WILEY-BLACKWELL

Cladistics 28 (2012) 645–654

Cladistics

10.1111/j.1096-0031.2012.00410.x

Endemicity analysis, parsimony and biotic elements: a formal comparison using hypothetical distributions

M. Dolores Casagranda*, Leila Taher† and Claudia A. Szumik

Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto Superior de Entomología, Facultad de Ciencias Naturales, Miguel Lillo 205, 4000, S.M. de Tucumán, Argentina

Accepted 2 May 2012

Abstract

There is as yet no general agreement regarding the most appropriate solution to the problem of identifying areas of endemism, not even in particular cases. In this study, we compared Endemicity Analysis (EA), Parsimony Analysis of Endemicity (PAE), and Biotic Elements Analysis (BE) based on their ability to identify hypothetical predefined patterns that represent nested, overlapping, and disjoint areas of endemism supported by species with different degrees of sympatry. We found that PAE performs poorly when applied to patterns that either overlap with each other or are supported by species with imperfect sympatry. BE exhibits a counterintuitive sensitivity to the degree of congruence among the distributions of endemic species, being unable to recognize areas of endemism supported by perfectly sympatric species. In contrast, in all cases examined we found that EA results in a high proportion of correctly identified distributional patterns. In addition to highlighting the strengths and limitations of these approaches, our results show how different methods can lead to seemingly conflicting conclusions and caution about the possibility of identifying distributional patterns that are merely methodological artefacts.

Current methods for identifying areas of endemism can be classified on the basis of whether they aim to determine (i) species patterns, i.e. groups of species with overlapping distributions, or (ii) geographical patterns, i.e. groups of area units with similar species composition. These approaches assess closely related but slightly different aspects of biogeographical data. Species patterns methods group species with similar distributions and result in clusters of species which may or may not define obvious spatial patterns, while geographical patterns methods are more related to the classical notion of area of endemism, resulting in geographical areas defined by species distributions. Despite this fundamental difference, the different methods are usually applied to address the same problem (e.g. Moline and Linder, 2006).

*Corresponding author:
E-mail address: dolores.casagranda@gmail.com
†Present address: Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

Here, we analyse the performance of three current methods to identify areas of endemism: Endemicity Analysis (EA; Szumik et al., 2002; Szumik and Goloboff, 2004), Parsimony Analysis of Endemicity (PAE; Morrone, 1994), and Biotic Elements Analysis (BE; Hausdorf and Hennig, 2003). While the first two methods are based on an area pattern approach, the last named follows a species pattern approach. We selected PAE as a representative of hierarchical methods because, although several alternatives have been proposed (see Linder, 2001; García-Barros et al., 2002), it remains the most widely used in empirical analyses (e.g. Cracraft, 1991; Geraads, 1998; De Grave, 2001; Aguilar-Aguilar et al., 2003; Contreras-Medina et al., 2007; Cabrero-Sañudo and Lobo, 2009). We decided not to include in this analysis the more recent network analysis method (NAM) (Dos Santos et al., 2008) because its theoretical and operational limitations have been described elsewhere (Casagranda et al., 2009a).

Endemicity analysis, PAE, and BE have been recently compared based on real distributional data (see Moline

and Linder, 2006; Carine et al., 2009; Casazza and Minuto, 2009); however, real datasets provide only a limited assessment of the differences among the procedures. Some characteristics of the distribution of species (e.g. geographical shape, number of records) affect pattern recognition in uncertain ways. Furthermore, sampling bias, which often affects available distributional data, causes problems in the identification of biogeographical patterns (see, for example, Hortal et al., 2007). As it is often difficult to distinguish whether the identified patterns result from singularities of the data or properties of the methods, an evaluation based on real datasets—or data simulated under realistic conditions—is insufficient to establish general conclusions on the performance of the methods.

The main purpose of this work is to highlight the differences between alternative methods in the analysis of areas of endemism and identify some of their practical limitations, in order to uncover potential pitfalls associated with their application. Here, we evaluate the performance of EA, PAE and BE on a collection of hypothetical species distributions designed to recreate specific patterns that might be observed in nature. These hypothetical distributions serve as a reference result and allow for a fair comparison of the methods. Furthermore, the use of schematic species distributions provides a standardized means to recognize and illustrate the specific problems of each method.

Methods for the identification of areas of endemism (compared)

Parsimony analysis of endemicity

Parsimony analysis of endemicity was the first method proposed to formally identify areas of endemism (Morrone, 1994). The input data for PAE consist of a binary matrix in which the presence of a given species (rows) in an area unit (columns) is coded as 1 and its absence as 0. Analogous to a cladistic analysis, PAE hierarchically groups area units (analogous to taxa) based on their shared species (analogous to characters) according to the maximum-parsimony criterion. Therefore, PAE attempts to minimize both "dispersion events" (parallelisms) and "extinctions" (secondary reversions) of species within a given area. Areas of endemism are defined from the most-parsimonious tree (or strict consensus) as groups of area units supported by two or more "synapomorphic species" (i.e. endemic species; see Morrone, 1994). In its most classical formulation, species that present reversions (i.e. are absent in any of the area units) and/or parallelisms (i.e. are present elsewhere) in their distributions are not considered endemic. Therefore, in contrast to the two other methods discussed here, PAE is especially strict when penalizing the absence of a species within an area, which makes it more likely to fail to detect a relatively large number of areas of endemism. To allow for a more equitable comparison, when using PAE we considered a species endemic to an area even if it is absent in up to 40% of its cells. The parsimony analyses presented here were performed with TNT (Goloboff et al., 2008).

Despite the known limitations of hierarchical classification models in the delimitation of areas of endemism (Szumik et al., 2002; Aagesen et al., 2009; Arias et al., 2010), PAE remains the most widely used method for describing biogeographical patterns (e.g. Pizarro-Araya and Jeréz, 2004; Contreras-Medina et al., 2007; Cabrero-Sañudo and Lobo, 2009).

Endemicity analysis

In 2004, Szumik and colleagues proposed an optimality criterion to identify areas of endemism by explicitly assessing the congruence among species distributions, implemented in NDM/VNDM (Szumik et al., 2002; Goloboff, 2004; Szumik and Goloboff, 2004). The congruence between a species distribution and a given area is measured by an Endemicity Index (EI) ranging from 0 to 1. The EI is 1 for species that are uniformly distributed in the area under study, and only within that area ("perfect endemism"), and decreases for species that are present elsewhere, and/or poorly distributed within the area. In turn, the endemicity value of an area (EIA) is calculated as the sum of the EIs of the endemic species included in the area. Therefore, two factors contribute to the EIA: the number of species included in the area and the degree of congruence (measured by the EI) between the species distributions and the area itself (for details see Szumik and Goloboff, 2004).

