ECOGRAPHY

*Research*

# A comparison of network and clustering methods to detect biogeographical regions


**Nathaniel J. Bloomfield, Nunzio Knerr and Francisco Encinas-Viso**

*N. J. Bloomfield, N. Kerr and F. Encinas-Viso (http://orcid.org/0000-0003-0426-2342) (francisco.encinas.viso@csiro.au), NCMI and Centre for Australian National Biodiversity Research, CSIRO, Canberra, Australia. Present address of NJB: Research School of Chemistry, Australian National Univ., Australia.*

**Ecography**
**41: 1–10, 2018**
doi: 10.1111/ecog.02596

Subject Editor: Michael Borregaard.
Editor-in-Chief: Miguel Araújo.
Accepted 23 November 2016

Bioregions are an important concept in biogeography, and are key to our understanding of biodiversity patterns across the world. The use of networks in biogeography to produce bioregions is a relatively novel approach that has been proposed to improve upon current methods. However, it remains unclear if they may be used in place of current methods and/or offer additional biogeographic insights. We compared two network methods to detect bioregions (modularity and map equation) with the conventional distance-based clustering method. We also explored the relationship between network and biodiversity metrics. For the analysis we used two datasets of iconic Australian plant groups at a continental scale, Acacia and eucalypts, as example groups. The modularity method detected fewer large bioregions produced the most succinct bioregionalisation for both plant groups corresponding to Australian biomes, while map equation detected many small bioregions including interzones at a natural scale of one. The clustering method was less sensitive than network methods in detecting bioregions. The network metric called participation coefficient from both network partition methods identified interzones or transition zones between bioregions. Furthermore, another network metric (betweenness) was highly correlated to richness and endemism. We conclude that the application of networks to biogeography offers a number of advantages and provides novel insights. More specifically, our study showed that these network partition methods were more efficient than the clustering method for bioregionalisation of continental-scale data in: 1) the identification of bioregions and 2) the quantification of biogeographic transition zones using the participation coefficient metric. The use of network methods and especially the participation coefficient metric adds to bioregionalisation by identifying transition zones which could be useful for conservation purposes and identifying biodiversity hotspots.

## Introduction

Biodiversity is being lost at an increasing rate, and conservation is more critical than ever before (Butchart et al. 2010). Biogeography – the study of taxa distributed across



NORDIC SOCIETY OIKOS

www.ecography.org

space and time – is emerging as key field in informing current conservation efforts (Whittaker et al. 2005). Its methods are important in identifying biodiversity hotspots, regions with rare and endemic taxa and bioregions, i.e. geographic areas which contain similar taxa. By highlighting these important areas and communities, this information helps design reserves which can protect biodiversity more efficiently.

Current methods in biogeography use presence datasets, either from collected specimens or range distributions, to assign taxa to grid cells. From these grid cells several measures of biodiversity can be calculated, such as richness (the total number of taxa within the cell), weighted endemism (WE) (the summed inverse range of all taxa, generally species, in a cell) and corrected weighted endemism (CWE) (the averaged inverse range of all taxa in a cell (Laffan et al. 2016)). This approach is used to identify biodiversity hotspots (Nagalingum et al. 2014, Daru and le Roux 2015). However, these measures are susceptible to sampling biases, as well collected regions tend to contain higher numbers of taxa, and conversely poorly collected taxa will tend to have smaller ranges and so contribute more weight towards WE and CWE (Schmidt-Lebuhn et al. 2012).

Clustering analyses have been commonly used to identify biogeographical regions and are still widely used today (Milligan and Cooper 1987, Gonzalez-Orozco et al. 2014b, Ebach et al. 2015). In the WPGMA (weighted pair-group method using arithmetic averages) clustering algorithm used in the Biodiverse software package (Laffan et al. 2010), a distance or turnover metric compares the composition of taxa between cells to create a similarity tree or dendrogram, with cells that are grouped within the same branch being more similar. This dendrogram can then be split at the first n number of branches to produce bioregions. Three commonly used similarity metrics include the Jaccard, Sorenson and Beta Simpson, and each have their advantages (Laffan et al. 2016). When two cells are tied in their similarity to a growing cluster according to a chosen metric, the algorithm selects one of the cells at random and the clustering will be different each time it is run. To avoid this a tiebreaker can be used, which selects one of the cells that maximizes or minimizes a second score, such as CWE (Gonzalez-Orozco et al. 2013). Hence the bioregions produced are not robust, as a different choice of metric, tiebreaker or split could change the result.

Recently, network methods have been applied to detect bioregions as an alternative to clustering methods (Carstensen and Olesen 2009, Thébault 2013, Vilhena and Antonelli 2015). In this approach the network is bipartite, with two sets of nodes; locations and taxa, with the taxa linked to locations in which they are present. Two nodes of the same type are not permitted to have a link between them. One commonly used network method is Netcarto. Netcarto is a method that finds the best network partition by optimizing the modularity metric (Newman and Girvan 2004) using a simulated annealing (SA) algorithm (Guimera and Amaral 2005). From here on, we will refer to it as modularity SA method. The modularity SA method has been used in several studies to delineate biogeographical regions (Carstensen and Olesen 2009, Carstensen et al. 2012, 2013a, Thébault 2013, Dalsgaard et al. 2014). This method is also widely used to detect modules or communities (i.e. groups of nodes) in ecological networks (Olesen et al. 2007, Encinas-Viso et al. 2016) and other biological networks (e.g. metabolic networks) (Sales-Pardo et al. 2007).

The modularity SA method has proved useful in detecting bioregions (also called biogeographical modules, i.e. clusters of areas and species that are associated with each other (Carstensen et al. 2012, 2013a)) and it has been mainly applied to the bioregionalisation of archipelagos (Carstensen and Olesen 2009, Carstensen et al. 2012, Dalsgaard et al. 2014, Kougioumoutzis et al. 2014). For example, Carstensen et al. (2012) have used the modularity SA method to understand the processes driving the assembly of island bird communities in Wallacea (Indonesia) and the West Indies (Caribbean).

