Automatic selection of the number of clusters using Bayesian clustering and sparsity-inducing priors.

Denis Valle[1*], Yusuf Jameel[1], Brenda Betancourt[2], Ermias T. Azeria[3], Nina Attias[4], Joshua Cullen[1]

[1] School of Forest Resources and Conservation, University of Florida, Gainesville, Florida, United States of America.

[2] Department of Statistics, University of Florida, Gainesville, Florida, United States of America.

[3] Alberta Biodiversity Monitoring Institute, University of Alberta, Edmonton, AB T6G 2E9, Canada

[4] Ecology and Conservation Graduate Program, Federal University of Mato Grosso do Sul, Campo Grande, Mato Grosso do Sul, Brazil.

* Corresponding author. 136 Newins-Ziegler Hall, Gainesville, Florida 32611. PO Box 110410. Fax: 352-392-1707. Telephone: 352-392-3806. Email: drvalle@ufl.edu.

Running head: Determining the number of clusters

Submitted to: Ecological Applications

Type of article: Article

Data availability statement

The Breeding Bird survey data are freely available at the USGS Patuxent Wildlife Research Center website (https://www.pwrc.usgs.gov/bbs/rawdata/). Vascular plant data from Alberta are freely available for download from the ABMI website (https://abmi.ca/home/data-analytics/da-top/da-product-overview/Species-Habitat-Data.html) by selecting "Vascular Plant" under "Species Variables" within "Terrestrial Variables". Average temperature and precipitation data are freely available at WorldClim (www.worldclim.org/version2) and can be found under "Historical climate data" and "5 minutes". Armadillo movement data from Brazil is available for download upon request at http://movebank.org. These data can be found under "Three banded armadillo Attias Serra do Amolar (clean burrow)" and "Tree banded armadillo Attias Caceres (clean burrow)".

Abstract

Clustering is a ubiquitous task in ecological and environmental sciences and multiple methods have been developed for this purpose. Because these clustering methods typically require users to *a priori* specify the number of groups, the standard approach is to run the algorithm for different numbers of groups and then choose the optimal number using a criterion (e.g., AIC or BIC). The problem with this approach is that it can be computationally expensive to run these clustering algorithms multiple times (i.e., for different numbers of groups) and some of these information criteria can lead to an overestimation of the number of groups.

To address these concerns, we advocate for the use of sparsity-inducing priors within a Bayesian clustering framework. In particular, we highlight how the truncated stick-breaking (TSB) prior, a prior commonly adopted in Bayesian nonparametrics, can be used to simultaneously determine the number of groups and estimate model parameters for a wide range of Bayesian clustering models without requiring the fitting of multiple models. We illustrate the ability of this prior to successfully recover the true number of groups for three clustering models (two types of mixture models, applied to GPS movement data and species occurrence data, as well as the Species Archetype model) using simulated data in the context of movement ecology and community ecology. We then apply these models to armadillo movement data in Brazil, plant occurrence data from Alberta (Canada), and bird occurrence data from North America.

We believe that many ecological and environmental sciences applications will benefit from Bayesian clustering methods with sparsity-inducing priors given the ubiquity of clustering and the associated challenge of determining the number of groups. Two R packages, EcoCluster and

bayesmove, are provided that enable the straightforward fitting of these models with the TSB prior.

Key-words: Bayesian nonparametrics, clustering, Biogeographic Region model, mixture model, movement ecology, Species Archetype model

Introduction

Clustering algorithms are commonly used across multiple disciplines to reduce data dimensionality by grouping data items with similar features, enabling the identification of the main latent structural characteristics of highly multivariate data (Berkhin, 2006; Jain et al., 1999; Legendre and Legendre, 2012). In environmental sciences and ecology, clustering approaches have been extensively used since at least the 1920's (Legendre and Legendre, 2012). Examples for biodiversity datasets include cluster analysis to define biogeographical regions (Azeria et al., 2007; Foster et al., 2017; Kreft and Jetz, 2010; Lyons et al., 2017), identify indicator species by grouping species that tend to co-occur (Azeria et al., 2009), identify microbial community patterns associated with sample origin and/or sampling time (Ramette, 2007), and cluster species that tend to have similar relationships with other species in food web studies (e.g., set of predator species that feed on the same set of prey species) (Baskerville et al., 2011). Cluster analysis has also been extensively used in other environmental science applications. For example, clustering has been used to classify water catchments in data-scarce regions (Auerbach et al., 2016) and to understand the spatial variation in the detection rate of pharmaceuticals in rivers across different regions (Jameel et al., 2020).