Biotic element analysis

Hausdorf (2002) considers areas of endemism in the context of the vicariance model, and argues for the use of "biotic elements" defined as "groups of taxa whose ranges are significantly more similar to each other than to those of taxa of other such groups" (Hausdorf, 2002, p. 651), rather than the more traditional areas of endemism (Hausdorf and Hennig, 2003). His method is implemented in the R package *Prabclus* (Hennig, 2003). *Prabclus* calculates a Kulczynski dissimilarity matrix (Shi, 1993) between pairs of species which is then reduced using a non-metric multidimensional scaling (NMDS; Kruskal, 1964). A Model-Based Gaussian clustering (MBGC) is applied to this matrix to identify clusters of species with similar distributions, or biotic elements. In spatial terms, a biotic element is equivalent to the spatial extent of the distributions of all species included in the cluster.

Finally, we recognize that using alternative implementations of the aforementioned methods may lead to

differences in results. However, for convenience and clarity, in our discussion and conclusions will refer simply to PAE, EA and BE.

Hypothetical distribution patterns

Although many patterns of sympatry between species are obvious to the naked eye, their description is often conceptually and computationally challenging. To observe differences in the performance of the methods under study, we designed three basic hypothetical classes of areas of endemism that could be found in nature, but are particularly difficult to identify using current computational methods. We constructed examples of nested (Fig. 1a), overlapping (Fig. 1b) and disjoint areas (Fig. 1c), which we refer to as cases 1, 2 and 3, respectively. As reference, we chose a non-conflictive pattern (case 0; Fig. 1d). Also, because sympatry among species within a given area of endemism is variable in nature—and it is unlikely to find species with strictly identical distributional ranges—we analysed areas of endemism defined by species with different degrees of sympatry as subcases of the above-mentioned examples. For each case, we then considered further subcases, where we systematically modified the distributions of the species by randomly adding presences outside, but adjacent to, the predefined area, eliminating presences within the area, or both. We refer to the incongruence between a species distribution and the area predefined in the examples as noise, and to the different subcases as internal, adjacent and mixed noise, respectively. Subcases with no noise were defined by areas of endemism supported by species distributions that are perfectly congruent to the corresponding areas (Fig. 2). To construct subcases with internal, adjacent, and mixed noise, we modified species distributions that were perfectly congruent to the area of endemism by randomly removing or adding from one to four cells. All examples were constructed using 18 species distributed on a spatial matrix of 100 cells (10 × 10). We considered hypothetical areas of endemism supported by different numbers of endemic species (see Table 1 for details).

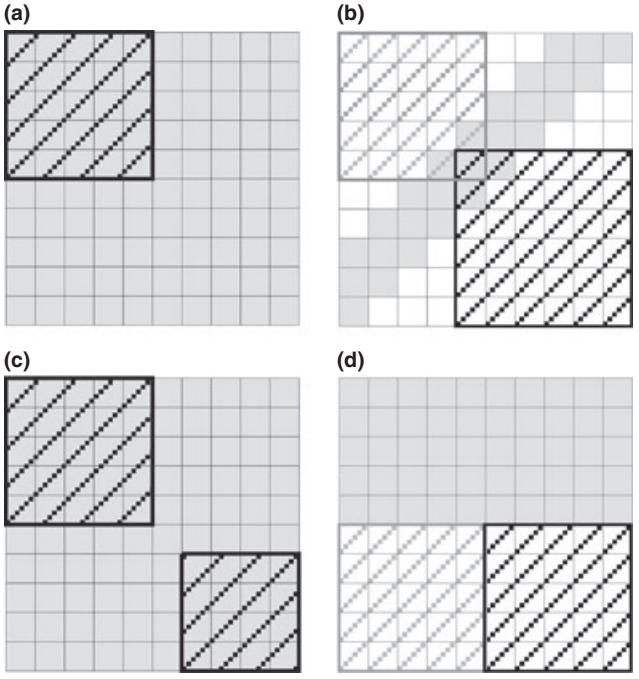

Fig. 1. Hypothetical cases of distribution. (a) Case 1, *Nested areas*: a small area of endemism defined by a unique set of endemic species (striped area) is nested in a bigger area defined by a different and singular group species (solid grey area). (b) Case 2, *Overlapping areas*: three spatially overlapping areas of endemism (in solid grey, striped-grey and striped-black), each defined by a particular set of endemic species. (c) Case 3, *Disjoint areas:* a disjoint area defined by a particular set of endemic species (striped area) is nested in a bigger area of endemism (in solid grey). (d) Case 0, *Non-conflictive patterns*: three contiguous areas of endemism (in solid grey, striped-grey and striped-black), each defined by it own group of endemic species.

Performance criteria

The performance of the different methods was evaluated using two indices that quantify the congruence between predefined and identified patterns, sensitivity (I^{Sens}) and specificity (I^{Spec}).

1. I^{Sens}: measures the proportion of cells within a predefined pattern that are correctly identified by the method, i.e.

$$I^{\text{Sens}} = \frac{\text{TP}}{\text{Predefined}},$$

where, TP (true positives) is the number of cells included in both the predefined and the identified pattern. Values for I^{Sens} vary between 1 (if the identified pattern includes all the cells in the predefined pattern) and 0 (if the identified pattern includes none of the cells in the predefined patterns).

2. I^{Spec}: measures the proportion of cells within an identified pattern that are included in a predefined pattern, i.e.

$$I^{\text{Spec}} = \frac{\text{TP}}{\text{Identified}}.$$

Values for I^{Spec} vary between 1 (if all cells within the identified pattern are included in the predefined pattern) and 0 (if the identified area includes none of the cells within a predefined pattern). Values of I^{Spec} also decrease if the identified area includes a large number of false positives (FP), that is cells present in the identified area but not in the predefined one (see Fig. 3).