Vilhena and Antonelli (2015) have recently proposed using a new network method to identify bioregions from taxon presence data. They utilized the map equation (ME) algorithm (Rosvall and Bergstrom 2008), and found it to outperform clustering methods using a dataset of all American plants in identifying the generally recognized biomes of the United States. They also applied ME on a worldwide dataset of amphibians, and found it to be successful. ME is a method based upon the 'flow' of information through the network (Rosvall and Bergstrom 2008), in contrast modularity SA method is a purely topological (and unsupervised) method that focuses on the structure of the network (Sales-Pardo et al. 2007). The ME and modularity SA (implemented in Quanbimo (Dormann and Strauss 2014)) methods can both incorporate edge weights into the analysis. This could potentially be useful in adding taxon range and phylogenetic information into the network.

The application of networks to biogeographical studies is relatively novel, and it is still unclear whether they provide a superior bioregionalisation to the currently used clustering techniques. More importantly, we ask: which network method (modularity SA or ME) is better for bioregionalisation? And, what additional insights could network methods provide to bioregionalisation analysis, especially for continental-scale data? Herein, we present a comparison of three methods for identifying bioregions – clustering, ME and modularity SA – in large, continental datasets of two Australian plant groups, the Acacia and eucalypts (Eucalyptus, Corymbia and Angophora). The comparison of the network methods was done without considering any weights for the links (i.e. using unweighted networks). We also examined the correlations between several network metrics and conventional biodiversity metrics (richness, weighted endemism (WE) and corrected weighted endemism (CWE)) using a principal component analysis (PCA). Network surrogates for endemism were also further

explored using a range-weighted network and analysed with ME.

## Material and methods

### Sample datasets

In this study we have applied each of the methods to two species level datasets of Australian plants, the genus Acacia (Mishler et al. 2014) and the eucalypts (Eucalyptus, Corymbia and Angophora) (unpublished). These contained a total of 132 295 and 209 396 records, with 508 and 684 species, respectively. These datasets were chosen as they both have continental coverage and a large number of collections, which are ideal for these biogeographical analyses. These datasets were transformed using the Australian Albers equal area projection (EPSG-3577).

### Clustering methods

We used Biodiverse (Laffan et al. 2010), a software package designed for biogeographic studies, to calculate a turnover matrix using the Simpson's Beta ($\beta_{sim}$) metric (also called S2) at $100 \times 100$ km grid cell sizes. Simpson's Beta metric best accounts for cases where cells contain different numbers of species (Tuomisto 2010). The $\beta_{sim}$ value between a pair of cells i and j is given by

$$\beta_{sim} = 1 - \frac{1}{a + \min(b, c)}$$

where a is the number of species common to both cells, b is the number found in i but not j, and c is the number found in j but not i. A low value of $\beta_{sim}$ (low turnover) means a large number of taxa are shared between the cells. The clustering was conducted using the $\beta_{sim}$ turnover matrix in Biodiverse, which uses the WPGMA (weighted pair-group method using arithmetic averages) algorithm. To guarantee a reproducible solution, CWE was chosen to be maximized in the event of a tiebreaker (Gonzalez-Orozco et al. 2013).

### Modularity SA

The degree of modularity M in a particular partition of a network is defined by

$$M = \sum_{S=1}^{N_m} \left[ \frac{l_S}{L} - \left( \frac{d_S}{2L} \right)^2 \right]$$

where $N_M$ is the number of modules, L is the number of links in the network, $l_s$ is the number of links between nodes in module s, and $d_s$ is the sum of the degrees of the nodes in module s. This equation follows the rationale that a good partition of a network must have many within-module links and few between-module links, while a partition in which the whole network is a module, or every node is a module has $M = 0$. The modularity SA method uses a stochastic simulated annealing procedure to find a partition that maximizes this value of M. In the simulated annealing, a new partition is accepted according to the cost ($C = -M$) with the probability p

$$p = \begin{cases} 1 & \text{if } C_f \leq C_i \\ e^{-\left( \frac{C_f - C_i}{T} \right)} & \text{if } C_f > C_i \end{cases}$$

where $C_i$ is the cost of the previous partition, $C_f$ is the cost of the new partition and T is the computation temperature, which starts high and iteratively decreases according to $T' = cT$, where c is the cooling factor. At each step, there is also an iteration factor that determines how much of the partition can be altered (Guimera and Amaral 2005). After the simulated annealing is complete, the algorithm runs a randomization test in which it generates a random network with the same connectivity distribution as the original network and calculates the modularity.

This test can be used to determine if the modular structure of the original network is significant or not (Guimera et al. 2004). The modularity SA method implemented in the program Netcarto (Guimera and Amaral 2005) also calculates two metrics per node: participation coefficient (PC) and within-module degree (WMD), which are based upon the partitioning of the network. PC measures the degree to which 142 a node is connected to other modules. The PC of a node, $P_i$, is given by

$$P_i = 1 - \sum_{s=1}^{N_m} \left( \frac{k_{is}}{k_i} \right)$$

where $k_{is}$ is the number of links of node i to nodes in module s, and $k_i$ is the total degree of node i. The more links a node has to different modules, the higher the PC of the node will be.

The WMD of a node, $z_i$, is given by

$$z_i = \frac{k_i - \overline{k_s}}{\sigma_{k_s}}$$

where $k_i$ is the number of links of node i to other nodes in its module, s, $\overline{k_s}$ is the average of k over all nodes in s, and $\sigma_{k_s}$ is the standard deviation of k in s. In this study, the WMD has been calculated only over the location nodes, according to Carstensen et al. (2013b). This measures the number of links a node has to its own module, compared with all other nodes in the module (Guimera et al. 2004).

The modularity SA analysis was run using the program NetCarto (Guimera and Amaral 2005) using an initial temperature $T_i = 5.0$, a cooling factor $c = 0.95$ and an iteration factor of 1.0, as these parameters lead to a partition with the highest modularity score. For both datasets 100 randomizations were used to test the significance of the bioregionalisation using the randomization test (Guimera and Amaral 2005).

