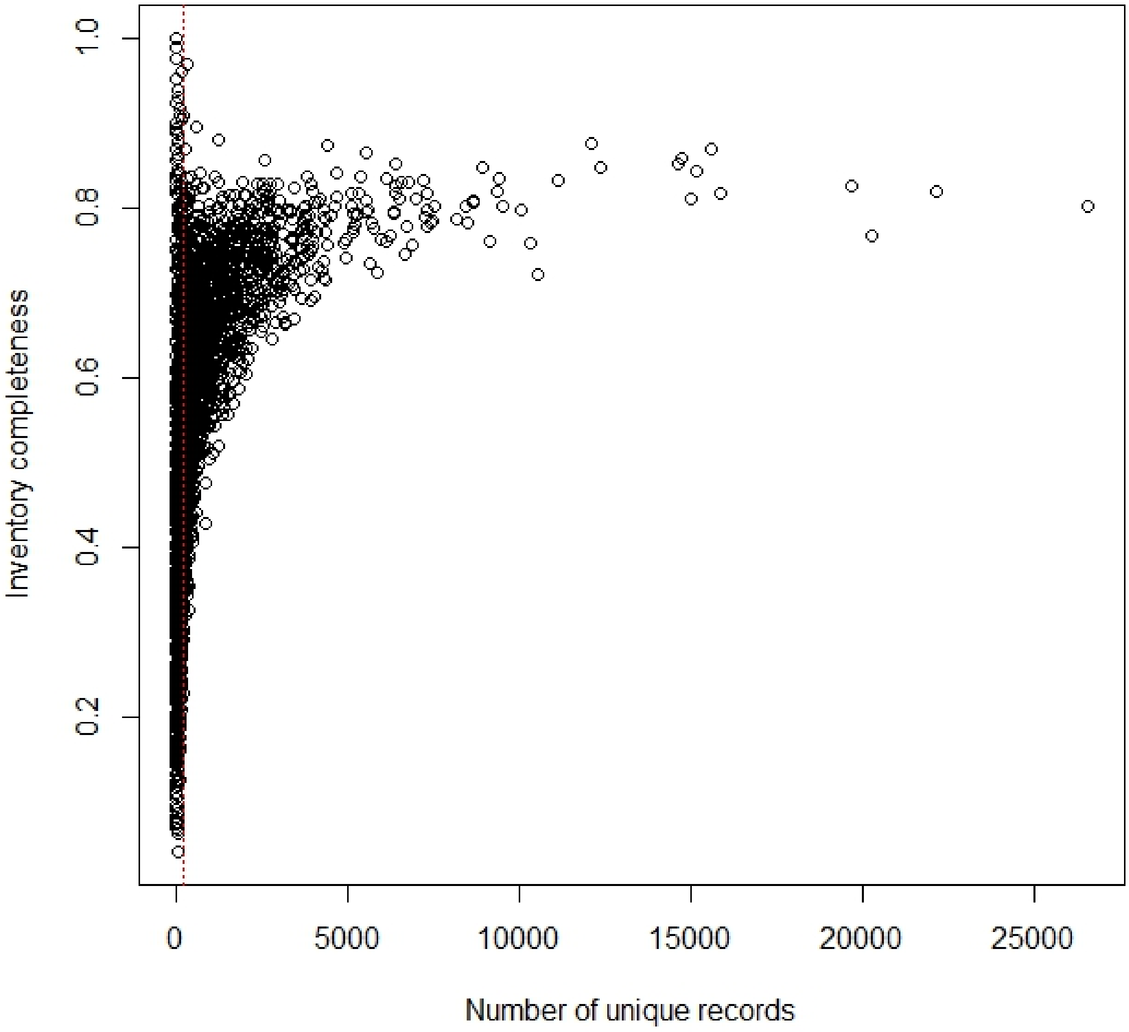
Webb J. B. (2003) *The botanical endeavour: journey towards a flora of Australia*. Surrey Beatty & Sons, New South Wales, Australia.

Willis C. G., Ellwood E. R., Primack R. B., Davis C. C., Pearson K. D., Gallinat A. S., Yost J. M., Nelson G., Mazer S. J., Rossington N. L., Sparks T. H. & Soltis P. S. (2017) Old Plants, New Tricks: Phenological Research Using Herbarium Specimens. *Trends in Ecology and Evolution* **32**, 531-546.

Yesson C., Brewer P. W., Sutton T., Caithness N., Pahwa J. S., Burgess M., Gray W. A., White R. J., Jones A. C. & Bisby F. A. (2007) How global is the global biodiversity information facility? *PLoS One* **2**, e1124.

SUPPLEMENTARY INFORMATION

Appendix S1. *Relationship between estimate of inventory completeness and number of unique records (i.e. unique combination of date, location of collection and species name); dashed line indicates 200 unique records. Each dot in the graph represents a sampling unit of 50 x 50 km. We found that a monotonic relationship exists above ~200 unique records. Therefore, we present estimates of inventory completeness only for sampling units that have more than 200 unique records*



CHAPTER FOUR

TAXONOMIC SHORTFALLS IN DIGITISED COLLECTIONS OF AUSTRALIA'S FLORA

ABSTRACT

Rapid growth in the digitisation of the world's natural history collections substantially simplifies scientific access to taxonomic and biogeographic information. Despite recent efforts to collate more than two centuries of biodiversity inventories into comprehensive databases, these collections suffer limitations across spatial, temporal and taxonomic dimensions. We assessed taxonomic shortfalls in preserved specimens from 296 plant families native to Australia, for which records have been collated into the Australasian Virtual Herbarium (AVH), specifically addressing the following questions: 1) Is a taxonomic bias apparent in this collection of preserved specimens? That is, based on the number of specimen records per species, are some plant families more, or less, likely to be collected? 2) To what extent does the distribution of collectors among plant families influence taxonomic bias? We found that the number of preserved specimens per family is not proportional to the family's known species richness. For 29% of Australia's plant families (i.e. 86), the number of digitised records constitutes < 50% of the number expected. Further, only 34% of families (100) have at least 20 specimens digitised for each species recorded in AVH. There is a strong positive correlation between the number of collectors sampling a family and the taxonomic bias of that family. A sound understanding of biodiversity is critical for megadiverse countries such as Australia, and identifying biases in digital inventories may help with establishing future sampling and digitisation strategies to enhance taxonomic representation.

INTRODUCTION

Human-induced global environmental change may be resulting in a sixth mass extinction of species (Ceballos *et al.* 2015). Preventing this event necessitates conservation decisions that are informed by the best available data (Geijzendorffer *et al.* 2016; Proença *et al.* 2017). Rapid growth in access to digital records of specimens housed in the world's natural history collections (NHC) offers considerable potential for expanding our understanding of biodiversity patterns across time and space, and facilitating conservation decision-making (Graham *et al.* 2004; Sullivan *et al.* 2017; Ward 2012). Despite recent efforts to collate more than two centuries of biodiversity inventories into comprehensive databases, these collections suffer biases and limitations in terms of their spatial, temporal and taxonomic scope – a set of problems largely known as the biodiversity knowledge shortfalls (Hortal *et al.* 2015). The credibility of the science based upon NHC data largely depends upon recognising and quantifying these shortfalls to minimize the inefficient use of limited conservation resources (Grand *et al.* 2007; Hortal *et al.* 2008).

Unlike spatio-temporal gaps in primary biodiversity data (Haque *et al.* 2017; Stropp *et al.* 2016), quantification of bias and gaps in taxonomic breadth have drawn less attention, although within conservation science taxonomic bias is well acknowledged (Clark and May 2002; Lawler *et al.* 2006). Taxonomic bias often stems from a selective focus on taxonomic uniqueness (e.g. endemics), rarity or economic value (Bonnet *et al.* 2002), or from societal preferences (Wilson *et al.* 2007). For example, taxonomic bias has been found in fauna collections within the Global Biodiversity Information Facility (GBIF, https://www.gbif.org/) where vertebrates, especially birds, are better represented than Arthropods (Troudet *et al.* (2017). However, to what extent such biases exist in botanical collections is poorly understood. Preserved plant specimens held in herbaria are an incredibly valuable source of verifiable, repeatable, sustainable and persistent data (Holmes *et al.* 2016). As the proportion of these specimens that are digitised and included in databases increases, so too does their usefulness for scientific research in terms of the range of projects these data can be used for. However, taxonomic biases may skew results of such studies.

The disparity in taxonomic knowledge in herbarium collections is often triggered by the collector's influence. For example, taxonomists tend to preferentially collect rare or unusual species, disregarding or under-representing common taxa (Garcillán and Ezcurra 2011). Even within a taxon, conspicuous or readily detectable species are more frequently recorded (Schmidt-Lebuhn *et al.* 2013), as are individuals that can easily be positioned onto a herbarium sheet (Zopfi 1993). The recent decline in the numbers of the efficient collectors (i.e. expert botanists) due to lack of training and funding may also contribute to uneven coverage of taxa in flora collections (Ahrends *et al.* 2011). Such taxonomic unevenness may create artificial inflations in species numbers for certain taxa, and therefore may influence decision-making regarding resource allocation and conservation actions (Farrier *et al.* 2007; Grand *et al.* 2007; Pillon and Chase Mark 2006; Walsh *et al.* 2012).

Australia is a megadiverse country with approximately 85% of its flowering plants being endemic (Chapman 2009). Presently, herbaria in Australia and New Zealand house more than 8,000,000 records, of which an estimated ~ 80% have been digitised, and incorporated into the Australasian Virtual Herbarium (AVH, https://avh.chah.org.au/). This database plays a pivotal role in plant conservation science (González-Orozco *et al.* 2016; Guerin *et al.* 2016; Silcock *et al.* 2015). However, as with other primary biodiversity databases, spatial and temporal biases and gaps occur among the records of the AVH (Haque *et al.* 2017; Haque *et al. in press*). Yet to date, no study has quantified the taxonomic shortfalls of this resource. Hence, here we specifically address the following questions: 1) Is a taxonomic bias apparent in the collection of preserved specimens from Australian native plant families digitised in the AVH? That is, based on the number of specimen records per species, are some plant families more, or less, likely to be collected? 2) To what extent does the distribution of collectors among plant families influence taxonomic bias?

METHODS

Data Source

We downloaded records contained within the Australasian Virtual Herbarium (AVH) (http://avh.chah.org.au/) accessed from the Atlas of Living Australia (ALA) (http://www.ala.org.au/; 1 January 2016). These comprised all records from the Australian native plant families recognised in the Australian Plant Census (www.anbg.gov.au/chah/apc/about-APC.html, APC). The APC provides accepted scientific names of Australian flora and is endorsed by the Council of Heads of Australasian Herbaria (CHAH).

We undertook a multi-step data cleaning process to exclude records that: were not preserved specimens; lacked geographic coordinates or had coordinates that placed them in the ocean, coastal waterways, or on offshore islands or were missing collection date; were not identified to species or lower level (i.e. consisted of a genus name and the epithet “sp.”); were labelled as “cultivated” or hybrids; or for which the collector was not identified or was flagged “unknown”. Records that represented duplicate specimens were then removed from the dataset, where duplicates were defined as two or more records containing the same species name, collected on the same date by the same collector, and from within 1 km of each other (to account for records with rounded or truncated geographic coordinates). The final dataset contained records of 2,430,220 specimens belonging to 296 families, both vascular and nonvascular.

Taxonomic bias

For each family, following Troudet *et al.* (2017) we calculated the number of specimens (I) that would be expected if there was no taxonomic bias as:

$$I = NB_{occ}*(N/N_{tot})$$

where N is the number of native species sampled for each family, N_{tot} is the total number of known native species and NB_{occ} is the number of specimens across all 296 families. We then measured the taxonomic bias as a difference (O-I) and the ratio (O/I) between the observed (O) and expected (I) specimens for each family.

To identify under-sampled families, we also calculated the proportion (p) of native species per family for which more than one ($p > 1$) and twenty ($p \geq 20$) unique preserved specimens have been digitised. We chose the threshold of 20 unique specimens as this is frequently the minimum number of records used to calibrate habitat suitability models which are tools that commonly make use of NHC data (Feeley and Silman 2011a; b).

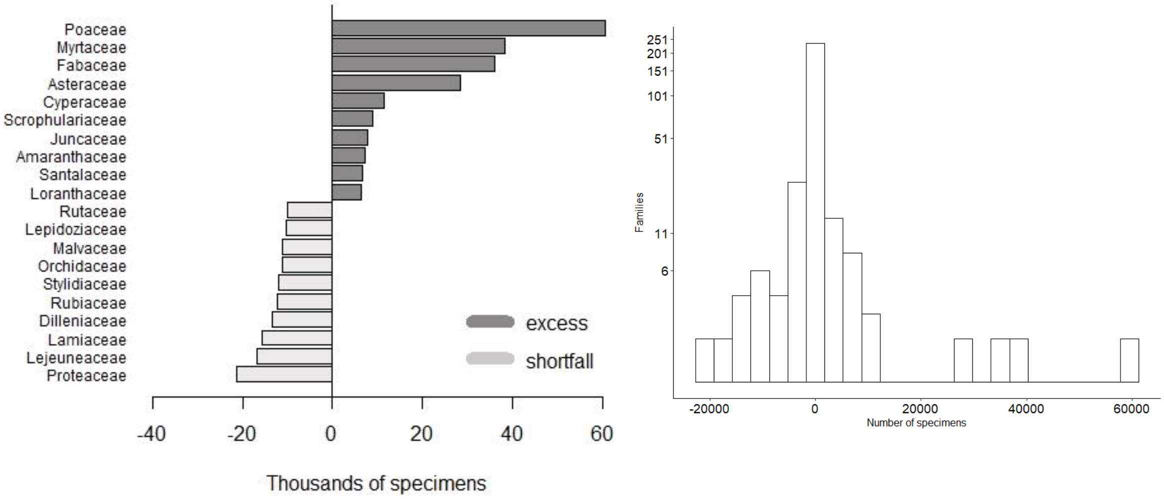
Distribution of collectors among families

We assessed the distribution of collectors among families and whether this influenced taxonomic bias. That is, we hypothesised that if a particularly active collector targeted a given taxon, this could contribute to taxonomic biases in the overall collection. We then counted the number of collectors per family and performed a Spearman rank correlation test between the number of collectors per family and taxonomic bias. We also calculated the total number of specimens collected per collector per family to assess the sampling intensity of the collector.

RESULTS

Taxonomic Bias

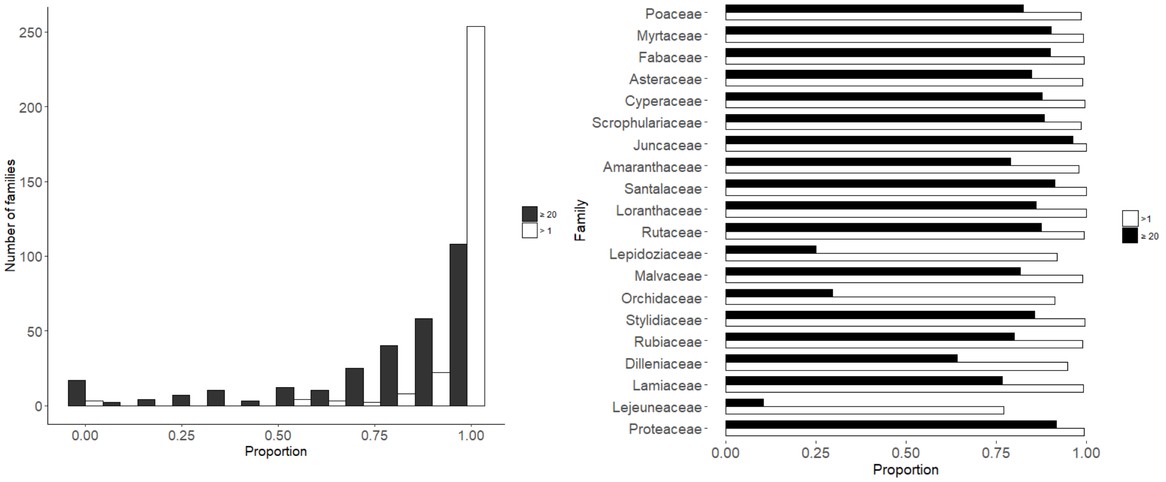
The number of preserved specimens per family is not proportional to families' known species richness, highlighting a taxonomic bias in the records of native plant families in the AVH. The family Proteaceae (887 species) is the most under-represented, in terms of number of records, lacking an estimated 21,337 records. This represents 17% of the records it should have if there was no taxonomic bias. This was followed by Lejeuneaceae (122 species, lacking 16,653 records, i.e. 45%), then Lamiaceae (376 species, lacking 15,715 records, i.e. 29%).