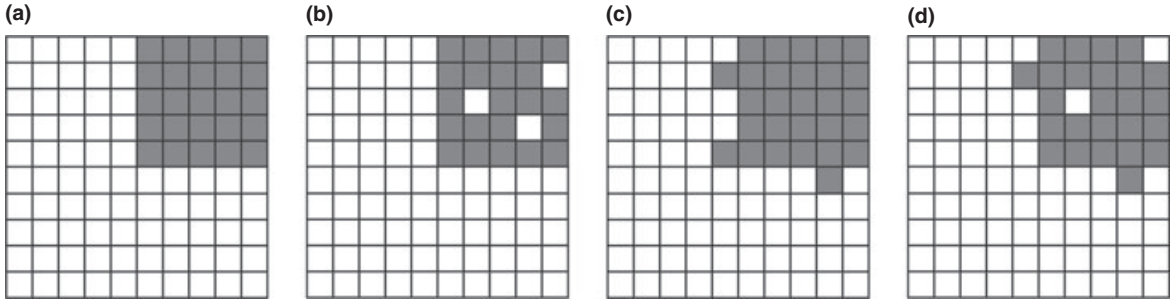

Fig. 2. Subcases. (a) Perfect species and ideal patterns: species whose distribution is perfectly congruent with a predefined area and present complete sympatry with other species. Noisy and ideal species patterns: (b) species with *inner* noise distributions: species is absent in some cells inside the area; (c) species with *adjacent* noise distributions: species is present outside the predefined area, in adjacent cells; (d) species with *mixed* noise distributions: species can be absent in some cells inside the area, and/or present in outside cells adjacent to the area.

Table 1
Details of hypothetical examples: cases, subcases, number of endemic species per area and total number of areas included by case

Case	Subcase	Number of endemic species per area	Sum of resulting areas by case
Case 0 Non-conflictive areas (3 areas: A, B and C)	No noise (total 21 areas)	A: 6, B: 6, C: 6; A: 10, B: 6, C: 2.; A:10, B: 2, C: 6; A: 6, B: 10, C: 2; A: 6, B: 2, C: 10; A: 2, B: 10, C: 6; A: 2; B: 6; C: 10	84 areas
	Internal noise (total 21 areas)	A: 6, B: 6, C: 6; A: 10, B: 6, C: 2.; A:10, B: 2, C: 6; A: 6, B: 10, C: 2; A: 6, B: 2, C: 10; A: 2, B: 10, C: 6; A: 2; B: 6; C: 10	
	Adjacent noise (total 21 areas)	A: 6, B: 6, C: 6; A: 10, B: 6, C: 2.; A:10, B: 2, C: 6; A: 6, B: 10, C: 2; A: 6, B: 2, C: 10; A: 2, B: 10, C: 6; A: 2; B: 6; C: 10	
	Mixed noise (total 21 areas)	A: 6, B: 6, C: 6; A: 10, B: 6, C: 2.; A:10, B: 2, C: 6; A: 6, B: 10, C: 2; A: 6, B: 2, C: 10; A: 2, B: 10, C: 6; A: 2; B: 6; C: 10	
Case 1 Nested areas (2 areas: A and B)	No noise (total 6 areas)	A: 9, B: 9; A: 16, B: 2; A: 2, B: 16	24 areas
	Inner noise (total 6 areas)	A: 9, B: 9; A: 16, B: 2; A: 2, B: 16	
	Adjacent noise (total 6 areas)	A: 9, B: 9; A:16, B: 2; A: 2, B: 16	
	Mixed noise (total 6 areas)	A: 9, B: 9; A: 16, B: 2; A: 2, B: 16	
Case 2 Overlapping areas (3 areas: A, B and C)	No noise (total 21 areas)	A: 6, B: 6, C: 6; A: 10, B: 6, C: 2.; A:10, B: 2, C: 6; A: 6, B: 10, C: 2; A: 6, B: 2, C: 10; A: 2, B: 10, C: 6; A: 2; B: 6; C: 10	84 areas
	Internal noise (total 21 areas)	A: 6, B: 6, C: 6; A: 10, B: 6, C: 2.; A:10, B: 2, C: 6; A: 6, B: 10, C: 2; A: 6, B: 2, C: 10; A: 2, B: 10, C: 6; A: 2; B: 6; C: 10	
	Adjacent noise (total 21 areas)	A: 6, B: 6, C: 6; A: 10, B: 6, C: 2.; A:10, B: 2, C: 6; A: 6, B: 10, C: 2; A: 6, B: 2, C: 10; A: 2, B: 10, C: 6; A: 2; B: 6; C: 10	
	Mixed noise (total 21 areas)	A: 6, B: 6, C: 6; A: 10, B: 6, C: 2.; A:10, B: 2, C: 6; A: 6, B: 10, C: 2; A: 6, B: 2, C: 10; A: 2, B: 10, C: 6; A: 2; B: 6; C: 10	
Case 3 Disjoint areas (2 areas: A and B)	No noise (total 6 areas)	A: 9, B: 9; A: 16, B: 2; A: 2, B: 16	24 areas
	Internal noise (total 6 areas)	A: 9, B: 9; A: 16, B: 2; A: 2, B: 16	
	Adjacent noise (total 6 areas)	A: 9, B: 9; A: 16, B: 2; A: 2, B: 16	
	Mixed noise (total 6 areas)	A: 9, B: 9; A: 16, B: 2; A: 2, B: 16	

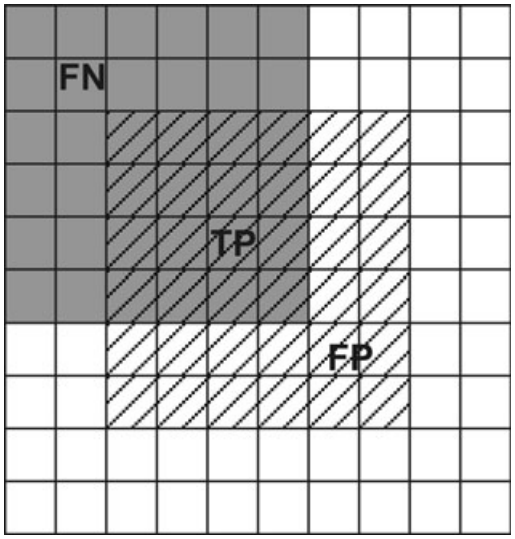

Fig. 3. True positives, false negatives and false positives. The predefined and identified areas are shown: the predefined area is defined by grey cells while the identified area is defined by striped cells. Cells included in both the predefined and the identified areas are named *true positives* (TP), while cells present in the predefined area but not in the predicted area are *false negatives* (FN). Cells present in the identified area but not included in the predefined one are *false positives* (FP).

