A legacy of sampling: Exploring spatial patterns among occurrence records in Australia’s Virtual Herbarium

A thesis submitted for the degree of Master of Research

October 2014

Md Mohasinul Haque

Department of Biological Sciences, Faculty of Science

Macquarie University, Sydney, Australia

ABSTRACT

Understanding biases and errors in Natural History Collections (NHCs) is of paramount importance for the validity of conservation and environmental studies. We explored spatial biases in Australia's Virtual Herbarium (AVH), a database containing more than 7,000,000 records of ~21,000 native species, recorded from across the continent. Specifically, we assessed spatial patterns of sampling and representativeness of the floristic composition of AVH data.

Location Australia

Methods Biases were explored by calculating a sampling allocation index (*SAI*, ratio of observed to expected records) and an index of inventory completeness (*C*-index) based on the ratio of observed to estimated species richness, at multiple spatial scales. Representativeness was determined by the spatial resolution at which the AVH data most closely approximates the composition of vegetation survey plot data gathered at the local scale (0.04 ha).

Results *SAI* indicated that sampling of Australia's native flora has been severely geographically biased, with comparatively few records from arid and subtropical regions. While the *C*–index demonstrated that resolution can significantly impact patterns of spatial bias, Tasmania and the Northern Territory generally retained high *C*–index values. Finally, we found that a 16 ha buffer surrounding the vegetation survey plots was required for AVH data to match 90% of the species known to occur at the plot level.

Main Conclusions Significant spatial biases exist within the AVH. Failure to account for these, when using this database may have serious ramifications for biogeographic studies and conservation planning. We suggest that studies similar to ours be used to assist in planning future systematic surveys and species inventories, and when identifying areas of conservation priority across the continent of Australia.

Keywords

Australia's virtual herbarium, Chao 1 estimator, inventory completeness, natural history collections, sampling allocation, sampling effort, sampling redundancy, spatial bias, spatial scale, species composition.

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. These collections began around 300 years ago, and it is estimated that there are 2.5 – 3 billion biological specimens in collections throughout the world (Pyke & Ehrlich, 2010). Such information has substantial intrinsic value with respect to genetic, phylogenetic, biogeographic, ecological and biographical data, and the specimens have formed the basis of a multitude of environmental and ecological studies (Lane, 1996).

NHCs are now at a turning point in their history. Digital databases that store information recorded on specimen labels began in the 1970s (Graham *et al.*, 2004; Thomas, 2009). Today, more than 528,000,000 digitized records have been incorporated into the largest publically accessible biodiversity distribution network, the Global Biodiversity Information Facility (GBIF, see http://www.gbif.org/). Sharing databases across multiple institutions through distributed networks, and connecting these online to create freely-accessible 'meta-collections', such as GBIF, increases not only the scientific value of these data but also the range of questions that can be explored (Paton, 2009; Balke *et al.*, 2013).

Ready access to NHCs data has enabled researchers to conduct inferential studies on the spatial distribution of biological diversity from global to continental and regional scales (Ter Steege *et al.*, 2006; Barthlott *et al.*, 2007; Ballesteros-Mejia *et al.*, 2013; Lavoie, 2013). In the era of global environmental change, these virtual databases have become an invaluable resource to assess the impact of climate change (Robbirt *et al.*, 2011; Feeley, 2012; Hart *et al.*, 2014); biological invasions (Beaumont *et al.*, 2014) and biodiversity status, and for developing conservation strategies(Ward, 2012) .

Innovations such as the integration of environmental variables with specimen records and introduction of state-of-the-art image-based digitization of information is creating new research scope in morphological, phenological, genetic and biogeographical studies (Bi *et al.*, 2013). The result is a data explosion ripe for scientific enquiry (Krishtalka & Humphrey, 2000).

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 & Ehrlich, 2010) a major concern remains: how comprehensive are these data across space and time? This is a vital question for information within biological collections to be scientifically useful, their errors and biases need to be understood.

Errors and biases in NHCs data

The validity of studies utilising records from these databases is strongly dependent upon the abundance and representativeness of records (Hijmans *et al.*, 2000; Yesson *et al.*, 2007; Santos *et al.*, 2010) and data quality (Soberón & Peterson, 2004). These factors, in turn, are influenced by the haphazard collection of most specimens (Pike & Ehrlich, 2010) and historical biases in sampling effort (Hortal *et al.*, 2008).

Errors and biases in biological collections can be classified as spatial, temporal and taxonomic. To summarise briefly, spatial errors can occur due to incorrect recording of collection locations, while biases result from the opportunistic approach of the collector. For instance, the spatial intensity of 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 under-sampled or not sampled at all (Nelson *et al.*, 1990) and frequently no information on collection effort or methods is recorded. Hence, non-representativeness of sampling associated with opportunistic collecting is one of the most difficult biases to identify (Graham *et al.*, 2004). A consequence is that incorrect conclusions may be drawn with regards to the spatial distribution of biodiversity (Soria-Auza & Kessler, 2008; Boakes *et al.*, 2010). Temporal biases may arise, for example, from collections being concentrated in a particular periods or seasons (e.g. spring and therefore excluding wintering flowering plants) (Rich, 2006). Taxonomic bias occurs due to the preferential collecting of particular species, and can result in artificial gaps in species distributions and community composition and over-estimates of relative abundance (Garcillán & Ezcurra, 2011).

What impact can biases in NHCs have on the outcomes of scientific studies?

Biases associated with NHCs can alter the conclusions of studies in a number of ways. Garcillan & Ezcurra (2011) found that when collectors actively sample *rare species*, and avoid common ones, the result is two-fold: firstly, relative abundance of species becomes artificially skewed such that true abundance cannot be predicted from herbarium collections; secondly, floristic lists compiled from herbarium collections are likely to be more complete than those gathered from field sampling. Similarly, Nelson *et al.* (1990) argued that some proposed centres of endemism in Brazilian Amazonia resulted from highly localised studies of flora rather than real uniqueness.

Observed changes in species richness may be an artefact of changes in sampling trends over time. Museum and herbaria collecting has declined markedly in recent decades, and comparisons of biodiversity trends at different time periods may suggest losses that have actually not occurred (Boakes *et al.*, 2010). Conversely, apparent latitudinal shifts of species range margins may reflect increases in collecting effort rather than establishment of new populations. For example, occurrence records from Australia's Virtual Herbarium (AVH) of seaweed populations gathered over a 20 year period indicated a latitudinal shift had occurred. Based on the assumption that data were free of opportunistic biases Wernberg *et al.* (2011) concluded that the distributions of numerous seaweed species have shifted southward due to climate change. This conclusion was contested by Huisman & Millar (2013) who questioned the completeness of the occurrence records. They argued that population extirpations cannot be determined from NHCs and that Wernberg *et al*. (2011) had made incorrect assumptions regarding collection effort. Huisman & 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. Detectability may also lead to collection biases, such that the distribution of a species with low detectability may be underestimated to a greater extent than an easily detected species (Guillera-Arroita *et al.*, 2010; Sheth *et al.*, 2012).

Australia's Virtual Herbarium

Australia's Virtual Herbarium (www.chah.gov.au/avh) is the digitised form of Australia's state herbaria and now contains specimen records from most of the country's major herbaria. This database is increasingly used in diverse studies associated with conservation and environmental gradients, and ecology and evolution. These include assessments of species distributions across geographic space and environmental gradients (Crisp *et al.*, 2001; Mellick *et al.*, 2011) or in response to climate change; identifying hotspots of invasive species richness (O'Donnell *et al.*, 2012; Duursma *et al.*, 2013); phytogeographical analyses (González-Orozco *et al.*, 2014); prioritizing regions for conservation (Colloff *et al.*, 2014; Lee & Mishler, 2014) and measuring evolutionary signals from phylogeny, taxonomy, endemism and genetic diversity (Laffan & Crisp, 2003; Bickford *et al.*, 2004; Rosauer *et al.*, 2009).

Yet, as with other NHC's, biases and errors are present within this collection, as demonstrated by the aforementioned case-study of seaweed collections (Wernberg *et al*. 2011; Huisman & Millar, 2013). Schemidt-Lebuhn *et al*. (2012) also found that sampling biases within the AVH may lead to erroneous perceptions of species diversity. Their analysis of Asteraceae records indicated that species richness was higher in central Australia, in comparison to other interior regions. This was a reflection of collection activity rather than a true biogeographic pattern. To date there has been no comprehensive study to assess the spatial biases across the entire AVH database. Given that this is the premier database for describing the flora of the Australian continent, identifying its limitations is of paramount interest for the validity of conservation and environmental studies (Hortal *et al.*, 2007; Beck *et al.*, 2014). Therefore, the objectives of this study are to explore, and map, spatial patterns of AVH records. Specifically, to:

1. Explore the spatial pattern of sampling allocation of AVH records to:
    a. Assess whether sampling allocation differs across biomes;
    b. Identify which bioregions are under- or over-allocated in terms of sampling effort;
    c. Assess the extent to which patterns in sampling allocation change with spatial resolution.
2. Explore the spatial pattern of completeness of species inventories from AVH records to:
    a. Assess whether inventory completeness differs across biomes;
    b. Identify how complete inventories are among bioregions;
    c. Assess how patterns of inventory completeness change with spatial resolution.
3. Assess the representativeness of species composition of AVH data at local spatial scales.

MATERIALS AND METHODS

Dataset

All native terrestrial plant specimen data within Australia's Virtual Herbarium (AVH) (http://avh.chah.org.au/) were downloaded from the Atlas of Living Australia (ALA) (http://www.ala.org.au/; accessed 1 January 2014). Individual data files for each plant family identified by the Australian Plant Census (APC) were accessed (*n* = 361 families) and aggregated into 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 were: (1) duplicates; (2) represented non-unique combinations of species, latitude and longitude; (3) not identified to species level (i.e. consisted of a genus name and the epithet "sp."); (4) lacked georeferencing information; (5) cultivated (e.g. in a garden or agricultural trial); (6) hybrids; (7) outside the geographic boundary of the Australian coastline (i.e. records that fall in the ocean, coastal waterways, or on offshore islands). After applying these filters, the final dataset incorporated 7,362,958 records belonging to 21,141 species and 301 families.

Spatial categorization of bioregions by biomes

The Interim Biogeographic Regionalisation of Australia (IBRA v 7.0, 2012) is a key tool used for national and regional planning frameworks. Within IBRA, Australia is divided into 89 terrestrial bioregions based on characteristic geology, landform, native vegetation and climate. The bioregions are also aggregated into seven biomes defined by Olson *et al.*(2001): tropical and subtropical grasslands, savannas and shrublands; deserts and xeric shrublands; temperate broadleaf and mixed forests; mediterranean forests, woodlands and shrublands; temperate grasslands, savannas and shrublands; tropical and subtropical moist broadleaf forests; montane grasslands and shrublands (IBRA, 2012). For ease of reading these will be referred to as: tropical savannas; desert shrublands; temperate forests; mediterranean forests; temperate savannas; tropical forests; and montane grasslands.

Shapefiles of IBRA bioregions and biomes were downloaded (http://www.environment.gov.au/metadataexplorer/explorer.jsp). 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 bioregions and biomes. This enabled us to calculate the number of occurrence records, species and families within each bioregion. Four bioregions (Coral Sea, Indian Tropical Islands, Pacific Sub-tropical Islands, and Sub-Antarctic Islands) were excluded from our analysis as they were beyond the Australian continent.

Of the seven biomes, tropical savannas contain the highest number of bioregions (23), although desert shrublands occupy the largest area (21 bioregions). While temperate forest spans only 7% of the continent, it contains 20 bioregions. Montane grassland and tropical forest are represented by one and two bioregions, respectively (Fig. 1, Table 1).

We further divided the continent into three equal area grids at spatial resolutions of 100 × 100 km (coarse scale: 770 grid cells), 50 × 50 km (medium scale: 3,077 grid cells) and 25 × 25 km (fine scale: 12,294 grid cells) (Table 1). For grid cells along coastlines, we retained only those for which land covered at least 50% of the cell.

The occurrence records, bioregion and biome shapefiles were projected from latitude and longitude (WGS 1984) to Australian Albers equal area grid using the packages "sp" and "raster" (Bivand *et al*., 2013; Hijmans *et al*., 2014) for R version 3.0.1(R Development Core Team 2013). We then overlaid these data with the equal area grids, and calculated the number of occurrence records, species and families within each grid cell at all three spatial resolutions, as well as the bioregion and biome within which the centre of the grid cell was located.

Table 1 Distribution of IBRA (Interim Biogeographic Regionalisation of Australia) bioregions at three spatial resolutions (Coarse [100 × 100 km], medium [50 × 50 km] and fine [25 × 25 km]) within their corresponding biomes.

Biome*	No. bioregions	% Australia occupied by bioregions	No. grid cell		
			Coarse (100 × 100 km)	Medium (50 × 50 km)	Fine (25 × 25 km)
TSG	23	28%	216	889	3,516
DXS	21	46%	359	1,424	5,718
TBM	20	7%	58	307	885
MFW	14	10%	85	226	1,257
TGS	4	7%	46	213	850
TSM	2	0.40%	3	14	53
MGS	1	0.10%	2	4	15

* TSG = Tropical and subtropical grasslands, savannas and shrublands; DXS = Deserts and xeric shrublands; TBM = Temperate broadleaf and mixed forests; MFW = Mediterranean forests, woodlands and shrublands; TGS = Temperate grasslands, savannas and shrublands; TSM = Tropical and subtropical moist broadleaf forests; MGS = Montane grasslands and shrublands.