1a) 1b)

Figure 1. *(a) Taxonomic bias in the records of Australian native plant voucher specimens digitised in the Australasian Virtual Herbarium (AVH). a) The vertical line at 0 depicts the 'expected' number of specimens per family based on the proportion of actual specimens to the number of native species per family. The graph identifies the ten families most over- (dark grey) and under-represented (light grey) compared to expected. b) The frequency distribution of taxonomic bias (as the number of records) across 296 families included in this study.*

Indeed, for 29% of Australia's plant families (i.e. 86), the number of digitised records constitutes < 50% of the number expected (see details in Appendix S1). In contrast, the most over-represented family within the AVH, in terms of number of records, is Poaceae (963 species) with 60,543 (143%) more records than expected, followed by Myrtaceae (1788 species, 38,341 (115%) more records) and Fabaceae (2088 species, 35,891 (112%) more records (Fig. 1). In addition, 10 families have in excess of 200% more records than expected, although eight of these families are species poor, with digitised records for < 10 species (Appendix S1). We note, however, that due to our data cleaning process, additional records may be available for each family that we have not included in this analysis.


2a)

2b)

Figure 2. *Proportion of species per family with > 1 or ≥ 20 digitised specimens per species a) across 296 Australian native plants families, and b) for the ten families most over- and under-represented within the Australasian Virtual Herbarium.*

Three-quarters of plant families (i.e. 222) have more than one specimen digitised for each species, with 34% of families (100) having at least 20 specimens digitised per species

(Fig. 2a). There were 11 families for which less than 10 unique specimens were collected, although each of these families have only 1-2 species with records in our final database. Among the ten most over-represented families in the AVH, those with the greatest proportion of species represented by at least 20 unique specimens were Juncaceae (96%), followed by Ericaceae (93% species), Santalaceae (91%), Fabaceae (90%) and Myrtaceae (90%). Conversely, among underrepresented families, only 10% of species belonging to Lejeuneaceae have ≥ 20 unique specimens within the AVH, followed by Lepidoziaceae (25%) and Orchidaceae (29%) (Fig. 2b and for details see Appendix S1).

Distribution of collectors among families

The number of collectors per family ranged from 1 to 14,553, with a median of 275 (IQR = 70.5–883) (Fig. 3a). Fabaceae has been sampled by the greatest number of collectors (14,553), followed by Myrtaceae (12,862), Asteraceae (8653) and Poaceae (8455). Five families have been sampled by only 1–2 collectors (Calypogeiaceae (1 collector), Jubulaceae (1), Trichotemnomataceae (1), Blepharidophyllaceae (2), Haplomitriaceae (2), Orobanchaceae (2)). The average number of specimens collected per collector per family ranged from 1 to 23 (Fig. 3b). For ~ 80% of families, individual collectors sampled on average < five specimens per family. The highest average number of specimens collected per collector were for Poaceae (23 ± 166 SD specimens per collector), Fabaceae (23 ± 144 SD), Myrtaceae (23 ± 193 SD), Asteraceae (18 ± 109 SD) and Cyperaceae (17 ± 107 SD).

We found a strong positive correlation between the number of collectors sampling a family and the taxonomic bias of that family (Spearman rank) (*rho* = 0.54, $p = < 0.0001$) indicating that collector preference influences taxonomic bias in the AVH (Fig. 4). For example, Hypericaceae (with three species) has a taxonomic bias of 3.11 (i.e. three times more specimens collected than expected) and has been sampled by 525 collectors. In contrast, Radulaceae (with 22 species) has a taxonomic bias of 0.16 (i.e. only 16% of specimens collected compared to expected) and has been sampled by only 69 collectors.

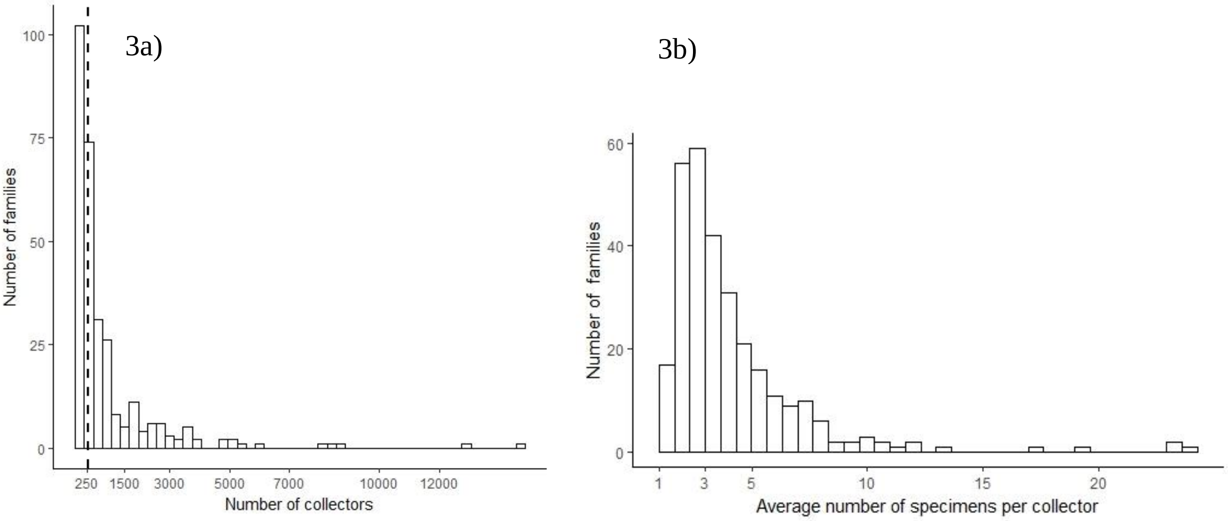


Figure 3. *(a) Frequency distribution of collectors per plant family. The black dashed line indicates the median number of (275) collectors per family. (b) Frequency distribution of average number of preserved specimens per collector per family (as digitised in the Australasian Virtual Herbarium).*

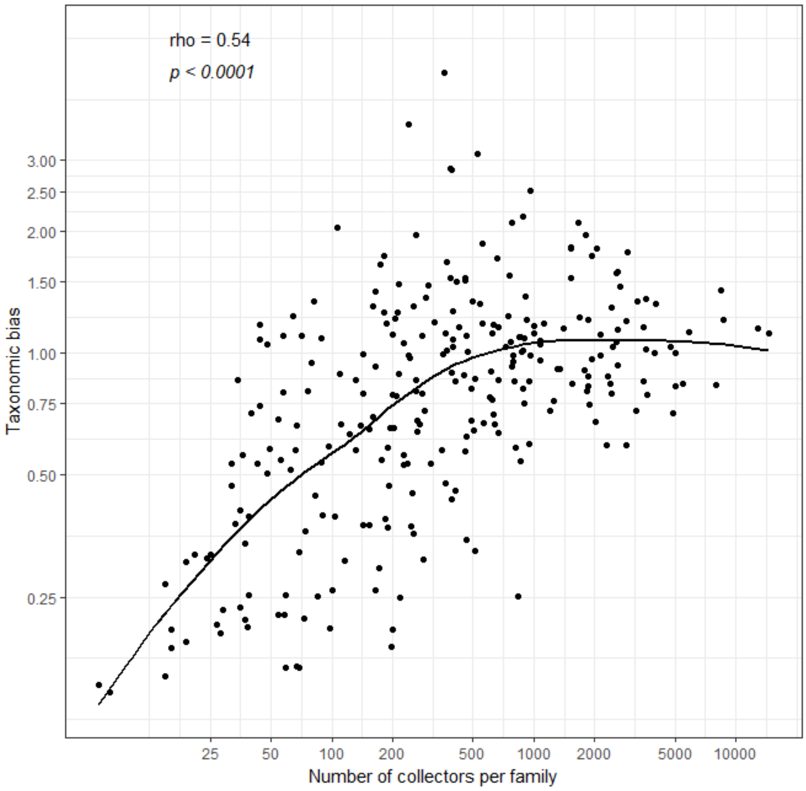


Figure 4. *Scatterplot showing taxonomic bias of plant families (ratio of observed to expected number of specimens digitised per family) and number of collectors per family, based on data in the Australasian Virtual Herbarium (AVH). The thick black line represents a locally weighted regression.*

DISCUSSION

In this study, we explored taxonomic bias in the digitisation of records of Australian native plant specimens that have been include in the Australasian Virtual Herbarium (AVH). We demonstrate that: a) there is considerable taxonomic bias in AVH data with the number of digitised records, for 29% of families, constituting < 50% of the number expected based on the families' species richness, b) only 34% of families have at least 20 specimens digitised for each species recorded in AVH, and c) there is a strong positive correlation between the number of collectors sampling a family and the taxonomic bias of that family.

Collectors play an essential role in determining the taxonomic representativeness of herbarium and other NHC collections (Bebber *et al.* 2012; Penn *et al.* 2018), but the extent to which the number of collectors influences taxonomic biases has rarely been tested. It has been argued that if many different botanists have been collecting specimens from the same area over many years, the inherent biases in collections would be lowered (Petersen *et al.* 2003). We have found limited support for this in the AVH. Families that are over-represented in the collection (i.e. positive taxonomic bias) are generally associated with a large number of collectors. However, the cumulative influence of collectors on taxonomic bias appears to decline beyond a certain number of collectors (~ 1500) (see Fig. 4). The underlying reason for this phenomenon may be that historically a fraction of collectors termed 'big hitters' made a disproportionally large contribution to the total collection (Bebber *et al.* 2012). Therefore, Bebber *et al.* (2012) emphasised the need for more training and funding of experts rather than merely increasing the number of lay collectors. Moreover, given that the number of expert taxonomists is declining (Halme *et al.* 2015), it is vital to identify the number of collectors required and ensure better training and coordination to sample poorly known families.

In this study, we have explored and quantified the over and under representation of families in the collections of AVH. The measurement of under or over-representation must be relative to some standard. A simple standard is a neutral model where all species have an equal chance of being represented in a collection. This is the standard against which taxonomic representativeness (reported by family) is measured in this thesis. This neutral model is, arguably, also what an ideal specimen collection should look like (that is, all species represented equally). This model is valuable in that assumptions are minimised. While range size, abundance, functional traits and phylogenetic uniqueness may all influence whether a species is collected or not, these factors are best used as predictors of representativeness rather than assumed in a neutral model of specimen collection. It should be emphasised that taxonomic representativeness is being measured in this thesis and thus a standard based on species counts is more appropriate than assuming specimens are collected proportional to abundance (which would be a model of ecological representativeness).

For studies requiring a number of unique occurrence records, such as those using habitat suitability models, only one-third of families could have all species assessed. In contrast, for 43 families, less than 50% of species could be assessed. For example, despite Lejeuneaceae, the largest family among the liverworts (Ahonen et al. 2003), having 121 species with digitised records, only 10% of these species have ≥ 20 records within the AVH. It is, however, important to note that we have removed specimens missing complete attribute information, i.e. lacking or dubious geographic co-ordinates, missing complete collection date and collector's name or duplicated specimens from our analysis. We ran a sensitivity analysis and found that this process resulted in 1–134,139 records per family (median 471, IQR = 109–1671) being discarded (see Appendix S2). The exclusion of these specimens is unlikely to have influenced the proportion of specimens per family retained for the measure of taxonomic bias (Appendix S3). However, this highlights the need for data collection and curating to follow adequate standards to maximise the utilisation of these records. It is also important to note that we have assessed taxonomic representativeness of the digitised specimens only. According to data on the webpages of each of the 28 Australian herbaria, it is estimated that around 20% of records across Australia's 28 herbariums remain to be digitised (see https://avh.chah.org.au/index.php/about/). Completing the digitisation of these collections will provide a greater understanding of the taxonomic representativeness of the Australian native plan families.

In conclusion, analyses of meta collections (such as the AVH) are revealing biases in our knowledge of the distribution of species (Meyer *et al.* 2016). Conservation science is at a critical stage with unprecedented rates of species extinction (Rands *et al.* 2010). As such, it is imperative that we overcome taxonomic shortfalls (Hortal *et al.* 2015), as the knowledge that comes with documenting of the distribution, habitats and abundance of species plays an integral role in conservation planning (Mace 2004).

ACKNOWLEDGEMENTS

We thank Dr Rachael Gallagher and Stuart Allen for their assistance in data retrieval from Australasian Virtual Herbarium (AVH) via Atlas of Living Australia (ALA).

REFERENCES

Ahonen I., Muona J. & Piippo S. (2003) Inferring the Phylogeny of the Lejeuneaceae (Jungermanniopsida): A First Appraisal of Molecular Data. *The Bryologist* **106**, 297-308.

Ahrends A., Rahbek C., Bulling M. T., Burgess N. D., Platts P. J., Lovett J. C., Kindemba V. W., Owen N., Sallu A. N., Marshall A. R., Mhoro B. E., Fanning E. & Marchant R. (2011) Conservation and the botanist effect. *Biological Conservation* **144**, 131-140.

Bebber D. P., Carine M. A., Davidse G., Harris D. J., Haston E. M., Penn M. G., Cafferty S., Wood J. R. I. & Scotland R. W. (2012) Big hitting collectors make massive and disproportionate contribution to the discovery of plant species. *Proceedings of the Royal Society B-Biological Sciences* **279**, 2269-2274.

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

Ceballos G., Ehrlich P. R., Barnosky A. D., García A., Pringle R. M. & Palmer T. M. (2015) Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances* **1**, e1400253.

Chapman A. D. (2009) *Numbers of Living Species in Australia and the World*. Australian Department of the Environment, Heritage and the Arts, Canberra, AU.

Clark J. A. & May R. M. (2002) Taxonomic Bias in Conservation Research. *Science* **297**, 191-193.

Dettmann M. E. & Jarzen D. M. (1998) The early history of the Proteaceae in Australia: the pollen record. *Australian Systematic Botany* **11**, 401-438.

Farrier D., Whelan R. & Mooney C. (2007) Threatened species listing as a trigger for conservation action. *Environmental Science & Policy* **10**, 219-229.

Feeley K. J. & Silman M. R. (2011a) The data void in modeling current and future distributions of tropical species. *Global Change Biology* **17**, 626-630.

Feeley K. J. & Silman M. R. (2011b) Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions* **17**, 1132-1140.

Garcillán P. P. & Ezcurra E. (2011) Sampling procedures and species estimation: testing the effectiveness of herbarium data against vegetation sampling in an oceanic island. *Journal of Vegetation Science* **22**, 273-280.

Geijzendorffer I. R., Regan E. C., Pereira H. M., Brotons L., Brummitt N., Gavish Y., Haase P., Martin C. S., Mihoub J. B., Secades C., Schmeller D. S., Stoll S., Wetzel F. T., Walters M. & Cadotte M. (2016) Bridging the gap between biodiversity data and policy reporting needs: An Essential Biodiversity Variables perspective. *Journal of Applied Ecology* **53**, 1341-1350.

González-Orozco C. E., Pollock Laura J., Thornhill Andrew H., Mishler Brent D., Knerr N., Laffan Shawn W., Miller Joseph T., Rosauer Dan F., Faith Daniel P., Nipperess David A., Kujala H., Linke S., Butt N., Külheim C., Crisp Michael D. & Gruber B. (2016) Phylogenetic approaches reveal biodiversity threats under climate change. *Nature Climate Change* **6**, 1110-1115.