We focused on four categories of identified patterns, which we defined according to their I^{Sens} and I^{Spec} values: correctly identified patterns ($I^{Sens} \geq 0.8$ and $I^{Spec} \geq 0.8$), partially identified patterns ($I^{Sens} \leq 0.5$ and $I^{Spec} \geq 0.8$), over-identified patterns ($I^{Sens} \geq 0.8$ and $I^{Spec} \leq 0.5$), and incorrectly identified patterns ($I^{Sens} \leq 0.5$ and $I^{Spec} \leq 0.5$) (see Fig. 4).

Results and discussion

Hypothetical distribution patterns

Case 0: non-conflictive patterns. This case illustrates three neighbouring but non-overlapping areas of endemism. From a total of 84 predefined areas, BE and PAE recognize 43 (51%) and 53 (63%) areas respectively, with only 33 (39%) and 43 (51%) being correctly identified (Fig. 5a). EA identifies all predefined areas, but also 18 additional areas that represent minor variation of these. In spite of the redundancy, all identified areas are correct in terms of species and cell composition. EA performs well across all subcases of non-conflictive patterns (no noise, internal, adjacent, and mixed noise), while PAE is ineffective at identifying noisy patterns, and, paradoxically, BE shows low success rates at identifying areas supported by perfectly sympatric species (Fig. 5b).

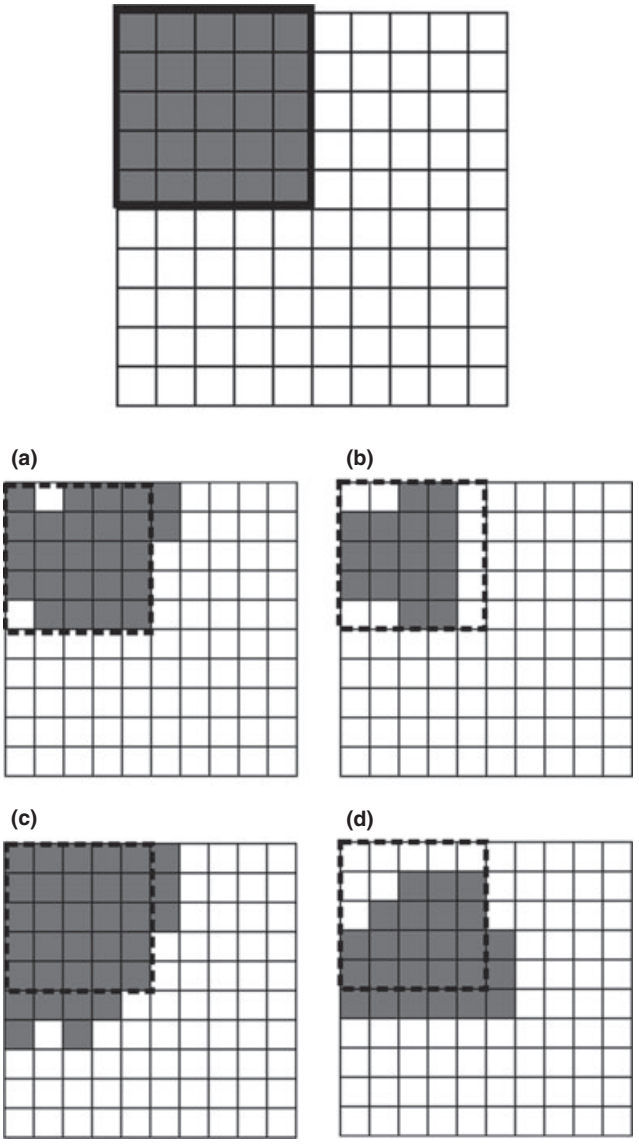

Fig. 4. Classification of identified areas according to their correspondence with a predefined area. The predefined area is shown at the top, and examples of identified areas are shown below: (a) correctly identified, (b) partially identified, (c) over-identified and (d) incorrectly identified.

Case 1: nested patterns. Endemicity analysis is the only method able to recover 100% of predefined areas. BE identifies 15 out of 24 predefined areas, and PAE only 10 out of 24 (see Fig. 6a). While EA is equally effective at recovering patterns defined by species with different degrees of sympatry, BE shows a paradoxical behaviour: it cannot identify areas defined by perfectly sympatric species. PAE recognizes a low percentage of nested patterns, showing a poor performance across all studied subcases (Fig. 6b).

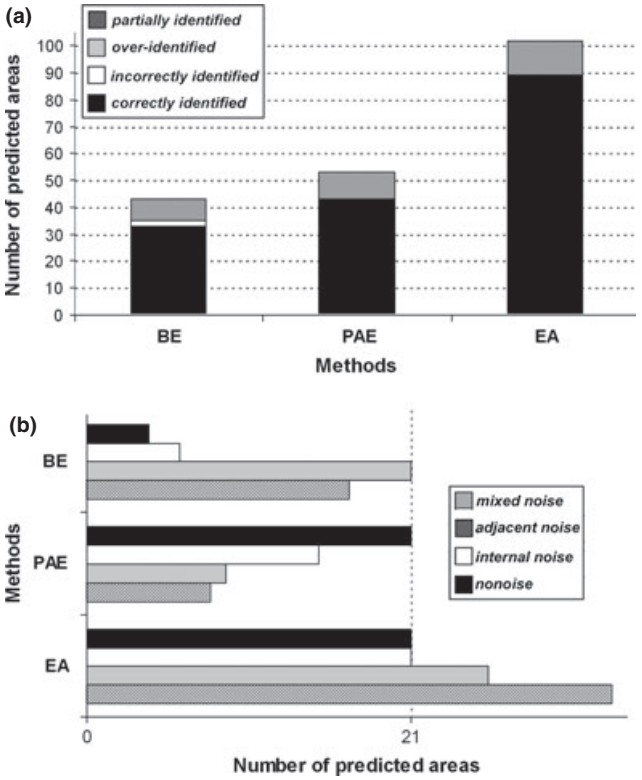

Fig. 5. Case 0: number of *Non-nonflictive* areas recognized by each method. (a) The number of correctly, partially, over-identified and incorrectly identified areas recovered by each method. (b) The number of areas recovered by subcase and by method. The dotted line indicates the total number of predefined areas.