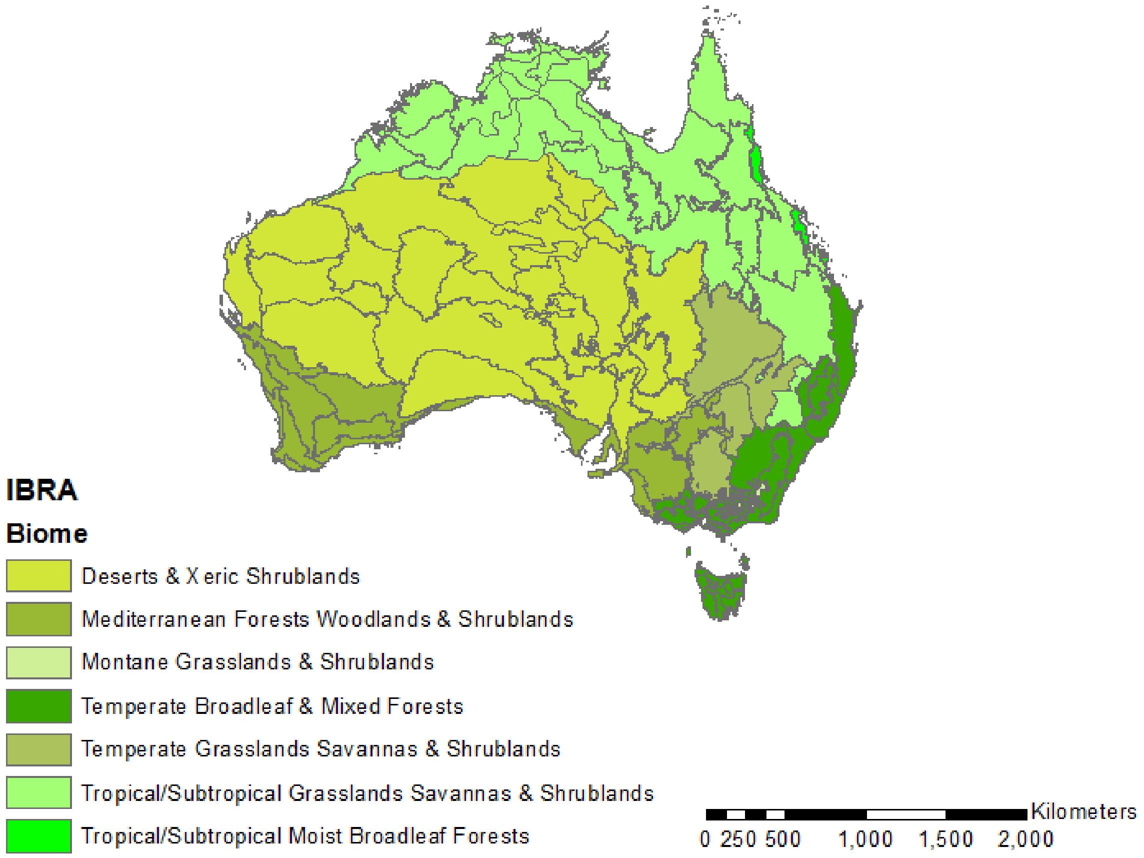
IBRA
Biome
Deserts & Xeric Shrublands
Mediterranean Forests Woodlands & Shrublands
Montane Grasslands & Shrublands
Temperate Broadleaf & Mixed Forests
Temperate Grasslands Savannas & Shrublands
Tropical/Subtropical Grasslands Savannas & Shrublands
Tropical/Subtropical Moist Broadleaf Forests
Kilometers
0 250 500 1,000 1,500 2,000

Figure 1 Spatial distribution of 85 IBRA (Interim Biogeographic Regionalisation of Australia) bioregions in Australia, within their corresponding biome.

Spatial Analysis

We explored non-randomness in sampling effort by assessing sampling allocation, which is a useful measure of the spatial intensity of sampling effort (Fotheringham *et al.*, 2000; Garcillán *et al.*, 2003).We used this metric, in conjunction with inventory completeness, to explore the spatial characteristics of the AVH data at a variety of spatial scales (bioregions and three equal area grids).

(i) Sampling allocation

We calculated sampling allocation for each bioregion and for three equal area grids (coarse, medium and fine scale: 100 × 100 km, 50 × 50 km and 25 × 25 km, respectively). We defined sampling allocation as the ratio of the observed number of occurrence records to the number expected if the spatial distribution of records was random. That is, if occurrence records were randomly distributed across the continent, then the expected number of records (R_{exp}) per spatial unit (*i*) (bioregion or grid cell) would be

$$R_{\mathrm{exp(i)}} = \sum R_{\mathrm{obs(i)}} / \sum A_{\mathrm{i}} \times A_{\mathrm{i}}$$

where, $R_{\mathrm{obs(i)}}$ = observed number of occurrence records for spatial unit *i*, and A_{i} = area of a spatial unit. From this, the Sampling Allocation Index (*SAI*) for each spatial unit (*i*) can then be calculated as

$$SAI_{\mathrm{(i)}} = R_{\mathrm{obs(i)}} / \mathrm{R}_{\mathrm{exp(i)}}$$

SAI values below 1 indicate an under-allocation of occurrence records, relative to a random allocation, while values above 1 indicate over-allocation. We conducted ANOVAs to assess whether *SAI* differs between biomes using bioregions as replicates, where number of replicates was 82.

(ii) Inventory completeness

Inventory completeness (*C*-index) can be defined as the ratio of observed species richness to estimated species richness in a given spatial unit (Soberón *et al.*, 2007). We analysed inventory completeness at the four spatial resolutions (bioregional level and three equal area grids), as choice of resolution can have a considerable effect on inventory completeness (Soberón *et al.*, 2007; Mora *et al.*, 2008; Hortal *et al.*, 2010).

As observed species richness is often a poor estimator of real species richness (Walther & Moore, 2005), we estimated this variable using the non-parametric Chao 1 estimator (Colwell & Coddington, 1994). The Chao 1 estimator calculates the total number of species present, including those species that were not sampled, by extrapolating the asymptote of a rarefaction curve. Small sample sizes (few records in a given spatial unit) often artificially inflates confidence in estimates of species richness, thereby creating artifactual completeness values (Soberón *et al.*, 2000; Chao *et al.*, 2009; Sousa-Baena *et al.*, 2014a). To reduce this effect, we estimated species richness for only those spatial units above a minimum threshold sample size, which we defined as 50% redundancy. Sampling redundancy is the mean number of occurrence records per species, per spatial unit (Garcillán *et al.,* 2003; González-Orozco *et al.*, 2014), therefore 50% redundancy equates to an average of two records per species.

The Chao 1 estimator was used because it is one of the most accurate non-parametric estimators across landscapes with varying biophysical conditions, and is more appropriate with the kind of dataset used in this study (i.e. presence-only records) (Brose *et al.*, 2003; Hortal *et al.*, 2006; Soria-Auza & Kessler, 2008; Schmidt-Lebuhn *et al.*, 2012; Ballesteros-Mejia *et al.*, 2013).

For a given spatial unit *i* (bioregion or grid cell) this statistic ($S_{\mathrm{est(i)}}$) can be calculated as

$$S_{\mathrm{est(i)}} = S_{\mathrm{obs(i)}} + (f_1^{\,2} / 2f_2)$$

where, $S_{\mathrm{obs}(i)}$ = observed species richness in spatial unit *i*, and f_1 and f_2 are the number of singletons (species represented by a single occurrence record) and doubletons (species represented by only two occurrence records), respectively, found in *i*. The completeness index (*C*) was then calculated as

$$C = S_{\mathrm{obs(i)}}/S_{\mathrm{est(i)}}$$

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

We also assessed the relationship between *SAI* and C-index, as inventories may be influenced by allocation of sampling effort (Soberón *et al.*, 2007). We calculated the Pearson correlation coefficient for the relationship. Significance was tested with a permutation test of 999 permutations to account for inflation of the correlation coefficient due to spatial autocorrelation. We conducted ANOVAs to assess whether *C* differs between biomes using bioregions as replicates, where number of replicates was 82

(iii) Assessing the representativeness of floristic composition of AVH data: comparisons with vegetation survey data

Occurrence data derived from herbarium collections is used widely in ecological and conservation applications, such as for generating models of species distribution. However, it remains unclear how well this type of occurrence data approximates known species richness at the site-level. For instance, knowledge of the representativeness of the AVH may be important for understanding the composition of vegetation in a national park or at a bush regeneration site, and may help inform conservation practice. In addition, it has been documented that collectors often focus their attention on species of their interest rather than taking a representative sample – a phenomenon known as the ‘botanist effect’(Ahrends *et al.*, 2011). Therefore, it is necessary to assess the representativeness of herbarium-derived data in terms of the floristic composition of sites, particularly while using these data for applied conservation purposes (Kricsfalusy & Trevisan, 2014).

Our goal was to determine the spatial resolution at which AVH data most closely approximates known floristic composition at a local spatial scale. The YETI 3.2 database, held by the New South Wales (NSW) Office of Environment and Heritage (www.environment.nsw.gov.au/research/VISplot), contains data from more than 50,000 native vegetation surveys conducted within 0.04 ha plots across the state. A portion of these data, representing remnant patches of native vegetation in the Hunter Valley region of NSW, had previously been obtained by Letten *et al.* (2013) who subsetted the data to identify plots that met the following criteria: (1) occurred in areas classified as ‘native vegetation’; (2) contained an inventory of vascular plants made by botanists in the field; (3) were of a standard size (0.04 ha); and (4) the location of the plot was georeferenced with latitude and longitude coordinates. In total, Letten *et al.* (2013) identified 2,490 plots with inventories spanning the time frame 1998 to 2010, and which contained 2,889 vascular plant species from 189 families.

In order to compare the two sources of species records, we deliberately focused on YETI survey plots with high species richness, which we defined as ≥ 30 species, and whose centres were at least 10 km from each other. This resulted in a subset of 192 plots. Plot locations were mapped in ArcGIS v. 10. A series of non-overlapping buffers of increasing size were created, based on the centre of each plot: 0.4, 4, 8, 16, 32, 64, 128, 256, and 512 ha. For each buffer we determined species composition and calculated richness based on the occurrence records from the AVH. Richness and composition were then compared to the known floristic composition of the YETI survey plots. The buffer approach was used to assess the spatial distance at which the AVH data best approximates known richness and composition in a plot. Representativeness of AVH data within a given buffer was calculated as

$$R = S / V$$

where, S = number of shared species listed in both the YETI and AVH dataset and V = number of species in the associated YETI plot.

RESULTS

Characteristics of Australia's Virtual Herbarium (AVH) data

Of the 85 IBRA (Interim Biogeographic Regionalisation of Australia) bioregions, the Sydney Basin has the highest number of occurrence records within the AVH (886,447). The Southern Eastern Queensland bioregion has the most species reported (4,244) while the highest number of families have been reported from the Wet Tropics (250). In contrast, the Gibson Desert has the fewest occurrence records (3,529), species (619) and families (64) reported than any other bioregion.

For each of the seven biomes, we calculated the mean number of occurrence records per bioregion. On average, fewest occurrence records were collected from bioregions in desert shrublands (mean per bioregion = 42,489; no. of bioregions = 21), even though this biome occupies the greatest spatial extent (Fig. 3, S1). This was followed by tropical savannas (mean per bioregion = 48,846; no. of bioregions = 23). In contrast, the greatest number of records were collected from bioregions within the temperate forest biome (mean per bioregion = 3,469,127; no. of bioregions = 21). Bioregions within this biome contained 47% of the total occurrence records despite occupying a relatively small spatial extent. Average species richness and number of families was highest in tropical forest bioregions (3,167 and 250, respectively).

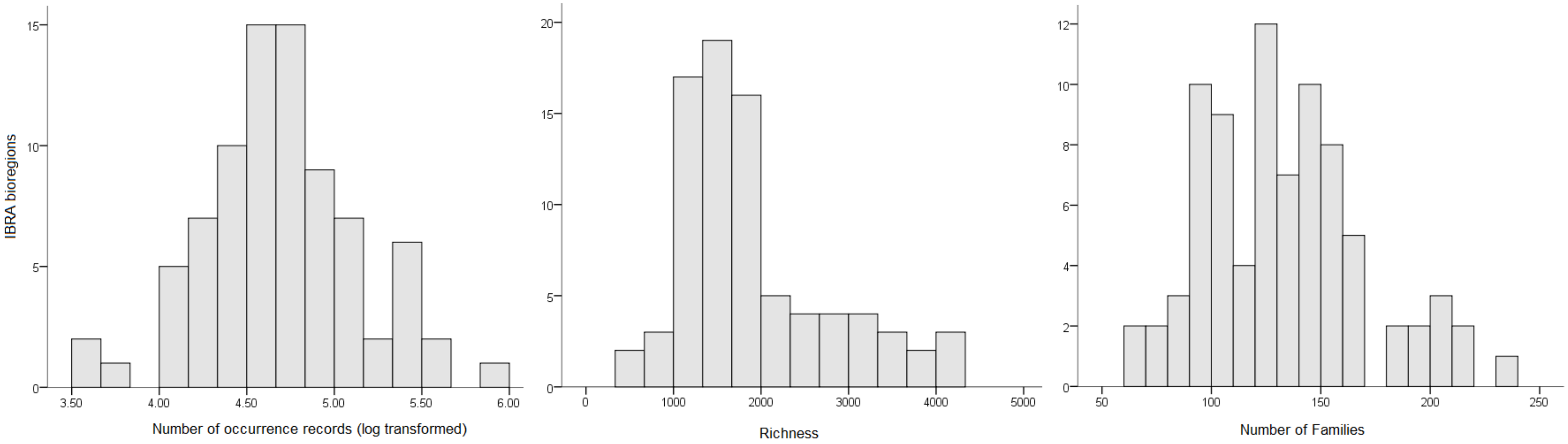

Figure 2 Frequency distribution of the number of occurrence records of native Australian plants, number of species (richness) and number of families, held within Australia's Virtual Herbarium, across 85 IBRA bioregions.