Graham C. H., Ferrier S., Huettman F., Moritz C. & Peterson A. T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution* **19**, 497-503.

Grand J., Cummings M. P., Rebelo T. G., Ricketts T. H. & Neel M. C. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecology Letters* **10**, 364-374.

Guerin G. R., Biffin E., Baruch Z. & Lowe A. J. (2016) Identifying centres of plant biodiversity in South Australia. *PLoS One* **11**, e0144779.

Halme P., Kuusela S. & Juslén A. (2015) Why taxonomists and ecologists are not, but should be, carpooling? *Biodiversity and Conservation* **24**, 1831-1836.

Haque M. M., Nipperess D. A., Gallagher R. V. & Beaumont L. J. (2017) How well documented is Australia's flora? Understanding spatial bias in vouchered plant specimens. *Austral Ecology* **42**, 690-699.

Haque M. M., Nipperess D. A., Baumgartner J. B. & Beaumont L. J. (*in press*) A journey through time: exploring temporal patterns among digitised plant specimens from Australia. *Systematics and Biodiversity (accepted on 24 April 2018). DOI:10.1080/14772000.2018.1472674*.

Holmes M. W., Hammond T. T., Wogan G. O., Walsh R. E., Labarbera K., Wommack E. A., Martins F. M., Crawford J. C., Mack K. L. & Bloch L. M. (2016) Natural history collections as windows on evolutionary processes. *Molecular Ecology* **25**, 864-881.

Hortal J., De Bello F., Diniz-Filho J. A. F., Lewinsohn T. M., Lobo J. M. & Ladle R. J. (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 J. F., Lobo J. M. & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* **117**, 847-858.

Lawler J. J., Aukema J. E., Grant J. B., Halpern B. S., Kareiva P., Nelson C. R., Ohleth K., Olden J. D., Schlaepfer M. A., Silliman B. R. & Zaradic P. (2006) Conservation science: a 20-year report card. *Frontiers in Ecology and the Environment* **4**, 473-480.

Mace G. M. (2004) The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **359**, 711-719.

Meyer C., Weigelt P. & Kreft H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**, 992-1006.

Penn M. G., Cafferty S. & Carine M. (2018) Mapping the history of botanical collectors: spatial patterns, diversity, and uniqueness through time. *Systematics and Biodiversity* **16**, 1-13.

Petersen F. T., Meier R. & Larsen M. N. (2003) Testing species richness estimation methods using museum label data on the Danish Asilidae. *Biodiversity & Conservation* **12**, 687-701.

Pillon Y. & Chase Mark W. (2006) Taxonomic exaggeration and its effects on orchid conservation. *Conservation Biology* **21**, 263-265.

Proença V., Martin L. J., Pereira H. M., Fernandez M., Mcrae L., Belnap J., Böhm M., Brummitt N., García-Moreno J., Gregory R. D., Honrado J. P., Jürgens N., Opige M., Schmeller D. S., Tiago P. & Van Swaay C. A. M. (2017) Global biodiversity monitoring: From data sources to Essential Biodiversity Variables. *Biological Conservation* **213**, 256-263.

Rands M. R. W., Adams W. M., Bennun L., Butchart S. H. M., Clements A., Coomes D., Entwistle A., Hodge I., Kapos V., Scharlemann J. P. W., Sutherland W. J. & Vira B. (2010) Biodiversity Conservation: Challenges Beyond 2010. *Science* **329**, 1298-1303.

Schmidt-Lebuhn A. N., Knerr N. J. & Kessler M. (2013) Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and Conservation* **22**, 905-919.

Sheahan M. (2010) Nitrariaceae. In: *Flowering Plants. Eudicots* Vol. 10 pp. 272-275. Berlin: Springer.

Silcock J. L., Healy A. J. & Fensham R. J. (2015) Lost in time and space: re-assessment of conservation status in an arid-zone flora through targeted field survey. *Australian Journal of Botany* **62**, 674-688.

Stropp J., Ladle R. J., Malhado M., Malhado A. C. M., Hortal J., Gaffuri J., H Temperley W., Skøien J. O. & Mayaux P. (2016) Mapping ignorance: 300 years of collecting flowering plants in Africa. *Global Ecology and Biogeography* **25**, 1085-1096.

Sullivan B. L., Phillips T., Dayer A. A., Wood C. L., Farnsworth A., Iliff M. J., Davies I. J., Wiggins A., Fink D., Hochachka W. M., Rodewald A. D., Rosenberg K. V., Bonney R. & Kelling S. (2017) Using open access observational data for conservation action: A case study for birds. *Biological Conservation* **208**, 5-14.

Ter Steege H., Haripersaud P. P., Banki O. S. & Schieving F. (2011) A model of botanical collectors' behavior in the field: never the same species twice. *American Journal of Botany* **98**, 31-37.

Troudet J., Grandcolas P., Blin A., Vignes-Lebbe R. & Legendre F. (2017) Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* **7**, 9132-9146.

Walsh J. C., Watson J. E. M., Bottrill M. C., Joseph L. N. & Possingham H. P. (2012) Trends and biases in the listing and recovery planning for threatened species: an Australian case study. *Oryx* **47**, 134-143.

Ward D. F. (2012) More than just records: analysing natural history collections for biodiversity planning. *PLoS One* **7**, e50346.

Wilson J. R., Procheş Ş., Braschler B., Dixon E. S. & Richardson D. M. (2007) The (bio)diversity of science reflects the interests of society. *Frontiers in Ecology and the Environment* **5**, 409-414.

Zopfi H.-J. (1993) Ecotypic variation in *Rhinanthus alectorolophus* (Scopoli) Pollich (Scrophulariaceae) in relation to grassland management: I. Morphological delimitations and habitats of seasonal ecotypes. *Flora* **188**, 15-39.

SUPPLEMENTARY INFORMATION

Appendix S1. *Characteristics of specimens from 296 native plant families available in Australasian Virtual Herbarium (AVH). Taxonomic bias calculated as the difference and ratio between observed and expected (0-E) and O/E respectively. p > 1 and p ≥ 20 represent the number of species with more than one and at least 20 specimens, respectively, within the corresponding family. Average number of specimens collected per collector and standard deviation is also included.*