### Map equation

The map equation (ME) algorithm (Rosvall and Bergstrom 2008, Rosvall et al. 2009) detects clusters or groups of nodes using a random walker, which randomly travels along links in the network. The network is then partitioned in an iterative procedure which attempts to minimize the length of a code which describes the movements of the walker (Rosvall and Bergstrom 2008). This method allows the links to be weighted, the links with higher weighting being wider and more likely to be traveled through compared with links of a smaller weight. The algorithm can also be applied hierarchically, in which it attempts to find submodules within the modules. The ME output also provides a flow measure, which measures how important a particular node is to the overall connectivity of the network (Bohlin et al. 2014). In this study, the ME algorithm as implemented in Infomap code (Rosvall and Bergstrom 2008) 168 was run with 100 000 iterations to ensure the ergodic solution had been found (Bohlin et al. 2014), as an undirected network using the two-level algorithm. The multilevel algorithm (i.e. to find sub-modules in the network) was also explored, but little hierarchical structure was found. In each of these 100 000 iterations the ME algorithm is applied, and at the end the iteration with the shortest code-length is output. Other settings in the algorithm were left at the default. We also calculated the PC and WMD metrics using the network partitions of ME to compare them with the modularity SA partitions as these metrics are independent of the network partition method used.

### Network creation, measures and visualization

To construct the networks, the presence data was exported from Biodiverse, and an R script was used to generate the adjacency matrix of the location-species bipartite network (R Core Team), as described in Vilhena and Antonelli (2015). The biodiversity metrics (richness, weighted endemism (WE) and corrected weighted endemism (CWE)) were calculated in Biodiverse, and the network centrality metrics (which estimates the importance of nodes in the network) betweenness, closeness, eigenvector centrality and alpha centrality were calculated using the R package igraph (Csardi and Nepusz 2006). The networks were visualized in the software Gephi using the Force Atlas 3D network layout (Bastian et al. 2009). All other plotting was done in R using ggplot2 (Wickham 2009). Furthermore, a PCA was conducted on the location nodes, using the R package vegan (Oksanen et al. 2015), to explore potential relationships between biodiversity and network metrics (including PC and WMD). We used the estimated biodiversity and network metrics from each network (i.e. the Acacia and eucalypts biogeographical networks) to do the PCA. We also included the calculated network metrics of a range-weighted network for both plant groups to test the importance of range in the correlation between biodiversity and network metrics. The analysis of range-weighted networks was done with map equation where the links between nodes were weighted by range (i.e. the number of location nodes a taxon occurs in; see Supplementary material Appendix 1 for more details). We also explored the hierarchical structure of the networks and the multiple network scales using an analysis of multiscale community detection (Schaub et al. 2012). The analysis did not show any strong multiscale network structure for the two study groups (Supplementary material Appendix 1).

### Data deposition

The Acacia data is available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.33kn3> and the eucalypt data is available as supporting information in Gonzalez-Orozco et al. (2014b).

## Results

### Modularity randomization test and map equation partition

We found that the Acacia and eucalypt biogeographical networks were highly modular. More specifically, the observed modularity of the Acacia ($M = 0.55$, $zscore = 253.3$, $p < 0.0001$) and the eucalypts ($M = 0.63$, $zscore = 244.4$, $p < 0.0001$) networks were highly significant. The map equation analyses show that the best partitions (i.e. shortest description length) for Acacia and eucalypts are described by $codelength = 9.765$ and $codelength = 8.487$ bits, respectively.

### Bioregions

The bioregions produced for each of the three methods are displayed in Fig. 1 for the Acacia and eucalypts datasets. For the Acacia dataset, all three methods produce a similar pattern; a south-west (Fig. 1C, region 1) and Euronotian cluster (Fig. 1C, region 2) (bioregion nomenclature following Gonzalez-Orozco et al. 2014a), in addition to an Eremean-south (Fig. 1C, region 3), Eremean-north (Fig. 1C, region 4) and a monsoonal zone (Fig. 1C, region 5). The network methods also identified a southern bioregion (Fig. 1C, region 6), which did not appear in the clustering analysis. The ME produced more bioregions than the modularity SA method (see regions 7 and 9, Fig. 1B). These extra regions appear at the edges between bioregions (interzones), both geographically and within the network. On the PC heat-map overlayed with the bioregion borders in Fig. 2, the additional bioregions inferred by the map equation (Fig. 2A, regions 7 and 9) are in locations of high PC. This is also evident from their locations on the network in Fig. 1B, and suggests that these may be important interzone regions. The bioregion borders in the clustering analysis (Fig. 1A) generally also lie within the regions of high PC in both the ME and modularity SA heatmaps (Fig. 2A, B), with the exception of the south-east being split into a core southern region