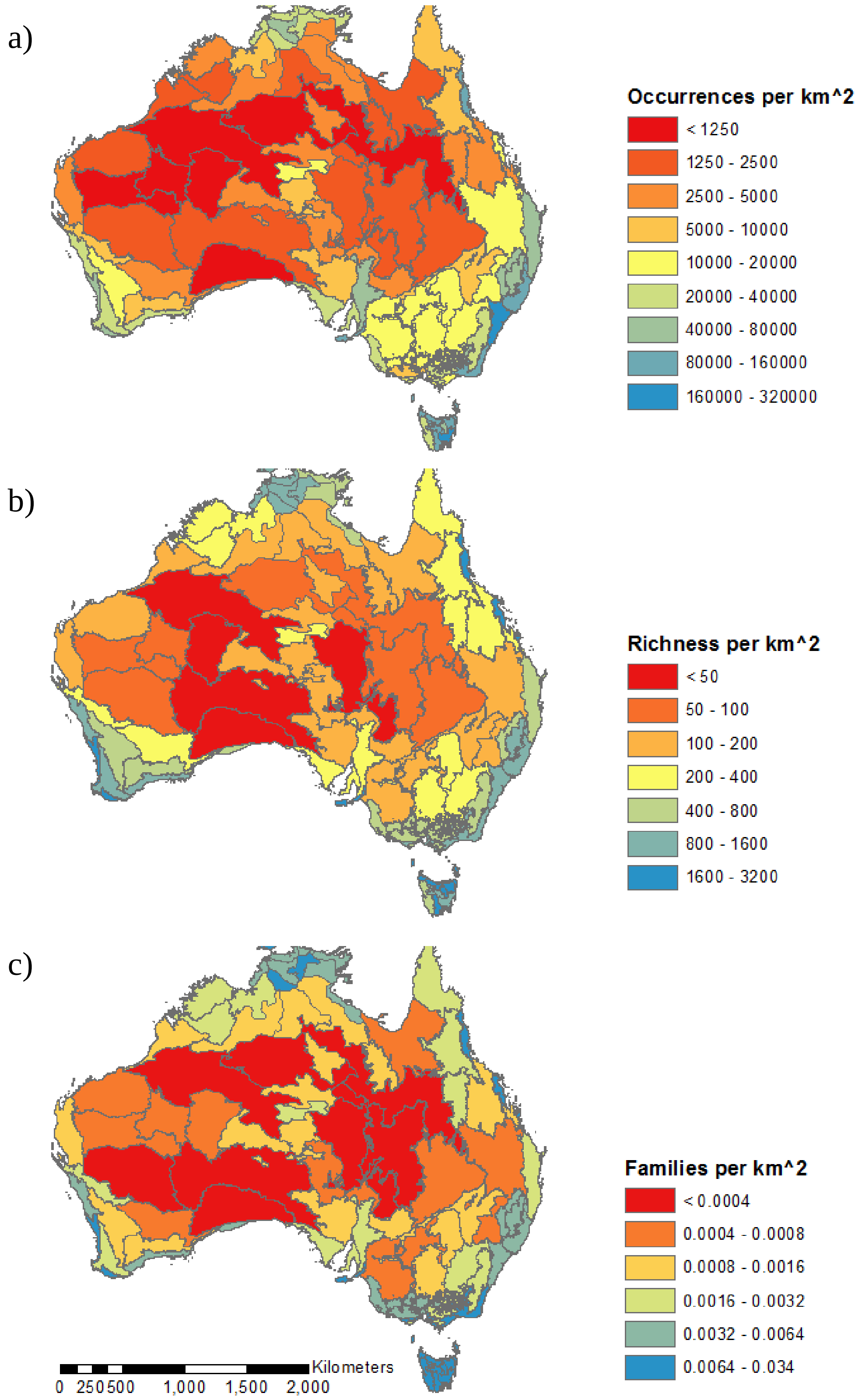

Figure 3 Spatial distribution of: (a) occurrence records, (b) species richness, and (c) no. of families based on data of native Australian flora held in Australia’s Virtual Herbarium. Data are plotted for each of 85 IBRA bioregions and are standardised for area.

(i) Sampling allocation:

We calculated the sampling allocation index (*SAI*) for bioregions and three equal area grids (coarse, medium and fine) to assess whether bioregions or equal area grids had been sampled with the same intensity, with respect to their sizes. Our analysis revealed that *SAI* differs significantly between biomes (ANOVA: $F = 6.13$; $p = 0.001$) (Fig. 4). Of the 85 bioregions, 42 (49.4 %) were under-allocated (*SAI* < 1) with these being mostly located in desert shrublands (19 bioregions) and tropical savannas (16 bioregions) (Fig. 5; see detail in S1). Those bioregions that were most poorly allocated (*SAI* ≤ 0.10) included the tropical savannah bioregion, Mitchell Grass Downs, and six desert shrubland bioregions: Gibson Desert, Little Sandy Desert, Great Sandy Desert, Gascoyne, Nullabor, and Tanami. In contrast, sampling allocation was much higher (*SAI* > 1) than expected for most bioregions in temperate forests, particularly the Sydney Basin (*SAI* = 25), and mediterranean woodlands. Bioregions located in the tropical forests (Wet Tropics and Central Mackay) and montane grasslands (Australian Alps) were also over-allocated (*SAI* > 1). In the temperate savannas, over-allocation was observed in the Riverina bioregion.

Analysis of *SAI* by grid cells revealed that as spatial resolution increased from coarse to fine, variance increased, and heavily over-allocated or under-allocated areas become more apparent across the biomes (Fig. 5 b-d). At the fine scale, ~ 87 % (10,695 out of 12,294 cells) of cells were found under allocated, while under-allocation occurred in 82 % (2,523 out of 3,077 cells) and 76 % (585 out of 770 cells) of cells at medium and coarse resolutions, respectively. Areas heavily over-allocated predominantly include the temperate forests of south-eastern Australia, and, to a lesser extent, south west Western Australia and the very centre of the continent. In contrast poorly allocated areas occur across much of the interior, which consists mostly of desert shrublands and tropical savannas biomes.

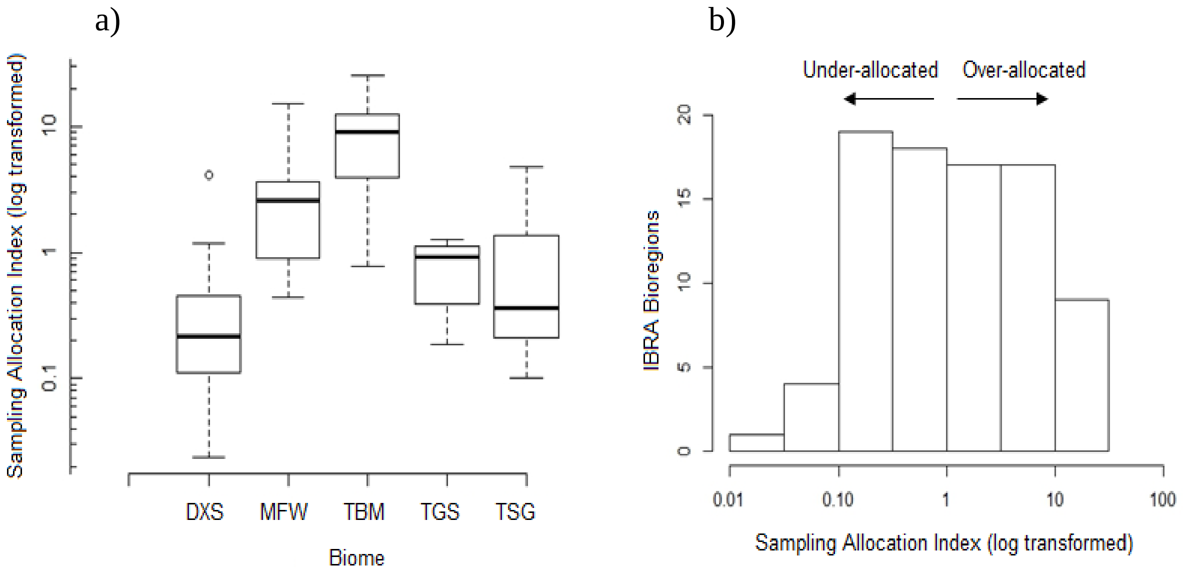

Figure 4 (a) Box-and-whisker plots showing the distribution of Sampling Allocation Index (*SAI*, ratio of observed records to expected records) values across 82 IBRA (Interim Biogeographic Regionalisation of Australia) bioregions within five terrestrial biomes. *SAI* Values < 1 indicate an under-allocation of sampling intensity while values > 1 indicate over-allocation. Biomes are: DXS, Deserts and xeric shrublands; MFW, Mediterranean forests, woodlands and shrublands; TBM, Temperate broadleaf and mixed forests; TGS, Temperate grasslands, savannas and shrublands; TSG , Tropical and subtropical grasslands, savannas and shrublands. The two biomes, TSM (Tropical and subtropical moist broadleaf forests) and MGS (Montane grasslands and shrublands) are not shown in this figure, as they contain only two and one bioregion(s), respectively. (b) Frequency distribution of *SAI* values across the 85 IBRA bioregions of Australia. In both figures, data have been log transformed.

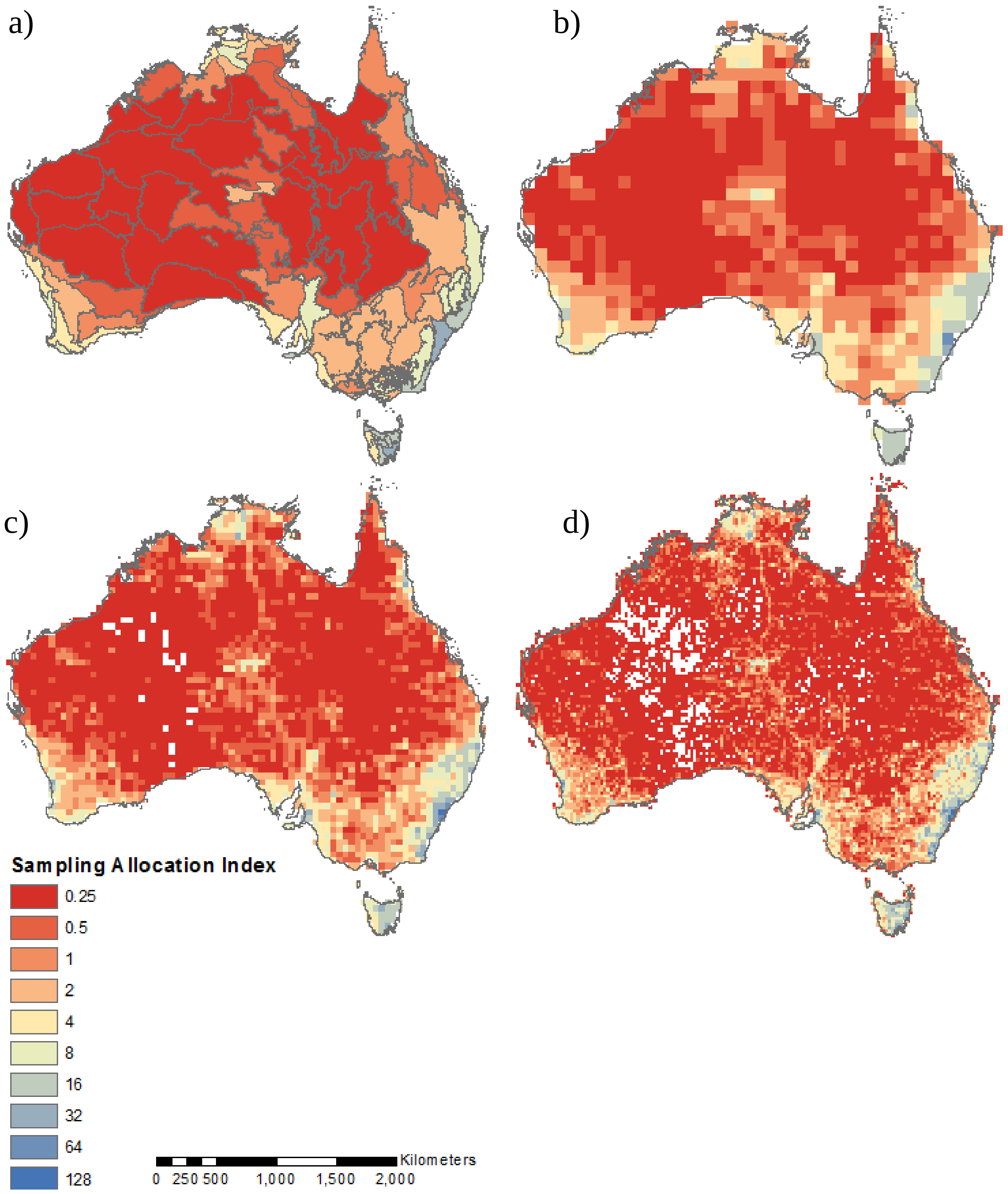

Figure 5 Spatial distribution of the Sampling Allocation Index (*SAI*) (ratio of observed records to expected records) for (a) 85 IBRA bioregions and for three spatial resolutions: (b) coarse (100 ×100 km), (c) medium (50 × 50 km), (d) fine (25×25 km). Values below 1 indicate an under-allocation of sampling and values over 1 indicates over-allocation of sampling. Note that white areas within the continental margins of Australia (maps (c) and (d)) indicate regions for which there are no occurrence records in the AVH.