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Acanthaceae	44	7305	6361	944	1.15	1.00	0.82	1392	5.24 (16.5)
Acrobolbaceae	12	450	1735	-1285	0.26	0.92	0.33	100	4.5 (8.17)
Adelanthaceae	10	244	1446	-1202	0.17	1.00	0.40	67	3.64 (8.54)
Aizoaceae	41	7279	5928	1351	1.23	0.98	0.90	1671	4.35 (11.86)
Akaniaceae	1	59	145	-86	0.41	1.00	1.00	35	1.68 (2.28)
Alismataceae	6	692	867	-175	0.80	1.00	0.67	281	2.46 (3.91)
Alseuosmiaceae	2	167	289	-122	0.58	1.00	1.00	66	2.53 (3.61)
Amaranthaceae	144	28,018	20,819	7199	1.35	0.98	0.79	3270	8.56 (31.15)
Amaryllidaceae	14	930	2024	-1094	0.46	1.00	0.71	411	2.26 (3.33)
Anacardiaceae	10	1936	1446	490	1.34	1.00	0.90	497	3.89 (9.57)
Anarthriaceae	8	1737	1157	580	1.50	1.00	1.00	415	4.18 (10.44)
Aneuraceae	21	637	3036	-2399	0.21	1.00	0.43	98	6.5 (13.82)
Annonaceae	44	2781	6361	-3580	0.44	1.00	0.80	393	7.07 (24.03)
Anthocerotaceae	4	49	578	-529	0.08	1.00	0.25	23	2.13 (2.65)
Aphanopetalaceae	2	268	289	-21	0.93	1.00	1.00	164	1.63 (1.28)
Apiaceae	97	14,923	14,024	899	1.06	1.00	0.92	2570	5.8 (20.5)
Apocynaceae	160	18,423	23,133	-4710	0.80	0.99	0.82	2424	7.6 (48.51)
Apodanthaceae	2	62	289	-227	0.21	1.00	0.50	27	2.29 (2.25)
Aponogetonaceae	5	181	723	-542	0.25	1.00	0.60	85	2.12 (3.26)
Aquifoliaceae	2	258	289	-31	0.89	0.50	0.50	109	2.36 (3.64)
Araceae	38	1791	5494	-3703	0.33	0.97	0.55	510	3.51 (6.84)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Araliaceae	110	13,917	15,904	-1987	0.88	0.99	0.83	2326	5.98 (23.32)
Araucariaceae	6	501	867	-366	0.58	0.83	0.83	131	3.82 (11.1)
Arecaceae	43	2151	6217	-4066	0.35	1.00	0.81	465	4.62 (11.37)
Argophyllaceae	5	285	723	-438	0.39	1.00	0.80	104	2.74 (3.97)
Aristolochiaceae	13	489	1880	-1391	0.26	1.00	0.69	164	2.98 (4.91)
Arnelliaceae	1	6	145	-139	0.04	1.00	0.00	4	1.5 (1)
Asparagaceae	139	23,368	20,096	3272	1.16	1.00	0.89	3463	6.74 (25.71)
Asphodelaceae	7	2126	1012	1114	2.10	1.00	1.00	774	2.74 (5.87)
Aspleniaceae	26	3791	3759	32	1.01	1.00	0.85	871	4.35 (11.67)
Asteliaceae	9	490	1301	-811	0.38	1.00	0.78	154	3.18 (4.93)
Asteraceae	940	164,362	135,904	28458	1.21	0.99	0.85	8653	18.99 (109.02)
Atherospermataceae	9	1322	1301	21	1.02	1.00	1.00	372	3.55 (11.09)
Austrobaileyaceae	1	81	145	-64	0.56	1.00	1.00	36	2.25 (2.68)
Aytoniaceae	9	490	1301	-811	0.38	0.89	0.33	142	3.45 (7.24)
Balanopaceae	1	170	145	25	1.18	1.00	1.00	44	3.86 (7.4)
Balanophoraceae	1	84	145	-61	0.58	1.00	1.00	49	1.71 (1.32)
Balantiopsidaceae	7	228	1012	-784	0.23	0.86	0.29	58	3.93 (5.4)
Bataceae	1	99	145	-46	0.68	1.00	1.00	54	1.83 (2.12)
Bignoniaceae	14	1163	2024	-861	0.57	1.00	0.71	455	2.55 (3.82)
Bixaceae	3	597	434	163	1.38	1.00	1.00	293	2.03 (2.23)
Blandfordiaceae	4	315	578	-263	0.54	1.00	0.75	176	1.78 (2.58)
Blechnaceae	30	5075	4337	738	1.17	0.93	0.83	999	5.08 (12.19)
Blepharidophyllaceae	2	3	289	-286	0.01	0.50	0.00	2	1.5 (0.7)
Boraginaceae	117	15,833	16,916	-1083	0.94	0.97	0.81	2583	6.12 (20.17)
Boryaceae	9	1313	1301	12	1.01	1.00	0.89	469	2.79 (5.72)
Brassicaceae	94	12,335	13,590	-1255	0.91	0.99	0.84	1770	6.96 (23.41)
Brevianthaceae	1	6	145	-139	0.04	1.00	0.00	3	2 (1.73)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Burmanniaceae	3	281	434	-153	0.65	1.00	1.00	152	1.84 (1.78)
Burseraceae	6	960	867	93	1.11	1.00	1.00	279	3.44 (8.21)
Byblidaceae	5	622	723	-101	0.86	1.00	1.00	260	2.39 (4.87)
Cabombaceae	1	45	145	-100	0.31	1.00	1.00	24	1.87 (1.96)
Calycanthaceae	1	124	145	-21	0.86	1.00	1.00	34	3.64 (6.4)
Calypogeiaceae	1	1	145	-144	0.01	0.00	0.00	1	NA
Campanulaceae	69	14,533	9976	4557	1.46	1.00	0.91	2667	5.44 (17.96)
Cannabaceae	7	1891	1012	879	1.87	1.00	0.86	558	3.38 (8.08)
Capparaceae	22	3574	3181	393	1.12	1.00	0.91	997	3.58 (8.38)
Caprifoliaceae	1	137	145	-8	0.95	1.00	1.00	79	1.73 (1.78)
Cardiopteridaceae	1	39	145	-106	0.27	1.00	1.00	15	2.6 (3.86)
Caryophyllaceae	41	5424	5928	-504	0.92	1.00	0.88	1303	4.16 (9.77)
Casuarinaceae	59	13,426	8530	4896	1.57	1.00	0.95	2561	5.24 (15.02)
Celastraceae	65	10,478	9398	1080	1.11	1.00	0.89	2147	4.88 (19.36)
Centrolepidaceae	29	4152	4193	-41	0.99	1.00	0.69	792	5.24 (13.31)
Cephalotaceae	1	57	145	-88	0.39	1.00	1.00	39	1.46 (1.09)
Cephaloziellaceae	6	387	867	-480	0.45	0.67	0.50	83	4.66 (11.52)
Ceratophyllaceae	2	193	289	-96	0.67	1.00	0.50	111	1.73 (1.53)
Chrysobalanaceae	2	504	289	215	1.74	1.00	1.00	182	2.76 (4.67)
Cleomaceae	12	2041	1735	306	1.18	0.92	0.58	634	3.21 (5.15)
Clusiaceae	13	1007	1880	-873	0.54	1.00	1.00	235	4.28 (14.47)
Colchicaceae	34	4512	4916	-404	0.92	1.00	0.94	1348	3.34 (9.33)
Combretaceae	37	5425	5349	76	1.01	1.00	0.97	883	6.14 (17.89)
Commelinaceae	28	4416	4048	368	1.09	0.93	0.79	891	4.95 (13.41)
Connaraceae	2	192	289	-97	0.66	1.00	1.00	67	2.86 (3.4)
Convolvulaceae	100	16,570	14,458	2112	1.15	0.98	0.85	2610	6.34 (20.14)
Cornaceae	1	295	145	150	2.04	1.00	1.00	106	2.78 (4.73)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Corsiaceae	1	15	145	-130	0.10	1.00	0.00	7	2.14 (2.19)
Corynocarpaceae	2	158	289	-131	0.55	1.00	1.00	56	2.82 (3.86)
Costaceae	2	73	289	-216	0.25	1.00	1.00	39	1.87 (1.96)
Crassulaceae	12	4385	1735	2650	2.53	1.00	0.83	964	4.54 (12.01)
Cucurbitaceae	35	3811	5060	-1249	0.75	1.00	0.80	900	4.23 (11.27)
Cunoniaceae	35	4323	5060	-737	0.85	1.00	0.94	942	4.58 (14.12)
Cupressaceae	20	5249	2892	2357	1.82	1.00	1.00	1520	3.45 (8.29)
Cyatheaceae	6	1032	867	165	1.19	1.00	1.00	323	3.19 (4.91)
Cycadaceae	17	919	2458	-1539	0.37	0.94	0.76	245	3.75 (7.95)
Cymodoceaceae	10	683	1446	-763	0.47	1.00	0.70	192	3.55 (7.69)
Cyperaceae	617	100,619	89,205	11414	1.13	1.00	0.88	5888	17.08 (107.73)
Dasypogonaceae	11	914	1590	-676	0.57	1.00	0.91	350	2.61 (6.12)
Davalliaceae	5	474	723	-249	0.66	1.00	1.00	194	2.44 (3.8)
Dennstaedtiaceae	12	1423	1735	-312	0.82	1.00	0.75	488	2.91 (6.76)
Dichapetalaceae	1	107	145	-38	0.74	1.00	1.00	44	2.43 (2.86)
Dicksoniaceae	4	853	578	275	1.47	1.00	1.00	302	2.82 (5.37)
Dilleniaceae	228	19,531	32,964	-13433	0.59	0.95	0.64	2883	6.77 (24.99)
Dioscoreaceae	4	887	578	309	1.53	1.00	0.75	387	2.29 (3.44)
Doryanthaceae	2	61	289	-228	0.21	1.00	0.50	38	1.6 (1.44)
Droseraceae	117	11,438	16,916	-5478	0.68	0.99	0.68	2029	5.63 (18.44)
Dryopteridaceae	23	3007	3325	-318	0.90	1.00	0.91	613	4.9 (10.82)
Dumortieraceae	1	18	145	-127	0.12	1.00	0.00	7	2.57 (1.9)
Ebenaceae	14	2678	2024	654	1.32	1.00	0.86	544	4.92 (14.17)
Ecdeiocoleaceae	3	346	434	-88	0.80	1.00	1.00	142	2.43 (6.34)
Elaeagnaceae	1	194	145	49	1.34	1.00	1.00	82	2.36 (3.19)
Elaeocarpaceae	76	9114	10,988	-1874	0.83	0.99	0.92	1851	4.92 (18.67)
Elatinaceae	11	1753	1590	163	1.10	1.00	0.91	463	3.78 (10.37)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Emblingiaceae	1	28	145	-117	0.19	1.00	1.00	19	1.47 (1.42)
Ericaceae	384	57,661	55,518	2143	1.04	1.00	0.93	4771	12.08 (67.12)
Erythroxylaceae	2	378	289	89	1.31	1.00	1.00	160	2.36 (3.92)
Escalloniaceae	9	831	1301	-470	0.64	1.00	1.00	263	3.15 (7.09)
Euphorbiaceae	247	28,253	35,711	-7458	0.79	1.00	0.79	3658	7.72 (43.9)
Eupomatiaceae	3	481	434	47	1.11	1.00	1.00	200	2.4 (3.72)
Fabaceae	2088	337,771	301,880	35891	1.12	0.99	0.90	14553	23.2 (144)
Flagellariaceae	1	533	145	388	3.69	1.00	1.00	238	2.23 (3.13)
Fossombroniaceae	30	512	4337	-3825	0.12	0.87	0.30	66	7.75 (23.19)
Frankeniaceae	35	2731	5060	-2329	0.54	1.00	0.74	857	3.18 (7.34)
Frullaniaceae	46	1962	6651	-4689	0.30	0.89	0.35	172	11.4 (43.94)
Gentianaceae	28	2375	4048	-1673	0.59	0.89	0.68	811	2.92 (5.75)
Geocalycaceae	3	15	434	-419	0.03	1.00	0.00	12	1.25 (0.45)
Geraniaceae	24	6351	3470	2881	1.83	0.96	0.79	1519	4.18 (12.86)
Gesneriaceae	4	516	578	-62	0.89	1.00	1.00	213	2.42 (4.15)
Gleicheniaceae	12	2708	1735	973	1.56	1.00	1.00	755	3.58 (8.1)
Goodeniaceae	345	50,015	49,880	135	1.00	1.00	0.87	4995	10.01 (38.04)
Grammitidaceae	3	223	434	-211	0.51	1.00	1.00	62	3.59 (6.07)
Gunneraceae	1	46	145	-99	0.32	1.00	1.00	21	2.19 (2.04)
Gymnomitriaceae	4	36	578	-542	0.06	1.00	0.00	17	2.11 (2.08)
Gyrostemonaceae	15	2620	2169	451	1.21	1.00	0.87	925	2.83 (5.1)
Haemodoraceae	77	8476	11133	-2657	0.76	1.00	0.91	1254	6.75 (36.57)
Haloragaceae	93	13,948	13,446	502	1.04	1.00	0.95	2463	5.66 (32.9)
Hamamelidaceae	3	147	434	-287	0.34	1.00	1.00	37	3.97 (6.19)
Hanguanaceae	1	55	145	-90	0.38	1.00	1.00	33	1.66 (1.4)
Haplomitriaceae	1	2	145	-143	0.01	1.00	0.00	2	1 (0)
Hemerocallidaceae	58	10,849	8386	2463	1.29	1.00	0.88	2420	4.48 (12.92)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD
Herbertaceae	3	13	434	-421	0.03	1.00	0.00	6	2.16 (1.83)
Hernandiaceae	4	612	578	34	1.06	1.00	1.00	227	2.69 (5.97)
Himantandraceae	1	160	145	15	1.11	1.00	1.00	57	2.8 (3.98)
Hydatellaceae	7	311	1012	-701	0.31	0.86	0.57	115	2.7 (3.65)
Hydrocharitaceae	28	2712	4048	-1336	0.67	1.00	0.79	637	4.25 (11.9)
Hydroleaceae	1	77	145	-68	0.53	1.00	1.00	43	1.79 (1.5)
Hymenophyllaceae	36	3233	5205	-1972	0.62	1.00	0.83	462	6.99 (19.35)
Hymenophytaceae	1	116	145	-29	0.80	1.00	1.00	57	2.03 (2.15)
Hypericaceae	3	1351	434	917	3.11	1.00	0.67	525	2.57 (4.91)
Hypoxidaceae	13	1214	1880	-666	0.65	0.92	0.62	505	2.4 (3.89)
Icacinaceae	6	747	867	-120	0.86	1.00	1.00	131	5.7 (13.82)
Iridaceae	23	3510	3325	185	1.06	1.00	0.87	1079	3.25 (7.08)
Isoetaceae	13	846	1880	-1034	0.45	0.92	0.46	249	3.39 (7.66)
Jackiellaceae	2	30	289	-259	0.10	1.00	0.50	14	2.14 (3.46)
Jamesoniellaceae	1	27	145	-118	0.19	1.00	1.00	16	1.68 (1.13)
Jubulaceae	1	1	145	-144	0.01	0.00	0.00	1	NA
Juncaceae	55	15,626	7952	7674	1.97	1.00	0.96	1794	8.71 (50.54)
Juncaginaceae	22	3486	3181	305	1.10	1.00	0.91	850	4.1 (10.46)
Jungermanniaceae	8	193	1157	-964	0.17	0.75	0.25	59	3.27 (5.27)
Lamiaceae	376	38,646	54,362	-15,716	0.71	0.99	0.77	4860	7.95 (31.21)
Lauraceae	135	18,866	19,518	-652	0.97	1.00	0.98	1943	9.7 (103.7)
Lecythidaceae	6	973	867	106	1.12	1.00	1.00	363	2.68 (4.33)
Lejeuneaceae	122	985	17,639	-16,654	0.06	0.77	0.11	90	10.94 (30.09)
Lentibulariaceae	58	4995	8386	-3391	0.60	1.00	0.71	944	5.29 (15.71)
Lepicoleaceae	1	44	145	-101	0.30	1.00	1.00	19	2.31 (1.79)
Lepidolaenaceae	3	95	434	-339	0.22	0.67	0.33	37	2.56 (2.97)
Lepidoziaceae	87	2378	12,578	-10,200	0.19	0.92	0.25	198	12.01 (40.58)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Limeaceae	6	626	867	-241	0.72	1.00	1.00	286	2.18 (2.59)
Linaceae	2	820	289	531	2.84	1.00	1.00	394	2.08 (3.23)
Linderniaceae	34	3483	4916	-1433	0.71	1.00	0.76	629	5.53 (12.5)
Lindsaeaceae	13	2112	1880	232	1.12	0.92	0.85	624	3.38 (6.21)
Loganiaceae	92	9968	13,301	-3333	0.75	0.98	0.82	1869	5.33 (15.29)
Lomariopsidaceae	8	647	1157	-510	0.56	1.00	0.75	228	2.83 (4.43)
Lophocoleaceae	52	1566	7518	-5952	0.21	0.83	0.25	201	7.79 (23.74)
Loranthaceae	73	16,788	10,554	6234	1.59	1.00	0.86	2609	6.43 (19.09)
Lunulariaceae	1	178	145	33	1.23	1.00	1.00	64	2.78 (4.93)
Luzuriagaceae	2	287	289	-2	0.99	1.00	1.00	142	2.02 (3.56)
Lycopodiaceae	13	2173	1880	293	1.16	0.92	0.77	667	3.25 (7.87)
Lygodiaceae	4	970	578	392	1.68	1.00	0.75	369	2.62 (3.72)
Lythraceae	24	3222	3470	-248	0.93	1.00	0.79	775	4.15 (8.4)
Malpighiaceae	5	163	723	-560	0.23	0.80	0.40	54	3.01 (4.64)
Malvaceae	449	53,768	64,916	-11148	0.83	0.99	0.82	4997	10.76 (44.52)
Marattiaceae	3	234	434	-200	0.54	0.67	0.67	89	2.62 (3.1)
Marchantiaceae	3	256	434	-178	0.59	1.00	1.00	96	2.66 (5.38)
Marsileaceae	9	2231	1301	930	1.71	1.00	0.78	656	3.4 (7.25)
Mastigophoraceae	1	21	145	-124	0.15	1.00	1.00	8	2.62 (3.15)
Melastomataceae	11	1422	1590	-168	0.89	1.00	0.91	389	3.65 (7.04)
Meliaceae	42	4957	6072	-1115	0.82	1.00	0.90	883	5.61 (24.13)
Menispermaceae	24	2695	3470	-775	0.78	1.00	0.88	607	4.43 (14.84)
Metzgeriaceae	11	512	1590	-1078	0.32	0.82	0.18	69	7.42 (18.54)
Molluginaceae	5	1110	723	387	1.54	1.00	0.80	458	2.42 (4.17)
Monimiaceae	28	2755	4048	-1293	0.68	1.00	0.86	490	5.62 (22.1)
Monocarpaceae	1	22	145	-123	0.15	1.00	1.00	7	3.14 (1.57)
Moraceae	50	7743	7229	514	1.07	0.98	0.90	1077	7.18 (25.49)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Musaceae	2	59	289	-230	0.20	1.00	0.50	28	2.1 (2.13)
Myristicaceae	4	732	578	154	1.27	1.00	1.00	210	3.48 (7.31)
Myrtaceae	1788	296,848	258,506	38342	1.15	0.99	0.90	12862	23.07 (193.37)
Nepenthaceae	3	158	434	-276	0.36	1.00	0.33	74	2.13 (2.89)
Nitrariaceae	1	713	145	568	4.93	1.00	1.00	360	1.98 (2.78)
Nothofagaceae	3	514	434	80	1.19	1.00	1.00	187	2.74 (6.32)
Nyctaginaceae	16	2467	2313	154	1.07	1.00	0.88	763	3.23 (6.2)
Nymphaeaceae	17	1313	2458	-1145	0.53	1.00	0.65	309	4.24 (9.52)
Ochnaceae	1	152	145	7	1.05	1.00	1.00	48	3.16 (3.63)
Olacaceae	14	1358	2024	-666	0.67	1.00	0.79	564	2.4 (3.36)
Oleaceae	28	4791	4048	743	1.18	1.00	0.93	1119	4.28 (14.08)
Oleandraceae	1	23	145	-122	0.16	1.00	1.00	15	1.53 (1.06)
Onagraceae	18	3594	2602	992	1.38	1.00	0.61	903	3.98 (11.9)
Ophioglossaceae	10	1279	1446	-167	0.88	1.00	1.00	450	2.84 (6.08)
Opiliaceae	2	478	289	189	1.65	1.00	1.00	174	2.74 (3.7)
Orchidaceae	104	3782	15,036	-11254	0.25	0.91	0.30	832	4.54 (17.84)
Orobanchaceae	2	73	289	-216	0.25	1.00	0.50	59	1.23 (0.59)
Osmundaceae	2	428	289	139	1.48	1.00	1.00	214	2 (2.76)
Oxalidaceae	8	2518	1157	1361	2.18	1.00	1.00	883	2.85 (6.3)
Pallaviciniaceae	9	288	1301	-1013	0.22	1.00	0.33	73	3.94 (6.96)
Pandanaceae	19	987	2747	-1760	0.36	1.00	0.68	253	3.9 (7.16)
Paracryphiaceae	5	570	723	-153	0.79	1.00	1.00	200	2.85 (5.33)
Passifloraceae	5	474	723	-249	0.66	1.00	0.80	202	2.34 (4.43)
Pedaliaceae	3	303	434	-131	0.70	1.00	0.67	160	1.89 (2.38)
Pennantiaceae	1	157	145	12	1.09	1.00	1.00	88	1.78 (1.8)
Petalophyllaceae	1	7	145	-138	0.05	1.00	0.00	4	1.75 (1.5)
Petermanniaceae	1	68	145	-77	0.47	1.00	1.00	32	2.12 (3.37)

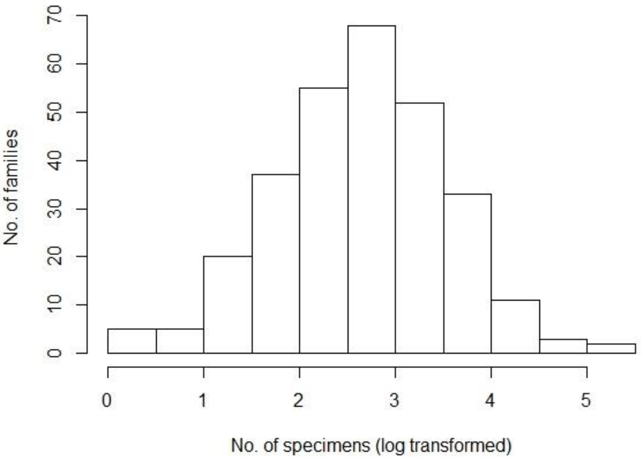
Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Philydraceae	5	752	723	29	1.04	1.00	1.00	399	1.88 (2.05)
Phrymaceae	18	2237	2602	-365	0.86	1.00	0.89	667	3.35 (9.1)
Phyllanthaceae	141	17,182	20,386	-3204	0.84	0.97	0.82	2401	7.15 (34.92)
Phytolaccaceae	1	77	145	-68	0.53	1.00	1.00	32	2.4 (4.67)
Picrodendraceae	29	4023	4193	-170	0.96	1.00	0.97	1072	3.75 (10.71)
Piperaceae	12	1162	1735	-573	0.67	1.00	1.00	271	4.28 (12.45)
Pittosporaceae	74	12,883	10,699	2184	1.20	1.00	0.95	2859	4.5 (13.57)
Plagiochilaceae	24	341	3470	-3129	0.10	0.83	0.21	80	4.26 (8.74)
Plantaginaceae	79	11,309	11,422	-113	0.99	1.00	0.86	2148	5.26 (17.66)
Pleuroziaceae	1	30	145	-115	0.21	1.00	1.00	16	1.87 (2.21)
Poaceae	963	199,773	139,229	60544	1.43	0.98	0.83	8455	23.62 (166.12)
Podocarpaceae	14	1746	2024	-278	0.86	1.00	1.00	509	3.43 (7.42)
Podostemaceae	2	68	289	-221	0.24	1.00	1.00	35	1.94 (2.05)
Polygalaceae	70	8175	10,120	-1945	0.81	0.99	0.80	1820	4.49 (12.16)
Polygonaceae	29	6453	4193	2260	1.54	1.00	0.93	1526	4.22 (12.73)
Polypodiaceae	37	4557	5349	-792	0.85	1.00	0.89	796	5.72 (16.83)
Pontederiaceae	4	573	578	-5	0.99	1.00	0.75	240	2.38 (2.93)
Porellaceae	1	156	145	11	1.08	1.00	1.00	44	3.54 (8.57)
Portulacaceae	59	7176	8530	-1354	0.84	0.98	0.75	1557	4.6 (12.81)
Posidoniaceae	8	767	1157	-390	0.66	1.00	1.00	139	5.51 (19.75)
Potamogetonaceae	17	3036	2458	578	1.24	1.00	0.94	746	4.06 (11.66)
Primulaceae	44	4586	6361	-1775	0.72	1.00	0.75	1207	3.79 (11.47)
Proteaceae	887	106,904	128,241	-21337	0.83	0.99	0.92	8024	13.32 (63.36)
Pseudolepicoleaceae	4	27	578	-551	0.05	0.75	0.00	16	1.68 (1.01)
Psilotaceae	7	1005	1012	-7	0.99	1.00	1.00	357	2.81 (4.63)
Pteridaceae	46	12,075	6651	5424	1.82	0.98	0.96	2040	5.91 (20.95)
Putranjivaceae	4	706	578	128	1.22	1.00	1.00	206	3.42 (7.88)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Radulaceae	22	534	3181	-2647	0.17	0.86	0.32	69	7.73 (20.76)
Ranunculaceae	58	10,145	8386	1759	1.21	1.00	0.91	1864	5.44 (22.57)
Restionaceae	111	14,119	16,048	-1929	0.88	1.00	1.00	1858	7.59 (40.63)
Rhamnaceae	200	24,741	28,916	-4175	0.86	1.00	0.90	3503	7.06 (24.28)
Rhizophoraceae	12	2625	1735	890	1.51	1.00	1.00	455	5.76 (14.41)
Ricciaceae	41	1481	5928	-4447	0.25	0.90	0.46	216	6.85 (20.78)
Riellaceae	2	7	289	-282	0.02	1.00	0.00	4	1.75 (1.5)
Ripogonaceae	5	424	723	-299	0.59	1.00	1.00	190	2.23 (3.63)
Rosaceae	20	3177	2892	285	1.10	1.00	0.85	859	3.69 (12.31)
Rousseaceae	2	412	289	123	1.42	1.00	1.00	165	2.49 (4.09)
Rubiaceae	308	32,138	44,530	-12392	0.72	0.99	0.80	3199	10.04 (45.01)
Ruppiaceae	4	756	578	178	1.31	1.00	1.00	255	2.96 (4.39)
Rutaceae	429	52,137	62,024	-9887	0.84	0.99	0.88	5444	9.57 (50.87)
Salicaceae	12	1187	1735	-548	0.68	1.00	1.00	263	4.51 (14.33)
Salviniaceae	2	567	289	278	1.96	1.00	1.00	262	2.16 (3.94)
Santalaceae	58	14,949	8386	6563	1.78	1.00	0.91	2898	5.15 (14.57)
Sapindaceae	208	30,153	30,072	81	1.00	1.00	0.89	3954	7.62 (36.99)
Sapotaceae	32	3559	4627	-1068	0.77	1.00	0.88	626	5.68 (21.35)
Scapaniaceae	13	172	1880	-1708	0.09	0.85	0.15	56	3.07 (3.93)
Schistochilaceae	10	74	1446	-1372	0.05	0.50	0.10	33	2.24 (2.07)
Schizaeaceae	7	1177	1012	165	1.16	1.00	0.71	424	2.77 (4.74)
Scrophulariaceae	192	36,658	27,759	8899	1.32	0.98	0.89	4027	9.1 (43.9)
Selaginellaceae	9	1111	1301	-190	0.85	0.89	0.67	406	2.73 (4.92)
Simaroubaceae	6	679	867	-188	0.78	1.00	1.00	207	3.28 (6.71)
Smilacaceae	6	940	867	73	1.08	1.00	0.83	396	2.37 (3.11)
Solanaceae	180	26,659	26,024	635	1.02	1.00	0.80	3571	7.46 (38.28)
Solenostomataceae	2	67	289	-222	0.23	1.00	1.00	29	2.31 (3.96)

