Syst. Biol. 69(1):76–90, 2020

DOI:10.1093/sysbio/syz033
Advance Access publication May 24, 2019

Statistical Modeling of Distribution Patterns: A Markov Random Field Implementation and Its Application on Areas of Endemism

Nelson R. Salinas[1,2,*] and Ward C. Wheeler[1]

[1]*Division of Invertebrate Zoology, American Museum of Natural History, New York City, NY 10024, USA; and* [2]*Instituto de Hidrología, Meteorología y Estudios Ambientales IDEAM, Calle 25D #96B–70, Bogotá D.C., Colombia*
**Correspondence to be sent to: Division of Invertebrate Zoology, American Museum of Natural History, New York City, NY 10024, USA; E-mail: nrsalinas@gmail.com.*

Received 17 May 2017; reviews returned 28 July 2018; accepted 4 May 2019
Associate Editor: Richard Glor

Abstract.—A statistical framework to infer areas of endemism from geographic distributions is proposed. This novel method is based on hidden Markov random fields (HMRFs), a type of undirected graph model commonly used in computer vision. This framework assumes areas of endemism are the states of the hidden layer of the model, whereas taxon distributions are emitted values in the observed layer. Taxon distributions are associated to the observed layer through a clustering procedure based on the extent of overlap. Observations are emitted by the hidden layer according to a Gaussian distribution, whereas the joint distribution of the hidden layer follows a Potts model. State and parameter inference of the *maximum a posteriori* configuration is performed through a modified version of the expectation-maximization algorithm. The optimal number of areas of endemism in the data set is estimated through the pseudolikelihood information criterion, a model selection procedure that uses an approximation to likelihood. The performance of the new algorithm was assessed on simulated data, and compared with the most popular methods for delimitation of areas of endemism: biotic element analysis, parsimony analysis of endemism, and endemicity analysis. HMRFs efficiently recovered the true pattern across a wide range of uncertainty values. The performance was also examined on empirical data: South African weevils (*Sciobius*) and Central American ground beetles and funnel-web tarantulas (Carabidae and Dipluridae, respectively). HMRFs uncovered six areas of endemism from the weevil data set, whereas eight were estimated for the Central American arthropods (compared with 3–5 and 3–14 from the other methods, respectively). [Areas of endemism; biogeography; geographic distributions; hidden Markov random fields.]

Areas of endemism have been a fundamental concept in systematic biology and have played an important role in the development of historical biogeography. They have been traditionally considered geographic regions in which extensive distributional congruence among two or more taxa exists (Nelson and Platnick 1981; Platnick 1991). Research on areas of endemism rose among systematists three decades ago, as they were largely deemed the logical operational units in vicariance biogeography (Henderson 1991; Platnick 1991). Ever since, their utility has been expanded beyond historical biogeography, and they have being used to design biogeographic regionalization schemes (Morrone 2014a), suggest regions of interest for conservation (Martínez-Hernández et al. 2015), or even study organism-climate dynamics (Gámez et al. 2014). Although there are still critical disputes regarding their conceptual definition and philosophical foundations (e.g., Casagranda et al. 2009; Crother and Murray 2013; Morrone 2014b), they are still a popular and effective way to describe spatial biological patterns.

Several computational methods have been proposed in the last 20 years to identify areas of endemism or analogous patterns of geographic distribution (Morrone 1994; Szumik et al. 2002; Hausdorf and Hennig 2003). Although they all have different assumptions and design, none can incorporate distributional uncertainty into their calculations. Such a limitation is very important, since geographic distributions are non-deterministic by nature. That is, they cannot be predicted with absolute certainty even if all possible information regarding the organism and the system associated with it were known (Real et al. 2017). A detailed description of uncertainty levels pertaining to ecological processes has been already presented by Regan et al. (2002), and it will not be repeated here. However, it is worth mentioning that some of the sources cited therein are particularly relevant regarding distributional data: measurement errors, either randomly distributed or systematically biased (e.g., insufficient sampling, taxonomic identification, geographic coordinates retrieval and manipulation); stochastic variation due to the interaction of deterministic processes that cannot be completely accounted for (e.g., individual-level interactions that affect population-level processes), or well-known mechanisms that are inherently random (e.g., molecular phenomena); as well as model misspecification.

The challenge of incorporating uncertainty can be overcome with the implementation of a statistical model that facilitates the inference of areas of endemism from imperfect distribution data. Here, we employ hidden Markov random fields (HMRFs) as an appropriate solution to this issue. We first present an introduction to this class of models, describing their architecture and probability computation techniques. This serves as the basis for the following section, where a computational framework to identify areas of endemism through HMRF is developed. Performance of this method is examined on both simulated and empirical data, and compared with that of popular algorithms for delimitation of areas of endemism. It is important

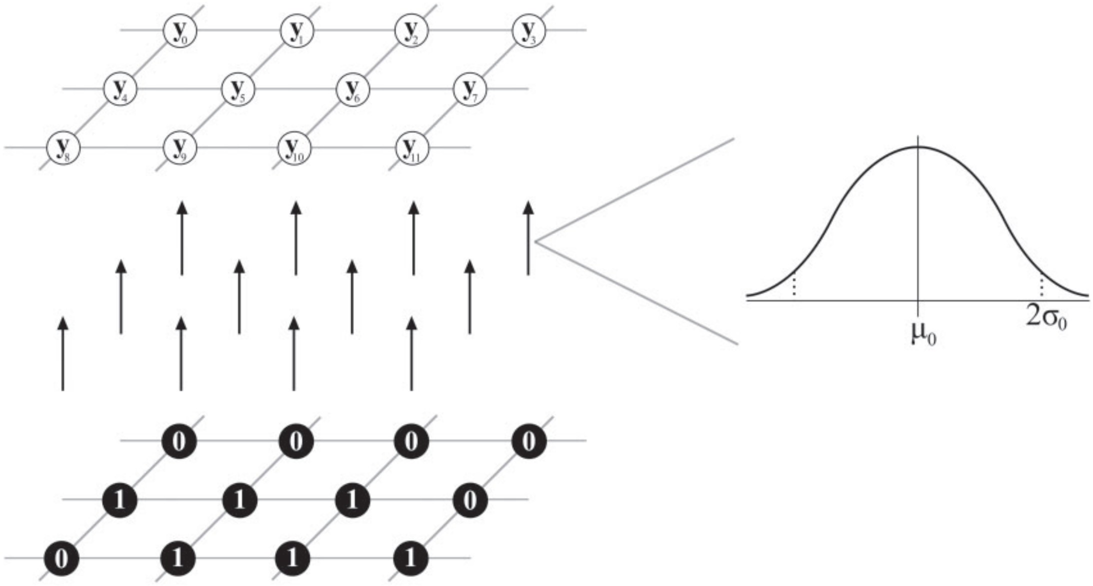

FIGURE 1. Architecture of a hidden Markov random field. Nodes from the hidden state layer (black nodes, 0 or 1) emit the nodes of the observed layer (white nodes, $y_0 \ldots y_{11}$) following a Gaussian function (emissions represented as arrows). Parameters of the Gaussian function (μ and σ) are specific to each hidden state (in the example, state 0).

to note that the new framework is proposed as an alternative to methods that require codification of the distributions into a grid or a set of subareas. Thus, it is not benchmarked against any method that uses raw geographic coordinates as input, such as network analysis (Dos Santos et al. 2008) or geographic interpolation of endemism (Oliveira et al. 2015).

Hidden Markov Random Fields: Generalities

HMRFs are acyclic, graphical statistical models widely used for computer vision processes, such as segmentation, classification, and noise reduction (Li 2009). In the domain of comparative biology, they have only been used for clustering genetic variants within populations (François et al. 2006). A thorough and didactic introduction to this class of models is presented by Blake et al. (2011).

A HMRF consists of a set of nodes distributed in two lattices of size T (Fig. 1). Nodes in the hidden layer are the model states, and their values are integer labels from the set $L=\{1,2,\ldots,l\}$:

$$X=\{x_1,x_2,\ldots,x_T \mid x_i \in L\}$$

Nodes in the observed layer are emissions, and their values are real numbers:

$$Y=\{y_1,y_2,\ldots,y_T \mid y_i \in \mathbb{R}\}$$

How can this model architecture be interpreted as an area of endemism? First, assume that the geographic distribution of each taxon is decoded as a grid of ones (presence) and zeros (absence). A collection of taxa grids (those that belong to the same area of endemism) can be summarized into a single grid, each cell bearing the average symbol recorded among all the taxa (Fig. 2). The latter grid contains real values in the range $[0.0-1.0]$ and can be considered the observed layer in the model. The area of endemism can be represented by the hidden layer, the component that generates the geographic distributions, a layer that is not observed but inferred. This layer contains two possible states, 0 (meaning that this cell does not belong to the area of endemism) or 1 (if it does). Interactions among and within the layers will be discussed next.