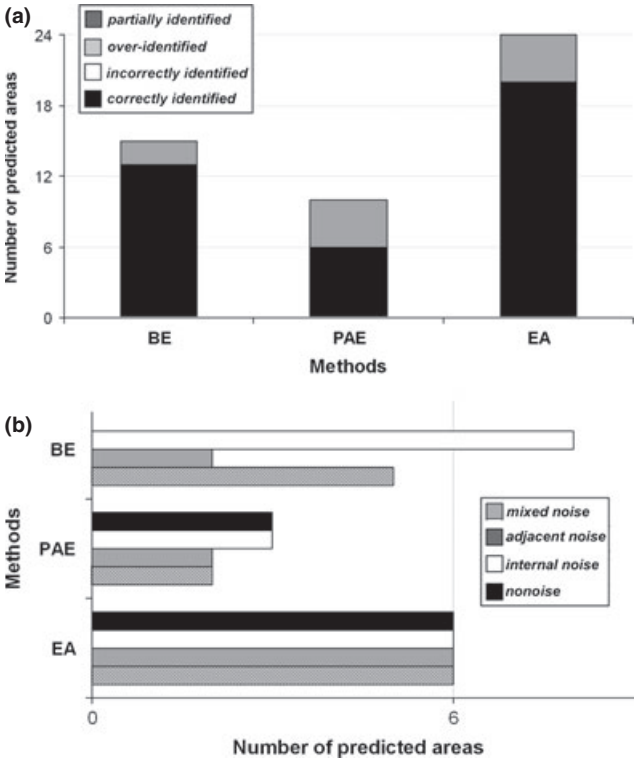

Fig. 6. Case 1: number of *Nested* areas recognized by each method. (a) The number of correctly, partially, over-identified, and incorrectly identified areas recovered by each method. (b) The number of areas recovered by subcase and by method. The dotted line indicates the total number of predefined areas.

Case 2: overlapping areas. Endemicity analysis recovers all 84 predefined areas, whereas BE and PAE recover only 48 (57%) and 25 (29%), with one out of 16 and one out of six areas being only partially identified, respectively. Similarly, as described for non-conflictive patterns, all areas identified by EA are correct. However, EA also identifies four extra areas that were not predefined, but constitute slight variations of others that were (see Fig. 7a). EA effectively recovers areas across all subcases, whereas BE fails—as in all other cases—to identify areas supported by perfectly sympatric species. PAE shows low effectiveness at recognizing areas in all subcases (see Fig. 7b).

Case 3: disjoint areas. Endemicity analysis correctly identifies 20 (83%) of 24 predefined areas, and partially identifies another four (17%; see Fig. 8a). BE identifies only 14 (58%) predefined areas, two of them only partially. PAE recovers only seven (29%) of predefined areas (Fig. 8a), exhibiting the worst performance. Also in this case, the performance of EA is not affected by noise in the species distributions, while PAE recognizes best patterns defined by perfectly sympatric species (no noise subcases). The performance of BE is not strongly affected by the increase of noise, recognizing areas with the same success in all subcases, except—as in all other cases—for those including no noise (see Fig. 8b).

Noise

Incongruence in species distributions has an evident effect on identification of areas of endemism. Paradoxically, the performance of BE decreases as the sympatry among species increases, and improves when the noise among species increases (see Fig. 9a). This is especially evident in cases 1 and 3, where this method recovered 0/6 and 1/6 areas, respectively, demonstrating a great deficiency in identifying ideal patterns, i.e. patterns defined by perfectly sympatric species, with no noise. As opposed to BE, the performance of PAE decreases with the increase of noise in the distribution of endemic species (Fig. 9b). This behaviour suggests possible limitations for identification of real patterns defined by real species with "imperfect" distributions. In contrast to the other methods, noise in distributions of species does not meaningfully affect the identification of areas of endemism when using EA (Fig. 9c). Indeed, EA correctly identified a high percentage

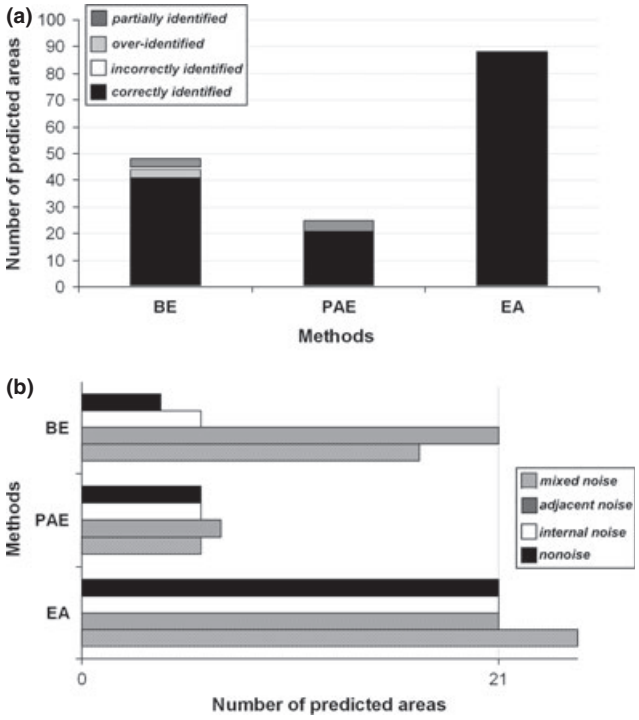

Fig. 7. Case 2: number of *Overlapping* areas recognized by each method. (a) The number of correctly, partially, over-identified and incorrectly identified areas recovered by each method. (b) The number of areas recovered by subcase and by method. The dotted line indicates the total number of predefined areas.

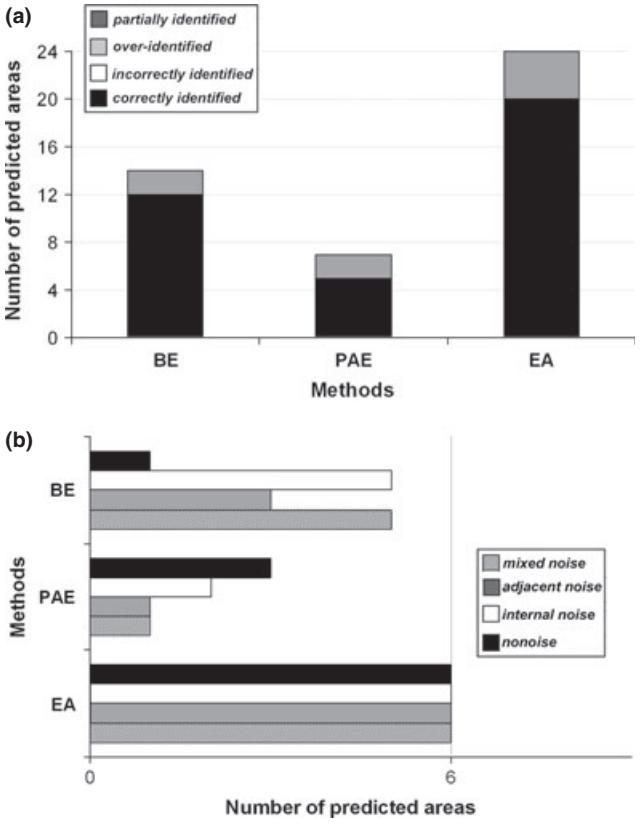

Fig. 8. Case 3: number of *Disjoint* areas recognized by each method. (a) The number of correctly, partially, over-identified and incorrectly identified areas recovered by each method. (b) The number of areas recovered by subcase and by method. The dotted line indicates the total number of predefined areas.