(ii) Inventory completeness

We assessed inventory completeness or *C*-index (ratio of observed to estimated species richness) at four spatial scales: bioregional level and at three equal area grids (coarse, medium and fine).

Analysis of *C*-index by biomes, indicated that significant differences occur (ANOVA: *F* = 3.45; *p* =0.012). Bioregions in temperate grasslands had substantially lower *C* values (median *C* = 0.80) than other biomes. However, for most bioregions (85 %), *C* values were > 0.80 (Fig 6). Individual bioregions with the highest values of completeness were mostly in tropical savannas (*C* > 0.90), while those with the lowest completeness (*C* < 0.80) were in desert shrublands or Mediterranean forests.

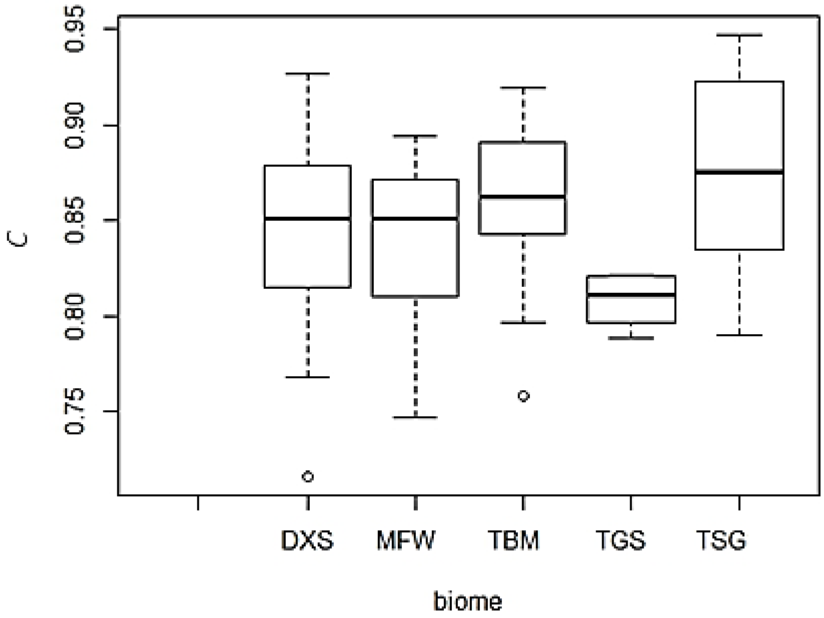

Figure 6 Box-and-whisker plots of index of inventory completeness (*C*) values for 82 IBRA (Interim Biogeographic Regionalisation of Australia) bioregions within their corresponding biome. DXS= Deserts and xeric shrublands; MFW = Mediterranean forests, woodlands and shrublands; TBM = Temperate broadleaf and mixed forests; TGS = Temperate grasslands, savannas and shrublands; TSG = Tropical and subtropical grasslands, savannas and shrublands. The two biomes, TSM = Tropical and subtropical moist broadleaf forests and Montane grasslands and shrublands, are not shown in the figure as they contain two and one bioregion(s), respectively.

As resolution changed from coarse to fine scale, the mean and variance of the C-index decreased (Fig. 7 b-d). Furthermore, the number of grid cells which were too poorly sampled to enable completeness to be estimated (i.e. sampling redundancy < 50%) increased with increasing spatial resolution (Fig. 8). These cells were primarily located within the desert shrublands biome, particularly in Western Australia, and in the tropical savannas of Queensland. At the fine resolution, completeness values could not be calculated for 62% of cells (7,623 out of 12,294 cells) due to low redundancy. This proportion decreased to 39% (1,158 out of 3,077) and 22% (94 out of 770) of cells at medium and coarse-scale resolutions.

Although completeness became more evenly distributed spatially as grid cell size increased, high values of *C* (e.g., *C* > 0.80) were mostly restricted to a few scattered sites. These locations included desert shrublands and tropical savannas regions of the Northern Territory (e.g. McDonald Ranges, Tanami, Gawler, Davenport Murchinson Ranges, Pine Creek, Arnhem Coast, Arnhem Plateau, Victoria Bonaparte, Darwin Coastal, Daly Basin, Gulf Fall and Uplands bioregions) as well as bioregions within the temperate forest biome in Tasmania.

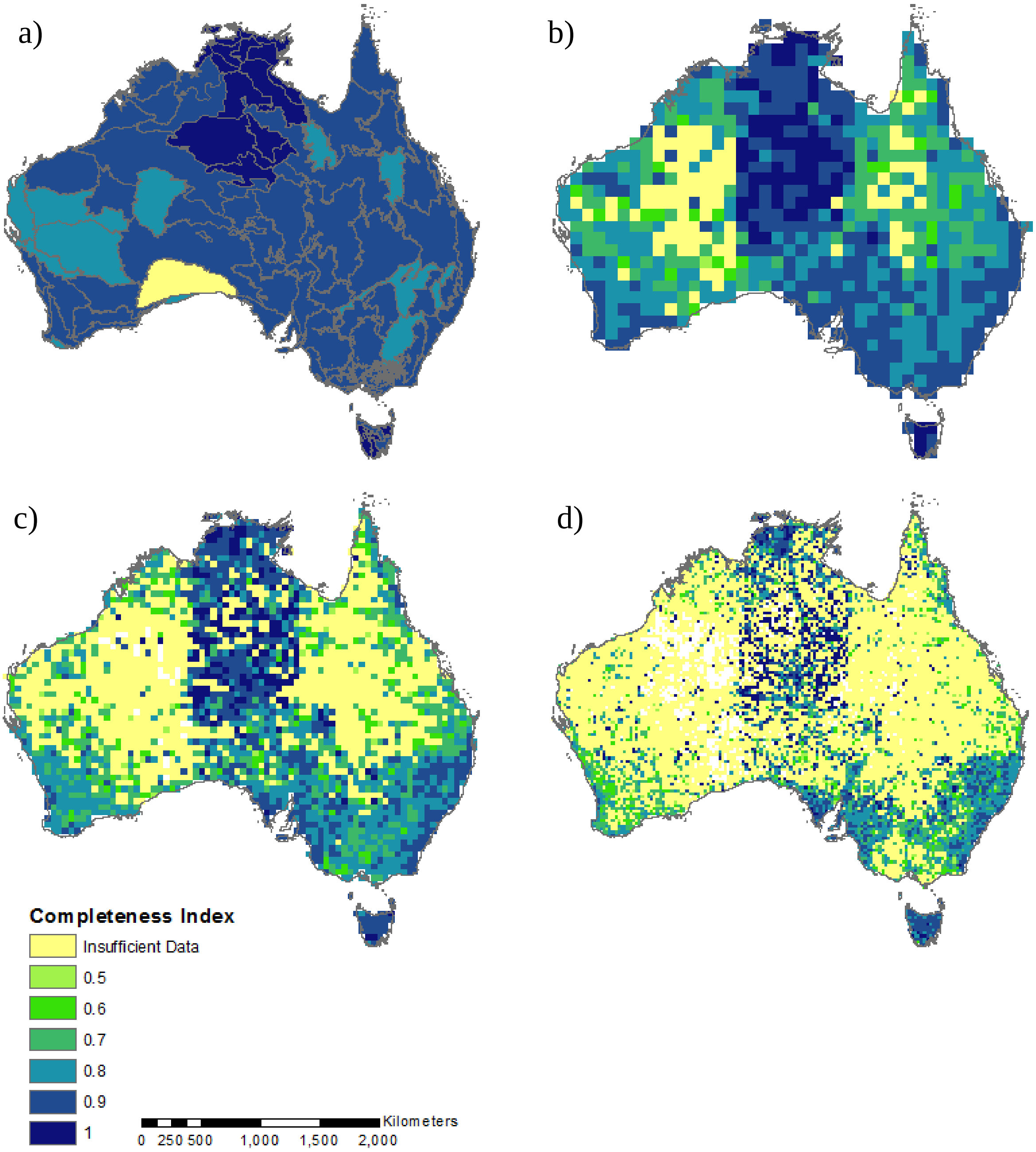

Figure 7 Spatial distribution of the completeness index, *C*, of occurrence records in Australia's Virtual Herbarium (based on Chao1 estimator of species richness) for a) 85 IBRA bioregions and for three spatial resolutions: (b) coarse (100 × 100 km), (c) medium (50 × 50 km), (d) fine (25 × 25 km). Note that white areas within the continental margins of Australia (maps (c) and (d)) indicate regions for which there are no occurrence records in the AVH.

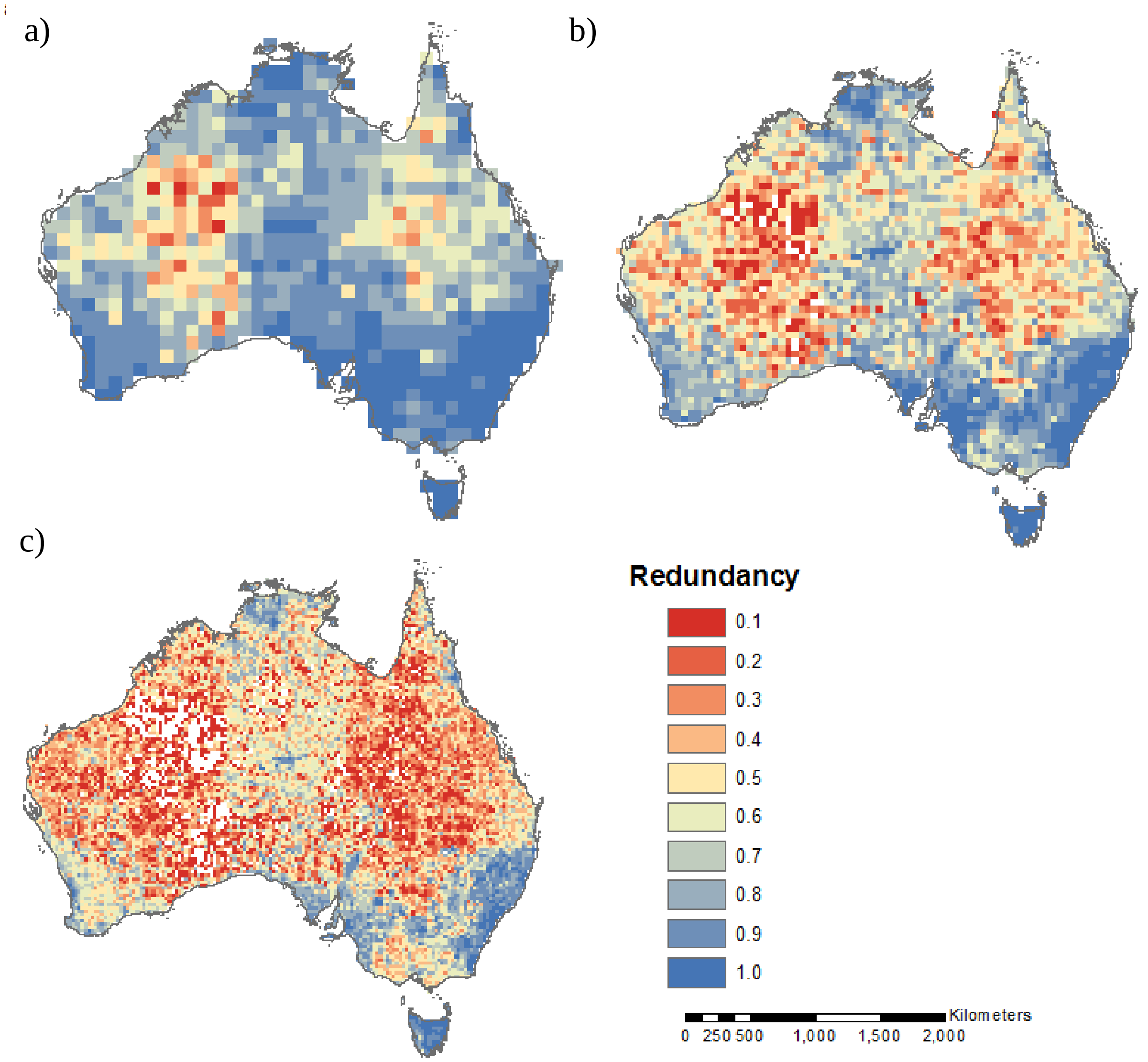

Figure 8 Spatial distribution of sampling redundancy within Australia's Virtual Herbarium for three spatial resolutions: (a) coarse (100 × 100 km), (b) medium (50 × 50 km) and (c) fine (25 × 25 km). Values approaching 1 indicate high redundancy while values close to 0 indicate low redundancy. Note that white areas within the continental margins of Australia (maps (b) and (c)) indicate regions for which there are no occurrence records in the AVH.