Given a state $x_i = l$ and its corresponding emission y_i, there is a conditional probability distribution:

$$p(y_i \mid x_i) = f(y_i;\Theta_l)$$

where Θ_l is the set of function parameters for the label l. Emissions of a HMRF are usually modeled as Gaussian processes, thus the conditional probability would be:

$$p(y_i \mid x_i) = \frac{1}{\sigma_l\sqrt{2\pi}}\exp\left(-\frac{(y_i-\mu_l)^2}{2\sigma_l^2}\right) \tag{1}$$

where μ and σ^2 are the mean and variance of the emission function for label l.

Within the hidden layer X, relations among its nodes are determined by a neighborhood system. Thus, for a node x_i, there is a set of neighbors

$$N_i = \{x_j \mid x \in X, distance(x_i,x_j) \le g\}$$

where g is the neighborhood size (often called neighborhood order). The architecture described above is a HMRF only if it complies with both the positivity (equation 2) and Markovianity (equation 3) conditions:

$$P(x) > 0, \forall x \in X \tag{2}$$

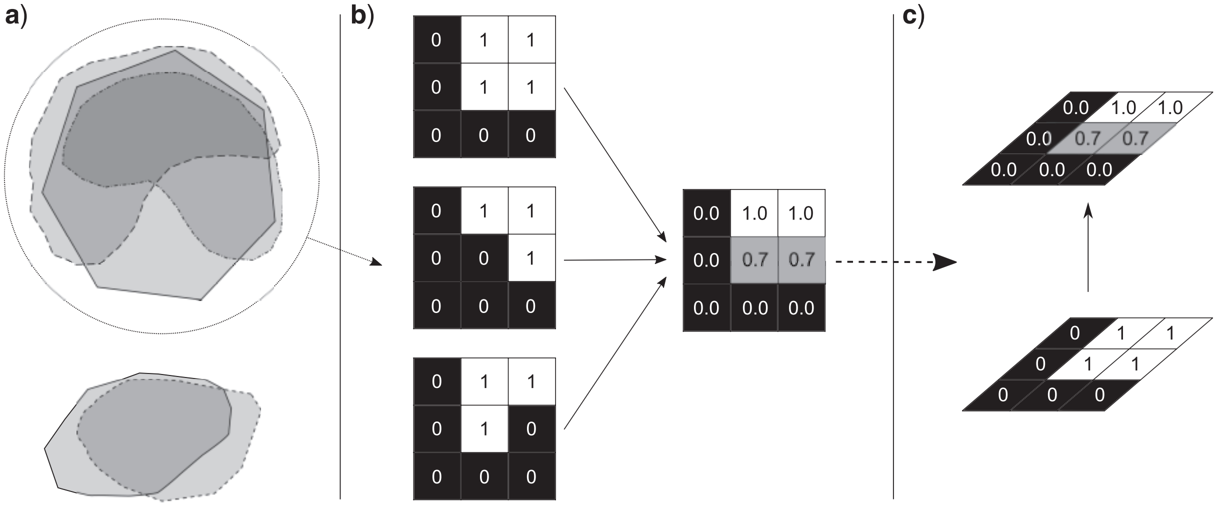

FIGURE 2. Overview of the proposed framework. a) Taxon distributions are clustered based on a distance measure. In this example, five taxon distributions are grouped in two clusters. b) Taxa distributions within a cluster (grids to the left) are summarized into a single grid (grid to the right) that contains average values of presence (1) and absence (0). c) The summary grid is incorporated into the model as the observed layer (top lattice), which is emitted by the hidden layer (bottom lattice), the representation of the area of endemism.

and

$$P(x_i \mid X \setminus x_i) = P(x_i \mid N_i) \tag{3}$$

In lay terms, the probability axioms introduced above indicate that 1) values of the taxa distribution grid are emitted by the states in the area of endemism following a Gaussian process (using Fig. 1 as reference, taxa distributions would be the observed layer, and the area of endemism would be the hidden layer), 2) there is spatial correlation among adjacent cells within the area of endemism, and 3) each area of endemism contains two Gaussian functions: one for each state (0 or 1). The Gaussian function associated to state 0 emits all the “absences” in the observed layer, thus its mean would be close to 0, whereas the state 1 function emits the “presences”. In both cases, the variance would determine the expected deviation from that mean in the emitted values.

Gaussian probability functions are here employed to model the relation between areas of endemism and taxa distributions on the grounds of the central limit theorem. Although it is unknown if these functions are appropriate descriptors of the area of endemism process, there is not information available to prefer an alternative statistical function. Later on, it would be shown that this choice does not affect the performance of the method.

The most important difference between HMRFs and other Markov processes popular in computational biology (such as Markov chains and hidden Markov models) is the lack of directionality. This property has a significant effect in the probability estimation procedures. For example, the probability of an individual state is conditional on all the other states in the lattice:

$$P(x_i) = P(x_i \mid \ldots, x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2}, \ldots) \tag{4}$$

In practical terms equation 4 implies that it is necessary to evaluate all possible configurations of the lattice to obtain $P(x_i)$. These conditional relations render the joint probability of the hidden layer—$P(X)$—intractable, even on small lattices. A solution to this problem is provided by the Hammersley–Clifford theorem (Besag 1974): the probability distribution $P(X)$ can be achieved through a Gibbs distribution with respect to subsets of nodes (neighborhoods) in X:

$$P(X) = \frac{\exp(-U(X))}{Z} \tag{5}$$

where $U(X)$ is the *energy function* and Z is the *partition constant*. The energy function is a measure of homogeneity of states distribution across the HMRF. The partition constant assures the distribution $P(X)$ sums up to 1—the sum of $\exp(U(X))$ over all possible configurations of states. Again, computing this Gibbs distribution seems infeasible, as the estimation of the partition constant necessarily entails the inspection of all possible configurations of X. This difficulty is usually overcome by using the logarithmic form of the function:

$$\ln(P(X)) \propto -U(X)$$

The energy function can take several forms, but in all cases it associates $P(x_i)$ exclusively with its neighbors (N_i), relaxing the conditional rule (Equation 4). For this application, the energy function is based on the Potts model energy function:

$$U(X) = \sum_{i=0}^{T} \gamma \sum_{n \in N_i} \begin{cases} 1, & \text{if } x_i = n \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where T is the number of nodes in the lattice, and γ is a parameter that modules space similarity within the hidden layer.

Maximum a Posteriori State Configuration

Finding the underlying state configuration of a HMRF can be seen as an optimization problem: given a set of observations Y, estimate the state configuration X that maximizes the probability of the system. This is usually achieved through Bayes rule:

$$P(X|Y)=\frac{P(Y|X)P(X)}{P(Y)}$$

For the purpose of *maximum a posteriori* state estimation, the marginal distribution $P(Y)$ can be ignored since it will remain constant. Thus, the alternative Bayes notation can be employed:

$$P(X|Y)\propto P(Y|X)P(X)$$

The likelihood, $P(Y|X)$, is the only term that can be easily estimated using the Gaussian emission probability function introduced above (equation 1). As mentioned before, the main difficulty of calculating $P(X)$ is the partition constant Z. However, this can be overcome by logarithmic simplification:

$$\ln(P(X|Y))\propto\ln(P(Y|X))+\ln(P(X)) \tag{7}$$

where

$$\ln(P(Y|X))=-\sum_{i=0}^{T}\left(\frac{(y_i-\mu)^2}{2\sigma^2}+\ln(\sigma)\right)$$

and $\ln(P(X))$ is estimated with equation 6. In statistical literature, terms in equation 7 are called *likelihood energy* $[\ln(P(Y|X))]$, *prior energy* $[\ln(P(X))]$, and *posterior energy* $[\ln(P(X|Y))]$.

Materials and Methods

General Description of the Framework

The framework presented here models areas of endemism as HMRFs, specifically as the hidden layer. As mentioned in the introduction, taxon distributions are assumed to be decoded as grids of ones (presence) and zeros (absence). During the first step, taxon distributions are clustered using a distance measure, which is a way to estimate the degree of overlap (Fig. 2a). Distributions within a group are summarized into a single grid by averaging the observed values in every cell (Fig. 2b). The resulting ensemble grid is then included in the statistical process as the observed layer, which is emitted by the hidden layer—the area of endemism (Fig. 2c). The optimal number of areas are chosen through a model selection procedure, the pseudolikelihood information criterion (PLIC).

Clustering Procedure

Through the first step of the method taxon distributions are clustered based on similarity. This clustering procedure is largely based on partition around medoids (PAM; Kaufman and Rousseeuw 1990). Medoids are a subset of elements of the input set ideally located at the center of each cluster. Thus, PAM seeks to iteratively find optimal medoids and cluster the remaining elements around the closest medoid. In the first cycle, medoids are randomly selected from the input set, but in the following cycles they are replaced by randomly selecting other members from their respective clusters. This is repeated until the composition of clusters reaches stability. Although it is fast, this algorithm is sensitive to outliers and can be easily stuck in local optima. There are several variants of PAM that aim to improve selection of initial medoids, or cluster composition after the initial groups are formed (Daiyan et al. 2012; Razavi Zadegan et al. 2013).