Clustering is an important task across scientific fields and, as a result, a rich assortment of methods and algorithms have been developed through time (Jain et al., 1999). These methods can be classified based on several dichotomies, such as whether a single partition (partitional) or a nested series of partitions (hierarchical) is created, if methods output hard (each data item can only belong to a single group) or fuzzy (each data item can have varying degree of membership to each group) groups, and if these methods are algorithmic or probabilistic (Berkhin, 2006;

Bouveyron and Brunet-Saumard, 2014; Jain et al., 1999; Legendre and Legendre, 2012; Saxena et al., 2017). A long-term challenge when using clustering algorithms consists of defining the appropriate number of clusters, which typically has to be *a priori* specified (Berkhin, 2006; Jain et al., 1999; Legendre and Legendre, 2012; Saxena et al., 2017). The standard approach for this task is to systematically vary the number of groups and run the algorithm once for each setting. Then, the optimal number of groups is determined using a performance metric (e.g., AIC, BIC, gap statistic, integrated classification likelihood, minimum message length) (Berkhin, 2006; Biernacki et al., 2000; Charrad et al., 2014; Daudin et al., 2008; Depraetere and Vandebroek, 2014; Fraley and Raftery, 2007; Hui and Warton, 2015; Hui et al., 2013; Lyons et al., 2017; Ter Braak et al., 2003; Tibshirani et al., 2001). This approach has been extensively used in the past but it can be computationally expensive and time consuming for large datasets and/or complex models. Importantly, large simulation studies have shown that no single performance metric is consistently better than the others (Depraetere and Vandebroek, 2014) and that some of these commonly adopted information criteria tend to favor models with a larger number of groups than warranted (Casella et al., 2014), even if the model faithfully mirrors the data generating mechanism (e.g., Pohle et al., 2017).

The generation of sparse solutions (i.e., where only a small fraction of the parameters are non-zero) is highly desirable for a range of modeling applications. For example, regularization (i.e., penalty terms added to the objective function) in statistical (e.g., regression) and machine learning (e.g., support vector machines) models is key to avoid overfitting, increase predictive skill, and improve interpretability of model results. Interestingly, many of the proposed regularization approaches can be interpreted as Bayesian models with very specific types of sparsity-inducing priors (Hahn and Carvalho, 2015; Hooten and Hobbs, 2015; Park and Casella,

2008; Wood, 2017). Likewise, the challenge of determining the number of clusters can also be tackled by defining Bayesian models with sparsity-inducing priors (i.e., priors that favor fewer clusters). In this paper we describe how different types of Bayesian clustering methods applied to ecological data, when used together with sparsity inducing priors, can automatically determine the number of clusters without requiring fitting multiple models. In particular, we focus on a specific type of sparsity-inducing prior, the truncated stick-breaking (TSB) prior (i.e., an approximation of the Dirichlet Process), that has been extensively used in Bayesian nonparametrics (Sethuraman, 1994) but that has seen relatively little application in ecological and environmental sciences.

To illustrate how this approach can be used for a range of models, we rely on three Bayesian clustering methods applied to ecological data: two types of mixture models, applied to movement and species occurrence data, and the Species Archetype (SA) model (Dunstan et al., 2013). To our knowledge, none of the three clustering methods with the TSB prior has been used in ecological applications. We apply these three clustering methods to simulated data to showcase the ability of the sparsity-inducing priors to successfully recover the true number of groups by fitting the model just once. We then perform an exploratory data analysis with these methods to reveal the latent structure in armadillo movements in the Pantanal wetlands (Brazil), plant occurrence in Alberta (Canada), and breeding bird occurrence from United States and Canada. We also provide two R packages (EcoCluster and bayesmove) that enable straight-forward fitting of these clustering models, which we expect will be of broad use for ecological and environmental science clustering tasks.