Relationship between sampling allocation and sampling completeness

We found a significant correlation between sampling allocation index (log-transformed) and completeness (*p* = 0.004, *r* = 0.30) at the bioregional level (Fig. 9). That is, bioregions with a high allocation of sampling tended to have high completeness values. Correlation tests were not undertaken at other spatial resolutions because excluding cells with missing completeness values would substantially skew the relationship.

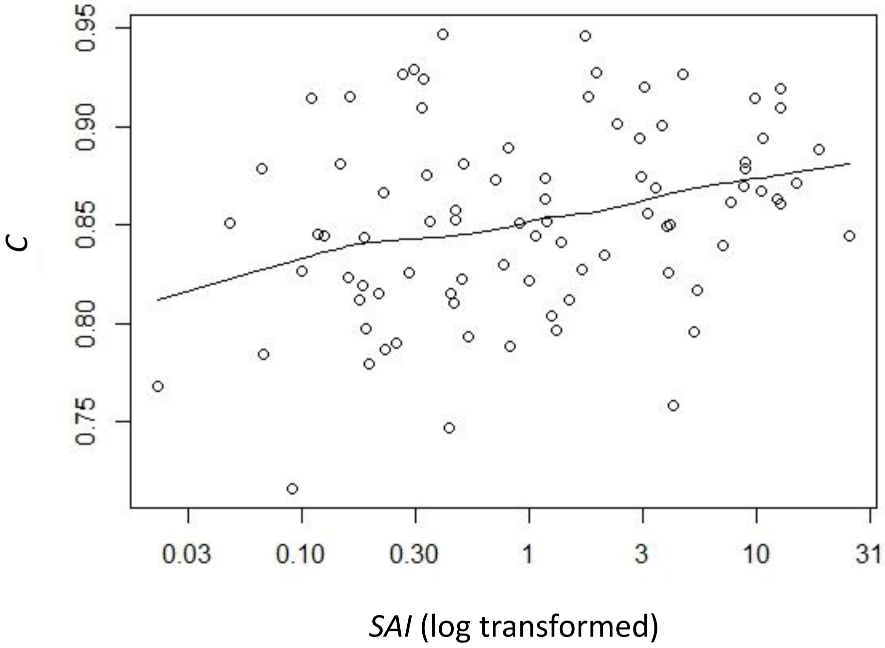

Figure 9 Correlation between C-index and log-transformed Sampling Allocation Index (*SAI*) values at the bioregional level, based on native species occurrence data in Australia's Virtual Herbarium. Line represents a locally weighted regression.

(iii) Representativeness of floristic composition of AVH data

Representativeness of floristic composition increased with increasing buffer size (Fig. 10, see detail in S2). At the smallest buffer size (0.04 ha), AVH occurrence records were available for less than 1 % of the species recorded within the 192 vegetation plots. This increased to 2.3 and 6.4% for 4 and 8 ha buffers, respectively. On average, a 16 ha buffer surrounding vegetation plots was required to achieve representativeness of 90%: further increases in buffer size did not result in higher representativeness.

.

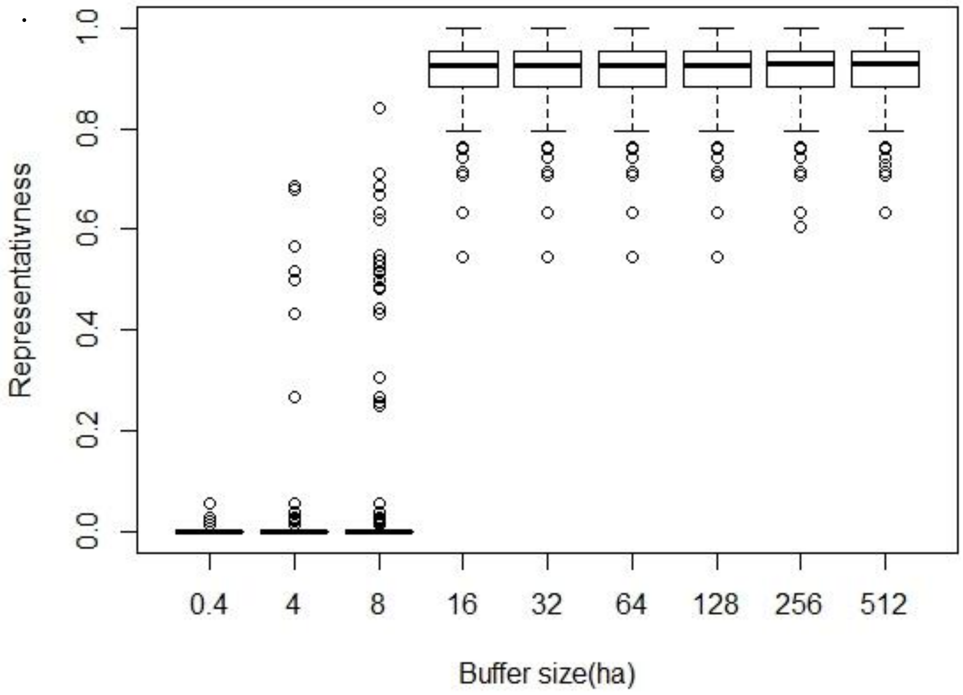

Figure 10 Similarity in composition (representativeness) of native flora recorded in 0.04 ha vegetation plots with data extracted from Australia's Virtual Herbarium at different spatial scales (based on circular buffers centred on the plots).

DISCUSSION

Non-randomness in spatial patterns across diverse spatial scales

Exploring non-randomness in natural history collections is of paramount importance to understanding spatial biases in sampling effort, and to identify geographic locations that have been under-sampled (Wieringa *et al.*, 2004). We assessed spatial patterns among occurrence records within Australia's Virtual Herbarium (AVH), at various spatial scales. Our analysis of sampling allocation index (*SAI*) revealed an overwhelming geographic bias in sampling effort at the bioregional level (Fig. 5a). Bioregions within the temperate forest biome were substantially more heavily sampled than other biomes, a pattern that was consistent at finer spatial scales. Historically, non-random patterns in specimen collections are often driven by human settlement and accessibility (Reddy & Dávalos, 2003; Aikio et al., 2010) and the temperate forest regions are the most densely populated and urbanized part of the continent (http://www.abs.gov.au). Despite covering a broader spatial extent, the desert shrublands biome has been very poorly sampled, particularly the interior of the arid zone. This area is generally inaccessible, sparsely populated, and remains relatively unexplored, a pattern that is consistent with other arid regions across the world (Newbold, 2010).

At finer resolutions, local pattern in sampling allocation become apparent across the biome (Fig. 5 c-d). For example the largest urban region in central Australia, Alice Springs, has been heavily sampled, in stark contrast to most of the arid zone. Non-randomness in sampling is driven by a number of factors, one of which is the roadmap effect (Hurlbert & Jetz; 2007, Kadmon et al., 2004, 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 from roads. Nonrandomness in collecting intensity may also be driven by the presence of hotspots of biodiversity or protected areas (Dennis & Thomas, 2000). For example, the south west corner (mediterranean forest biome), the northern top (tropical savannas), and north-east (tropical forests) of Australia found over allocated (*SAI* > 1) are the global hot spots, rich in endemism and taxonomic diversity (Myers et al., 2000; Crisp et al., 2001).

Inventory completeness: Influence of spatial scales and non-randomness in sampling effort

We assessed inventory completeness of AVH data based on the Chao1 estimator (Colwell & Coddington, 1994) at multiple scales and found that inventory completeness (*C*-index) is very scale dependent and is substantially biased by non-randomness in sampling effort. Though values of the *C*-index differed significantly between biomes, at the bioregion scale completeness was high. For most bioregions *C* values were above 0.80, a threshold often regarded as a well-collected sample (Soberón *et al.*, 2007; Mora *et al.*, 2008; Schmidt-Lebuhn *et al.*, 2012).

However, high *C* values may result from sampling artefacts. Extrapolation of species richness at a coarse scale often produces spurious species densities even in poorly known areas, where overestimation may result from a few well sampled areas (Rahbek, 2005; Sousa-Baena *et al.*, 2014a). For example, in the desert shrublands, we found that the Macdonald Ranges bioregion is very well sampled, while the adjacent bioregions remain very poorly sampled (Fig. 7a, see detail in S1).

We assessed inventory completeness of the AVH data across three equal area grids. With the change of resolution (from coarse to fine) greater variation in *C* values became apparent, indicating that completeness may decline at finer resolutions (Fig. 7d). At a coarse scale (100 km × 100 km) *C* values were found to be more evenly distributed and with more well sampled cells ($C > 0.80$) indicating that species in AVH data are well known at a coarser resolution (Fig. 7b). However, we found this database to be particularly vulnerable below a resolution of 100 km. At medium (50 × 50 km) and fine (25 × 25 km) scales, completeness values could only be calculated for 39% and 22% cells, respectively, indicating that knowledge of native species are insufficiently understood in AVH data with finer resolutions. These poorly sampled areas are mostly found in desert shrublands, tropical savannas and temperate savannas biomes for which no reasonable estimate based on the Chao 1 estimator could be given to calculate realistic *C* value.

The Chao 1 estimator is known to perform poorly in degenerated cells where non-parametric estimators are sensitive to cells with low sample coverage (Brose *et al.,* 2003). However, this estimator can be advantageous in that it precludes a false estimation of completeness values. Schmidt-Lebuhn *et al.* (2012) found the Chao 1 estimator to be more appropriate at the continental scale while assessing collection effort of specimen data on Asteraceae plants in Australia.

We also found a strong correlation between inventory completeness and the sampling allocation index (*SAI*) (Fig. 9). Within a biome, bioregions with higher *SAI* tend to have higher levels of completeness. These patterns are also consistent across the three equal area grids. The geographic patchiness in inventory completeness indicates that species richness within the AVH data base is geographically biased (Hortal *et al.*, 2007).

Although bioregions within the eastern temperate and southwest mediterranean forest biomes have been heavily sampled, the degree of completeness tends to be low, and declines from coarse to fine resolution (Fig 7 b-d). Conversely, Australia's Northern Territory, which is mostly desert shrublands and tropical savannas, is more thoroughly and evenly sampled regardless of resolution, and there is an obvious decline in completeness in adjacent regions.

Unevenness and spatial bias in sampling effort may affect the perception of species diversity (Romo *et al.*, 2006; Soria-Auza & Kessler, 2008). According to Gotelli & Graves (1994) greater evenness of individuals among species will lead to species being detected more quickly compared to a community with a long tail of rare species. Hence, the rarefaction curve will rise more quickly to an asymptote. In Australia, the south west mediterranean forest is a global hot spot of biodiversity and collectors often targeting endemic species while ignoring more common ones (Nelson *et al.*, 2013). Such unevenness in inventory completeness suggests that using this database to discern spatial patterns, especially in describing true diversity patterns of Australian flora, should be done with caution since large scale diversity patterns are built upon species data recorded at final resolution (Mora *et al*., 2008).

As the amount of data available online has increased so too has concern about completeness, quality and, bias of biodiversity data (Yang *et al.*, 2013; Beck *et al.*, 2014; Sousa-Baena *et al.*, 2014b). Our results emphasized the importance to assess the spatial quality and extent of the primary database at diverse spatial scales before applying them for practical conservation purposes and understanding biodiversity patterns. Proper assessments are essential since many ecological processes and biodiversity patterns are scale-dependent (Whittaker *et al.*, 2001). Inappropriate choice of resolution may significantly influence decision makers as to which areas are prioritized (Whittaker *et al.*, 2005) as seen where hotspots have been misidentified due to the choice of spatial scale (Hartley & Kunin, 2003; Hurlbert & Jetz, 2007).

Representativeness of species composition of AVH data

The capacity to identify species accurately in a particular area or habitat is important for planning, implementing and monitoring conservation activities (Paton, 2009), but taxonomic representativeness is the most difficult bias to identify in herbarium collections (Graham *et al.*, 2004). Both parametric and non-parametric models have been developed to extrapolate species richness, but these models cannot identify individual species (Garcillán & Ezcurra, 2011).

To determine how representative its occurrence records are, we compared AVH data to known species composition reported across a series of 192 survey plots, 0.04 ha in size (YETI database). The plots were situated within temperate forests of the Hunter Valley region of New South Wales, an area with high values of the *C*-index. Our analysis indicated that AVH data extracted from a 16 ha buffer surrounding the vegetation plots is required to achieve representativeness of 90% (Fig. 10), that is, for the AVH to contain occurrence records for 90% of species reported in the survey plot.

This result cannot necessarily be extrapolated to other regions, as we would expect the representativeness of AVH data to vary due to spatial variation in the *C*-index and *SAI*. For example, in the desert shrublands values of the C-index were typically very poor: to achieve high levels of representativeness of AVH data with YETI survey plots in this region would likely require buffer sizes larger than 16 ha. The method developed in this study can be an effective way to assess the representativeness of species composition of herbarium data at a local site level. We note, however, that the application of such a method depends on the availability and quality of representative vegetation survey plot data.