The clustering routine used here includes multiple modifications from the simple PAM procedure presented above. First, an additional parameter (cohesion value c) is included. This parameter sets the maximum distance allowed between medoids and their linked elements. It also determines which elements are outliers: if the distance between an element and all medoids is greater than the cohesion parameter, the element is appended to the outlier array. Second, medoids are not elements from the input list but HMRFs modeled upon them—namely, the hidden layer (Fig. 2c). Third, HMRFs are retrieved through a presampling process. This process involves modeling a new field from each element, finding all the elements within a distance c, and finally optimizing the field with the values from the associated elements.

The similarity statistic employed in this clustering step is the Kulczynski distance (Hubálek 1982) as implemented by Hausdorf and Hennig (2003):

$$d_K(A_1,A_2)=1-\frac{1}{2}\left(\frac{|A_1\cap A_2|}{|A_1|}+\frac{|A_1\cap A_2|}{|A_2|}\right) \tag{8}$$

where A_1 and A_2 are two input areas, $|A_x|$ their size, and $|A_1\cap A_2|$ the number of cells their distributions intersect. This function provides a good approximation to the overlap among distributions because it is estimated upon shared cell values across the grid. This distance measure has been successfully used in the context of identification of areas of endemism (Hausdorf and Hennig 2003).

Maximum a Posteriori State Configuration Estimation Through the Expectation-Maximization Algorithm

Once a valid clustering hypothesis is attained—that is, a set of presumptive areas of endemism (HMRF's hidden layers) and their respective associated taxon distributions (HMRF's observed layers)—state configuration and parameter values of the model are estimated through the expectation-maximization algorithm (EM).

In general sense, EM is an iterative algorithm composed of two basic routines: estimation of state probabilities throughout the state path given the current parameters (expectation), and update of the model parameters and state configuration (maximization). In the first cycle of the algorithm, parameter and state path values are chosen randomly, a property that usually does not affect its efficiency to converge into the optimal solution.

Although this algorithm has been widely used to estimate parameters in hidden Markov models (Rabiner 1989)—the one-dimensional counterpart of HMRFs—its full implementation in HMRFs is infeasible given the impossibility to estimate conditional probabilities in this class of models. However, several approximations to EM have been proposed for HMRFs. Here, we follow the proposal of Zhang et al. (2001), which is based on the relaxation of conditional probabilities and minimization of energy functions, an approach that avoids the computation of posterior probabilities. This approach can be described in the following steps:

1. Initialize the parameters (means and standard deviations). To speed up the estimation, initial state means are set to values close to 0 and 1.

2. State configuration is computed through the iterated conditional nodes algorithm (Besag 1986). This algorithm selects the state configuration that minimizes the posterior energy (equation 7).

3. Estimate the posterior distribution for both states at every cell. For each cell i in the grid S and each state label l in $L=\{0,1\}$ compute:

$$P(l\,|\,y_i)=\frac{P(y_i\,|\,l)P(l\,|\,N_i)}{\sum_{l\in L}P(y_i\,|\,l)P(l\,|\,N_i)}$$

   where $P(y_i\,|\,l)$ corresponds to the equation 1 and $P(l\,|\,N_i)$ to equation 5.

4. Update model parameters using state expectations. Therefore, the mean and standard deviation update rules at time $t+1$ for the label l are:

$$\mu_l^{t+1}=\frac{\sum_{i\in S}P^t(l\,|\,y_i)y_i}{\sum_{i\in S}P^t(l\,|\,y_i)}$$

   and

$$\sigma^{t+1}=\left(\frac{\sum_{i\in S}P^t(l\,|\,y_i)(y_i-\mu_l)^2}{\sum_{i\in S}P^t(l\,|\,y_i)}\right)^{0.5}$$

Therefore, two different outputs of the EM algorithm will help to interpret its underlying hypothesis of area of endemism: the state configuration in the hidden layer, and the posterior probabilities calculated at every hidden node. The former shows the geographic extent of the area of endemism, and the latter the probability that each cell belongs to the area of endemism.

Potts Model: Gamma Parameter

Although the EM algorithm can efficiently estimate the mean and standard deviation of the model, the gamma parameter is not updated. Therefore, a method is required to estimate the spatial autocorrelation within taxa geographic ranges and adjust the gamma parameter accordingly. The observed correlation distribution is estimated by randomly sampling cells among the input observations and calculating the homogeneity with reference to its neighbors. Subsequently, the gamma value is selected through least squares. Candidate values should be positive figures, as they indicate correlation. For the purpose of the experiments, we employed integers in the interval $[1-30]$, which worked accurately through the development stage of the project.

Model Selection

Models with higher number of fields—and parameters—naturally will have greater likelihood values than simpler ones. Therefore, a model selection procedure is required to avoid over-parameterization. Popular model selection procedures (Akaike information criterion, Bayesian information criterion, etc.) cannot be used under the HMRF framework, as they all require estimation of the likelihood. The Pseudolikelihood Information Criterion (PLIC) (Stanford and Raftery 2002) is an excellent alternative; instead of estimating the actual model likelihood it relies on pseudolikelihood values, a tractable approximation especially suitable for graph models (Besag 1975; Qian and Titterington 1992):

$$PL(Y\,|\,X)=\prod_i f(y_i\,|\,x_i)P(x_i\,|\,N_i) \tag{9}$$

In equation 9, the first term within the product is the Gaussian conditional probability introduced earlier (equation 1), and the second term is a conditional distribution based upon the Potts function:

$$P(x_i=m\,|\,N_i,\gamma)=\frac{\exp(\gamma U(N_i,m)}{\sum_{l\in L}\exp(\gamma U(N_i,l))}$$

where the function U is the Potts function used to estimate the prior energy (equation 6).

PLIC weights can be estimated using the pseudolikelihood of the model:

$$PLIC(K)=2\ln(PL(Y\,|\,X))-d_K\ln(N)$$

where K is the number of HMRFs, d_K is the number of parameters in the model K, N is the total number of grid cells in the model K, and PL is the pseudolikelihood. The behavior of pseudolikelihood function may differ from the likelihood function when the number of parameters are substantially different from the true model (Stanford and Raftery 2002). Therefore, pairwise comparisons of PLIC values are carried out progressively, from models bearing a single field (one area of endemism) to models with $N/2$ fields, the maximum number of areas allowed in any data set. A model m_i (containing i HMRFs) is selected when has a higher PLIC value than the next model in terms of parameter complexity, m_{i+1}.

An important caveat is the possibility that the EM could select model parameters that imply high pseudolikelihoods but are conceptually meaningless under the present biogeographic framework. For example, a very noisy set of observations can lead the EM algorithm to output identical means for both states (e.g., 0.5) or extremely wide variances. In either case, the pseudolikelihood value of the model would be high, but it would not be useful to interpret the data, as all observations could be emitted by any symbol and all cells could be part of the area of endemism (1) or not (0). Therefore, an additional term is included, the Bhattacharyya weight, a coefficient that measures the distance between two probability distributions (Bhattacharyya 1946). The form of the Bhattacharya coefficient used here was proposed by Coleman and Andrews (1979) to estimate the distance between two Gaussian distributions:

$$B=\frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\sigma_0^2}{\sigma_1^2}+\frac{\sigma_1^2}{\sigma_0^2}+2\right)\right)+\frac{1}{4}\left(\frac{(\mu_0-\mu_1)^2}{\sigma_0^2+\sigma_1^2}\right)$$

where μ_x and σ_x are the mean and standard deviation of the Gaussian distribution x, respectively.

Experimental Validation: Simulated Data Sets

The performance of the new framework was tested on simulated data, following an approach similar to Casagranda et al. (2012). Two sets of experiments were conducted: in the first one distributions were simulated by simply producing the absence–presence grid necessary for the analysis; but in the second set datapoints in a plane coordinate system were directly simulated, then distributions were coded into an absence–presence grid. Besides these differences in the simulation process, all experiments were very similar regarding the analytical and postprocessing steps.