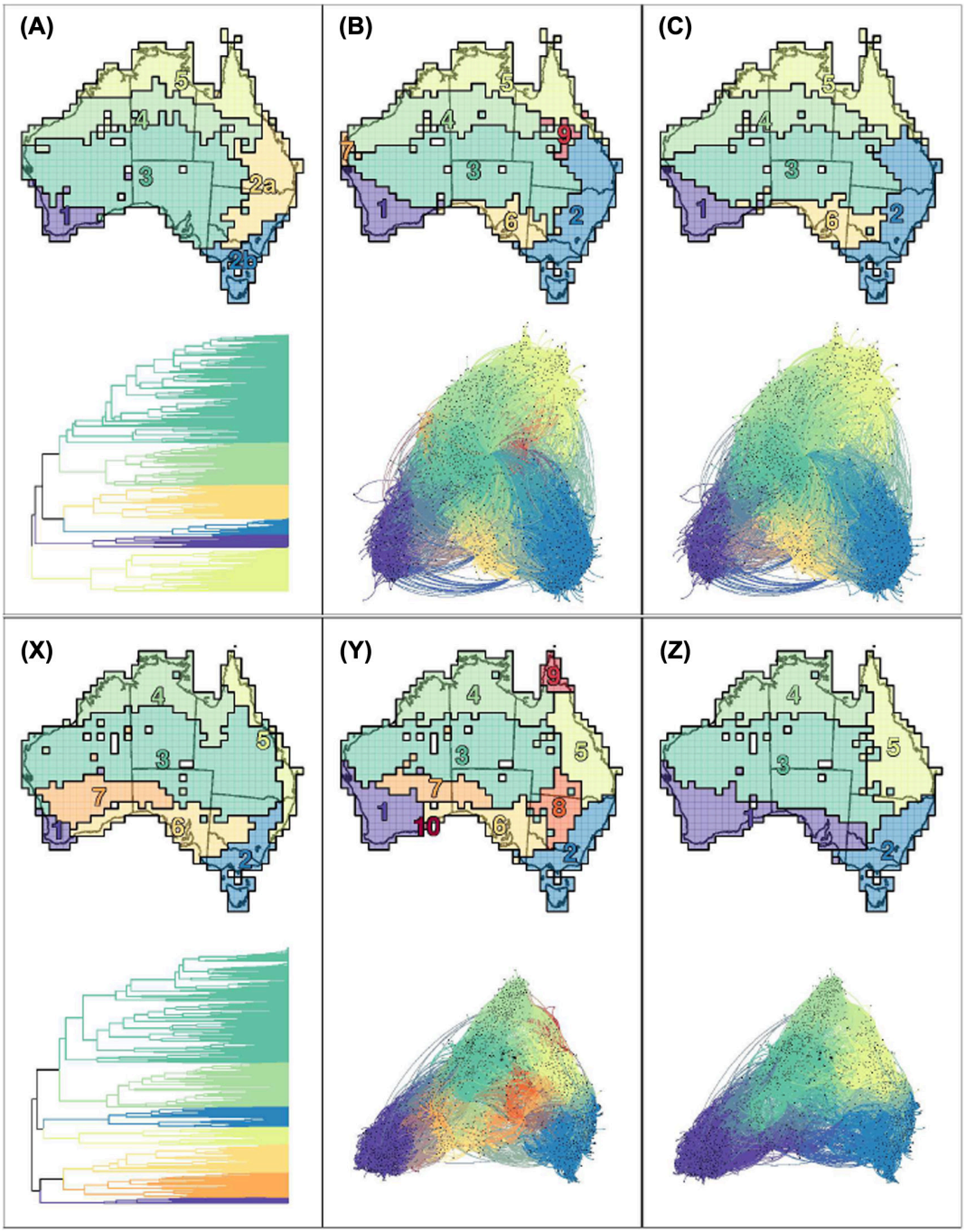


Figure 1. Analysis of the Acacia (top panel) and eucalypts (bottom panel) dataset comparing the three different bioregionalization methods. (A, X) show the result of the S2 ($\beta_{sim}$) clustering and dendrogram from which the bioregions were obtained. (B, Y) show the results of the map equation analysis and (C, Z) of the modularity analysis. The networks are colored according to the bioregions, and shown using the Force Atlas 2 layout.

(Fig. 1A, region 2b) and a northern region (Fig. 1A, region 2a) with greater PC. Although the species composition in Fig. 1A (region 2a) is a mix of Euronotian, arid and monsoonal species (relatively high PC), the clustering analysis identified it as a different cluster rather than grouping it into either one of these three regions. A different species composition generated a different bioregion in the clustering analysis.

The results are similar for the eucalypts dataset, with each method producing a similar bioregionalisation; a south-east (Fig. 1Z, region 2) and south-west (Fig. 1Z, region 1) region, an arid (Fig. 1Z, region 3) and monsoonal zone (Fig. 1Z, region 4), with Queensland (Fig. 1Z, region 5) also being identified as a bioregion. The two network methods produced similar results with the map equation identifying several interzones (Fig. 1Y, regions 6, 7, 8, 9, 10). The