Implication for conservation priorities

Virtual primary data on plants are increasingly being used in setting conservation priorities (Wulff *et al.*, 2013; Kricsfalusy & Trevisan, 2014; Mokany *et al.,* 2014). Inherent spatial biases, haphazard collections and lack of fine scale data are the major impediment in identifying and conserving areas with high priorities such as hotspots, protected areas and reserves (Brooks *et al.*, 2006; Grand *et al.*, 2007). Identification of these inherent limits of the primary data can improve decision making in prioritizing areas, and hence guide efforts to collect additional data (Funk & Richardson, 2002; Reddy & Dávalos, 2003; Sousa-Baena *et al.*, 2014b). Future conservation priorities can take these limitations into account for managing and prioritising actions.

We have found that within a given biome few bioregions are well sampled. A particularly striking example is the desert shrublands biome where most of the bioregions are very poorly sampled, except for a few located in the central range (Fig. 7). The anomalously high richness of these areas, compared to surrounding regions, represents a sampling artefact rather than a real biogeographic pattern. Lack of knowledge regarding these sampling biases may lead to incorrect assumptions and poor decision making (Nelson *et. al.*1990). A more detailed assessment is required to prioritise these areas for conservation.

We have also found that the AVH database is more useful for assessing biodiversity at coarse resolutions. This presents difficulties, as prioritizing areas to protect often requires data at a finer scale (Ferrier, 2002; Bombi *et al.*, 2012). Nevertheless assesments at a finer resolution are also important to identify knowledge gaps for species at risk of extinction (Hartley & Kunin, 2003). Australia is one of the most biodiverse continents and rich in enedemic flora (Crisp *et al*., 2001) where many plants are threatened due to climate change (Keith *et al.*, 2014). Accurate spatial knowledge of species occurrences is necessary for effectively conserving Australia's native flora.Our method to asses the representativness of species composition may also help in conservation practice especially if AVH data is used to inform the selection and design of protected areas at local level (e.g national park).

Improving the spatial knowledge gap in AVH

Comprehensive and reliable information on species occurrences is required to conduct effective research and implement better conservation strategies (Graham *et al.*, 2004; Pyke & Ehrlich, 2010). From the continental perspective it may be an overambitious goal to conduct comprehensive surveys across this sparsely populated continent, as data acquisition is critically dependent on the availability of funds, taxonomic expertise and research facilities, all of which are often limited (Gioia, 2010; Hardisty & Roberts, 2013; Vos *et al.*, 2014).

Minimum strategic sampling effort as well as choice of sampling resolution is important in establishing priority based sampling effort (Hermoso *et al.*, 2014). We propose that grid based completeness summaries be used to identify areas for which spatial knowledge of flora is poorly understood. In better sampled temperate or mediterranean forest bioregions, future sampling effort can be based on finer resolution (25 × 25 km). For poorly sampled areas, such as the mid interior arid zone bioregions and Queensland tropical savannas, coarse (100 ×100 km) or medium (50 × 50 km) resolution surveys can be conducted.

Spatial knowledge gaps within the AVH database can also be improved through best data management practices by reducing geographical co-ordinate errors and by proper identification of specimens to the species level. Approximately 2,000 genera contained specimens which could not be unaccounted for as they were not identified to species level.

Prioritising the digitising of records from poorly represented geographic locations could also rapidly fill the data void in the AVH.

ACKNOWLEDGEMENTS

This research paper first acknowledges the enormous contributions of my principal supervisor, Dr. Linda Beaumont, as well as my Associate-Supervisors Dr. Rachael Gallagher, and Dr. David Nipperess. Also, my special thanks to Andrew D. Letten for providing the YETI vegetation plot data. I would like to thank Toni Mizerek, Benjamin Ofori, Anup Kumar Roy, and Swapan Kumar Sarker for their valuable comments on this research paper.

REFERENCES

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. (2011) Conservation and the botanist effect. *Biological Conservation*, **144**, 131-140.

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.

Balke, M., Schmidt, S., Hausmann, A., Toussaint, E., Bergsten, J., Buffington, M., Häuser, C.L., Kroupa, A., Hagedorn, G. & Riedel, A. (2013) Biodiversity into your hands-A call for a virtual global natural history'metacollection'. *Frontiers in Zoology*, **10**, 55-63.

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.

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.

Beaumont, L.J., Gallagher, R.V., Leishman, M.R., Hughes, L. & Downey, P.O. (2014) How can knowledge of the climate niche inform the weed risk assessment process? A case study of Chrysanthemoides monilifera in Australia. *Diversity and Distributions*, **20**, 613-625.

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.

Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R. & Moritz, C. (2013) Unlocking the vault: next-generation museum population genomics. *Molecular Ecology*, **22**, 6018-32.

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.

Bombi, P., Salvi, D. & Bologna, M.A. (2012) Cross-scale predictions allow the identification of local conservation priorities from atlas data. *Animal Conservation*, **15**, 378-387.

Brooks, T.M., Mittermeier, R.A., da Fonseca, G.A., Gerlach, J., Hoffmann, M., Lamoreux, J.F., Mittermeier, C.G., Pilgrim, J.D. & Rodrigues, A.S. (2006) Global biodiversity conservation priorities. *Science*, **313**, 58-61.

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.

Chao, A., Colwell, R.K., Lin, C.-W. & Gotelli, N.J. (2009) Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, **90**, 1125-1133.

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.

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series 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.

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

Feeley, K.J. (2012) Distributional migrations, expansions, and contractions of tropical plant species as revealed in dated herbarium records. *Global Change Biology*, **18**, 1335-1341.

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

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2000) *Quantitative geography: perspectives on spatial data analysis*. Sage, London.

Funk, V. & Richardson, K. (2002) Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*, **51**, 303-316.

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.

Garcillán, P.P., Ezcurra, E. & Riemann, H. (2003) Distribution and species richness of woody dryland legumes in Baja California, Mexico. *Journal of Vegetation Science*, **14**, 475-486.

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.

Gotelli, N. J. & Graves, G. R. (1994) *Null models in ecology*. Washington, DC: Smithsonian Institution Press.

Graham, C., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. (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.

Guillera-Arroita, G., Ridout, M.S. & Morgan, B.J. (2010) Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, **1**, 131-139.

Hardisty, A. & Roberts, D. (2013) A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology*, **13**, doi: 10.1186/1472-6785-13-16

Hart, R., Salick, J., Ranjitkar, S. & Xu, J. (2014) Herbarium specimens show contrasting phenological responses to Himalayan climate. *Proceedings of the National Academy of Sciences*, **111**, 10615-10619.

Hartley, S. & Kunin, W.E. (2003) Scale dependency of rarity, extinction risk, and conservation priority. *Conservation Biology*, **17**, 1559-1570.

Hermoso, V., Kennard, M.J. & Linke, S. (2014) Evaluating the costs and benefits of systematic data acquisition for conservation assessments. *Ecography*, doi: 10.1111/ecog.00792

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.

Hijmans, R.J. & van Etten, J. (2014) raster: raster: Geographic data analysis and modeling. *R package version*, 2.2-12.

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., 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.

Hortal, J., Roura-Pascual, N., Sanders, N. & Rahbek, C. (2010) Understanding (insect) species distributions across spatial scales. *Ecography*, **33**, 51-53.

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. (2013) Australian seaweed collections: use and misuse. *Phycologia*, **52**, 2-5.

Hurlbert, A.H. & Jetz, W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences*, **104**, 13384-13389.

IBRA (2012) An interim biogeographic regionalisation for Australia, version 7. Australian Government Department of Sustainability, Environment, Water, Population and Communities.Availableat:http://www.environment.gov.au/parks/nrs/science/bioregion framework/ibra/index.html.

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

Keith, D.A., Mahony, M., Hines, H., Elith, J., Regan, T.J., Baumgartner, J.B., Hunter, D., Heard, G.W., Mitchell, N.J. & Parris, K.M. (2014) Detecting Extinction Risk from Climate Change by IUCN Red List Criteria. *Conservation Biology*, **28**, 810-819.

Kricsfalusy, V.V. & Trevisan, N. (2014) Prioritizing regionally rare plant species for conservation using herbarium data. *Biodiversity and Conservation*, **23**, 39-61.

Krishtalka, L. & Humphrey, P.S. (2000) Can natural history museums capture the future? *BioScience*, **50**, 611-617.

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.

Lane, M.A. (1996) Roles of natural history collections. *Annals of the Missouri Botanical Garden*, **83**, 536-545.

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**

Letten, A.D., Ashcroft, M.B., Keith, D.A., Gollan, J.R. & Ramp, D. (2013) The importance of temporal climate variability for spatial patterns in plant diversity. *Ecography*, **36**, 1341-1349.

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.

Mokany, K., Westcott, D. A., Prasad, S., Ford, A. J. & Metcalfe, D. J. (2014) Identifying Priority Areas for Conservation and Management in Diverse Tropical Forests. *PloS One*, **9,** e89084

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.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853-858.

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.

Nelson, W.A., Dalen, J. & Neill, K.F. (2013) Insights from natural history collections: analysing the New Zealand macroalgal flora using herbarium data. *PhytoKeys*,**30** 1-21.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.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2013) Package 'vegan,'. http://cran.rproject.org/web/packages/vegan/index.htm.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E. & Morrison, J.C. (2001) Terrestrial Ecoregions of the World: A New Map of Life on Earth A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, **51**, 933-938.

Paton, A. (2009) Biodiversity informatics and the plant conservation baseline. *Trends in Plant Science*, **14**, 629-37.

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*, **85**, 247-66.

Rahbek, C. (2005) The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology Letters*, **8**, 224-239.

R Development Core Team. (2013) R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2012. In. ISBN 3-900051-07-0. Available from: http://www. R-project. Org.

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

Rich, T.C. (2006) Floristic changes in vascular plants in the British Isles: geographical and temporal variation in botanical activity 1836–1988. *Botanical Journal of the Linnean Society*, **152**, 303-330.

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.

Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio, (2013) *Applied spatial data analysis with R*, Second edition. Springer, NY. http://www.asdar-book.org

Romo, H., García-Barros, E. & Lobo, J.M. (2006) Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography*, **29**, 873-885.

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.

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.

Sheth, S.N., Lohmann, L.G., Distler, T. & Jiménez, I. (2012) Understanding bias in geographic range size estimates. *Global Ecology and Biogeography*, **21**, 732-742.

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.

Soberón, J., Jiménez, R., Golubov, J. & Koleff, P. (2007) Assessing completeness of biodiversity databases at different spatial scales. *Ecography*, **30**, 152-160.

Soberón, J.M., Llorente, J.B. & Oñate, L. (2000) The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. *Biodiversity & Conservation*, **9**, 1441-1466.

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. (2014a) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions*, **20**, 369-381.

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.

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.

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

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.

Walther, B.A. & Moore, J.L. (2005) The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, **28**, 815-829.

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

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**, 1828-32.

Whittaker, R.J., Willis, K.J. & Field, R. (2001) Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography*, **28**, 453-470.

Whittaker, R.J., Araújo, M.B., Jepson, P., Ladle, R.J., Watson, J.E. & Willis, K.J. (2005) Conservation biogeography: assessment and prospect. *Diversity and Distributions*, **11**, 3-23.

Wieringa, J., Poorter, L., Bongers, F., Kouamé, F. & Hawthorne, W. (2004) Biodiversity hotspots in West Africa: patterns and causes. *Biodiversity of West African forests* (ed. by L. Poorter,F. Bongers, F. Kouame and W.D. Hawthorne), pp. 61–72. CABI Publishing, Oxford.

Wulff, A. S., Hollingsworth, P. M., Ahrends, A., Jaffré, T., Veillon, J.-M., L’huillier, L. & Fogliani, B. (2013) Conservation priorities in a biodiversity hotspot: analysis of narrow endemic plant species in new caledonia. *PLoS One*, **8,** e73371.

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. & Culham, A. (2007) How global is the global biodiversity information facility? *PLoS One*, **2**, e1124.

SUPPLEMENTARY MATERIALS

Table S1 Characteristics of data held in Australia's Virtual Herbarium (AVH), summarised for 85 IBRA bioregions. For each bioregion, the biome within which it occurs is given, as is the number of occurrence records in the AVH, species richness, and number of families. Also given are two measures of spatial bias: *SAI* (sampling allocation index) and C-index (index of inventory completeness). *TSG = Tropical and subtropical grasslands, savannas and shrublands; DXS = Deserts and xeric shrublands; TBM = Temperate broadleaf and mixed forests; MFW = Mediterranean forests, woodlands and shrublands; TGS = Temperate grasslands, savannas and shrublands; TSM = Tropical and subtropical moist broadleaf forests; MGS = Montane grasslands and shrublands.