Simulations: experimental set 1. Every simulated data set comprised 9–30 taxon distributions, and each

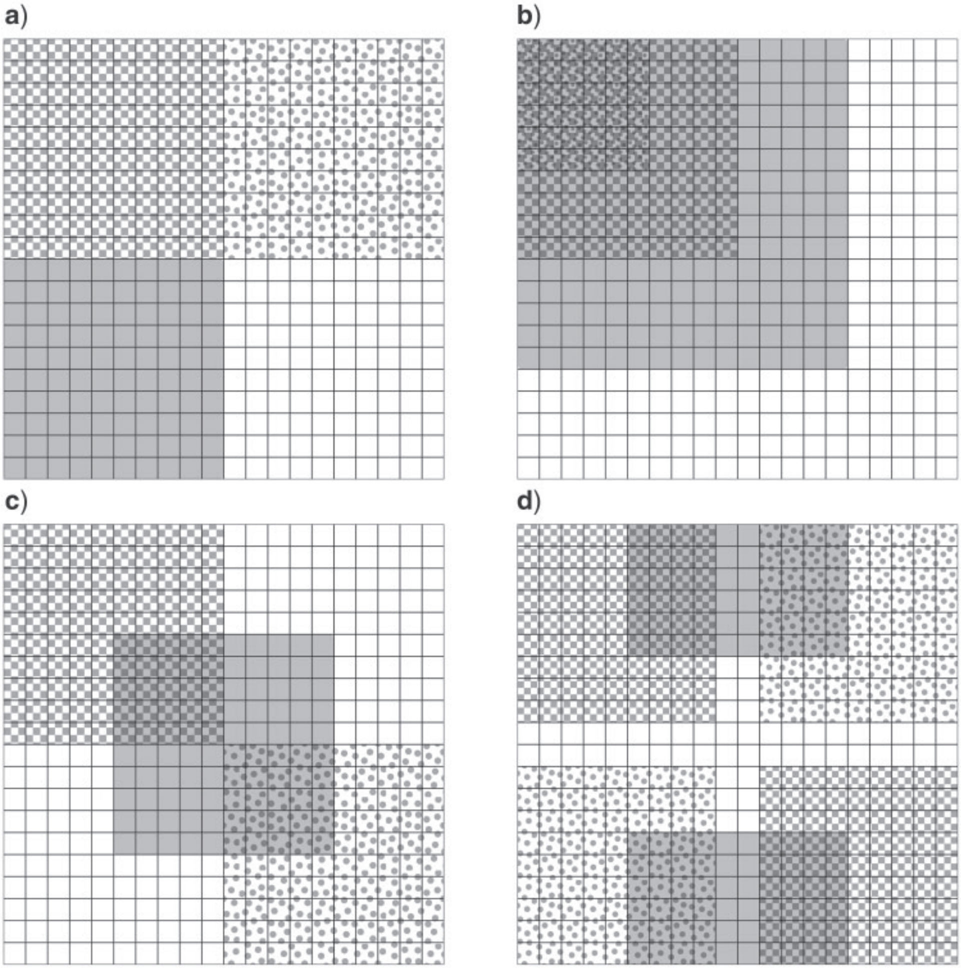

FIGURE 3. Simulated cases of areas of endemism (indicated as group of cells under some grade of shading). a) Non-problematic areas (case 0), b) nested areas (case 1), c) overlapping areas (case 2), and d) disjunct areas (case 3).

distribution was a grid composed by 400 cells (20 columns × 20 rows). Values within a grid were ones (presence) or zeros (absence). Simulated data sets were meant to be explained by the presence of three areas of endemism; that is, taxon distributions were intended to be clustered in three groups, and each group should have at least two components. Every time a taxon distribution was simulated the underlying area of endemism was used as template, then a randomly selected fraction of its cells were recoded as absences ("0") and another group outside its limits as presences ("1"). The fraction of cells undergoing this modification was set by a predefined uncertainty value. This means that—besides the data sets simulated with no uncertainty—all taxon distributions within an area of endemism were not fully overlaid.

Four parameters were used to simulate the complexity of areas of endemism and their associated distributions:

Uncertainty: the probability a given distribution would not conform to the symbol distribution of its underlying area of endemism. Probability values were sampled from a uniform statistical distribution. Two kinds of uncertainty were considered depending on their placement relative to the area of endemism: *internal*—within its limits—and *external*—outside, but within two cells of distance.

Non-clustering distributions: Distributions that do not belong to any area of endemism. In clustering terminology, outliers. Non-clustering distributions were distributional grids filled with absence symbols (zeros), except on either a column or a row randomly chosen from a uniform distribution.

Distributions per area: Number of taxon distributions simulated by each area of endemism.

Area pattern: Spatial arrangement of areas of endemism inside the grid (as proposed by Casagranda et al. (2012)). Four types are possible: non-problematic, overlapping, nested, and disjunct (Fig. 3).

Three different experiments were conducted on simulated data sets (Table 1). The aim of the first experiment was to assess the impact of different uncertainty values in the accuracy of the method. The second experiment was designed to study the effect of different levels of distribution congruence (measured as the number of taxa supporting an area of endemism). The purpose of the last experiment was to examine the impact of noisy data sets (data sets with distribution data that do not belong to any area of endemism) in the algorithm performance. Ten data sets were simulated for every combination of parameter values. HMRF analyses were executed under a cohesion value of 0.3 and exhaustive combinatorial enumeration.

Besides the HMRF framework, each data set was also analyzed under parsimony analysis of endemism (PAE) [as perform in TNT (Goloboff and Catalano 2016)], endemicity analysis (EA) [using NDM (Goloboff 2002)], and biotic element analysis (BAE) [as implemented in the R package prabclus (Hausdorf and Hennig 2003)]. TNT searches were conducted on ten starting Wagner trees, followed by 200 iterations of each ratchet and drift. Areas of endemism were recovered as clades supported by shared taxon acquisition events (0 → 1 synapomorphies) in the consensus tree.

NDM analyses were conducted using default settings (0.5 as factor for inferred presences, 0.5 as factor for external records, 1.0 as acceptance factor, absences were inferred if three surrounding cells were unoccupied, presences if seven were occupied, new solutions were proposed by deleting cells only, one cell was swapped at a time, suboptimal solutions were not stored, however, redundant areas were swapped and replaced as score improves, areas were narrowly initialized, and one replicate was executed per search). No consensus areas were estimated from the NDM analysis, instead every output solution was examined.

BAEs were executed using both Kulczynski distance (BEA) and Geco coefficient (BEA-GECO). The latter was introduced by Hennig and Hausdorf (2006) as an improvement to classic similarity distance estimation. It incorporates spatial correlation explicitly, a property that helps to overcome the problem caused by missing data in clustering tasks (Hennig and Hausdorf 2006). Besides distance measure parameters, all other BEA variables were set to default values. It is important to note that BEA do not recover areas of endemism directly, but clusters of taxa with similar distributions. However, it is possible to infer the underlying geographic pattern of such clusters by extracting the set of cells that contain more than 50% of the taxa belonging to that cluster. Here, we consider areas of endemism such sets of cells.

Accuracy was estimated as the frequency of detecting the underlying areas of endemism on each data set. The most objective way to measure success is to count how often the area of endemism was exactly recovered by the algorithm, but such a strict criterion would not let the reader examine the performance of BEA, PAE,

TABLE 1. Parameter settings of experiments conducted on simulated data sets

Experiment	Uncertainty level	Distributions per area	Non-clustering distributions
1	0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.0	3	0
2	0.05	(2,2,5), (2,5,5), (5,5,5), (2,2,10), (2,5,10), (2,10,10), (5,5,10), (5,10,10), (10,10,10)	0
3	0.05	3	0, 2, 4, 5, 8, 10

and EA, since they only recovered the exact area in rare occasions. Therefore, a result was accepted as positive if its Kulczynski distance to the real area was < 0.2.

Simulations: experimental set 2. An additional set of *in silico* experiments were conducted simulating the distributional datapoints directly. The shape and spatial arrangement of areas of endemism were the same as above; that is, we kept the four scenarios design proposed by Casagranda et al. (2012). Each area of endemism was modeled as a couple of Gaussian distributions (one for each plane axis) in which both means matched the center of the area, and the standard deviations a quarter of the area extent. We simulated 2–10 geographic distributions for each area of endemism, and 10–1000 datapoints for each geographic distribution. Datapoints were later coded into a presence–absence grid, using 1 × 1 degree cells. Finally, data sets were submitted to the same analytical procedures mentioned in the previous paragraph. For the purpose of these experiments, accuracy was measured as the frequency of recovering all the distributions associated to each area of endemism within the same cluster.

Two experiments were conducted with these data sets. The first analysis was aimed to study the effect of point density per distribution in the accuracy of each method. We assumed that point density was an approximate measurement of uncertainty, as distributions simulated with higher densities imply better consistency with the underlying area of endemism. In this analysis, three geographic distributions were simulated for each area of endemism, but different number of datapoints were drawn: 10, 50, 100, 500, and 1000 per distribution. The second experiment was focused on the effect of distribution congruency in the delimitation performance: the 500 datapoints were drawn for each distribution, but different number of distributions per area of endemism were simulated: 2, 5, and 10.

Empirical Tests

The algorithm was tested on two empirical data sets: 1) South African weevil genus *Sciobus* Schoenherr [2-degree matrix taken from Morrone (1994), with the modifications from Mast and Nyffeler (2003)], and 2) Central American beetles Carabidae Latreille, and spiders Dipluridae Simon [data set accessed from Szumik and Goloboff (2004)]. Performance on the South African data set was compared with results reported by Mast and Nyffeler (2003) and Hausdorf and Hennig (2003). Given that the analysis of this data set presented by Szumik et al. (2002) did not included the changes suggested by Mast and Nyffeler (2003), a reanalysis of the corrected matrix under EA was executed with the program NDM, using default settings. Custom PAE and BEA analyses were also conducted on the Central American arthropod data using TNT and prabclus, respectively.