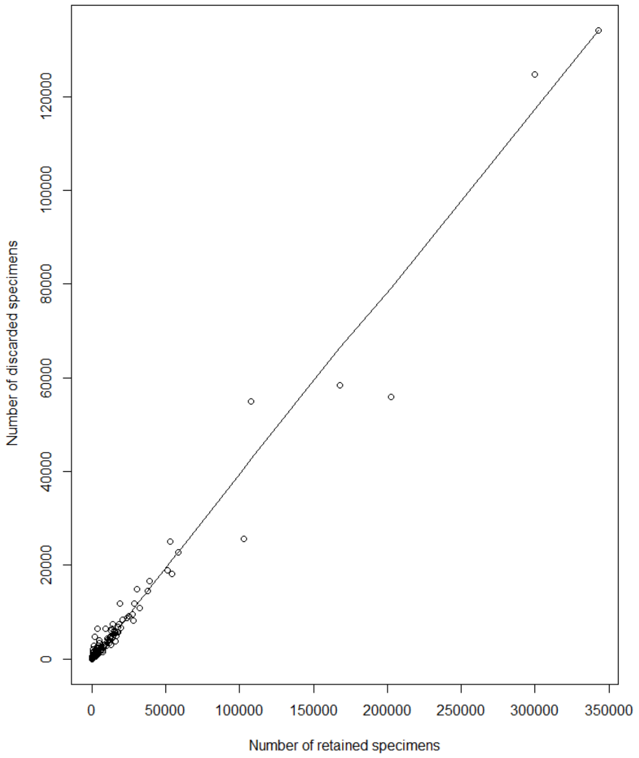
Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD)
Sparganiaceae	1	117	145	-28	0.81	1.00	1.00	76	1.53 (0.99)
Sphaerocarpaceae	1	6	145	-139	0.04	1.00	0.00	5	1.2 (0.44)
Sphenocleaceae	1	73	145	-72	0.50	1.00	1.00	48	1.52 (1.27)
Stemonaceae	4	230	578	-348	0.40	1.00	1.00	90	2.55 (3.03)
Stylidiaceae	205	17,604	29,639	-12035	0.59	1.00	0.86	2299	7.65 (23.81)
Surianaceae	5	916	723	193	1.27	1.00	1.00	395	2.31 (2.54)
Symplocaceae	12	678	1735	-1057	0.39	1.00	0.92	183	3.7 (9.02)
Taccaceae	2	364	289	75	1.26	1.00	1.00	182	2 (2.19)
Targioniaceae	1	46	145	-99	0.32	1.00	1.00	25	1.84 (1.88)
Tectariaceae	7	815	1012	-197	0.81	0.86	0.86	259	3.14 (5.97)
Thelypteridaceae	19	1314	2747	-1433	0.48	1.00	0.74	366	3.59 (6.56)
Thismiaceae	3	33	434	-401	0.08	1.00	0.33	24	1.37 (1.17)
Thymelaeaceae	102	20,132	14,747	5385	1.37	1.00	0.93	3583	5.61 (17.95)
Treubiaceae	2	5	289	-284	0.02	0.50	0.00	3	1.66 (0.57)
Trichocoleaceae	5	64	723	-659	0.09	0.80	0.20	35	1.82 (1.31)
Trichotemnomataceae	1	1	145	-144	0.01	0.00	0.00	1	NA
Trimeniaceae	1	103	145	-42	0.71	1.00	1.00	40	2.57 (3.9)
Triuridaceae	2	25	289	-264	0.09	1.00	0.50	11	2.27 (1.1)
Typhaceae	2	827	289	538	2.86	1.00	1.00	387	2.13 (3.95)
Urticaceae	23	3285	3325	-40	0.99	1.00	0.78	965	3.4 (7.94)
Verbenaceae	3	423	434	-11	0.98	1.00	1.00	243	1.74 (1.7)
Violaceae	21	6376	3036	3340	2.10	1.00	0.90	1665	3.82 (9.14)
Vitaceae	34	4695	4916	-221	0.96	1.00	0.91	786	5.97 (18.65)
Winteraceae	11	1883	1590	293	1.18	1.00	0.91	557	3.38 (6.93)
Woodsiaceae	10	769	1446	-677	0.53	0.90	0.90	227	3.38 (4.94)
Xanthophyllaceae	2	319	289	30	1.10	1.00	1.00	71	4.49 (8.3)
Xanthorrhoeaceae	27	2479	3904	-1425	0.64	1.00	1.00	664	3.73 (9.3)

Family	Species	Observed Specimens (O)	Expected specimens (E)	Bias (O-E)	Bias (0/E)	Proportion of species (p>1)	Proportion of species (p ≥ 20)	No. of Collectors	Average specimens collected per collectors (SD
Xyridaceae	23	3440	3325	115	1.03	1.00	0.87	723	4.75 (14.86)
Zamiaceae	25	1117	3614	-2497	0.31	0.96	0.76	285	3.91 (8.95)
Zingiberaceae	12	646	1735	-1089	0.37	1.00	0.75	188	3.43 (8.74)
Zosteraceae	4	364	578	-214	0.63	1.00	0.75	122	2.98 (4.57)
Zygophyllaceae	51	12,870	7374	5497	1.75	1.00	0.92	1937	6.64 (26.45)

Appendix S2. *Frequency histogram of number of specimen records (x-axis) discarded from each of the 296 native families (y-axis) of Australian flora available in Australasian Virtual Herbarium from our analysis. These records were discarded due to lack of complete attribute information, i.e. either lacking or dubious geographic co-ordinates, missing complete collection date, collector's name or represented a duplicated specimen.*



Appendix S3. *Scatter plot showing relationship between number of specimens retained (x axis) and the number of specimens discarded (y axis) from each of the 296 families of Australian flora available in Australasian Virtual Herbarium. The thick black line represents a locally weighted regression. The linear relationship between discarded and retained specimens per family indicates that the exclusion of these specimens is unlikely to have influenced the proportion of specimens per family retained for the measure of taxonomic bias (Appendix S2). These records were discarded due to lack of complete attribute information, i.e. either lacking or dubious geographic co-ordinates, missing complete collection date, collector's name or represented a duplicated specimen.*



CHAPTER FIVE

FILLING THE GAP: HOW QUICKLY DO AUSTRALIA’S HERBARIA DIGITISE THEIR VOUCHERED SPECIMENS?

ABSTRACT

Digitisation of plant specimen records held within Australia’s herbaria began in the mid-1980s. To date, more than 5,500,000 specimens have been digitised by herbaria across the country, substantially increasing accessibility to primary occurrence data for scientific research. However, there has been little analysis of the patterns of digitisation across the eight key state and territory herbaria, which hold more than 7,000,000 records, combined. In this study, we assessed the digitisation effort of these eight major herbaria over the last three decades (1986–2015) by asking: 1) To what extent is there a time lag between the date of collection of Australian native plant specimens and the subsequent digitisation of that record? and 2) How consistent has been the digitisation effort over time, both within and between herbaria? We obtained information on the date of collection and digitisation for almost 600,000 specimens collected over the period 1986–2015. We found that only 30% of specimens were digitised within a year of collection, while the digitisation lag for 19% exceeded a decade. Of the three decadal periods, the highest average rate of digitisation typically occurred during the mid-decade (1996–2005), while substantially smaller lag times occurred during the most recent period (2006–2015), with the median lag ranging from 0.23 years (Northern Territory Herbarium) to 1.44 years (Western Australia Herbarium). Based on estimated numbers of specimens held in each herbarium, ~ 30% of specimens remain to be digitised. A host of impediments are likely to contribute to a digitisation lag for herbaria, including lack of consistent funding, poor documentation of specimens during curation process and lack of efficient workforce. If Australia is to succeed in reaching its target of being the first OECD (Organisation for Economic Co-operation and Development) country to fully document its

biodiversity, considerably more resources will need to be placed into digitising the remainder of these specimens.

INTRODUCTION

Herbarium collections constitute a permanent record of the distribution of taxa through space and time. Digitisation of specimen records within herbaria began in the 1970s and involves electronically databasing the information contained on the specimen label, particularly the scientific name, collector's name, date of collection, locality description, and geographic coordinates (if known) (Crovello 1972; Morris and Glen 1978). In addition, other attributes such as images (if available), habitat description, and global unique identification number are also assigned against a digitised specimen. Digitization of herbaria records has now become a global enterprise, with data and images being captured by myriad institutions around the world (Ellwood et al. 2018).

Innovations in database design and the creation of data aggregators, such as the Global Biodiversity Information Facility (GBIF, https://www.gbif.org/), the Atlas of Living Australia (ALA, https://www.ala.org.au/), and Integrated Digitized Biocollections (iDigBio, https://www.idigbio.org/), permit access to millions of species records held in institutions that would otherwise be physically inaccessible to the researcher. As a result, researchers are now able to visualize and analyze patterns of biodiversity in novel and exciting ways, increasing the breadth and scale of research questions (Graham et al. 2004; Soltis 2017). For example, meta collections of digitised records can be applied to: develop 'maps of ignorance', highlighting locations with biodiversity knowledge shortfalls (Haque et al. 2017; Stropp et al. 2016); undertake species distribution modelling (Loiselle et al. 2008); aid with conservation planning (Greve et al. 2016); and identify hotspots of invasive species (O'Donnell et al. 2012).

The extent to which information on preserved specimens is digitally accessible to the broader community is a critical question (Yesson et al. 2007), as undoubtedly accessibility will greatly increase the potential for collections to aid research and conservation. Yet, how long does it take for this material to be digitised, and how much material is currently awaiting this process? Presently, there are around 3400 herbaria worldwide housing an estimated 380,000,000 specimens of plants, algae and fungi (Thiers 2017). Meyer *et al.* (2016) calculated that only 17% of 120,000,000 terrestrial herbarium specimens collected prior to 2014 are digitally accessible via GBIF, the largest aggregator of NHC data. GBIF presently contains > 213,000,000 records for Kingdom Plantae (as of 27 March 2018). Of these, approximately 66,500,000 represent preserved specimens while the basis of an additional 10,200,000 is unknown, indicating that around 19% of plant specimens are now digitised and accessible. However, estimates of the total number of specimens in herbaria are difficult to determine (Ariño 2010). Furthermore, institutions in non-western regions, especially south-east Asia, Africa and Brazil, have large numbers of specimens remaining to be digitised (Meyer et al. 2015; Sousa-Baena et al. 2014; Stropp et al. 2016; Yang et al. 2014) or incorporated into the major data aggregators such as GBIF. Hence, delays in the digitisation of existing collections, as opposed to the collection of new records (Meyer et al. 2015; Vollmar et al. 2010), is a major factor limiting the spatial, temporal and taxonomic coverage of digital natural history collections (Meyer et al. 2016).

Australia was an early proponent in the digitisation of herbarium specimens, which began in the Australian National Herbarium and Western Australian Herbarium in the mid-1980s. Today, the electronic databases of herbaria across the country are accessible via the Australasian Virtual Herbarium (AVH, https://avh.chah.org.au/). The AVH is one of the largest virtual herbariums in the world, containing more than 8,000,000 records from major herbaria in Australia and New Zealand. As such, the AVH is the main database for describing the Australian flora, and this site is widely used for numerous purposes ranging from research to citizen science (Cantrill 2018). Although the usage of the AVH is expanding, to date there has been no analysis of potential lags in the period between a specimen being collected and when it is digitised and thus made available to the broader scientific community. Knowledge of this lag may help to prioritise future digitisation and sampling effort, which will be key if Australia is to realise its goal of being the first OECD (Organisation for Economic Co-operation and Development) country to fully document its biodiversity (ASC, 2018). As such, this study evaluated digitisation efforts of the eight major herbaria in Australia by asking the following questions:

1. To what extent is there a time lag between the date of collection of Australian native plants specimens and their subsequent digitisation?
2. How consistent has been the digitisation effort over time?