IBRA Bioregion	**Area (SQ_KM)**	***Biome code**	**Occurrence records**	**Species Richness**	**Families**	***SAI***	**C-index**
Arnhem Coast	33,356	TSG	56,215	1,787	163	1.76	0.95
Arnhem Plateau	23,060	TSG	43,385	1,721	158	1.96	0.93
Australian Alps	12,330	MGS	83,799	1,686	137	7.1	0.84
Avon Wheatbelt	95,171	MFW	107,121	4,210	125	1.18	0.87
Brigalow Belt North	136,745	TSG	39,110	2,992	199	0.3	0.83
Brigalow Belt South	272,198	TSG	277,193	4,082	203	1.06	0.84
Ben Lomond	6,575	TBM	79,518	1,342	142	12.63	0.92
Broken Hill Complex	56,354	DXS	24,348	1,014	85	0.45	0.82
Burt Plain	73,797	DXS	24,024	1,128	94	0.34	0.91
Carnarvon	84,302	DXS	18,895	1,476	103	0.23	0.79
Central Arnhem	34,624	TSG	10,290	1,174	138	0.31	0.93
Central Kimberley	76,756	TSG	13,230	1,387	126	0.18	0.81
Central Ranges	101,640	DXS	35,438	1,199	89	0.36	0.85
Channel Country	304,094	DXS	47,082	1,465	101	0.16	0.82
Central Mackay Coast	14,642	TSM	21,073	2,265	203	1.5	0.81
Coolgardie	129,122	MFW	57,938	2,605	98	0.47	0.81
Cobar Peneplain	73,853	TGS	71,290	1,504	117	1.01	0.82
Cape York Peninsula	122,565	TSG	83,578	3,118	209	0.71	0.87
Daly Basin	20,922	TSG	36,536	1,483	137	1.82	0.92
Darwin Coastal	28,432	TSG	66,509	1,989	163	2.44	0.9
Dampierland	83,609	TSG	15,120	1,351	125	0.19	0.84
Desert Uplands	69,411	TSG	17,364	1,588	130	0.26	0.79
Davenport Murchison Ranges	58,051	DXS	15,539	991	86	0.28	0.93
Darling Riverine Plains	106,998	TGS	84,523	1,759	116	0.82	0.79

Einasleigh Uplands	116,257	TSG	56,500	3,375	217	0.51	0.82
Esperance Plains	29,213	MFW	85,048	3,318	129	3.04	0.89
Eyre Yorke Block	61,204	MFW	183,496	1,779	109	3.13	0.87
Finke	72,674	DXS	32,973	1,083	92	0.47	0.86
Flinders Lofty Block	66,158	DXS	259,855	2,487	141	4.1	0.83
Furneaux	5,375	TBM	65,504	1,429	153	12.72	0.86
Gascoyne	180,753	DXS	11,809	1,354	93	0.07	0.78
Gawler	120,029	DXS	58,839	1,510	102	0.51	0.88
Geraldton Sandplains	31,421	MFW	64,959	2,949	123	2.16	0.83
Gulf Fall and Uplands	118,479	TSG	39,047	1,851	133	0.34	0.92
Gibson Desert	156,289	DXS	3,529	619	64	0.02	0.77
Great Sandy Desert	394,861	DXS	25,435	1,459	97	0.07	0.88
Gulf Coastal	27,117	TSG	10,868	1,145	127	0.42	0.95
Gulf Plains	220,418	TSG	25,033	2,079	162	0.12	0.85
Great Victoria Desert	422,466	DXS	50,799	1,647	90	0.13	0.84
Hampton	10,882	MFW	4,624	451	68	0.44	0.75
Jarrah Forest	45,091	MFW	142,185	3,768	144	3.29	0.86
Kanmantoo	8,124	MFW	117,001	1,664	124	15.04	0.87
King	4,256	TBM	36,606	1,067	144	8.98	0.88
Little Sandy Desert	110,899	DXS	5,138	876	75	0.05	0.85
MacDonnell Ranges	39,294	DXS	44,446	1,351	112	1.18	0.86
Mallee	73,976	MFW	64,117	3,223	103	0.91	0.85
Murray Darling Depression	199,584	MFW	227,551	2,453	139	1.19	0.85
Mitchell Grass Downs	334,688	TSG	32,123	1,790	121	0.1	0.83
Mount Isa Inlier	67,783	TSG	12,422	1,132	103	0.19	0.8
Mulga Lands	251,883	TGS	45,094	1,720	118	0.19	0.82
Murchison	281,206	DXS	53,278	2,162	100	0.2	0.78
Nandewar	27,020	TBM	111,294	1,995	151	4.3	0.76
Naracoorte Coastal Plain	24,582	MFW	85,021	1,646	128	3.61	0.87
New England Tablelands	30,022	TBM	222,407	2,760	183	7.74	0.86
NSW North Coast	39,966	TBM	403,857	3,408	213	10.55	0.87
Northern Kimberley	84,201	TSG	28,880	1,839	154	0.36	0.88
NSW South Western Slopes	86,811	TBM	110,452	2,547	152	1.33	0.8
Nullarbor	197,228	DXS	17,148	843	71	0.09	0.72
Ord Victoria Plain	125,407	TSG	27,535	1,592	126	0.23	0.87
Pine Creek	28,518	TSG	128,695	2,099	161	4.71	0.93
Pilbara	178,231	DXS	37,037	1,580	102	0.22	0.81
Riverina	97,045	TGSS	116,522	1,990	126	1.25	0.8
South East Coastal Plain	17,492	TBMF	28,420	1,937	159	1.7	0.83
South East Corner	25,321	TBMF	299,483	2,723	185	12.35	0.86

South Eastern Highlands	83,760	TBMF	322,905	3,605	198	4.03	0.85
South Eastern Queensland	78,049	TBMF	313,340	4,244	233	4.19	0.85
Simpson Strzelecki Dunefields	279,843	DXS	39,456	1,246	93	0.15	0.88
Stony Plains	131,664	DXS	59,826	1,292	94	0.47	0.85
Sturt Plateau	98,575	TSG	15,502	1,083	107	0.16	0.91
Southern Volcanic Plain	24,403	TBMF	18,166	1,672	142	0.78	0.83
Swan Coastal Plain	15,258	MFW	80,061	3,093	134	5.48	0.82
Sydney Basin	36,296	TBMF	886,447	3,759	209	25.5	0.84
Tanami	259,973	DXS	27,385	1,282	97	0.11	0.91
Tasmanian Central Highlands	7,678	TBMF	71,790	1,332	141	9.76	0.91
Tiwi Cobourg	10,106	TSG	31,192	1,257	146	3.22	0.92
Tasmanian Northern Midlands	4,154	TBM	35,519	1,098	128	8.93	0.88
Tasmanian Northern Slopes	6,231	TBM	63,543	1,271	144	10.65	0.89
Tasmanian South East	11,318	TBM	204,388	1,842	160	18.86	0.89
Tasmanian Southern Ranges	7,572	TBM	92,211	1,499	153	12.72	0.91
Tasmanian West	15,651	TBM	57,228	1,263	146	3.82	0.9
Victoria Bonaparte	73,012	TSGS	57,121	2,093	153	0.82	0.89
Victorian Midlands	34,698	TBM	46,049	2,146	148	1.39	0.84
Warren	8,448	MFW	42,941	1,903	131	5.31	0.8
Wet Tropics	19,891	TSM	166,834	4,069	250	8.76	0.87
Yalgoo	50,876	MFW	26,546	1,677	98	0.54	0.79

Table S2 Representativeness of species data within Australia's Virtual Herbarium (AVH) with vegetation data recorded across 192 YETI survey plots of 0.04 ha. AVH data were extracted from buffers of increasing size (from 0.4 ha to 512 ha) surrounding the YETI plots to identify the spatial resolution at which all species reported by the YETI survey were also present within the AVH database.

YETI **vegetation plot identification code**	**Buffer size**								
	0.4 ha	**4 ha**	**8 ha**	**16 ha**	**32 ha**	**64 ha**	**128 ha**	**256 ha**	**512 ha**
ABD03P4M	0.000	0.000	0.000	0.919	0.919	0.919	0.919	0.919	0.919
AUB04N5M	0.000	0.000	0.000	0.897	0.897	0.931	0.931	0.931	0.931
BAR06	0.000	0.000	0.000	0.936	0.936	0.936	0.936	0.936	0.936
BAR09	0.000	0.000	0.000	0.925	0.925	0.925	0.925	0.925	0.925
BAR12	0.000	0.000	0.000	0.929	0.929	0.929	0.929	0.929	0.929
BAR21	0.000	0.000	0.000	0.833	0.833	0.833	0.833	0.833	0.833
BAR28	0.000	0.000	0.000	0.923	0.923	0.923	0.923	0.923	0.923
BAR36	0.000	0.000	0.000	0.902	0.902	0.902	0.902	0.902	0.902
BAR41	0.000	0.000	0.000	0.881	0.881	0.881	0.881	0.881	0.881
BAR48	0.000	0.000	0.000	0.905	0.905	0.905	0.905	0.905	0.905

BGE02B3U	0.000	0.000	0.000	0.968	0.968	0.968	0.968	0.968	0.968
BGE11T2M	0.000	0.000	0.000	0.886	0.886	0.886	0.886	0.886	0.886
BLG48N5R	0.000	0.000	0.000	0.950	0.950	0.950	0.950	0.950	0.950
BNB26T5L	0.000	0.000	0.000	0.952	0.952	0.952	0.952	0.952	0.952
BNB38M2M	0.000	0.000	0.000	0.794	0.794	0.794	0.794	0.794	0.794
BRB90N7U	0.000	0.000	0.000	0.882	0.882	0.882	0.882	0.882	0.882
BRS51P8M	0.000	0.000	0.000	0.894	0.894	0.894	0.894	0.915	0.915
BRSD5Q6F	0.000	0.000	0.000	0.973	0.973	0.973	0.973	0.973	0.973
BULS2US4	0.000	0.269	0.269	0.962	0.962	0.962	0.962	0.962	0.962
CAR03C5M	0.000	0.000	0.000	0.891	0.891	0.891	0.891	0.891	0.891
CLH16H2C	0.000	0.000	0.000	0.958	0.958	0.958	0.958	0.979	0.979
CLH17H4U	0.000	0.000	0.000	0.980	0.980	0.980	0.980	0.980	0.980
CLN58P0L	0.000	0.000	0.000	0.900	0.900	0.900	0.920	0.920	0.920
CLN70N8U	0.000	0.000	0.000	0.941	0.941	0.941	0.941	0.941	0.941
CLR14C8V	0.000	0.000	0.000	0.816	0.816	0.816	0.816	0.816	0.816
CMB06P1U	0.000	0.000	0.000	0.855	0.855	0.855	0.855	0.855	0.855
CMB74Q3V	0.000	0.000	0.000	0.909	0.909	0.909	0.909	0.909	0.909
CRC16N8M	0.000	0.000	0.000	0.953	0.953	0.953	0.953	0.953	0.953
CRN52N3M	0.000	0.000	0.000	0.936	0.936	0.936	0.936	0.936	0.936
CRN63N3U	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
CRR03C6M	0.000	0.000	0.000	0.872	0.872	0.872	0.872	0.872	0.872
CSN10P5U	0.000	0.000	0.000	0.857	0.857	0.857	0.857	0.857	0.857
CSNF2P6C	0.000	0.000	0.000	0.971	0.971	0.971	0.971	0.971	0.971
CSNF6P3L	0.000	0.000	0.444	0.972	1.000	1.000	1.000	1.000	1.000
CTH48N3U	0.000	0.020	0.020	0.939	0.939	0.939	0.939	0.939	0.939
CTHA5N4U	0.000	0.025	0.025	0.950	0.975	0.975	0.975	0.975	0.975
DNM36N2U	0.000	0.000	0.000	0.947	0.947	0.947	0.947	0.947	0.947
DRD09N7U	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
DRL40N3M	0.000	0.000	0.000	0.938	0.938	0.938	0.938	0.938	0.938
DRLC4N6M	0.000	0.000	0.000	0.971	0.971	0.971	0.971	0.971	0.971
DRLD5N4M	0.000	0.500	0.500	0.944	0.944	0.944	0.944	0.944	0.944
DWS11C8U	0.000	0.000	0.000	0.900	0.900	0.900	0.900	0.900	0.900
DWS17C3M	0.000	0.000	0.019	0.907	0.907	0.907	0.907	0.907	0.907
DYL04N8V	0.000	0.000	0.000	0.857	0.857	0.857	0.857	0.857	0.857
ELD06C5M	0.000	0.000	0.527	0.891	0.891	0.891	0.891	0.891	0.891
GGL22N6U	0.000	0.000	0.500	0.889	0.889	0.889	0.889	0.889	0.889
GLC31C1M	0.000	0.516	0.548	0.839	0.839	0.839	0.839	0.839	0.839
GLC52P3R	0.000	0.000	0.000	0.860	0.860	0.860	0.860	0.860	0.860
GNG19N4F	0.000	0.000	0.000	0.959	0.959	0.959	0.959	0.959	0.959
GNG25N5U	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
GNG32N1V	0.000	0.000	0.000	0.882	0.882	0.882	0.882	0.882	0.882
GRT06P8V	0.000	0.000	0.000	0.955	0.955	0.955	0.955	0.955	0.955
GRT19P2V	0.000	0.000	0.000	0.840	0.840	0.840	0.840	0.840	0.840
GRW19N7M	0.000	0.000	0.000	0.917	0.917	0.917	0.917	0.917	0.917
GRW22N7L	0.000	0.000	0.000	0.912	0.912	0.912	0.912	0.912	0.912
GSF20H7U	0.000	0.000	0.000	0.977	0.977	0.977	0.977	0.977	0.977
GSP24N5V	0.000	0.000	0.000	0.946	0.946	0.946	0.946	0.946	0.946