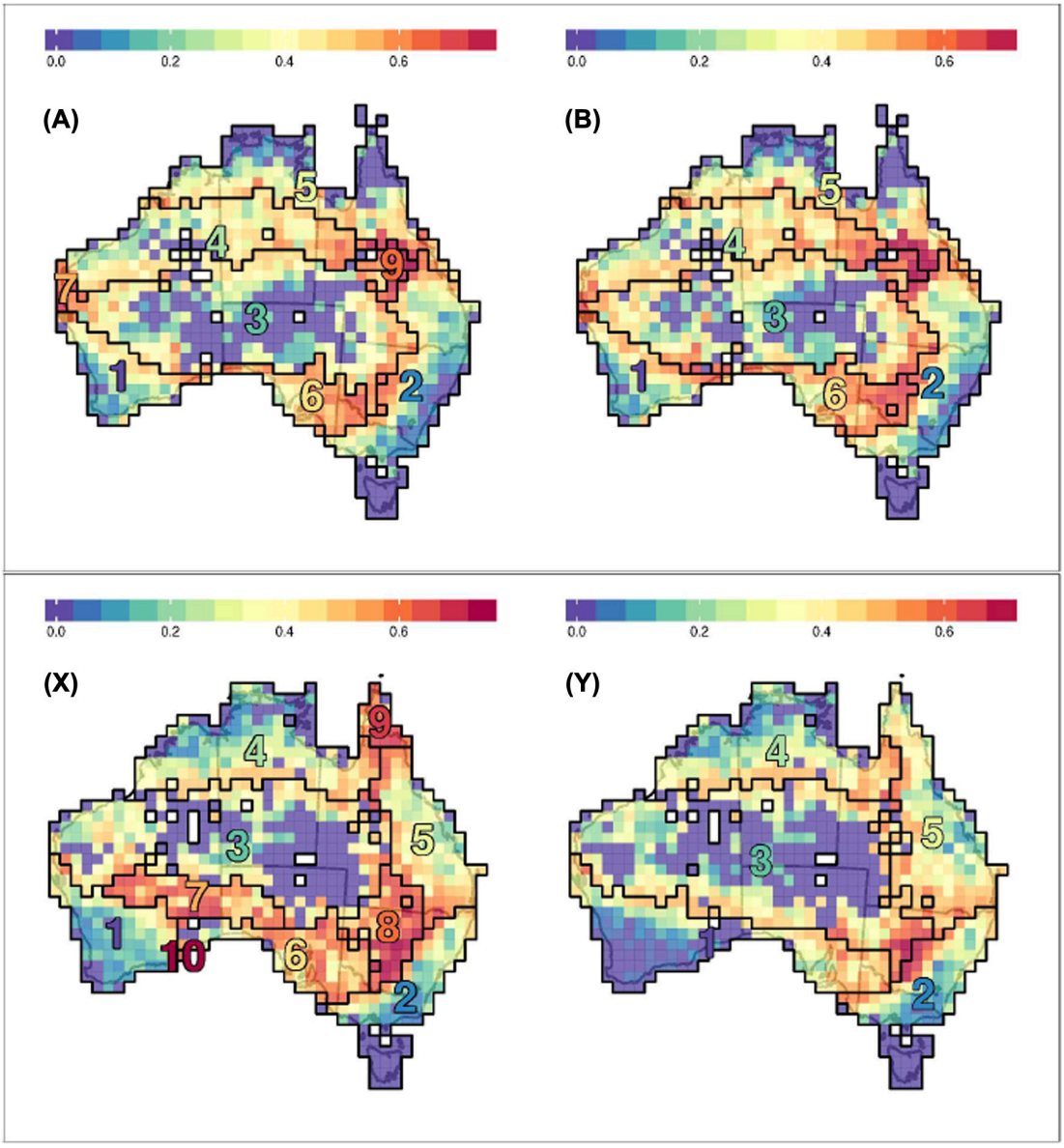


Figure 2. Participation coefficient (PC) heatmaps from the map equation (A, X) and modularity SA (B, Y) analysis for the Acacia (top panel) and eucalypts (bottom panel), overlayed with the bioregion borders. High PC areas show the transition zones between bioregions.

PC heat-map (Fig. 2) confirmed this, with these interzone regions from the map equation located in high PC areas. The bioregions from the clustering analysis (Fig. 1X) also appear to have borders that mostly overlap along regions of high PC in the network heatmaps (Fig. 2X, Y), particularly in the north-east, and the three southwestern clusters appear to be approximately distributed in a core region (Fig. 1X, region 1) with two zones with higher PC (Fig. 1X, regions 6, 7).

## Relationships between biodiversity and network metrics

A PCA analysis was conducted using several network centrality and biodiversity metrics for the Acacia and eucalypts networks (Fig. 3). The first two components explained 70% and 69% of the variance for Acacia and eucalypts, respectively (see Supplementary material Appendix 1, Table 1, 5). As expected, in both datasets a one to one correlation was found between ME flow and richness, and ME flow in the range weighted network and weighted-endemism. Betweenness was identified as a potential estimator of endemism because it was highly correlated with weighted endemism (WE) ($r = 0.79$) and richness ($r = 0.78$), along with within-module degree (WMD) ($r = 0.75$) (Supplementary material Appendix 1, Fig. A2, A6). Other network centrality metrics investigated (alpha-centrality, eigen-centrality and closeness) did not appear to have a significant correlation to any of the biodiversity metrics investigated here (Supplementary material Appendix 1, Fig. A1–A8).

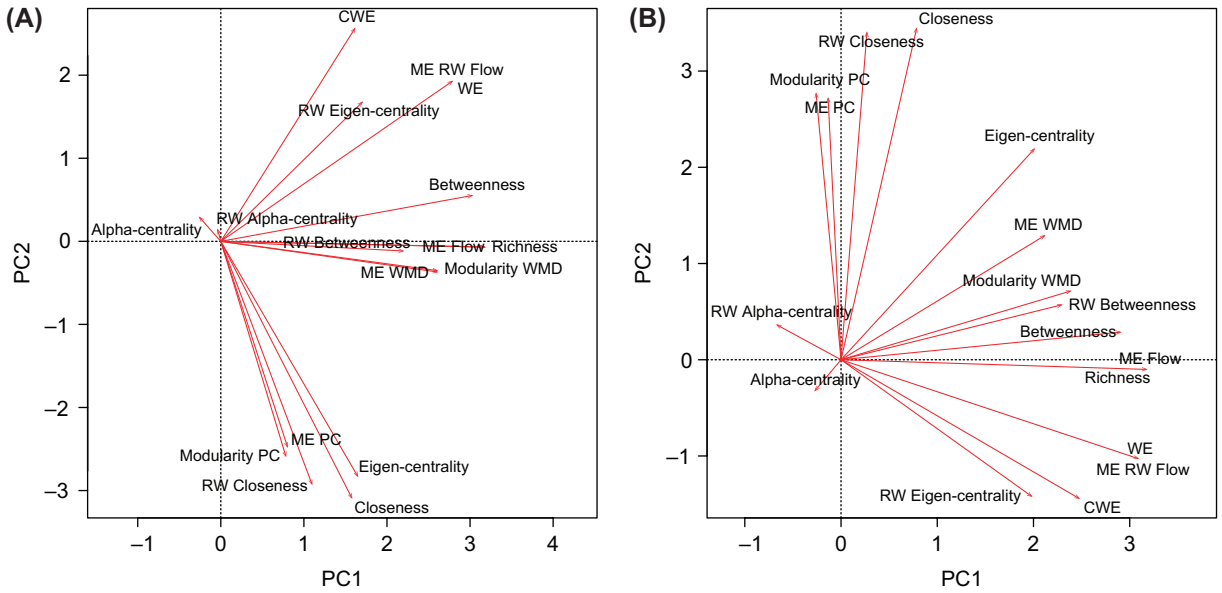


Figure 3. Principal component analysis (PCA) of several key measures from the Acacia (A) and eucalypts (B) datasets using the location nodes only. The first two axes have been plotted, and these describe 70% and 69% percent of the variation respectively. RW: range-weighted. ME: map equation. PC: participation coefficient. WMD: within-module degree. CWE: corrected-weighted endemism. WE: weighted-endemism.

## Discussion

### Bioregionalisation of Acacia and the eucalypts

The bioregionalisation of Acacia and the eucalypts have been previously studied (Gonzalez-Orozco et al. 2013, 2014b). Gonzalez-Orozco et al. (2013) conducted a similar clustering analysis to produce the bioregionalisation shown in Fig. 4A. They found their results to be mostly congruent with a previous bioregionalisation of Australian flora by Crisp et al. (2004) (Fig. 4C), with the exception of their identification of a northern Eremean zone. This zone was also identified in our analysis by all three methods (Fig. 1A, B, C, region 4), which suggests that within Acacia, this floristic zone is distinct. However, the network methods showed that this zone has very high participation coefficient (PC) (Fig. 2A, B) (i.e. the region contains species with many connections or geographical overlaps with other bioregions), which is likely due to taxa from the monsoonal (Fig. 1A, B, C, region 5) and Arid zones (Fig. 1 A, B, C, region 3) extending into this region. This suggests that for Acacia this northern Eremean zone may actually be a large scale interzone between the monsoonal tropics and the arid center. The same could also be true of the southern bioregion (Fig. 1B, C, region 6), which is also identified as a bioregion with a high PC.