Code Implementation

The method here described is implemented in the Python program "Gloria" (*G*eographic *L*ocation-Hidden Mark*O*v *R*andom f*I*eld *A*nalysis). It is released under GNU General Public License version 3 and available at https://github.com/nrsalinas/gloria. The program accepts a list of geographic coordinates as input, and returns two files: a log file including some statistics of the optimal solution (pseudolikelihood values, approximate posterior probabilities, etc.) and a GeoJSON file per area to facilitate visualization. A more detailed explanation

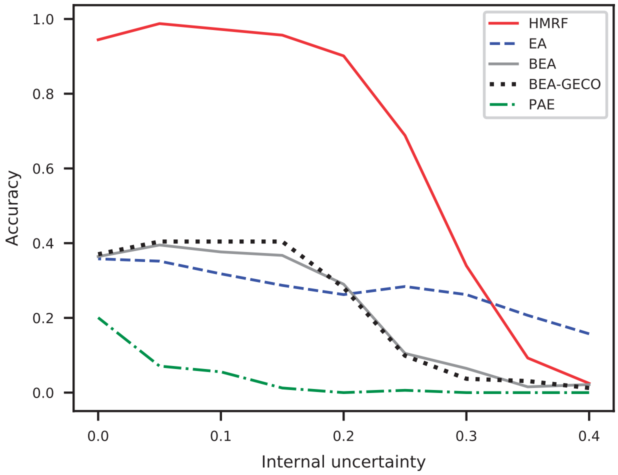

FIGURE 4. Accuracy of several methods under different values of internal uncertainty. HMRF = hidden Markov random fields; EA = endemicity analysis; BEA = biotic element analysis; BEA-GECO = biotic element analysis based on Geco distance coefficients; PAE = parsimony analysis of endemism.

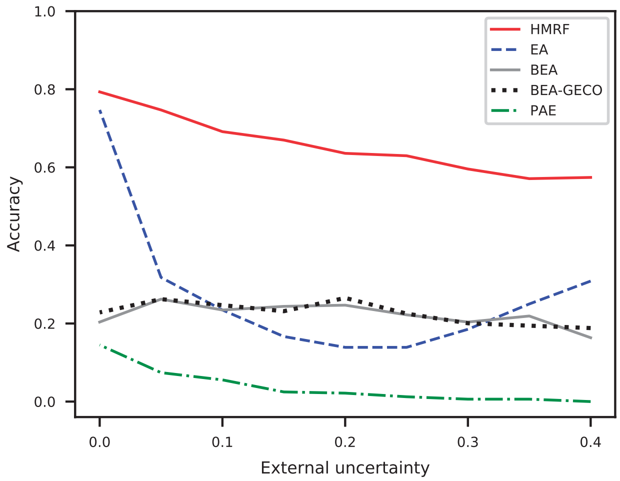

FIGURE 5. Accuracy of several methods under different values of external uncertainty. HMRF = hidden Markov random fields; EA = endemicity analysis; BEA = biotic element analysis; BEA-GECO = biotic element analysis based on Geco distance coefficients; PAE = parsimony analysis of endemism.

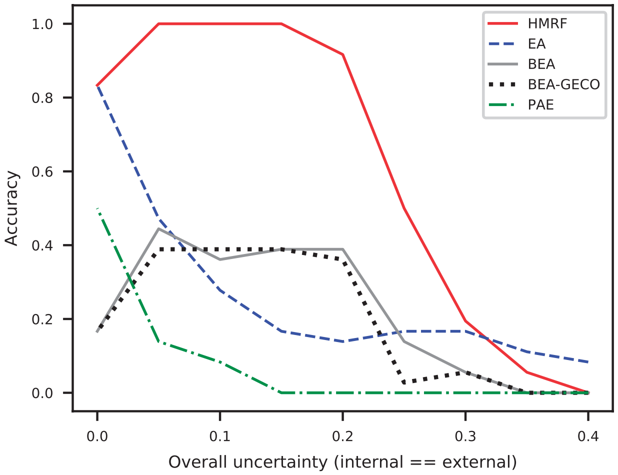

FIGURE 6. Accuracy of several methods when internal and external values are the same. HMRF = hidden Markov random fields; EA = endemicity analysis. BEA = biotic element analysis; BEA-GECO = biotic element analysis based on Geco distance coefficients; PAE = parsimony analysis of endemism.

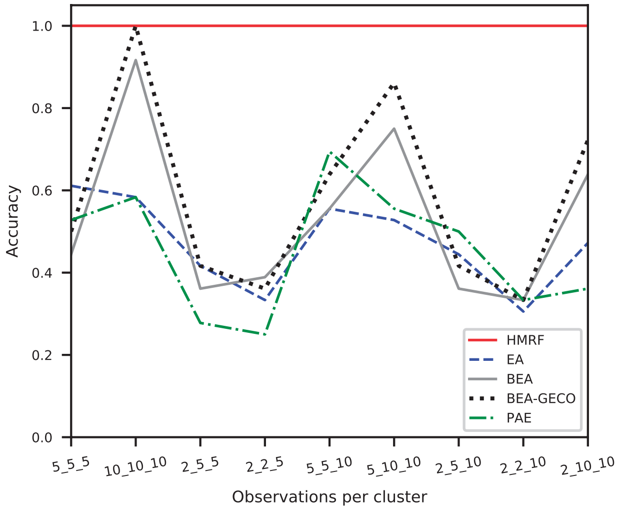

FIGURE 7. Accuracy of several methods when areas of endemism have fluctuating observation composition. HMRF = hidden Markov random fields; EA = endemicity analysis; BEA = biotic element analysis; BEA-GECO = biotic element analysis based on Geco distance coefficients; PAE = parsimony analysis of endemism.

about input format and analysis settings is provided in the program website.

RESULTS

Simulated Data: Experimental Set 1

HMRFs accurately recovered the underlying areas of endemism throughout the different treatments of uncertainty. The new framework performed reasonably well under values of internal uncertainty < 0.2, reaching an accuracy rate above 80% (Fig. 4). Nevertheless, the method was more sensitive to external uncertainty as its accuracy was $< 80\%$ in all cases (Fig. 5). The best performance was recorded in simulations that kept internal and external uncertainty at the same value, reaching an accuracy of 80–100% when the overall uncertainty level was less than 0.25 (Fig. 6).

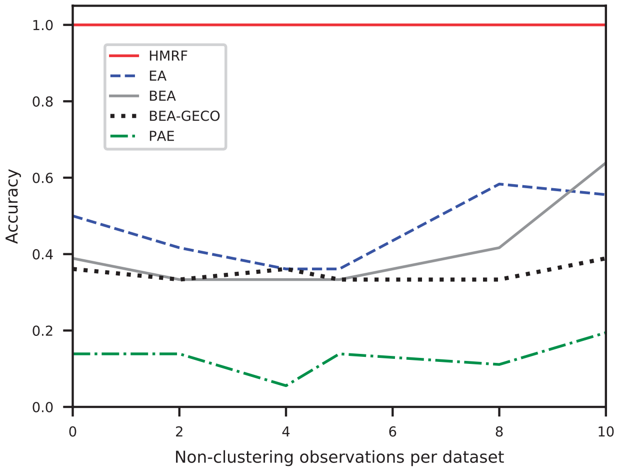

FIGURE 8. Accuracy of several methods on data sets with solitary distributions. HMRF = hidden Markov random fields; EA = endemicity analysis; BEA = biotic element analysis; BEA-GECO = biotic element analysis based on Geco distance coefficients; PAE = parsimony analysis of endemism.

Simulating a different number of distributions per area did not affect the accuracy of HMRFs. The new method nearly always recovered the three areas of endemism in the set correctly (Fig. 7). EA and BEA were only $> 80\%$ successful when most of the areas were supported by 10 distributions each; otherwise their accuracy (and that of PAE under all configurations) was 25–80%. PAE, EA, and BEA generally resulted in the same performance pattern across the parameter space.

HMRFs achieved perfect accuracy when non-clustering observations were included in the simulations, they recovered all the areas in every data set (Fig. 8). EA and BEA were negatively affected by these noisy data sets, and their accuracy only reached 30–60%. PAE only recovered 10–20% of the areas of endemism in simulated data sets.

Simulated Data: Experimental Set 2

When the analytical benchmark was submitted to data sets of simulated datapoints, the results were highly congruent with the first set of experiments. HMRFs were more efficient than competing methods, followed by EA (Fig. 9). When only few datapoints were simulated per distribution (10–50), all methods failed consistently and their accuracy did not reached 20%. However, when that parameter was increased to 100 or more, HMRF and EA greatly improved their performance ($> 50\%$). In contrast, the accuracy of PAE and BEA never was greater than 20%. Increasing the number of distributions per area only changed notoriously the efficiency of BEA, the other methods kept a more or less stable performance (Fig. 10).