METHODS

Data source and preparation

Using the Atlas of Living Australia web portal (http://ala.org.au), we downloaded digitised records in the Australasian Virtual Herbarium (AVH) (http://avh.chah.org.au/) that were collected between 1986–2015 (http://www.ala.org.au/; accessed 1 January 2016). We chose this time period because digitisation of specimens in Australia began in the 1980s (Cantrill 2018). Note, however, that digitisation began at different time periods across the various herbaria (as stated below). Our dataset includes AVH records of 598,975 preserved specimens from eight state and territory herbaria – State Herbarium of South Australia (Adelaide [AD], digitisation began 1990), Queensland Herbarium (Brisbane [BRI], digitisation began 2001), Australian National Herbarium (Canberra [CANB], digitisation began 1987), Tasmanian Herbarium (Hobart [HO], digitisation began 1996), National Herbarium of Victoria (Melborne [MEL], digitisation began 1993), National Herbarium of New South Wales (Sydney [NSW], digitisation began 1990), Western Australian Herbarium (Perth [PERTH], digitisation began 1986), Northern Territory Herbarium (Darwin, [DNA], digitisation began 2009) – that met the following criteria: specimens were not cultivated (i.e. specimens taken from gardens and agricultural trials were excluded); specimens were identified as native taxa at the species level (i.e. hybrid, non-native, and genus-level identifications were excluded); the geographic coordinates of sampling were recorded and located within the geographic boundary of the Australian coastline (i.e. records from offshore islands or erroneously placed over water were excluded); and the sampling date was recorded. The date of digitisation, vital for this project, is not available in ALA. However, the National Herbarium of Victoria had previously compiled a dataset containing digitisation dates for each specimen in the ALA. Hence, we merged the two datasets via the unique code for each specimen (i.e. a combination of the herbarium code and specimen identification number).

Statistical Analysis

We calculated the average rate of digitisation (i.e. average number of records digitised per year) for each herbarium for three consecutive time periods – 1986–1995, 1996–2005, 2006–2015. We then used methods of survival analysis and non-parametric techniques to examine the dynamics of specimen digitisation and assess how digitisation patterns varied among herbaria and over time (i.e. the entire period 1986–2015). Following Bebber et al. (2010), in the context of our study a specimen's "birth day" corresponded to the date when it was collected, its "death day" corresponded to the date when it was digitised, and its "lifespan" corresponded to the time elapsed between the specimen's birth and death day. We refer to the specimen lifespan as the digitisation lag (DL) because it corresponds to the time lag between a specimen's time of collection and processing at a herbarium. We then used the Kaplan-Meier estimator (Kaplan and Meier 1958) to calculate survival curves, with 95% confidence intervals (CIs), for each state herbarium for the full 30-year time period (1986–2015), and for the three consecutive decades encompassing this period (1986–1995, 1996–2005, 2006–2015). Finally, we conducted non-parametric Kruskal Wallis tests, followed by post hoc pairwise Wilcoxon rank-sum tests with Holm-adjusted *p* values, to assess the significance of differences in median DL among herbaria for different time periods.

RESULTS

Digitisation rate among herbaria

We calculated the average number of specimens digitised per year by each herbaria, for the three decades. Generally, greater numbers of specimens were digitised per year during the mid-decade (i.e. 1996–2005) (see Table 1) (note, however, that there are no data on the total number of specimens collected during any period). For the mid-decade, the Western Australian Herbarium digitised more records per year than any of the other herbaria (average = 8798; 95% CI: 6309–11,287) while the Tasmanian Herbarium digitised the fewest (average = 1301; 95% CI: 855–1748). In this decade, the Australian National Herbarium experienced a substantial increase in the number of specimens digitised, with an average of 2871 per year (95% CI: 1207–4535) compared to the previous decade, when an average of only 19 specimens per year (95% CI: 1–37) were digitised. We could not calculate the average number of specimens digitised in the early decade (i.e. 1986-1995), for the Queensland, Northern Territory and

Tasmania Herbaria, or during the mid-decade (1996-2005) for Northern Territory Herbarium as digitisation had did not begin until after these time periods. In the recent decade, the Queensland Herbarium digitised the highest average number of specimens per year (average = 5372; 95% CI: 2924–7819), the Tasmanian Herbarium averaged 710 specimens per year (95% CI: 371–1049).

Table 1. *Average number of native Australian plant specimens digitised per year by each of the eight state herbaria, across three decadal periods since digitisation of records begin in Australia in mid-1980s. 95% confidence interval is given in brackets. NA = Lag count not be calculated as digitisation had not yet begun at these institutions (see Methods for the date when digitisation began.*

Herbaria	First decade (1986–1995)		Second decade (1996–2005)		Third decade (2006–2015)	
	Average (95% CI)	Total	Average (95% CI)	Total	Average (95% CI)	Total
AD	2102 (703-3502)	12614	6946 (1522-12,369)	69,457	3866 (580-7152)	38,657
BRI	NA	NA	5565 (4250-6881)	27,827	5372 (2924-7819)	53,718
CANB	19 (1-37)	172	2871 (1207-4535)	28,713	2768 (2082-3453)	27,678
DNA	NA	NA	NA	NA	2011 (1175-2847)	14,078
HO	NA	NA	1301 (855-1748)	13,014	710 (371- 1049)	7098
MEL	2093 (793-3393)	6280	6196 (3845-8548)	61,964	2114 (1732-2496)	21,139
NSW	2750 (1088- 4413)	16502	3699 (2368-5031)	36992	887 (510-1264)	8869
PERTH	1326 (565-2086)	13255	8798 (6309-11,287)	87,980	5297 (3329-7265)	52,968

Estimates of Digitisation Lag

1986-2015

For the full dataset, comprising 598,975 records collected from eight herbaria, the median value for DL was 2.41 years (95% CI: 2.40–2.43 years). Approximately 30% of specimens (177,451) were digitised within a year of collection (DL < 1 year), while ~19% of specimens (113,589) had lags exceeding 10 years. The median DL over the full time period varied significantly among the eight herbaria (Kruskal-Wallis χ^2 = 89,642, df = 7, $p < 10^{-15}$), and was highest for the National Herbarium of Victoria (6.47 years) and lowest for the Northern Territory Herbarium (0.30 years) (see Figure 1 and for details see Table 1). Post hoc pairwise comparisons indicated that the median DL of all herbaria differed significantly from one another ($p < 0.05$, Holm-corrected p values) with the exception of National Herbarium of Victoria and State Herbarium of South Australia ($p = 0.86$).

1986–1995

During the first decade of digitisation, the herbaria digitised a total of 206,078 records. Median DL was considerably higher at 10.2 years (95% CI: 10.1–10.2 years) than the following two decades: half of the specimens collected during this period took more than a decade to be digitised (see Figure 1 and Table 1). DL also varied significantly among the herbaria during this time (χ^2 = 89,642, df = 7, $p < 10^{-15}$), and the greatest median lag occurring in the Northern Territory Herbarium (18.91 years for the 292 specimens collected during this decade, as digitisation did not begin in this herbarium until 2009) and the lowest in the Western Australian Herbarium (4.40 years) followed by National Herbarium of New South Wales (5.36 years).

1996–2005

During the period 1996–2005, 253,277 records were digitised with a median lag of only 1.71 years (95% CI: 1.69–1.72 years) across the state herbaria, although the time from collection to digitisation still took more than a decade for 33 % of specimens. DL also significantly varied among the herbaria (Kruskal-Wallis χ^2 = 38,755, df = 7, p < 10-15). The shortest lag occurred in the Tasmanian Herbarium (0.79 years), followed by the Western Australian Herbarium (1.00 year): the greatest lag was the Northern Territory Herbarium (9.49 years) followed by State Herbarium of Adelaide (5.25 years).

2006–2015

From 2006–2015, 139,620 records were digitised. Over this period, DL declined dramatically with a median of less than one year among six of the eight state herbaria, although there remained significant differences in DL among most institutions (see Table 1). In both the Northern Territory Herbarium and the National Herbarium of New South Wales, 50% of specimens were digitised in < 0.23 years (95% CI: 0.23–0.24 years) and 0.28 years (95% CI: 0.27–0.29 years) of being collected, respectively. The pairwise Wilcoxon test revealed that there was no significant difference in median DL in between these two institutions ($p = 0.46$). The highest median DL was recorded for the Western Australian Herbarium (1.44 years) followed by the Queensland Herbarium (1.34 years).

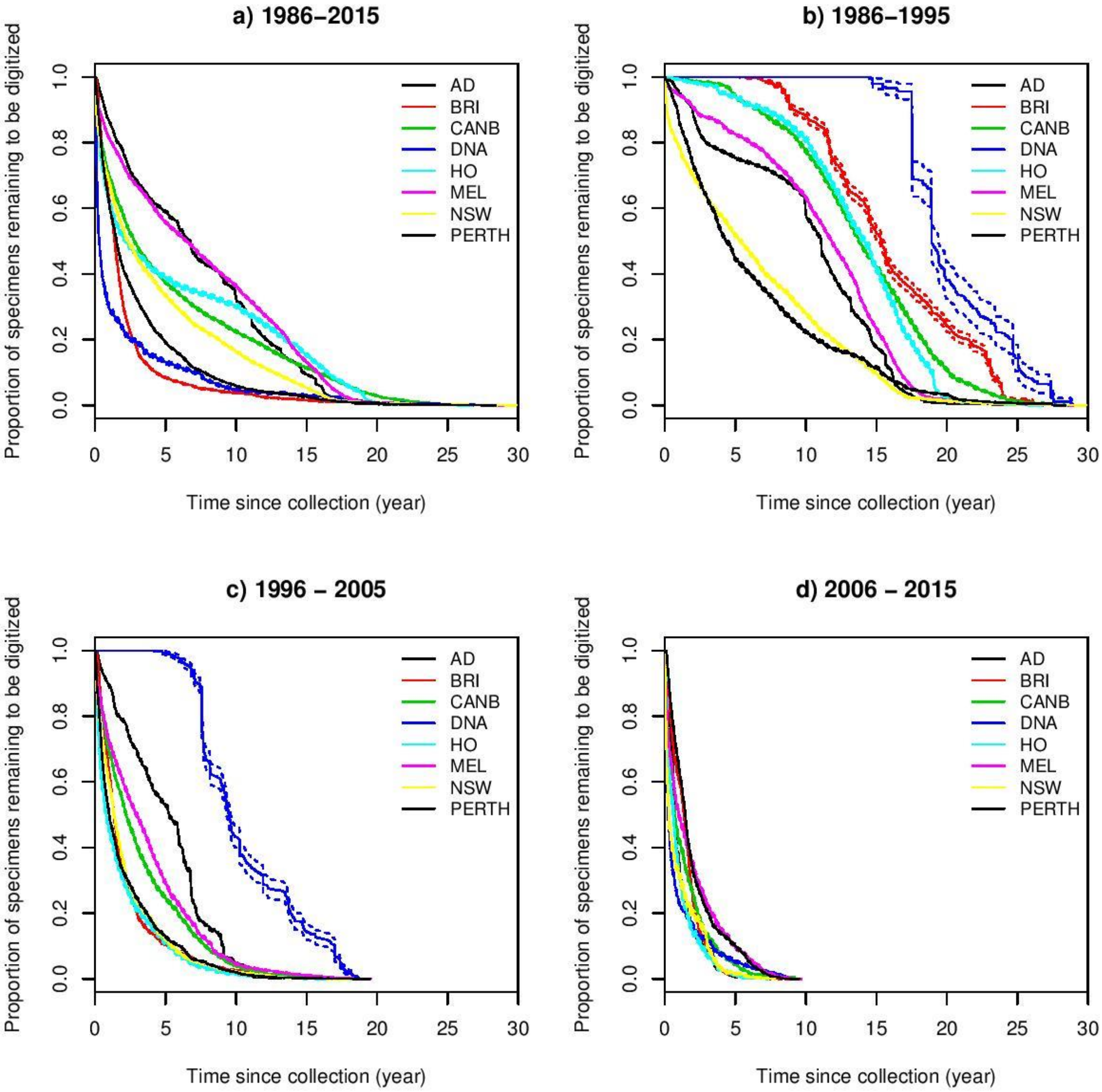


Figure 1. *The fraction of native Australian plant specimens that remaining to be digitised against collection time, for eight Australian state herbaria [South Australia (Adelaide [AD]), Queensland Herbarium (Brisbane [BRI]), Australian National Herbarium (Canberra [CANB]), Tasmanian Herbarium (Hobart [HO]), National Herbarium of Victoria (Melborne [MEL]), National Herbarium of New South Wales (Sydney [NSW]), Western Australian Herbarium (Perth [PERTH]), Northern Territory Herbarium (Darwin, [DNA]) For each series, the central line shows the median while the dashed lines show the 95% confidence limits (Kaplan-Meier estimators).*

Table 2. *Median digitisation lag (DL) and corresponding 95% confidence intervals (Kaplain-Meier estimator) for native Australian plant specimens sent to eight Australian state herbaria [Adelaide (AD), Brisbane (BRI), Canberra (CANB), Darwin (DNA), Hobart (HO), Melbourne (MEL), New South Wales (NSW) and Perth (PERTH)] over the time period 1986 and 2015. Kruskal-Wallis tests indicate highly significant differences in median DL among herbaria for all time periods (p < 10-15). See text for year in which digitisation begin within each institution.*

Herbaria	No. of events (Specimen	Median lag in digitisation (DL)	χ^2*(df), p*
		All-years (1986–2015)	
AD	120,728	6.75 (6.70-6.76)	89,642(7), $< 10^{-15}$
BRI	81,545	1.38 (1.37-1.39)	
CANB	56,563	2.78 (2.74-2.83)	
DNA	14,078	0.30 (0.28-0.31)	
HO	20,112	2.21 (2.09-2.31)	
MEL	89,383	6.47 (6.38-6.57)	
NSW	62,363	2.33 (2.28-2.36)	
PERTH	154,203	1.52 (1.50-1.53)	
		Early-decade (1986–1995)	
AD	63,359	11.08 (11.07-11.08)	36,103(7), $<10^{-15}$
BRI	2333	15.33 (15.18-15.50)	
CANB	15,031	13.82 (13.72-13.89)	
DNA*	292	18.91 (18.90-19.41)	
HO	7372	14.14 (13.97-14.35)	
MEL	49,276	11.82 (11.76-11.87)	
NSW	34,900	6.30 (6.20-6.38)	
PERTH	33,515	4.40 (4.28-4.49)	
		Mid-decade (1996–2005)	
AD	44,094	5.25 (5.23-527)	38,755(7), $< 10^{-15}$
BRI	39,310	1.34 (1.32-1.35)	
CANB	25,962	2.20 (2.17-2.20)	
DNA	886	9.49 (9.33-9.63)	
HO	7654	0.79 (0.74-0.83)	
MEL	27,187	2.96 (2.87-3.02)	
NSW	22,368	1.35 (1.32-1.37)	
PERTH	85,816	1.00 (0.99-1.01)	
		Recent-decade (2006–2015)	
AD	13,275	0.728 (0.71-0.74)	13,151(df), p $< 10^{-15}$
BRI	39,902	1.34 (1.33-1.36)	
CANB	15,570	0.72 (0.68-0.73)	
DNA	12,900	0.23 (0.23-0.24)	
HO	50,860	0.59 (0.57-0.61)	
MEL	12,920	0.99 (0.95-1.03)	
NSW	5095	0.28 (0.27-0.29)	
PERTH	34,872	1.44 (1.43-1.46)	

DISCUSSION

In this study, we assessed the digitisation effort of eight state/territory herbaria in Australia over the last three decades (1986–2015). We found that relatively older specimens took longer to be digitised compared to more recent collections, with the digitisation rate of individual herbaria differing across the three decades. There has also been a considerable difference in the initiation of digitisation efforts across these herbaria: while digitisation began in 1986 in the Western Australian Herbarium, it was not until 2009 that records were first digitised in the Northern Territory Herbarium.

A host of factors act as impediments to the digitisation of older specimens, including poor documentation and lack of resources (Vollmar et al. 2010). Older specimens are often prone to poor documentation, i.e., illegible handwriting, poor description of localities and lack of geographic coordinates, which may cause delays in the digitisation process. Heterogeneous practices in the data curation process may also impede the digitisation of older specimens (Vollmar et al. 2010). Curation involves navigation through space and collections with greater continuity of curation throughout their history can be far more effectively and efficiently digitized (Vollmar et al. 2010).