HWS53N4C	0.000	0.000	0.000	0.907	0.907	0.907	0.907	0.907	0.907
HWS56N7C	0.000	0.000	0.516	0.935	0.935	0.935	0.935	0.935	0.935
ILF01N8C	0.000	0.000	0.000	0.927	0.927	0.927	0.927	0.927	0.927
ING05C2M	0.000	0.000	0.000	0.962	0.962	0.962	0.962	0.962	0.962
ING11C7V	0.000	0.000	0.000	0.941	0.941	0.941	0.941	0.941	0.941
JRP03P1M	0.000	0.000	0.000	0.926	0.926	0.926	0.926	0.926	0.926
JV_DB160	0.000	0.000	0.000	0.873	0.873	0.873	0.873	0.873	0.873
JV_DB164	0.000	0.000	0.000	0.911	0.911	0.911	0.911	0.911	0.911
JV_DB169	0.000	0.000	0.000	0.760	0.760	0.760	0.760	0.760	0.760
JV_DB171	0.000	0.000	0.000	0.905	0.905	0.952	0.952	0.952	0.952
JV_DB184	0.000	0.000	0.000	0.885	0.885	0.885	0.885	0.885	0.885
JV_DB199	0.000	0.000	0.000	0.903	0.903	0.903	0.903	0.903	0.903
JV_DB201	0.000	0.000	0.000	0.793	0.793	0.793	0.793	0.793	0.793
JV_DB202	0.000	0.000	0.000	0.857	0.857	0.857	0.857	0.857	0.857
JV_DB228	0.000	0.000	0.000	0.906	0.906	0.906	0.906	0.906	0.906
JV_DB230	0.000	0.000	0.000	0.914	0.914	0.914	0.914	0.914	0.914
JV_DB236	0.000	0.000	0.000	0.911	0.911	0.911	0.911	0.911	0.911
KLN36H6U	0.000	0.000	0.000	0.905	0.905	0.905	0.905	0.905	0.905
KLN58H8U	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
KND25N3G	0.000	0.000	0.000	0.879	0.879	0.879	0.879	0.879	0.879
KND35N1M	0.000	0.000	0.000	0.881	0.881	0.881	0.881	0.881	0.881
KRB43Q5F	0.000	0.000	0.057	1.000	1.000	1.000	1.000	1.000	1.000
KRB45A4V	0.000	0.000	0.000	0.879	0.879	0.879	0.879	0.879	0.879
KRH37P1U	0.000	0.000	0.000	0.957	0.957	0.957	0.957	0.957	0.957
LPR15H0C	0.000	0.000	0.000	0.976	0.976	0.976	0.976	0.976	0.976
LTH76M7F	0.000	0.000	0.000	0.800	0.800	0.800	0.800	0.800	0.800
MER3003T	0.000	0.000	0.842	0.947	0.947	0.947	0.947	0.947	0.947
MER3006C	0.000	0.000	0.000	0.884	0.884	0.884	0.884	0.907	0.930
MER3014T	0.000	0.000	0.000	0.938	0.938	0.938	0.938	0.938	0.938
MER3022	0.000	0.000	0.000	0.873	0.873	0.873	0.873	0.873	0.889
MER3041C	0.000	0.000	0.632	0.632	0.632	0.632	0.632	0.632	0.632
MER3053C	0.000	0.000	0.000	0.875	0.875	0.875	0.875	0.875	0.875
MER3053T	0.000	0.000	0.000	0.848	0.848	0.848	0.848	0.848	0.848
MER3065T	0.000	0.000	0.000	0.907	0.907	0.907	0.907	0.907	0.907
MER3067	0.000	0.000	0.000	0.706	0.706	0.706	0.706	0.706	0.706
MER3068C	0.000	0.686	0.686	0.886	0.886	0.886	0.886	0.886	0.886
MER3069C	0.000	0.000	0.000	0.906	0.906	0.906	0.906	0.906	0.906
MER3074	0.000	0.000	0.000	0.714	0.714	0.714	0.714	0.714	0.714
MER3076C	0.000	0.000	0.484	0.935	0.935	0.935	0.935	0.935	0.935
MER3084C	0.000	0.000	0.000	0.857	0.857	0.857	0.857	0.857	0.857
MLG41H2U	0.000	0.000	0.000	0.951	0.951	0.951	0.951	0.951	0.951
MMR19T7R	0.000	0.000	0.000	0.833	0.833	0.833	0.833	0.833	0.833
MMR31A2F	0.000	0.000	0.000	0.912	0.912	0.941	0.941	0.941	0.941
MMS02B6L	0.000	0.000	0.000	0.953	0.953	0.953	0.953	0.953	0.953
MMS27N7C	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
MNB24N5L	0.000	0.000	0.000	0.938	0.938	0.938	0.938	0.938	0.938
MNG32H1U	0.000	0.000	0.000	0.958	0.958	0.958	0.958	0.958	0.958

MNH01N5U	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
MNN21A2S	0.000	0.000	0.256	0.923	0.923	0.923	0.923	0.923	0.923
MNN28N6U	0.000	0.000	0.000	0.933	0.933	0.933	0.933	0.933	0.933
MP2SF073	0.000	0.000	0.000	0.971	0.971	0.971	0.971	0.971	0.971
MPM17A6V	0.057	0.057	0.057	0.830	0.830	0.830	0.830	0.830	0.830
MRB49H2V	0.000	0.000	0.000	0.903	0.903	0.903	0.903	0.903	0.903
MRN33Q4C	0.000	0.000	0.621	0.897	0.897	0.897	0.897	0.897	0.897
MRS46N4U	0.000	0.000	0.000	0.972	0.972	0.972	0.972	0.972	0.972
MRW01J7M	0.000	0.000	0.000	0.808	0.808	0.808	0.808	0.808	0.808
MRY05N3M	0.000	0.000	0.000	0.940	0.940	0.940	0.940	0.940	0.940
MSW07P5M	0.000	0.000	0.000	0.935	0.935	0.935	0.935	0.935	0.935
MTL12C7M	0.000	0.431	0.431	0.902	0.902	0.902	0.902	0.902	0.902
MTY34B5F	0.000	0.000	0.000	0.974	0.974	0.974	0.974	0.974	0.974
MTY35N0F	0.000	0.000	0.000	0.846	0.846	0.846	0.846	0.846	0.846
MUR11N5L	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
MYALLS04	0.000	0.567	0.667	0.867	0.867	0.867	0.867	0.867	0.867
NAB4CGC	0.000	0.000	0.000	0.960	0.960	0.960	0.960	0.960	0.960
NBFF1268	0.000	0.000	0.000	0.880	0.880	0.880	0.880	0.880	0.880
NBFF1457	0.000	0.000	0.000	0.897	0.897	0.897	0.897	0.931	0.931
NCPP0068	0.000	0.000	0.000	0.828	0.828	0.828	0.828	0.828	0.828
NCPP0087	0.000	0.026	0.026	0.846	0.846	0.846	0.846	0.846	0.846
NCPP0207	0.000	0.000	0.000	0.762	0.762	0.762	0.762	0.762	0.762
NCPP0210	0.000	0.000	0.000	0.871	0.871	0.871	0.871	0.871	0.871
OLN03A8F	0.000	0.000	0.000	0.966	0.966	0.966	0.966	0.966	0.966
PRK17P2M	0.000	0.000	0.000	0.942	0.942	0.942	0.942	0.942	0.942
PRN36A6V	0.000	0.000	0.000	0.925	0.925	0.925	0.925	0.925	0.925
PRN59N7U	0.000	0.000	0.000	0.850	0.850	0.850	0.850	0.850	0.850
PTY46N6U	0.000	0.000	0.000	0.968	0.968	0.968	0.968	0.968	0.968
PTY57N8R	0.000	0.000	0.000	0.891	0.891	0.891	0.891	0.891	0.891
QRB44N3M	0.000	0.000	0.000	0.838	0.838	0.838	0.838	0.838	0.838
QRBB3P3C	0.000	0.000	0.000	0.939	0.939	0.939	0.939	0.939	0.939
RCH04C5M	0.000	0.000	0.538	0.897	0.897	0.897	0.897	0.897	0.897
RCH18C6U	0.000	0.000	0.000	0.913	0.913	0.913	0.913	0.913	0.913
REV12806	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
REV13805	0.000	0.000	0.000	0.878	0.878	0.878	0.878	0.878	0.878
REV13812	0.000	0.000	0.000	0.938	0.938	0.938	0.938	0.938	0.938
RKH23N4U	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
SCN11P1M	0.000	0.000	0.000	0.964	0.964	0.964	0.964	0.964	0.964
SHL46N1L	0.000	0.000	0.000	0.978	0.978	0.978	0.978	0.978	0.978
SHL58P8M	0.016	0.016	0.016	0.934	0.934	0.934	0.934	0.934	0.934
SLR06Q3V	0.000	0.000	0.000	0.938	0.938	0.938	0.938	0.938	0.938
SNG05P6F	0.000	0.000	0.000	0.952	0.952	0.952	0.952	0.952	0.952
SNG07P6L	0.000	0.000	0.000	0.929	0.929	0.929	0.929	0.929	0.929
SNG18P1V	0.000	0.000	0.000	0.886	0.886	0.886	0.886	0.886	0.886
STA38H3U	0.000	0.000	0.000	0.945	0.945	0.945	0.945	0.945	0.945
STA43H3U	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
SWN47N6C	0.000	0.024	0.024	0.976	0.976	0.976	0.976	0.976	0.976

SWN76Q2C	0.000	0.000	0.306	0.972	0.972	0.972	0.972	0.972	0.972
SXB24Q4F	0.000	0.000	0.000	0.944	0.944	0.944	0.944	0.944	0.944
SXB32N2U	0.000	0.000	0.000	0.861	0.861	0.861	0.861	0.861	0.889
TKL32Q1L	0.000	0.000	0.000	0.545	0.545	0.545	0.545	0.606	0.727
TLB21N1U	0.000	0.500	0.500	0.867	0.867	0.867	0.867	0.867	0.867
TLB25N5L	0.028	0.028	0.028	0.944	0.944	0.944	0.944	0.944	0.944
UTR18M7R	0.000	0.000	0.000	0.893	0.893	0.893	0.893	0.893	0.893
WAT001	0.000	0.000	0.000	0.900	0.900	0.900	0.900	0.900	0.900
WDD29B7C	0.000	0.000	0.481	0.741	0.741	0.741	0.741	0.741	0.741
WF02	0.000	0.000	0.000	0.925	0.925	0.925	0.925	0.925	0.925
WF09	0.000	0.000	0.000	0.944	0.944	0.944	0.944	0.944	0.944
WF17	0.000	0.000	0.000	0.857	0.857	0.857	0.857	0.857	0.857
WF29	0.000	0.000	0.000	0.950	0.950	0.950	0.950	0.950	0.950
WF32	0.000	0.000	0.000	0.938	0.938	0.938	0.938	0.938	0.938
WF47	0.000	0.000	0.000	0.923	0.923	0.923	0.923	0.923	0.923
WF50	0.000	0.000	0.000	0.833	0.833	0.833	0.833	0.833	0.833
WF58	0.000	0.000	0.513	0.974	0.974	0.974	0.974	0.974	0.974
WF62	0.022	0.022	0.022	0.957	0.957	0.957	0.957	0.957	0.957
WF76	0.000	0.040	0.040	0.920	0.920	0.920	0.920	0.920	0.920
WF81	0.000	0.000	0.000	0.852	0.852	0.852	0.852	0.852	0.852
WF85	0.000	0.000	0.000	0.952	0.952	0.952	0.952	0.952	0.952
WF86	0.022	0.022	0.022	0.935	0.935	0.935	0.935	0.935	0.935
WLM03A8F	0.000	0.000	0.000	0.966	0.966	0.966	0.966	0.966	0.966
WLM04A2F	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
WLM05N7U	0.000	0.000	0.000	0.865	0.865	0.865	0.865	0.865	0.865
WLM10N6U	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
WLN36N8L	0.000	0.000	0.000	0.881	0.881	0.881	0.881	0.881	0.881
WLR14N7U	0.000	0.677	0.710	0.968	0.968	0.968	0.968	0.968	0.968
WLR20N1V	0.000	0.000	0.000	0.947	0.947	0.947	0.947	0.947	0.947
WLS67P4M	0.000	0.000	0.000	0.927	0.927	0.927	0.927	0.927	0.927
WLS72P4L	0.000	0.000	0.000	0.944	0.944	0.944	0.944	0.944	0.944
WLT81Q8C	0.000	0.000	0.250	0.972	0.972	0.972	0.972	0.972	0.972
WLTG7Q4F	0.000	0.000	0.000	0.931	0.931	0.931	0.931	0.931	0.931
WPDLB025	0.000	0.000	0.000	0.966	0.966	0.966	0.966	0.966	0.966
WRB19H3R	0.000	0.000	0.000	0.946	0.946	0.946	0.946	0.946	0.946
WRB21Q-F	0.000	0.000	0.000	0.881	0.881	0.881	0.881	0.881	0.881
WRB27N3L	0.000	0.000	0.000	0.957	0.957	0.957	0.957	0.957	0.957
WSDLB023	0.000	0.000	0.000	0.966	0.966	0.966	0.966	0.966	0.966
WTDLB009	0.000	0.000	0.000	0.857	0.857	0.857	0.857	0.857	0.857
WVR04C2M	0.000	0.000	0.000	0.933	0.933	0.933	0.933	0.933	0.933
WYN06N4C	0.000	0.000	0.000	0.933	0.933	0.933	0.933	0.933	0.933