A similar clustering study of the eucalypts was conducted by Gonzalez-Orozco et al (2014b). Their bioregionalisation (Fig. 4B) correlates well with our own clustering analysis (Fig. 1X) with the exception of a southern Eremean zone, which in our case was split into two regions (Fig.1X, regions 6 and 7). This zone did not appear in the modularity SA analysis (Fig. 1Z), with the south-western zone instead extending further east and past Adelaide (south Australia), while in the results of the map equation, this region could be potentially reconstructed by combining regions 6 and 7 of Fig. 1Y. These two ME zones also have high PC (Fig. 2X, Y), potentially showing transition zones. These results suggest that the southern Eremean is a large interzone for the eucalypts, with species mixing between the arid center, south-west and south-east bioregions.

In both plant groups, the analyses clearly show that each bioregion detected by the three methods more or less correspond to a specific Australian biome (Crisp et al. 2004) and the transitions zones detected by the PC metric correspond to areas of changes of biomes. Transition zones between different biomes correspond to changes in climatic and soil conditions (Bui et al. 2014, Gonzalez-Orozco et al. 2013, 2014a). For example, the transition zone between the southern-eastern temperate and the monsoonal tropics (see Fig. 4) seems to be highly driven by changes in annual mean temperature and precipitation seasonality (Gonzalez-Orozco et al. 2013). Bui et al. (2014) also showed that high pH soils in the Eremean south (more specifically at the Nullarbor Plain and Eyre Peninsula) correlate with the presence of several Acacia species (anceps clade). This area was detected as a separate bioregion (Fig. 1B6, C6) with high PC (Fig. 2A6, B6) in both network methods. This highlights that the network methods are more sensitive than the conventional clustering in defining bioregions and more importantly, identifying transition zones.

### Utility of network approaches for bioregionalisation

Undeniably, clustering methods have been shown to be a successful approach (Kreft and Jetz 2010, Schmidt-Lebuhn

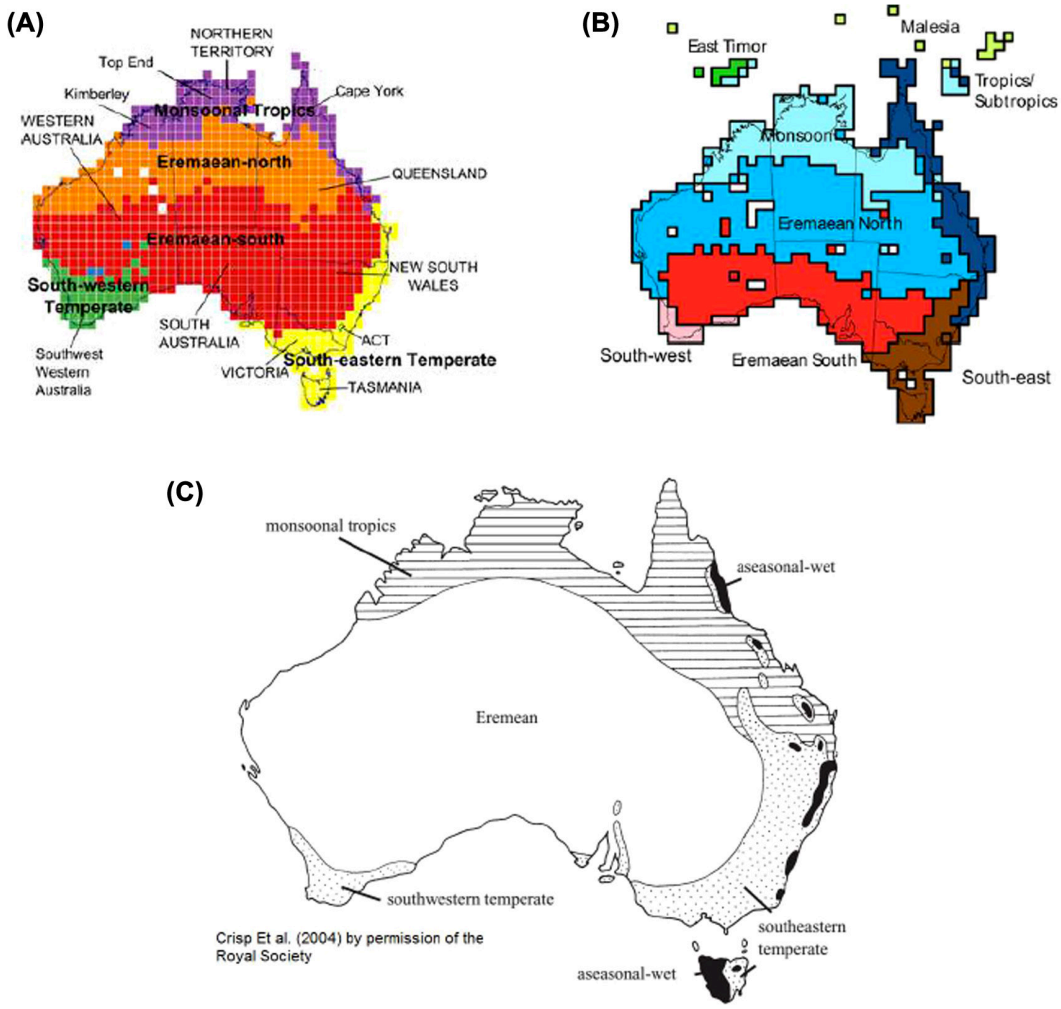


Figure 4. Previously published bioregionalisations of Acacia (A), eucalypts (B) and Australian plants (C). (A) Clustering based bioregionalisation of Acacia produced by Gonzalez-Orozco et al. (2013). (B) Clustering based bioregionalisations of the eucalypts by Gonzalez-Orozco et al. (2014b). (C) Opinion based bioregionalisation of Australian plants produced by Crisp et al. (2004). In (A) and (B) bioregions are represented by different colours.

et al. 2012, Gonzalez-Orozco et al. 2013). However, network methods seem to be very efficient at detecting bioregions and providing new insights (Carstensen et al. 2012, 2013a, Thébault 2013, Vilhena and Antonelli 2015, Economo et al. 2015). Vilhena and Antonelli (2015) recently found that the map equation (ME) method performs better than clustering methods when compared to well-accepted biogeographic breaks; Wallace's and Weber's line for the amphibian dataset, and the American biomes for the dataset of American plants. Although delineations of Australian biomes and phytogeographical regions have been previously performed (Crisp et al. 2004, Ebach et al. 2015), our datasets only reflected a single plant group and so are not directly comparable.