of predefined areas across all subcases, although in some particular cases—mixed noise subcase for cases 0 and 2—EA tended to predict a large number of redundant areas, i.e. areas that mainly constitute variations of one another. This occurs because equally optimal combinations of cells increase with noise, often resulting in twin areas, which are slight variations of a particular area.

General performance of methods

Parsimony analysis of endemicity. Our results confirm previously reported limitations of using hierarchical methods to detect areas of endemism (see Szumik et al., 2002). In particular, PAE performs poorly at identifying overlapping and disjoint patterns. In all cases considered here, PAE is able to recover areas that have been predefined using perfectly sympatric species (perfect areas), but its performance decreases as the noise in the species distribution increases—especially in subcases involving "adjacent" and "mixed" noise (Fig. 9b). Taking into account that in nature overlapping and disjoint patterns are relatively common and that, in general, sympatry between species varies widely, PAE is probably not the most suitable method to describe areas of endemism based on real distributional data.

Biotic elements analysis. BE is very sensitive to the degree of congruence among the distributions of the species that determine the area, although in a counterintuitive manner: while the method cannot recognize patterns defined by perfectly sympatric species, its performance improves with increasing levels of noise in the species distributions (see Figs 4b, 5b, 6b and 7b). Biotic elements are determined by a MBGC technique on Kulczynski distances between pairs of species assuming a Gaussian distribution among species distances. However, Kulczynski distances between species with identical ranges of distribution are distributed according to a Gaussian distribution with variance zero, a case presumably not considered by the model. Although this ideal case is not frequently observed at the spatial scale used for most biogeographical analyses, the inability to identify a perfect case of the pattern which the method intends to describe is questionable. The method produced further seemingly counterintuitive results, often reporting multiple distinct biotic elements for species which actually have very similar distributions (Fig. 10a), as well as reporting a single biotic element including species with completely

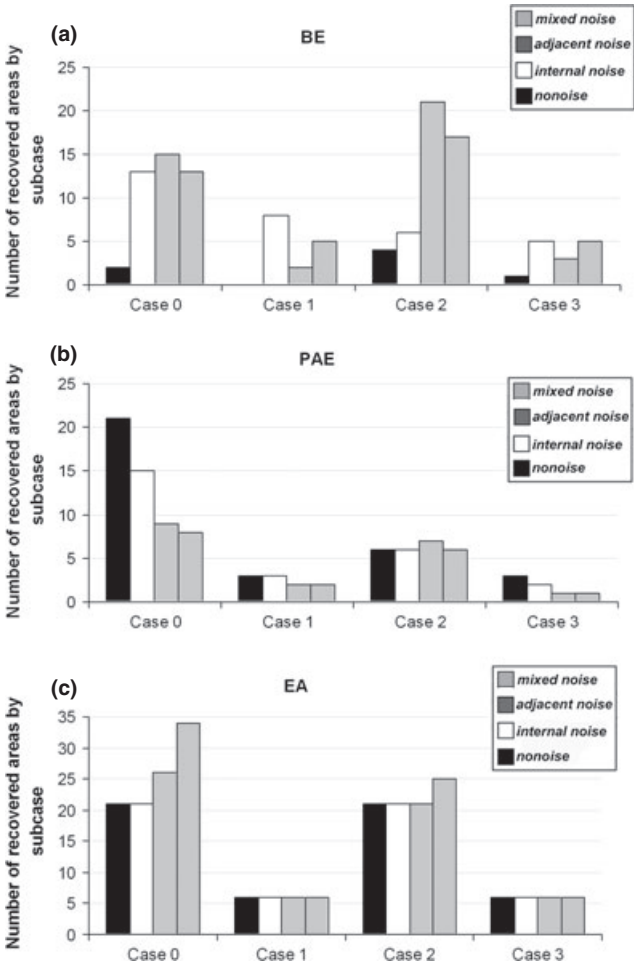

Fig. 9. Noise effects on identification of areas of endemism. The number of areas identified by each method, for different subcases: (a) results using BE, (b) results using PAE and (c) results using EA.

allopatric distributions (Fig. 10b). These examples show discordance between the theoretical basis of the approach (Hausdorf, 2002) and its practical implementation. Together, these limitations suggest the users should exercise caution when interpreting the results generated by this method.

Endemicity analysis. Endemicity analysis shows a high percentage of success in the recovery of predefined areas (Figs 4–7). Flexibility in recognition of areas displayed by EA is associated with the fact that, in contrast to the other methods considered here, EA utilizes both the number of species and the overlap between their distributions as optimality criteria to search for areas of endemism. The main problem with EA is the frequent report of redundant, ''twin'' areas that have only slight differences in spatial structure and/or in their species composition. This problem can be solved by constructing consensus areas (for some details see Aagesen et al., 2009), which merge similar areas to simplify the analysis of the results.

In this sense, this comparison shows that EA (in conjunction with consensus areas) is the best available option for endemicity analyses, despite other studies indicating that EA is rather sensitive to certain aspects of the data, such as spatial gaps of information (Arias et al., 2010). The advantages of EA over other methods are related to considering spatial information during the identification of areas, as well as using the classical definition of area of endemism as the basis for the analysis (''[an area of endemism] is identified by the congruent distributional boundaries of two or more species, where congruent does not demand complete agreement on those limits at all possible scales of mapping, but relatively extensive sympatry is a prerequisite''; Platnick, 1991).