Consistency in the digitisation workflow is critical for increasing the digitisation rate, which can be deterred by the lack of a continuous stream of funding and efficient staff. Failure of constant funding mechanisms often causes partially-populated databases for which the principal investigator has moved on to something more fundable, leaving the stakeholders at a loss as to how to resource the additional cost of populating the database with the data required to render it useful (Lughadha and Miller 2009). A survey of more than 200 people, from multiple institutions globally, associated with specimen digitisation found that issues related to funding was the primary barrier to digitising the collection (Vollmar et al. 2010). Across

Australia, the average number of records digitised during the decade 2006-2015 was lower for most herbaria than the previous decade, likely due to declines in budget and staff numbers at these institutions (Barker 2012), although decreases in the quantity of material collected (ACS 2018) may also play a role.

Further challenges emerge as electronic resources and their user communities grow. In the endeavour to provide greater online access to NHC data, we should not forget that consistent and high data quality are paramount to maximise the usability of data. In a recent study, Haque et al. (in press) found that one-fourth of digitised records in the Australasian Virtual Herbarium (AVH) were missing attribute information (i.e. lacking collection date, geographical coordinates or taxonomic identity at the species level), although it is likely that these attributes were also missing from the associated specimen label. Regardless, a viable workflow is necessary for institutions to maintain quality while increasing their digitisation rate. Recent innovations in automation approaches may be quite effective for large-scale digitisation (Sweeney et al. 2018; Tegelberg et al. 2012). For example, Sweeney *et al.* (2018) outlined an object-to-image-to-data workflow that enabled them to image and transcribe data for close to 350,000 specimens at a rate three times faster than that typically achieved using traditional approaches.

Where automation is not an option due to lack of resources, especially for small herbaria (i.e. university/regional herbaria), a viable workflow can be established through citizen science and crowdsourcing (Ellwood et al. 2015; Harris and Marsico 2017). Indeed, a number of NHCs within Australia have established citizen science initiatives. Via Digivol (https://volunteer.ala.org.au/#/expeditionList), an initiative of ALA, volunteers can capture data from specimen labels or transcribe text from the diaries and notes of early natural historians. To date, more than 919,000 tasks have been completed by 3211 volunteers transcribing data from 49 institutions (https://volunteer.ala.org.au/institution/list, accessed 17 June 2018).

Completing the digitisation of existing collections will provide a greater understanding of bias and gaps in our knowledge of species' distributions across spatial, temporal and taxonomic space. This information can, in turn, guide future sampling efforts to ensure the optimisation of limited financial and personnel resources (Dalton 2003). As such, how much information remains to be digitised is a key consideration and varies substantially across institutions. For instance, although digitisation did not begin in the Northern Territory Herbarium until 2009, almost all records have now been digitised (see https://collections.ala.org.au/public/show/co25). Similarly, all records in the Western Australian Herbarium, and 99% of records in the Queensland Herbarium have been digitised (see Supplementary Information Table 1). In contrast, 29% of records within the Australian National Herbarium, for which digitisation began in 1987, remain to be digitised. In sum, across the eight institutions an estimated 1,900,000 records, representing 28% of specimens, are yet to be digitised (Supplementary Information Table 1). Hence there remains considerable scope for our understanding of patterns in the distribution of Australia's flora to shift once all herbaria records have been digitised. It is also important to note that it was not possible to quantify the number of specimens remaining to be digitised or how backlogs are prioritised for digitisation, with much of this information being anecdotal rather than numerical.

.

ACKNOWLEDGMENT

I thank Niels Klazenga from National Herbarium of Victoria for helping with retrieval of specimen digitisation date.

REFERENCES

ACS. (2018) *Discovering Biodiversity: A Decadal Plan for Taxonomy and Biosystematics in Australia and New Zealand 2018–2027*. Australian Academy of Science (ACS), Australia.

Ariño A. H. (2010) Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics* **7**, 81-92.

Barker B. T., K., Breitwieser, I. (2012) Sustaining Australasian plant systematics at a time of major achievements. *Australasian Systematic Botany Society Newsletter*, 31-32.

Bebber D. P., Carine M. A., Wood J. R. I., Wortley A. H., Harris D. J., Prance G. T., Davidse G., Paige J., Pennington T. D., Robson N. K. B. & Scotland R. W. (2010) Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences* **107**, 22169-22171.

Cantrill D. J. (2018) The Australasian Virtual Herbarium: Tracking data usage and benefits for biological collections. *Applications in Plant Sciences*, e1026.

Crovello T. J. (1972) Computerization of specimen data from the Edward Lee Greene Herbarium (ND-G) at Notre Dame. *Brittonia* **24**, 131-141.

Dalton R. (2003) Natural history collections in crisis as funding is slashed. Nature Publishing Group.

Ellwood E. R., Dunckel B. A., Flemons P., Guralnick R., Nelson G., Newman G., Newman S., Paul D., Riccardi G., Rios N., Seltmann K. C. & Mast A. R. (2015) Accelerating the Digitization of Biodiversity Research Specimens through Online Public Participation. *Bioscience* **65**, 383-396.

Ellwood E. R., Kimberly P., Guralnick R., Flemons P., Love K., Ellis S., Allen J. M., Best J. H., Carter R., Chagnoux S., Costello R., Denslow M. W., Dunckel B. A., Ferriter M. M., Gilbert E. E., Goforth C., Groom Q., Krimmel E. R., Lafrance R., Martinec J. L., Miller A. N., Minnaert-Grote J., Nash T., Oboyski P., Paul D. L., Pearson K. D., Pentcheff N. D., Roberts M. A., Seltzer C. E., Soltis P. S., Stephens R., Sweeney P. W., Von Konrat M., Wall A., Wetzer R., Zimmerman C. & Mast A. R. (2018) Worldwide Engagement for Digitizing Biocollections (WeDigBio): The Biocollections Community's Citizen-Science Space on the Calendar. *Bioscience* **68**, 112-124.

Graham C. H., Ferrier S., Huettman F., Moritz C. & Peterson A. T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution* **19**, 497-503.

Greve M., Lykke A. M., Fagg C. W., Gereau R. E., Lewis G. P., Marchant R., Marshall A. R., Ndayishimiye J., Bogaert J. & Svenning J.-C. (2016) Realising the potential of herbarium records for conservation biology. *South African Journal of Botany* **105**, 317-323.

Haque M. M., Nipperess D. A., Gallagher R. V. & Beaumont L. J. (2017) How well documented is Australia's flora? Understanding spatial bias in vouchered plant specimens. *Austral Ecology* **42**, 690-699.

Haque M. M., Nipperess D. A., John B. B. & Beaumont L. J. (*in press*) A journey through time: exploring temporal patterns among digitised plant specimens from Australia. *Systematics and Biodiversity (accepted on 24 April 2018). DOI:10.1080/14772000.2018.1472674*.

Harris K. M. & Marsico T. D. (2017) Digitising specimens in a small herbarium: A viable workflow for collections working with limited resources. *Applications in Plant Sciences* **5**, 1600125.

Kaplan E. L. & Meier P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.

Loiselle B. A., Jørgensen P. M., Consiglio T., Jiménez I., Blake J. G., Lohmann L. G. & Montiel O. M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography* **35**, 105-116.

Lughadha E. N. & Miller C. (2009) Accelerating global access to plant diversity information. *Trends in Plant Science* **14**, 622-628.

Meyer C., Kreft H., Guralnick R. & Jetz W. (2015) Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6**, 8221-8229.

Meyer C., Weigelt P. & Kreft H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**, 992-1006.

Morris J. & Glen H. (1978) PRECIS, the National Herbarium of South Africa (PRE) computerized information system. *Taxon* **27**, 449-462.

O'donnell J., Gallagher R. V., Wilson P. D., Downey P. O., Hughes L. & Leishman Michelle R. (2012) Invasion hotspots for non-native plants in Australia under current and future climates. *Global Change Biology* **18**, 617-629.

Soltis P. S. (2017) Digitization of herbaria enables novel research. *American Journal of Botany* **104**, 1281-1284.

Sousa-Baena M. S., Garcia L. C., Peterson A. T. & Brotons L. (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions* **20**, 369-381.

Stropp J., Ladle R. J., Malhado M., Ana C., 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.

Sweeney P. W., Starly B., Morris P. J., Xu Y. M., Jones A., Radhakrishnan S., Grassa C. J. & Davis C. C. (2018) Large-scale digitization of herbarium specimens: Development and usage of an automated, high-throughput conveyor system. *Taxon* **67**, 165-178.

Tegelberg R., Haapala J., Mononen T., Pajari M. & Saarenmaa H. (2012) The development of a digitising service centre for natural history collections. *ZooKeys* **209**, 75-86.

Thiers B. (2017) Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/science/ih/.

Vollmar A., Macklin J. A. & Ford L. (2010) Natural History Specimen Digitization: Challenges and Concerns. *2010* **7**, 93-112.

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.

Yesson C., Brewer P. W., Sutton T., Caithness N., Pahwa J. S., Burgess M., Gray W. A., White R. J., Jones A. C., Bisby F. A. & Culham A. (2007) How global is the global biodiversity information facility? *PLoS One* **2**, e1124.

SUPPLEMENTARY INFORMATION

Table S1. *Eight state or territory herbaria in Australia with their estimated number of specimens and estimated number of specimens digitised (data from www.ala.org.au).*

Herbaria (Code)	Estimated number of specimens	Estimated number of specimens digitised (%)
State Herbarium of South Australia (AD)	1,030,000	723,000 (70.2)
Queensland Herbarium (BRI)	849,023	839,023 (98.8)
Australian National Herbarium (CANB)	1,142,785	810,768 (70.9)
Northern Territory Herbarium (DNA)	287,000	288000 (99.6)
Tasmanian Herbarium (HO)	285,700	177,000 (62.0)
National Herbarium of Victoria (MEL)	1,386,403	868,232 (62.6)
National Herbarium of New South Wales (NSW)	1,425,000	720,000 (50.5)

CHAPTER SIX

DISCUSSION

An overview of key findings

Species occurrence data represent primary information on the three basic dimensions that characterise species' distributions—taxonomy, space and time—and are fundamental to biodiversity research, natural resource management and conservation policy. However, biases, gaps and uncertainties or errors in these data remain a central problem, potentially leading to incorrect ecological inferences and inefficient use of resources for conservation.

The overarching goal of my thesis was to explore and assess information on Australia's flora accessible via the Australasian Virtual Herbarium (AVH). In Chapter 2, I explored spatial patterns among preserved specimens recorded in the AVH, through an assessment of sampling effort and inventory completeness at the bioregional scale. This chapter extended research that I undertook for a Master of Research, by focusing on vouchered specimens, assessing the relationship between spatial patterns of sampling and human influence, and identifying regions likely to harbour previously unrecorded species. I found considerable biases in sampling effort with the density of specimen records across Australia's bioregions ranging from 0.02 records/km^2 in Gibson Desert to 8.37 records/km^2 in the Wet Tropics. The spatial pattern in sampling effort is strongly influenced by human settlement and accessibility as well as the presence of known biodiversity hotspots—a pattern consistent with the other Natural History Collections (NHC) (Nelson *et al.* 1990; Ubirajara *et al.* 2016). While inventory completeness of Australia's flora is relatively high at the bioregional scale, I found that the probability of sampling a species previously unrecorded in a bioregion ranges from 0.33% (Wet Tropics) to 11.7% (Arnhem Coast).

In Chapter 3, I assessed temporal patterns in the collection of flora specimens, sampled over the last ~200 years (1800–2009). There was a clear temporal bias in collecting, with 80%

of records collected between 1970–1999. South-eastern Australia, the Wet Tropics in north-east Queensland, and parts of Western Australia have received the most consistent sampling effort over time, whereas much of central Australia has had low temporal consistency. I also mapped potential areas of lost and future opportunies. Areas where we have lost the opportunity to capture previous patterns in flora are those with a poor record of historical sampling and are mainly located in the northern region of Western Australia and upper Queensland (see Figures 1 & 2, Chapter 3). Areas that represent future sampling opportunities are those that experienced considerable historical sampling but from which few records have been collected recently. This includes the western region of South Australia and central parts of the Northern Territory. I have also identified regions for which we can make reliable inferences about biodiversity patterns of Australia's flora, i.e. regions with a long history of persistent and intensive sampling. These areas primarily lie in the south-east and south-west of the continent.

The detail and accuracy of information recorded on specimen labels is rarely quantified but constitutes an important step in assessing the quality of data held in NHCs (Boakes *et al.* 2010; Meyer *et al.* 2016; Stropp *et al.* 2016). In Chapter 3, I also found that incomplete attribute information on specimen labels (i.e. lacking geographic coordinates or a complete date) were most common among older records, although ~30% of specimens collected in the last three decades (1980–2009) also had incomplete information.

In Chapter 4, I assessed taxonomic shortfalls in preserved specimens from 296 plant families native to Australia, for which records have been collated into the AVH. I found that the number of preserved specimens per family is not proportional to the family's known species richness, highlighting a taxonomic bias in the records of native plant families in the AVH. For 29% of Australia's plant families (i.e. 86), the number of digitised records constitutes < 50%

of the number expected. Further, only 34% of families (100) have at least 20 specimens digitised for each species recorded in the AVH. There is a strong positive correlation between the number of collectors sampling a family and the taxonomic bias of that family.

Explicitly quantifying inventory completeness across spatio-temporal space for individual families was beyond the scope of this thesis, although this is a logical next step for future studies. For example, we have identified that less than a quarter of the continent has been reliably sampled (i.e. has both high completeness and high temporal consistency), with most of these locations lying in south-eastern Australia, north-eastern Queensland, and south-western Australia (Chapter Three, Fig. 3). Widespread families such as Poaceae, Myrtaceae, Fabaceae and Asteraceae are most likely to be sampled more comprehensively across space and time around these locations. We have also identified areas (i.e. mainly remotely located) with a relatively higher degree of inventory completeness but which lack historical temporal coverage. In these areas, spatially restricted families might have a good taxonomic coverage due to the recent sampling activities. For example, the Wet Tropics has a concentration of small-ranged families with relatively few species and was also explored relatively late in European settlement. It is now of considerable scientific and conservation interest due to these families. Thus, we have in this region an interesting combination of high completeness, good taxonomic coverage but poor temporal coverage (many families not discovered until relatively late because they are only found in the Wet Tropics).

The extent to which information on preserved specimens is digitally accessible to the broader community is a critical question (Yesson *et al.* 2007), as accessibility increases the potential for collections to aid research and conservation. Hence, delays in the digitisation of existing collections, in addition to the collection of new records (Meyer *et al.* 2015; Vollmar *et al.* 2010), is a major factor limiting the spatial, temporal and taxonomic coverage of digital

NHCs (Meyer *et al.* 2016). In Chapter 5, I assessed the digitisation effort of Australia's eight major state and territory herbaria over the last three decades (1986–2015). Using survival analysis, I have calculated the time lag between the date of collection of a plant specimen and the subsequent digitisation of that record. Only 30% of specimens are digitised within a year of collection, and this lag exceeds a decade for 19% of specimens. For most herbaria, the highest average rate of digitisation occurred during the mid-decade (1995–2005), while the shortest lag occurred in the most recent decade (2006-2015).

Implications of my research

Australia is one of the world's 17 biologically megadiverse countries, which together support more than 70% of the world's biological diversity on 10% of its surface area (Mittermeier *et al.* 2011). Most plant taxa in Australia are endemic and many are globally important for understanding the evolutionary history of the planet. In an era of accelerated habitat modification and species extinction rates, a comprehensive understanding of biodiversity is necessary to meet international, national and regional targets for conservation and sustainable development.

The recent decadal plan for biodiversity and systematics (2018-2027) of Australia and New Zealand aims to improve the knowledge of the taxa in the Australasia region through accelerating sampling and digitisation effort (ACS 2018). The ACS (2018) report revealed that since the 1990s there has been a decline in the databasing of plants specimen collections and the annual rate of naming new species. This decadal report also highlighted the potential consequences of poor knowledge of taxa, which includes compromising the effectiveness of biosecurity and research into diverse areas such as the effects of climate change and other environmental stresses on biodiversity.