Nevertheless, it is clear that in both of the datasets tested and at the natural scale of one, modularity SA detected fewer large modules showing more concise bioregions. More importantly, one of the main advantages of the modularity SA method, implemented in Netcarto, is that it evaluates the importance of the different nodes in terms of network connectivity within and between modules through the calculation of WMD and PC. We calculated these metrics for ME as well and found that the PC metric can detect efficient boundaries or interzones between bioregions, which is an important problem that the conventional clustering method could not properly resolve (Vilhena and Antonelli 2015). It is important to highlight that PC and WMD are calculated by Netcarto, however they are independent metrics that could be applied to any network partition method, as we did here with ME. ME could detect some of those interzones as bioregions according to the values of PC in the region. Thus, the ME method detected a larger number of small bioregions, which also include those 'interzone bioregions' at the network scale studied (i.e. natural scale of one). We think that for our two data sets the modularity SA method is more informative than map equation at the continental-scale because bioregions detected by the modularity SA method corresponds better with the Australian biomes and phytogeographical regions described by Crisp et al. (2004). Interestingly, the estimated

PC values using the ME partition overlaps pretty well with those from the modularity SA analysis. This shows that both network methods are relatively good at outlining the transition zones. Finally, the clustering method produced bioregions that expanded along regions of high PC, or produced new bioregions (clusters) within these locations. Thus, for the two datasets tested the modularity SA appears to produce the best bioregionalisation.

Our study showed that network centrality metrics are poor estimators of biodiversity metrics. However, betweenness (i.e. quantifies the number of times a node acts as a bridge along the shortest path between two other nodes) and within-module degree (WMD) seem to be a good proxy for richness and weighted endemism (WE). On one hand, this is expected since nodes with high betweenness values tend to have high degree (high number of connections or links), which means for a 'location' node to have a high number of species present or for a 'species' node to have a high number of locations (i.e. species with large ranges). On the other hand, high WMD values indicate nodes with a high number of connections within a module or bioregion. Thus, WMD is a measure of the relative number of species associated with the module that occur at a particular location. This means that WMD shows to what extent endemic taxa are present within a module and hence it is highly correlated with WE and CWE. We think these network metrics (betweenness and WMD) could be used as indirect estimators of richness and endemism in biogeographical studies. Clustering analysis is fast and convenient and has also been widely used and validated as a biogeographic tool. However, there are many choices that need to be made when using this approach, which affect the final bioregionalisation. It also currently lacks the ability to group taxa within bioregions, which is one of the most notable advantages of the network methods (Carstensen et al. 2012, Economo et al. 2015). Our results highlight the usefulness of the PC metric, which to our knowledge for the first time can quantitatively identify transition zones, and describe the soft edges between bioregions where their taxa mix. We think both network methods outperform clustering methods in the detection of biodiversity hotspots because they are more precise and efficient at detecting bioregions and quantifying species turnover across large geographical scales. This is obviously important to highlight areas of conservation interest and in the design of conservations areas.

### The network-scale problem

Network community-detection methods, such as modularity-based methods or map equation tend to suffer from a resolution limit (i.e. a scale that determines a minimum size below which communities cannot be detected) (Fortunato and Barthélemy 2007, Lancichinetti and Fortunato 2009, Kawamoto and Rosvall 2015). Specifically, the resolution limit depends on the total number of links in the system for modularity and on the number of links between modules for the map equation (Kawamoto and Rosvall 2015). In the case of the two continental-scale data sets studied here we found that there is not a specific network scale that defines the number of communities for both network methods (Supplementary material Appendix 1). More specifically, in the case of modularity at a natural scale of one, there were fewer large modules. However, at the same scale map equation detected many small modules (Supplementary material Appendix 1). Although, we do not find a specific network scale that stands out for any of the methods, different results could be expected when using other data sets. For example, map equation could detect a more instructive scale for the bioregionalisation of other taxa. We recommend evaluating the network scale problem when applying and comparing network community-detection algorithms.

### Conclusions

We conclude that network methods, particularly in conjunction with PC, improve current methods of bioregionalisation for continental-scale data and ultimately our understanding of biodiversity. Our network bioregionalisation analysis of Acacia and eucalypts have confirmed the general outcomes of previous studies (Crisp et al. 2004, Gonzalez-Orozco et al. 2013, 2014b), but more importantly, it has generated new insights about Australian biodiversity patterns. Future directions in the application of network methods to biogeography should include the possibility of considering evolutionary history (i.e. phylogenetic information) (Rosauer et al. 2009) to better reflect spatial patterns of biodiversity.

*Acknowledgements* – We would like to thank the CSIRO Summer Scholar program, of which this work was a part. FEV would like to thank the CSIRO OCE Postdoctoral Fellowship program for funding support. We would also like to thank Daniel Carstensen for insightful comments that greatly improved the manuscript, as well as Carlos E. González-Orozco, Alexander Schmidt-Lebuhn, Joe Miller, Andrew Thornhill and Shawn Laffan for insightful discussions.

## References

Bastian, M. et al. 2009. Gephi: an open source software for exploring and manipulating networks. – International AAAI Conference on Weblogs and Social Media.

Bohlin, L. et al. 2014. Community detection and visualization of networks with the map equation framework. – In: Ding, Y. et al. (eds), Measuring scholarly impact. – Springer.

Bui, E. N. et al. 2014. Salt- and alkaline-tolerance are linked in Acacia. – Biol. Lett. 10: 20140278.

Butchart, S. H. M. et al. 2010. Global biodiversity: indicators of recent declines. – Science 328: 1164–1168.