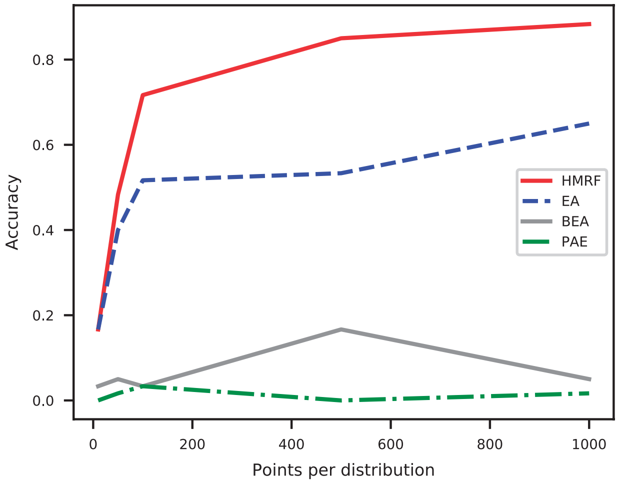

FIGURE 9. Accuracy of several methods on point-simulated data sets when several values of record density are considered. HMRF = hidden Markov random fields; EA, endemicity analysis; BEA = biotic element analysis; PAE = parsimony analysis of endemism.

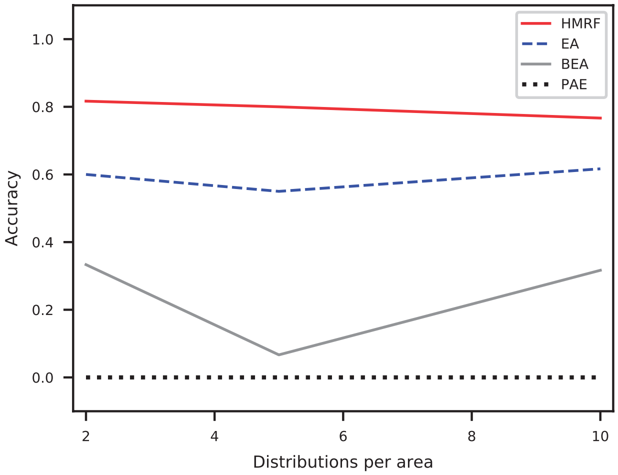

FIGURE 10. Accuracy of several methods on point-simulated data sets when the number of distributions per area are different. HMRF = hidden Markov random fields; EA = endemicity analysis; BEA = biotic element analysis; PAE = parsimony analysis of endemism.

Sciobius Data set

The new framework identified six areas of endemism from the *Sciobius* data set (Fig. 11). Supporting taxa per area of endemism are shown in Table 3. Half of the areas—areas 1, 2, and 3—were not recovered by the other algorithms, whereas only two—areas 4 and 5—were consistently identified by all of the algorithms (Table 2). Our area 1 is reminiscent of the element 1 in Hausdorf and Hennig (2003), and area 2 is similar to an area recovered from EA [analogous to area 3 in Szumik et al. (2002)].

The other methods resulted in different numbers of areas of endemism. According to Mast and Nyffeler (2003), PAE recovered three areas (five if nested areas are recognized), whereas BEA identified four elements (Hausdorf and Hennig 2003). A custom EA analysis on the corrected data set retrieved five areas. Species composition between areas recovered by both HMRF and the other algorithms was variable, ranging from 0.0 to 0.77 Jaccard of similarity.

Carabidae and Dipluridae Data set

The new framework identified eight areas of endemism in the Carabidae and Dipluridae data set (Fig. 12). Most of the areas are supported by just two endemic species (Table 4). Only one area was also retrieved by another algorithm: area 7, which was recovered by BEA. However, species composition of this area between both analyses was very dissimilar, as only one species was shared.

Although HMRFs did not recover any area presented by Szumik and Goloboff (2004), several areas of endemism are similar between the two analyses: our areas 2 and 7 resemble their sets 2 and 0, respectively. Species composition is also similar: the species supporting an area under the HMRF framework also support the analogous EA area.

This data set has previously been reported to contain 14 areas of endemism, as identified through EA (Szumik and Goloboff 2004). Custom analyses with BEA and PAE returned three areas each.

DISCUSSION

The new method using HMRFs is a useful framework for uncovering areas of endemism as shown by both simulated and empirical experiments. It was particularly efficient on simulated data sets with internal uncertainty, often times recovering the true areas more efficiently than competing algorithms. This does not mean that the new method performs without errors across the parameter landscape. It fails consistently under two circumstances: 1) when uncertainty is greater than 0.3, and 2) in data sets without internal uncertainty. The former case corresponds to extremely noisy distributions, in which the signal of areas of endemism has vanished to an extent difficult to overcome. The failure of the latter case is paradoxical; the reader would intuitively believe that a method performing well under cases with moderate uncertainty should perform even better on "perfect" data sets, devoid of any noise. A detailed examination of the algorithm and the simulated data sets revealed that HMRF only failed consistently on data sets with nested areas—case 1 *sensu* Casagranda et al. (2012). In such cases, the initial clustering algorithm fails to model a HMRF for every area of endemism. In the absence of uncertainty, the Kulczynski distance between nested areas decreases below the cohesion parameter (0.3), and only one HMRF is modeled. This artifact can be eliminated by decreasing the cohesion value; however, using a low cohesion value throughout the simulations would reduce the accuracy of the method under more realistic values of uncertainty.

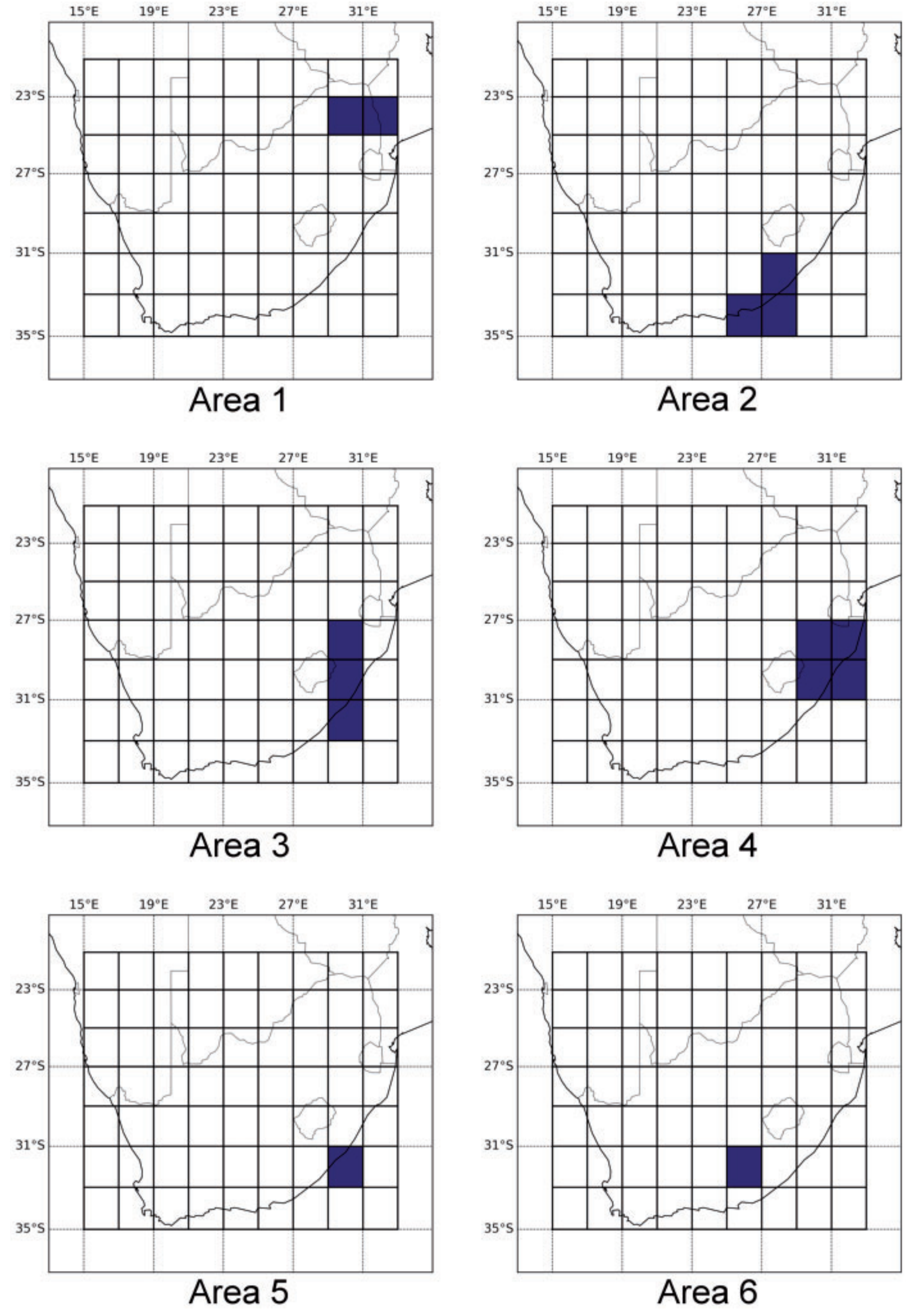

FIGURE 11. Areas of endemism recovered from the *Sciobius* data set.