An important step necessary to advance our understanding of biodiversity is to identify critical gaps in knowledge (Hortal *et al.* 2015). Moreover, only by knowing where we should trust (or doubt) our knowledge of species' occurrences will we be able to accurately discern biodiversity patterns and know where best to allocate limited resources to improve the quality and coverage of species' occurrence data (Rocchini *et al.* 2011). The findings of my thesis have direct implications for establishing future strategies to improve our knowledge of Australian flora, and will also guide current users of the AVH data to incorporate the magnitude of bias and gaps at spatial, temporal and taxonomic space for future studies.

The findings of Chapter 2 would assist the decision maker to prioritise IBRA bioregions for future sampling. These bioregions are management units frequently used for describing species diversity patterns and for developing national conservation strategies (Mackey *et al.* 2008; Polak *et al.* 2015; Williamson *et al.* 2011). There is considerable variation in the density of records in the AVH collations across bioregions, with sampling effort strongly influenced by human activities.

Spatial gaps and bias in AVH can lead to erroneous perceptions on the spatial distribution of Australian native plants. Therefore, there is a need to be cautious in interpreting the results of studies based on the AVH, given the magnitude of bias and gaps in the collections. For example, in central Australia (i.e. Alice Springs), the inventory completeness is relatively higher in stark contrast to most arid regions (see Chapter Two, Figure 1). Such high inventory completeness reported for the area mostly an artefact of high sampling activity compared with the surrounding arid regions.

The spatio-temporal assessment of data quality, consistency and inventory completeness of Australian flora at a relatively fine scale (50 x 50 km^2 grid cells) (Chapter 3) identified areas where we can be more, or less, confident in our ability to discern biodiversity patterns. This may help to make reliable inferences in characterising species distributions and biodiversity patterns across the continent, and for establishing a temporal baseline for long-term monitoring activities. Furthermore, future studies could incorporate finer temporal resolutions (i.e. daily/monthly/seasonal occurrence information) to conduct phenological studies (Lavoie and Lachance 2006) for regions with a long history of persistent and intensive sampling. Failing to understand the spatio-temporal biases and gaps in AVH collections may hinder implementation of effective conservation strategies, particularly when habitat destruction and climate change have led to conspicuous reductions in biological diversity (Ceballos *et al.* 2015). Further, temporal patterns to species composition will be difficult to discern due to high temporal variability across the collection.

To calculate inventory completeness, we estimated species using the Chao1 estimator. The Chao1 estimator is widely used and accepted to estimate species richness for presence-only data, particularly at the continental scale (Schmidt-Lebuhn et al. 2012; Sousa-Baena et al. 2014; Stropp et al. 2016). However, it should also be noted that this estimator was originally developed for random ecological sampling data, and the conceptual basis of Chao1 is the ‘stopping rule’ used in biodiversity sampling, i.e. additional species are unlikely to be found when all species in a sample are represented by at least two individuals (or samples) (Gotelli and Colwell 2011). Furthermore, species richness estimations based on rarefaction curves, such as Chao1, may underestimate richness due to the higher presence of rare species in the these collections compared to ecologically sampling data, as taxonomic collectors do not collect proportionally to abundance (Garcillán and Ezcurra 2011; Guralnick and Van Cleve 2005). One way to correct the richness estimation is to compare the Natural History Collections (i.e. presence only data) with vegetation-based plot data. However, given the volume of data analysed in this thesis (more than three million vouchered specimens) and its spatial extent (i.e. continental scale), it was beyond our scope to test the effectiveness of the Chao1 in estimating species richness in comparing with vegetation plot data. To date, few non-parametric estimators have emerged as an alternative to the Chao1 estimator (Alroy 2017), but their effectiveness in estimating species richness for presence-only data at a larger scale is yet to be tested. Therefore, we recommend, in future studies, to estimate species richness at a local scale or for targeted taxa, and that it would be preferable to compare the taxonomic collections with ecological sampling data if available.

Australia is a megadiverse country with approximately 85% of its flowering plants being endemic. Taxonomic unevenness may create artificial inflations in species numbers for certain taxa (i.e. chage in the number and proportion of speceis recorded in a given area), and therefore may influence decision-making regarding resource allocation and conservation actions (Farrier *et al.* 2007; Grand *et al.* 2007; Pillon and Chase Mark 2006; Walsh *et al.* 2012). My findings on taxonomic shortfalls (Chapter 4) in Australian native flora families may help to establish future sampling and digitisation strategies to enhance taxonomic representation, especially for species of those families we have very little knowledge. Of the ~7,800,000 records accounted for in the 28 Australian herbaria listed in Table 1, Chapter 1, around 5,500,000 have been digitised, leaving 28.9% remaining. Completing the digitisation of these collections will provide a greater understanding of bias and gaps in our knowledge of species distribution across spatial, temporal and taxonomic space, which can then guide future sampling efforts to ensure the optimisation of limited financial and personnel resources (Dalton 2003). As such, institutions should evaluate their capacity for digitisation of existing material and take the necessary steps to maximise their digitisation effort.

In addition to specimens held within Australian herbaria, many specimens dating to the late 18th and 19th centuries remain held overseas institutions, although some material has since been returned to Australia (Webb 2003). Some of these data held in overseas may already be integrated in the GBIF. Repeating my study with data from GBIF may help to elucidate the extent to which specimens sent overseas, particularly those by early collectors, may distort estimates of inventory completeness and patterns in species' distributions.

It is crucial to scrutinize records before they are used and explicitly communicate their related spatial and temporal errors to avoid erroneous inferences. The quality assessment of species occurrence information resulted in ~ 25% of digitised records being excluded from analyses due to incomplete attribution information (i.e. missing or error in event date, geographical coordinates), or because specimens were recorded as taxa above the species level (e.g. genus). While these records may have been unsuitable for analysis in the present study, they may still be appropriate for other purposes (e.g. those missing dates but which have geographic coordinates may still be useful for some analyses of spatial patterns). Furthermore, the duplicate records I omitted (~13%) may have also suffered from incorrect or incomplete labelling. I excluded these from my dataset due to methodological concerns that their use could introduce noise into our results.

Completeness of the digitisation of NHCs would not only help us to understand the bias and gaps in species' distributions across spatial, temporal and taxonomic space but can also guide us in prioritising future strategic sampling effort under constrained economic budgets (Dalton 2003). With the growing demand for access to AVH data (Cantrill 2018), it is necessary for herbaria to evaluate their institutional capacity to maximise their digitisation effort. In Chapter 5, I was unable to calculate the period of time required for remaining specimens to be digitised, based on current digitisation rates, as to do so requires information on when undigitised specimens were collected. However, exploring other approaches to forecasting required digitisation rates would be useful for prioritisation of budgets.

Recent advances in citizen and crowd source science offer a considerable potential to engage the public in collections-based systematics (Bonney *et al.* 2009; Gioia 2010; Harris and Marsico 2017; Jun *et al.* 2015). For instance, by using Notes from Nature, citizen scientists can help with transcription tasks by transcribing and annotating digital images of biodiversity records held in NHCs (Hill *et al.* 2012). Within Australia, citizen scientists can use DigiVol (https://digivol.ala.org.au/) to transcribe data on plants and animals derived from specimen records and diaries or notes of early natural historians. Furthermore, given the limitation on time and resources, public engagement through citizen science can be applied for target sampling, such as in areas that experienced considerable historical sampling but from which few records have been collected recently. In Australia, such as study could target the western region of South Australia and central parts of the Northern Territory.

Innovations in new technology to foster digitisation efforts of existing collections will also enable us to improve knowledge of biodiversity. For example, by using an automation approach for large scale digitisation it is possible that we will be able to rapidly digitise vast quantities of NHC data (Hudson *et al.* 2015; Sweeney *et al.* 2018). However, to recognise the potential of these innovations it is imperative that the collections-based systematics community trains future systematists with appropriate skills and educate both the public and decision-makers on the value of digital collections (Jun *et al.* 2015).

Conclusion

A sound understanding of biodiversity is critical, particularly as we seek to achieve both environmental and economic sustainability for mega-diverse regions such as Australasia. The knowledge that comes with species occurrence data on distribution, habitat, abundance or rarity, plays a vital role in conservation planning for species and areas. Taxa that are undocumented are more likely to be lost and lost without knowledge of their loss (Ceballos *et al.* 2015). Moreover, conservation science is at a critical stage with unprecedented rates of species extinction (Rands *et al.* 2010). As such, it is imperative that we overcome biodiversity knowledge shortfalls (Hortal *et al.* 2015). The first and fundamental step to achieving this enormous task is to quantify and document our ‘map of ignorance’. This thesis represents the most thorough assessment of biases of the AVH to date and will serve as a springboard to help accelerate and document patterns in Australia’s flora.

REFERENCES

Alroy J (2017) Effects of habitat disturbance on tropical forest biodiversity. Proceedings of the National Academy of Sciences **114**:6056-6061.

Boakes E. H., McGowan P. J., Fuller R. A., Chang-Qing D., Clark N. E., O'connor K. & Mace G. M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology* **8**, e1000385.

Bonney R., Cooper C. B., Dickinson J., Kelling S., Phillips T., Rosenberg K. V. & Shirk J. (2009) Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *Bioscience* **59**, 977-984.

Cantrill D. J. (2018) The Australasian Virtual Herbarium: Tracking data usage and benefits for biological collections. *Applications in Plant Sciences* **6**, e1026.

Ceballos G., Ehrlich P. R., Barnosky A. D., García A., Pringle R. M. & Palmer T. M. (2015) Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances* **1**, e1400253.

Dalton R. (2003) Natural history collections in crisis as funding is slashed. Nature Publishing Group.

Farrier D., Whelan R. & Mooney C. (2007) Threatened species listing as a trigger for conservation action. *Environmental Science & Policy* **10**, 219-229.

Gioia P. (2010) Managing biodiversity data within the context of climate change: towards best practice. *Austral Ecology* **35**, 392-405.

Grand J., Cummings M. P., Rebelo T. G., Ricketts T. H. & Neel M. C. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecology Letters* **10**, 364-374.

Garcillán PP, Ezcurra E (2011) Sampling procedures and species estimation: testing the effectiveness of herbarium data against vegetation sampling in an oceanic island. Journal of Vegetation Science 22:273-280 doi:https://doi.org/10.1111/j.1654-1103.2010.01247.x

Gotelli NJ, Colwell RK (2011) Estimating species richness. Biological diversity: frontiers in measurement and assessment, Second Edition edn., Oxford University Press, Oxford, UK.

Guralnick R, Van Cleve J (2005) Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches. Diversity and Distributions 11:349-359 doi:https://doi.org/doi:10.1111/j.1366-9516.2005.00164.x

Harris K. M. & Marsico T. D. (2017) Digitizing specimens in a small herbarium: A viable workflow for collections working with limited resources. *Applications in Plant Sciences* **5**, 1600125.

Hill A., Guralnick R., Smith A., Sallans A., Gillespie R., Denslow M., Gross J., Murrell Z., Conyers T., Oboyski P., Ball J., Thomer A., Prys-Jones R., De La Torre J., Kociolek P. & Fortson L. (2012) The notes from nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys* **209**, 219-233.

Hortal J., De Bello F., Diniz-Filho J. A. F., Lewinsohn T. M., Lobo J. M. & Ladle R. J. (2015) Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46**, 523-549.

Hudson L. N., Blagoderov V., Heaton A., Holtzhausen P., Livermore L., Price B. W., Van Der Walt S. & Smith V. S. (2015) Inselect: Automating the Digitization of Natural History Collections. *PLoS One* **10**, e0143402.

Jun W., Ickert-Bond S. M., Appelhans M. S., Dorr L. J. & Funk V. A. (2015) Collections-based systematics: Opportunities and outlook for 2050. *Journal of Systematics and Evolution* **53**, 477-488.

Lavoie C, Lachance D (2006) A new herbarium-based method for reconstructing the phenology of plant species across large areas. American Journal of Botany **93**:512-516 doi:https://doi.org/doi:10.3732/ajb.93.4.512

Mackey B. G., Berry S. L. & Brown T. (2008) Reconciling approaches to biogeographical regionalization: a systematic and generic framework examined with a case study of the Australian continent. *Journal of Biogeography* **35**, 213-229.

Meyer C., Kreft H., Guralnick R. & Jetz W. (2015) Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* **6**, 8221-8229.

Meyer C., Weigelt P. & Kreft H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**, 992-1006.

Mittermeier R. A., Turner W. R., Larsen F. W., Brooks T. M. & Gascon C. (2011) Global biodiversity conservation: the critical role of hotspots. In: *Biodiversity hotspots* pp. 3-22. Springer.

Nelson B. W., Ferreira C. A., Da Silva M. F. & Kawasaki M. L. (1990) Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature* **345**, 714-716.

Pillon Y. & Chase Mark W. (2006) Taxonomic Exaggeration and Its Effects on Orchid Conservation. *Conservation Biology* **21**, 263-265.

Polak T., Watson J. E., Fuller R. A., Joseph L. N., Martin T. G., Possingham H. P., Venter O. & Carwardine J. (2015) Efficient expansion of global protected areas requires simultaneous planning for species and ecosystems. *Royal Society Open Science* **2**, 150107.

Rands M. R. W., Adams W. M., Bennun L., Butchart S. H. M., Clements A., Coomes D., Entwistle A., Hodge I., Kapos V., Scharlemann J. P. W., Sutherland W. J. & Vira B. (2010) Biodiversity Conservation: Challenges Beyond 2010. *Science* **329**, 1298-1303.

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

Schmidt-Lebuhn AN, Knerr NJ, González-Orozco CE (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. Journal of Biogeography **39**:2072-2080.

Sousa-Baena MS, Garcia LC, Peterson AT, Brotons L (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. Diversity and Distributions 20:369-381 doi:https://doi.org/10.1111/ddi.12136

Stropp J., Ladle R. J., Malhado M., Ana C., 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.

Sweeney P. W., Starly B., Morris P. J., Xu Y. M., Jones A., Radhakrishnan S., Grassa C. J. & Davis C. C. (2018) Large-scale digitization of herbarium specimens: Development and usage of an automated, high-throughput conveyor system. *Taxon* **67**, 165-178.

Ter Steege H, Haripersaud PP, Banki OS, Schieving F (2011) A model of botanical collectors' behavior in the field: never the same species twice. American Journal of Botany **98**:31-37 doi:https://doi.org/10.3732/ajb.1000215.

Ubirajara O., Pereira P. A., Brescovit A. D., de Carvalho C. J. B., Silva D. P., Rezende D. T., Sa Fortes Leite F., Batista J. A. N., Barbosa J. P. P. P., Stehmann J. R., Ascher J. S., de Vasconcelos M. F., De Marco Jr P., Lowenberg-Neto P., Dias P. G., Ferro V. G. & Santos A. J. (2016) The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. *Diversity and Distributions* **22**, 1232-1244.

Vollmar A., Macklin J. A. & Ford L. (2010) Natural History Specimen Digitization: Challenges and Concerns. *2010* **7**, 93-112.

Walsh J. C., Watson J. E. M., Bottrill M. C., Joseph L. N. & Possingham H. P. (2012) Trends and biases in the listing and recovery planning for threatened species: an Australian case study. *Oryx* **47**, 134-143.

Webb J. B. (2003) *The Botanical Endeavour: Journey Towards a Flora of Australia*. Surrey Beatty & Sons, New South Wales, Australia.

Williamson G. J., Christidis L., Norman J., Brook B. W., Mackey B. & Bowman D. M. (2011) The use of Australian bioregions as spatial units of analysis to explore relationships between climate and songbird diversity. *Pacific Conservation Biology* **17**, 354-360.

Yesson C., Brewer P. W., Sutton T., Caithness N., Pahwa J. S., Burgess M., Gray W. A., White R. J., Jones A. C., Bisby F. A. & Culham A. (2007) How global is the global biodiversity information facility? *PLoS One* **2**, e1124.