Carstensen, D. W. and Olesen, J. M. 2009. Wallacea and its nectarivorous birds: nestedness and modules. – J. Biogeogr. 36: 1540–1550.

Carstensen, D. W. et al. 2012. Biogeographical modules and island roles: a comparison of Wallacea and the West Indies. – J. Biogeogr. 39: 739–749.

Carstensen, D. et al. 2013a. Introducing the biogeographic species pool. – Ecography 36: 1310–1318.
Carstensen, D. et al. 2013b. The functional biogeography of species: biogeographical species roles of birds in Wallacea and the West Indies. – Ecography 36: 1097–1105.
Crisp, M. et al. 2004. Radiation of the Australian flora: what can comparisons of molecular phylogenies across multiple taxa tell us about the evolution of diversity in present-day communities? – Phil. Trans. R. Soc. B 359: 1551–1571.
Csardi, G. and Nepusz, T. 2006. The igraph software package for complex network research. – InterJournal Complex Syst. 1695.
Dalsgaard, B. et al. 2014. Determinants of bird species richness, endemism, and island network roles in Wallacea and the West Indies: is geography sufficient or does current and historical climate matter? – Ecol. Evol. 4: 4019–4031.
Daru, B. H. and le Roux, P. C. 2015. Marine protected areas are insufficient to conserve global marine plant diversity. – Global Ecol. Biogeogr. 25: 324–334.
Dormann, C. F. and Strauss, R. 2014. A method for detecting modules in quantitative bipartite networks. – Methods Ecol. Evol. 5: 90–98.
Ebach, M. C. et al. 2015. A revised area taxonomy of phytogeographical regions within the Australian bioregionalisation atlas. – Phytotaxa 208: 264–277.
Economo, E. P. et al. 2015. Breaking out of biogeographical modules: range expansion and taxon cycles in the hyperdiverse ant genus Pheidole. – J. Biogeogr. 42: 2289–2301.
Encinas-Viso, F. et al. 2016. Plant-mycorrhizal fungus co-occurrence network lacks substantial structure. – Oikos 125: 457–467.
Fortunato, S. and Barthélemy, M. 2007. Resolution limit in community detection. – Proc. Natl Acad. Sci. USA 104: 36–41.
Gonzalez-Orozco, C. E. et al. 2013. A biogeographical regionalization of Australian Acacia species. – J. Biogeogr. 40: 2156–2166.
Gonzalez-Orozco, C. E. et al. 2014a. Quantifying phytogeographical regions of australia using geospatial turnover in species composition. – PloS One 9: e92558.
Gonzalez-Orozco, C. E. et al. 2014b. Biogeographical regions and phytogeography of the eucalypts. – Divers. Distrib. 20: 46–58.
Guimera, R. and Amaral, L. A. N. 2005. Functional cartography of complex metabolic networks. – Nature 433: 895–900.
Guimera, R. et al. 2004. Modularity from fluctuations in random graphs and complex networks. – Phys. Rev. E 70: 025101.
Kawamoto, T. and Rosvall, M. 2015. Estimating the resolution limit of the map equation in community detection. – Phys. Rev. E 91: 012809.
Kougioumoutzis, K. et al. 2014. Network biogeographical analysis of the central Aegean archipelago. – J. Biogeogr. 41: 1848–1858.
Kreft, H. and Jetz, W. 2010. A framework for delineating biogeographical regions based on species distributions. – J. Biogeogr. 37: 2029–2053.
Laffan, S. et al. 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. – Ecography 33: 643–647.
Laffan, S. W. et al. 2016. Range-weighted metrics of species and phylogenetic turnover can better resolve biogeographic transition zones. – Methods Ecol. Evol. 7: 580–588.
Lancichinetti, A. and Fortunato, S. 2009. Community detection algorithms: a comparative analysis. – Phys. Rev. E 80: 056117.
Milligan, G. W. and Cooper, M. C. 1987. Methodology review: clustering methods. – Appl. Psychol. Meas. 11: 329–354.
Mishler, B. D. et al. 2014. Phylogenetic measures of biodiversity and neo- and paleo-endemism in Australian Acacia. – Nat. Commun. 5: 4473.
Nagalingum, N. S. et al. 2014. Overlapping fern and bryophyte hotspots: assessing ferns as a predictor of bryophyte diversity. – J. Plant Syst. 17: 383–392.
Newman, M. E. J. and Girvan, M. 2004. Finding and evaluating community structure in networks. – Phys. Rev. E 69: 026113.
Oksanen, J. et al. 2015. vegan: community ecology package. – R package ver. 2.3-2.
Olesen, J. M. et al. 2007. The modularity of pollination networks. – Proc. Natl Acad. Sci. USA 104: 19891–19896.
Rosauer, D. et al. 2009. Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. – Mol. Ecol. 18: 4061–4072.
Rosvall, M. and Bergstrom, C. T. 2008. Maps of random walks on complex networks reveal community structure. – Proc. Natl Acad. Sci. USA 105: 1118–1123.
Rosvall, M. et al. 2009. The map equation. – Eur. Phys. J. 178: 13–23.
Sales-Pardo, M. et al. 2007. Extracting the hierarchical organization of complex systems. – Proc. Natl Acad. Sci. USA 104: 15224–15229.
Schaub, M. T. et al. 2012. Markov dynamics as a zooming lens for multiscale community detection: non clique-like communities and the field-of-view limit. – PLoS One 7: e32210.
Schmidt-Lebuhn, A. N. et al. 2012. Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. – J. Biogeogr. 39: 2072–2080.
Thébault, E. 2013. Identifying compartments in presence–absence matrices and bipartite networks: insights into modularity measures. – J. Biogeogr. 40: 759–768.
Tuomisto, H. 2010. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. – Ecography 33: 2–22.
Vilhena, D. A. and Antonelli, A. 2015. A network approach for identifying and delimiting biogeographical regions. – Nat. Commun. 6: 6848.
Whittaker, R. J. et al. 2005. Conservation biogeography: assessment and prospect. – Divers. Distrib. 11: 3–23.
Wickham, H. 2009. ggplot2: elegant graphics for data analysis. – Springer.

Supplementary material (Appendix ECOG-02596 at <www.ecography.org/appendix/ecog-02596>). Appendix 1.