This phenomenon highlights the importance of selecting an appropriate cohesion value for the analysis, since unexpected interactions between grid extension and size of areas of endemism can lead to spurious results.

It is recommended to conduct a sensitivity analysis with several candidate cohesion values (0.0–1.0) before the final analysis is executed, particularly when geographic distributions have significant size differences. This should be coupled with a thorough exploration of grid parameters (offset values and cell size). If the biogeographic signal in the data set is strong, areas of endemism shape and their supporting species configuration will not change notoriously along a gradient of the aforementioned parameters. It is also advisable to check the set of geographic distributions that support each area of endemism and confirm that they actually overlap.

The new framework was robust under scenarios with significant uncertainty, and it was also accurate across several configurations of taxa diversity: few or numerous taxa per area of endemism, and even or unequal taxa distributions per area. Although the latter scenario is a common property of real data sets, the other

TABLE 2. Areas recovered by HMRF on the *Sciobius* data set, and their correspondence to results from other algorithms

HMRF	BEA	PAE	EA
1 (2)	—	—	—
2 (3)	—	—	—
3 (5)	—	—	—
4 (13)	2 (17, 12 shared)	1 (6, all shared)	1
5 (7)	4 (10, 7 shared)	3 (11, 7 shared)	—
6 (3)	—	2′ (3, all shared)	—

Note: Numbers outside the parentheses indicate area indexes used by authors in their publications, numbers inside the parentheses are the number of species supporting a given area. BEA results retrieved from Hausdorf and Hennig (2003), PAE from Mast and Nyffeler (2003), and EA from a custom analysis.

TABLE 3. Supporting species by area of endemism identified in the *Sciobius* data set by HMRF analysis

Area of endemism	Species
1	*Sciobius angustus*, *S. vittatus*
2	*Sciobius asper*, *S. capeneri*, *S. scapularis*
3	*Sciobius aciculatifrons*, *S. arrowi*, *S. granosus*, *S. panzanus*, *S. thompsonii*
4	*Sciobius barkeri*, *S. bistrigicollis*, *S. brevicollis*, *S. cognatus*, *S. cultratus*, *S. dealbatus*, *S. holmi*, *S. marginatus*, *S. marshalli*, *S. prasinus*, *S. spatulatus*, *S. tenuicornis*, *S. wahlbergii*
5	*Sciobius endroedyi*, *S. granipennis*, *S. lateralis*, *S. planipennis*, *S. pondo*, *S. scholtzi*, *S. transkeiensis*
6	*Sciobius minusculus*, *S. nanus*, *S. schoenlandi*

TABLE 4. Supporting species by area of endemism identified in the Carabidae and Dipluridae data set by HMRF analysis

Area of endemism	Species
1	*Platynus rotundulatus*, *Elliptoleus balli*
2	*Platynus nitidulus*, *Platynus rugulellus*
3	*Platynus pygmaeus*, *Elliptoleus zapotecorum*
4	*Elliptoleus whiteheadi*, *Euagrus gus*
5	*Platynus degallieri*, *Platynus flavomarginatus*
6	*Platynus aeneipennis*, *Elliptoleus vixtriatus*
7	*Platynus machetellus*, *Elliptoleus luteipes*, *Elliptoleus curtulus*, *Euagrus mexicanus*
8	*Elliptoleus crepericornis*, *Euagrus pristinus*

algorithms were only accurate when simulated data sets contained several taxa per area of endemism (namely, 10 per area).

Several reasons can explain the lack of accuracy of the other methods. In the case of PAE, areas of endemism are not properly optimized in the consensus tree. As the uncertainty level increases, the resulting consensus tree is mostly unresolved and contains polytomies. On such trees, taxon acquisitions (apomorphic state transformations) tend to be optimized across multiple shallow branches (parallel gains), sometimes even as autoapomorphies. This pattern departs widely from the expected scenario, in which areas of endemism are represented by clades subtended by uncontroverted synapomorphies.

Experiments from simulated datapoints (experimental set 2) support the previous discussion. Probably the only interesting difference is that accuracy of HMRF never reached 100% (85% in the most favorable combination of parameters). We believe this is a natural consequence of the experimental design. Given that datapoints were drawn from joint Gaussian distributions, areas of endemism do not have clear boundaries. Therefore, the extent of the simulated distributions greatly varied and did not necessarily overlap, even if thousands of datapoints are drawn. Moreover, grid cells located around the border zone are prone to be coded as a presence, since it is only required to harbor a single datapoint to be interpreted that way.

The lack of efficiency of EA seems to be rooted in the way the heuristic search is executed. The program NDM (Szumik and Goloboff 2004) starts the search using one of the input distributions as a template for the initial solution, then moves to new solutions by changing one or two cells in the border every cycle. It is not clear if internal cells within the current solution are modified to guide future evaluations, but if this is not a recurrent move the algorithm could have been in disadvantage to find the optimal area in the experiments as half of the uncertainty was simulated within the limits of the actual area.

BEA finds areas of endemism through a four-step clustering procedure: 1) a nonmetric multidimensional scaling (NMDS) is executed on the distribution distance matrix, 2) a mixture Gaussian model is fitted on the n-dimensional vectors retrieved from step 1, 3) the optimal number of clusters (biotic elements) in the model are selected via Bayesian Information Criterion, and 4) areas of endemism are retrieved from taxa distributions within each cluster from the optimal solution. Part of the limitation of this approach probably lays on how the mixture model is optimized. Usually half of the distributions in the simulated data sets were wrongly classified as noise (either clustered in the "noise" component or in "clusters" of a single element). This artifact logically implies that chances of identifying the correct set of clusters are reduced.

However, this behavior decreases as more distributions per area are simulated. As demonstrated in the third experiment (Table 1), BEA was extremely efficient uncovering the true areas if data sets contained 10 taxa per area (Fig. 7). This may indicate that the optimization of the mixture model could be sensitive to the number of samples. Most of the simulated data sets had three distributions per area of endemism, which means that only three data points per component were employed to estimate the parameters of the model. The HMRF framework seems less prone to this drawback since the sample units for parameter fitting are cells in the grid, not distributions. Given the size of the grid (20 × 20 cells), an area of endemism with only two distributions would provide at least ca. 70 samples for this task (the smallest simulated area of endemism contained 36 cells).