In summary, the problems encountered with the three methods can be classified into two broad categories: those derived from the choice of an inappropriate mathematical formulation, and those resulting from multiple solutions to the optimization problem. The former are intrinsic to the methods and therefore are more difficult to solve. Simply put, the method relies on an algorithm that is ineffective for its intended purpose. PAE, for example, is a hierarchical method implying that each cell is included in at most one area of endemism—the one supported by the largest number of endemic species. Consequently, PAE cannot describe overlapping patterns, such as nested areas. Additionally, the maximum-parsimony criterion aims to minimize the number of homoplasies, resulting in PAE hardly identifying any disjoint areas. BE suffers from the same problem; BE's model-based inference requires a series of distributional assumptions which, if not satisfied, may lead to unreliable, or simply erroneous conclusions. Thus, even if in theory a biotic element is defined as a ''group of taxa whose ranges are significantly more similar to each other than to those of taxa of other such groups'', *Prabclus* may both group totally allopatric species in a single biotic element, and fail to recognize biotic elements defined by perfectly sympatric species (for an example see Fig. 10). Finally, an inescapable consequence of the application of an optimality criterion is that multiple hypotheses (in the case of EA, the ''twin'' areas representing small variations in the addition or deletion of single cells) may be obtained in an analysis. The ambiguity in the input data often results in multiple ''best'' solutions according to an optimality criterion. Although the reported alternative and equally optimal patterns often force the researcher to more conservative interpretations, this result can also be helpful in designing sampling strategies to improve subsequent analysis.

Final comments

The influence of methodology on the outcome of different studies is well explored in diverse areas of

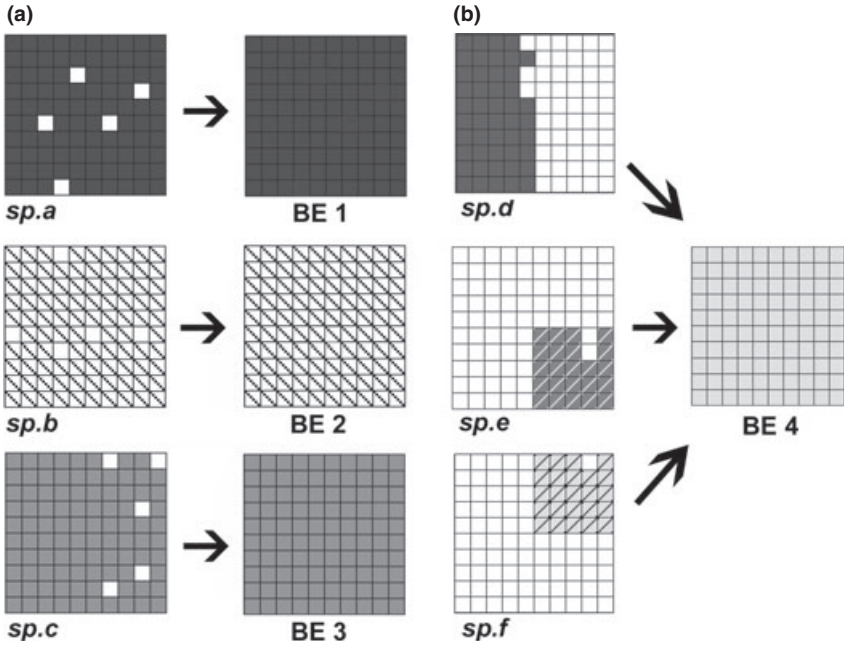

Fig. 10. Special results found by biotic elements. (a) Three species with similar distributions (*sp.a*, *sp.b*. and *sp.c*) are separated in different biotic elements (BE 1, BE 2 and BE 3); (b) three species with completely allopatric distributions (*sp.d*, *sp.e*. and *sp.f*) are grouped in the same biotic element (BE 4).

biology (see, for example, Hall, 2005; : Kosakovsky Pond and Frost, 2005; Peterson et al., 2007; Bombi et al., 2011) but, except for a few papers, it is poorly discussed in biogeography (Morrone and Carpenter, 1994; Brooks and McLennan, 2001; Moline and Linder, 2006; Casazza and Minuto, 2009; Buerki et al., 2011). The comparison conducted here shows how the application of different analytical methods can lead to identification of different areas of endemism, and reveals some undesirable effects produced by methodological idiosyncrasies in the description of these patterns. Our results caution about the possibility of finding distributional patterns which are merely methodological artefacts (for example see Fig. 10), highlighting the importance of methodological choice when analysing data. Although several aspects of the methods for identification of areas of endemism remain poorly explored, here we have summarized their main traits and underlined some strengths and weakness, to help provide an adequate methodological decision.

Some properties of distributional data (shape and extension of distributions, sample size, sample bias, etc.) can influence the identification of biogeographical patterns (e.g. Hortal et al., 2007; Casagranda et al., 2009b). Yet, it remains challenging to estimate how exactly properties of real distributional data—or distributional data simulated under realistic conditions—will affect the identification of biogeographical patterns. Consequently, the use of such data to compare the performance of methods to identify biogeographical patterns may introduce bias unrelated to the actual methods, invalidating the conclusions. Alternatively, simple examples constitute a powerful option to address the complex problem of recognizing areas of endemism, where a large number of factors can influence the results. Using this approach allows us to avoid the potential confusion caused by simultaneous effects of several variables and helps to analyse the impact of choice of method.

Analyses similar to those performed here and aimed at exploring other properties of the data, such as the shape of the distributions, the number of endemic species or sample bias, would lead us to a deeper understanding of the behaviour of proposed algorithms and the consequences of their application on data of different configurations. Given the increasing information on species distributions, such a contribution would be fundamental to methodological improvements for more robust and realistic descriptions of biogeographical patterns. This should be a main issue of current biogeography: to contribute to efficient decision-making in the conservation of biological diversity.

Acknowledgements

This research was supported by CONICET and FONCyT (PICT07 1314). We thank INSUE for provision of

workspace and equipment. A preliminary version of this study was presented at the VII Reunión Argentina de Cladística y Biogeografía (San Isidro, 2007); we would like to thank the organizers and our colleagues for helpful discussions and comments. Special thanks to J.M. Morales, P. Goloboff, and S. Catalano for their useful and constructive remarks. M.D.C. also thanks S. Arias and M. Mirande for discussion and support. We thank the Willi Hennig Society for subsidizing the program TNT and making it freely available. Finally, we greatly appreciate the comments from three anonymous reviewers which substantially improved the manuscript.