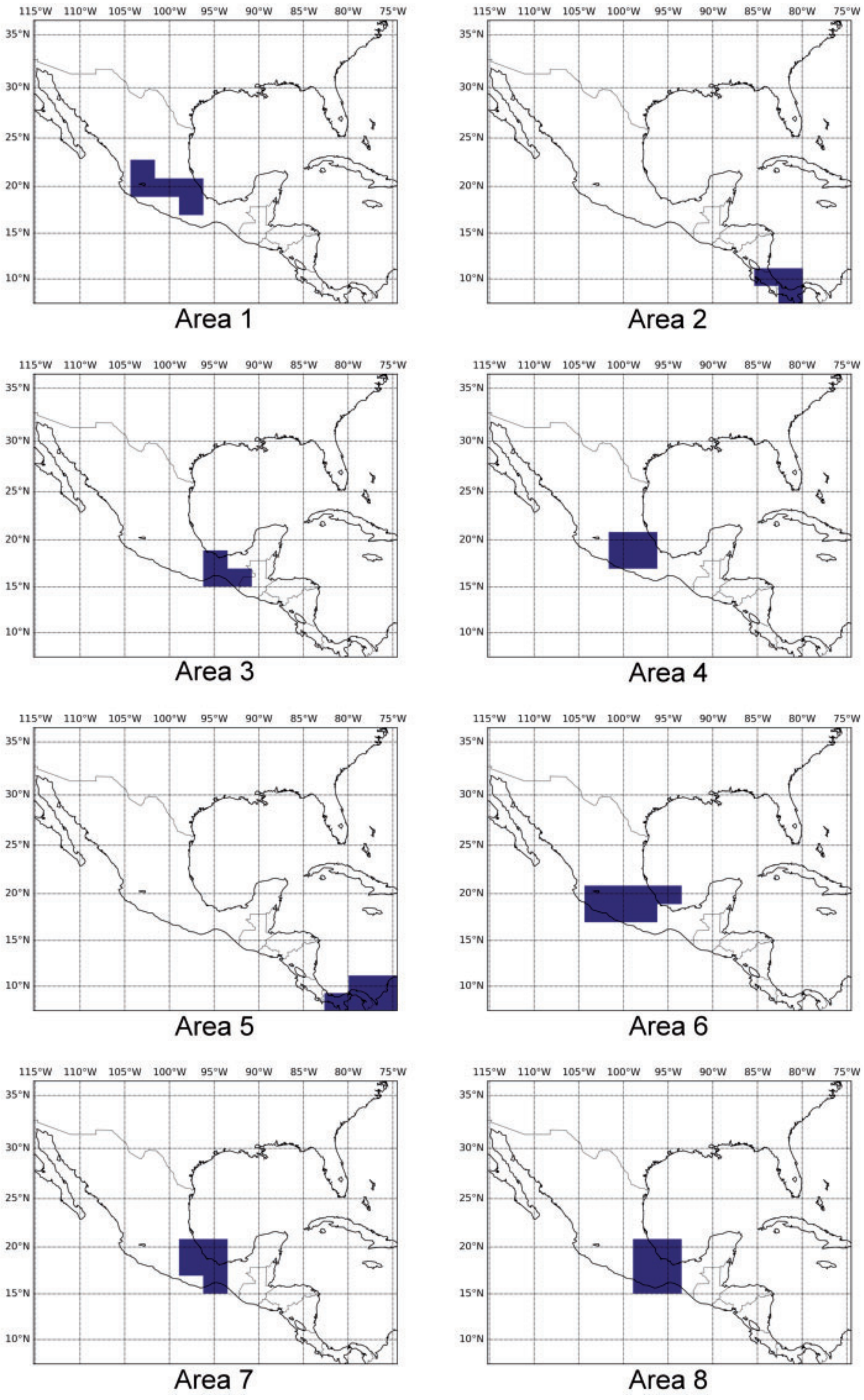

FIGURE 12. Areas of endemism recovered from the Carabidae and Dipluridae data set.

The experiments on empirical data were congruent with the basic assumptions of an area of endemism and validate the HMRF approach as a method to uncover such a kind of geographic patterns. First, all areas of endemism—in both data sets—were supported by species of sympatric geographic distributions. Second, most of the areas (except area 4 from the *Sciobius* data set and area 8 from the Carabidae and Dipluridae data set) were supported by taxa geographically restricted within the boundaries of the area. In these cases, however, the section of the range outside the area of endemism was always smaller than the section inside.

The HMRF framework is a promising alternative for modeling and representing geographic distribution patterns; however, there are some caveats that should be considered. First, this framework assumes that the geographic distribution of taxa supporting an area of endemism should conform to a Gaussian distribution on each cell of the grid. It is unknown whether this assumption is appropriate regarding areas of endemism, but this condition did not seem to affect the performance of the new framework. Even though the observations simulated for the experiments were sampled from a uniform distribution, the new method remained accurate. Furthermore, areas of endemism were simulated with few distributions (max. = 10), making it impossible to statistically test their cell-wise distribution.

Another caveat is that analyses using the new framework are more time intensive than other programs, especially with difficult—noisy—data sets. This behavior is partially due to the work overload generated by the combinatorial optimization that the search routine is based on. Currently, the program "Gloria" offers an optimization shortcut that reduces the set of candidate HMRFs to those with the highest pseudolikelihoods. This can dramatically reduce the computation on fuzzy data sets with many candidate HMRFs, but does not guarantee the best solution will be found. Other possible solutions to the time overload are still in development, and they include the implementation of multiprocessing and resampling (e.g., bootstrap aggregating) routines.

There are several ways that this model can be expanded or used to incorporate geographic uncertainty into related comparative analyses. For example, the model can be extended to accept cell values different from present or absent, such as real numbers in the range $[0-1]$. Under this premise, taxa distributions could be probabilities taken from an analysis that assesses the potential distribution of the taxa (e.g., niche modeling probabilities).

Funding

This work was supported by the Gerstner Scholars Fellowship and the Gerstner Family Foundation, and the Richard Gilder Graduate School, American Museum of Natural History.

Acknowledgments

To Jessica Allen, Jenna Dorey, John Flynn, Eric Ford, Sean McKenzie, Rick Ree, and two anonymous reviewers for their comments. To the Willi Hennig Society for distributing TNT free of charge.

References

Besag J. 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soci. Series B Stat. Methodol. 36:192–236.

Besag J. 1975. Statistical analysis of non-lattice data. J. R. Stat. Soc. Series D Stat. 24:179–195.

Besag J. 1986. On Statistical Analysis of Dirty Pictures. J. R. Stat. Soc. Series B Stat. Methodol. 48:259–302.

Bhattacharyya A. 1946. On a measure of divergence between two multinomial populations. SankhyÄ Indian J. Stat. (1933-1960) 7:401–406.

Blake A., Kohli P., Rother C. 2011. Markov random fields for vision and image processing. Cambridge (MA): MIT Press.

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

Casagranda M.D., Brako L., Szumik C.A. 2012. Endemicity analysis, parsimony and biotic elements: a formal comparison using hypothetical distributions. Cladistics. 28:645–654.

Coleman G.B., Andrews H.C. 1979. Image segmentation by clustering. Proc. IEEE. 67:773–785.

Crother B.I., Murray C.M. 2013. Parsimony analysis of endemism under the areas of endemism as individuals thesis. Cladistics. 29:571–573.

Daiyan G.M., Abid F.B.A., Khan M.A.R., Tareq A.H. 2012. An efficient grid algorithm for faster clustering using K medoids approach. 2012 15th International Conference on Computer and Information Technology (ICCIT): 1–3.

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.

François O., Ancelet S., Guillot G. 2006. Bayesian clustering using hidden Markov random fields in spatial population genetics. Genetics. 174:805–816.

Gámez N., Escalante T., Espinosa D., Eguiarte L.E., Morrone J.J. 2014. Temporal dynamics of areas of endemism under climate change: a case study of Mexican *Bursera* (Burseraceae). J. Biogeogr. 41:871–881.

Goloboff P. 2002. NDM and VNDM: programs for the identification of areas of endemism. Software distributed by the authors.

Goloboff P.A., Catalano S.A. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. Cladistics. 32:221–238.

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

Henderson I.M. 1991. Biogeography without area? Aust. Syst. Bot. 4:59–71.

Hennig C., Hausdorf B. 2006. A robust distance coefficient between distribution areas incorporating geographic distances. Syst. Biol. 55:170–175.

Hubálek Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biol. Rev. 57:669–689.

Kaufman L., Rousseeuw P.J. 1990. Finding groups in data: an introduction to cluster analysis. Hoboken (NJ): John Wiley & Sons (Wiley series in probability and mathematical statistics).

Li S.Z. 2009. Markov random field modeling in image analysis. 3rd ed. London: Springer Science & Business Media.

Martínez-Hernández F., Mendoza-Fernández A.J., Pérez-García F.J., Martínez-Nieto M.I., Garrido-Becerra J.A., Salmerón-Sánchez E., Merlo M.E., Gil C., Mota J.F. 2015. Areas of endemism as a conservation criterion for Iberian gypsophilous flora: a multi-scale test using the NDM/VNDM program. Plant Biosyst. 149:483–493.

Mast A.R., Nyffeler R. 2003. Using a null model to recognize significant co-occurrence prior to identifying candidate areas of endemism. Syst. Biol. 52:271–280.

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

Morrone J.J. 2014a. Biogeographical regionalisation of the Neotropical region. Zootaxa. 3782:1–110.

Morrone J.J. 2014b. Parsimony analysis of endemicity (PAE) revisited. J. Biogeogr. 41:842–854.

Nelson G., Platnick N.I. 1981. Systematics and biogeography. New York: Columbia University Press.

Oliveira U., Brescovit A.D., Santos A.J. 2015. Delimiting areas of endemism through kernel interpolation. PLoS One. 10:e0116673.

Platnick N.I. 1991. On areas of endemism. Aust. Syst. Bot. 4:xi–xii.

Qian W., Titterington D.M. 1992. Stochastic relaxations and EM algorithms for Markov random fields. J. Stat. Comput. Simul. 40:55–69.

Rabiner L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. 77:257–286.

Razavi Zadegan S.M., Mirzaie M., Sadoughi F. 2013. Ranked k-medoids: a fast and accurate rank-based partitioning algorithm for clustering large datasets. Knowl. Based Syst. 39:133–143.

Real R., Barbosa A.M., Bull J.W. 2017. Species distributions, quantum theory, and the enhancement of biodiversity measures. Syst. Biol. 66:453–462.

Regan H.M., Colyvan M., Burgman M.A. 2002. A Taxonomy and Treatment of Uncertainty for Ecology and Conservation Biology. Ecol. Appl. 12:618–628.

Stanford D., Raftery A. 2002. Approximate Bayes factors for image segmentation: the pseudolikelihood information criterion (PLIC). IEEE Trans. Pattern Anal. Mach. Intell. 24:1517–1520.

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

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

Zhang Y., Brady M., Smith S. 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20:45–57.