References

Aagesen, L., Szumik, C., Zuloaga, F.O., Morrone, O., 2009. Quantitative biogeography in the South America highlands—recognizing the Altoandina, Puna and Prepuna through the study of Poaceae. Cladistics 25, 295–310.

Aguilar-Aguilar, R., Contreras-Medina, R., Salgado-Maldonado, G., 2003. Parsimony analysis of endemicity (PAE) of Mexican hydrological basins based on helminth parasites of freshwater fishes. J. Biogeogr. 30, 1861–1872.

Arias, J.S., Casagranda, M.D., Diaz Gómez, J.M., 2010. A comparison of NDM and PAE using real data. Cladistics 26, 204 (abstract).

Bombi, P., Luiselli, L., D'Amen, M., 2011. When the method for mapping species matters: defining priority areas for conservation of African freshwater turtles. Divers. Distrib. 17, 581–592.

Brooks, D.R., McLennan, D.A., 2001. A comparison of a discovery-based and an event-based method of historical biogeography. J. Biogeogr. 28, 757–767.

Buerki, S., Forest, F., Alvarez, N., Nylander, J.A.A., Arrigo, N., Sanmartin, I., 2011. An evaluation of new parsimony-based versus parametric inference methods in biogeography: a case study using the globally distributed plant family Sapindaceae. J. Biogeogr. 38, 531–550.

Cabrero-Sañudo, F.J., Lobo, J.M., 2009. Biogeography of Aphodiinae dung beetles based on the regional composition and distribution patterns of genera. J. Biogeogr. 36, 1474–1492.

Carine, M.A., Humphries, C.J., Guma, I.R., Reyes-Betancort, J.A., Santos Guerra, A., 2009. Areas and algorithms: evaluating numerical approaches for the delimitation of areas of endemism in the Canary Islands archipelago. J. Biogeogr. 36, 593–611.

Casagranda, M.D., Arias, J.S., Goloboff, P.A., Szumik, C., Taher, L.M., Escalante, T., Morrone, J.J., 2009a. Proximity, interpenetration, and sympatry networks: a reply to Dos Santos et al. Syst. Biol. 58, 271–276.

Casagranda, M.D., Roig-Juñent, S., Szumik, C., 2009b. Endemismo a diferentes escalas espaciales: un ejemplo con Carabidae (Coleoptera: Insecta) de América del Sur austral. Rev. Chil. Hist. Nat. 82, 17–42.

Casazza, G., Minuto, L., 2009. A critical evaluation of different methods for the determination of areas of endemism and biotic elements: an Alpine study. J. Biogeogr. 36, 2056–2065.

Contreras-Medina, R., Luna Vega, I., Morrone, J.J., 2007. Application of parsimony analysis of endemicity to Mexican gymnosperm distributions: grid-cells, biogeographical provinces and track analysis. Biol. J. Linn. Soc. 92, 405–417.

Cracraft, J., 1991. Patterns of diversification within continental biotas: hierarchical congruence among the areas of endemism of Australian vertebrates. Aust. Syst. Bot. 4, 211–227.

De Grave, S., 2001. Biogeography of Indo-Pacific Potoniinae (Crustacea, Decapoda): a PAE analysis. J. Biogeogr. 28, 1239–1254.

Dos Santos, D.A., Fernández, H.R., Cuezzo, M.G., Domínguez, E., 2008. Sympatry inference and network analysis in biogeography. Syst. Biol. 57, 432–448.

García-Barros, E., Gurrea, P., Lucañez, M., Cano, J., Munguira, M., Moreno, J., Sainz, H., Sanz, M., Simón, J.C., 2002. Parsimony analysis of endemicity and its application to animal and plant geographical distributions in the Ibero-Balearic region (western Mediterranean). J. Biogeogr. 29, 109–124.

Geraads, D., 1998. Biogeography of circum-Mediterranean Miocene-Pliocene rodents; a revision using factor analysis and parsimonious analysis of endemicity. Palaeogeogr. Palaeoclimatol. Palaeoecol. 137, 273–288.

Goloboff, P.A. 2004. NDM/VNDM. Programs for identification of areas of endemism. Program and documentation. Available at: http://www.zmuc.dk/public/phylogeny/endemism/

Goloboff, P.A., Farris, J.S., Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

Hall, B.G., 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol. Biol. Evol. 22, 792–802.

Hausdorf, B., 2002. Units in biogeography. Syst. Biol. 51, 648–651.

Hausdorf, B., Hennig, C., 2003. Biotic element analysis in biogeography. Syst. Biol. 52, 717–723.

Hennig, C., 2003. *Prabclus Package*, test for clustering of presence-absence data. Available at: http://cran.r-project.org/src/contrib/Descriptions/prabclus/.html

Hortal, J., Lobo, J.M., Jiménez-Valverde, A., 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. Conserv. Biol. 21, 853–863.

Kosakovsky Pond, S.L., Frost, S.D., 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22, 1208–1222.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a Nonmetric hypothesis. Psychometrika 29, 1–27.

Linder, H.P., 2001. On areas of endemism, with an example from the African Restionaceae. Syst. Biol. 50, 892–912.

Moline, P.M., Linder, H.P., 2006. Input data, analytical methods and biogeography of *Elegia* (Restionaceae). J. Biogeogr. 33, 47–62.

Morrone, J.J., 1994. On the identification of areas of endemism. Syst. Biol. 43, 438–441.

Morrone, J.J., Carpenter, J.M., 1994. In the search of a method for cladistic biogeography: an empirical comparison of component analysis, Brooks parsimony analysis, and three-area statements. Cladistics 10, 99–153.

Peterson, T., Papes, M., Eaton, M., 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30, 550–560.

Pizarro-Araya, J., Jerez, V., 2004. Distribución geográfica del género Gyriosomus Guérin-Méneville, 1834 (Coleoptera: Tenebrionidae): una aproximación biogeográfica. Rev. Chil. Hist. Nat. 77, 491–500.

Platnick, N.I., 1991. On areas of endemism. Austral. Syst. Bot. 4, i–ii.

Shi, G.R., 1993. Multivariate data analysis in palaeoecology and palaeobiogeography – a review. Palaeogeogr. Palaeoclimatol. Palaeoecol. 105, 199–234.

Szumik, C., Goloboff, P.A., 2004. Areas of endemism: an improved optimality criterion. Syst. Biol. 53, 968–977.

Szumik, C., Cuezzo, F., Goloboff, P.A., Chalup, A., 2002. An optimality criterion to determine areas of endemism. Syst. Biol. 51, 806–816.