Material and methods

1. Truncated stick-breaking prior

Clustering methods (also referred to as mixture models (McLachlan and Peel, 2000)) explicitly or implicitly contain multiple latent variables z_i, i=1,…,n. The latent variable z_i indicates the cluster membership of unit i and can take on any integer value between 1 and K, where K is the number of clusters defined *a priori* by the user. Depending on the specific application, this unit can consist of individual forest plots, rivers, species, pharmaceuticals, sampling points, etc. In probabilistic clustering approaches, it is typically assumed that the latent variable z_i follows a categorical distribution:

$$z_i \sim Cat(\boldsymbol{\theta})$$

where the vector $\boldsymbol{\theta}$ is of size K (i.e., the number of clusters) and contains probabilities that sum to one, indicating the likelihood that unit i is assigned to individual clusters. This categorical distribution is used because it is assumed each unit can only belong to a single group.

Finding the optimal number of groups K by fitting the model multiple times (once for each K value) and choosing K using a model selection criterion such as AIC or BIC can be a prohibitive approach if fitting each model is computationally expensive. Furthermore, past research has suggested that some of these information criteria tend to favor models with a larger number of groups than warranted (Casella et al., 2014), even if the model faithfully mirrors the data generating mechanism (e.g., Pohle et al., 2017). The approach proposed here avoids these problems by relying on the truncated stick-breaking (TSB) prior, a particular type of prior for $\boldsymbol{\theta}$ that favors sparseness (i.e., a smaller number of groups). With this prior, the user is only required to specify the maximum number of groups K and the algorithm chooses the number of groups ($k \leq K$) that best clusters the sample data.

159 The stick-breaking prior has a long tradition in Bayesian nonparametric models. This prior
160 arises from the Dirichlet process (DP), which is arguably the most popular Bayesian
161 nonparametric model used for clustering applications (Ferguson, 1973). The distribution over
162 random partitions induced by the DP is commonly known in the machine learning community as
163 the Chinese restaurant process (CRP) (Teh, 2011) and is equivalent to the Ewens sampling
164 formula used to describe a distribution over partitions in population genetics that was introduced
165 before the DP (Crane, 2016; Ewens, 1972). Another definition of the DP is the Pólya urn
166 representation, which essentially describes the same distribution from the CRP (Blackwell and
167 MacQueen, 1973). Here, we focus on the alternative definition of the Dirichlet process known as
168 the "stick-breaking" construction (Sethuraman, 1994). This definition of the DP model is
169 considerably simpler and more general than the previously mentioned representations. This
170 approach has been extremely useful for the development of novel statistical models as well as
171 new Markov Chain Monte Carlo (MCMC) inference algorithms (Ishwaran and James, 2001).
172 Notice that, despite similar names, this "stick-breaking" prior is not associated with the "broken
173 stick model" for species abundance described in MacArthur (1957).
174 As commonly done in Bayesian nonparametrics (Ishwaran and James, 2001), we adopt a
175 truncated version of this prior by defining a maximum number of groups (hereafter referred to as
176 the Truncated Stick-Breaking [TSB] prior). To define the maximum number of groups, the
177 standard advice is to choose a truncation point such that the results would be indistinguishable
178 from what would have been obtained with a larger number of groups (Ishwaran and James,
179 2001). More practically, as long as most of the posterior mass is concentrated on the initial
180 components, then the actual value for the maximum number of groups should have no effect on
181 model results (Manrique-Vallier, 2016). On the other hand, if the maximum number of clusters is

reached, then the standard advice is to increase the maximum number of possible groups to avoid an incorrect approximation. Notice that, because the TSB prior is typically viewed as an approximation to the DP, some researchers actually prefer to work directly with the DP to avoid any potential approximation errors (e.g., MacEachern and Muller, 1998; Papaspiliopoulos and Roberts, 2008).

Instead of viewing the TSB prior solely as an approximation to the DP, our perspective is that the truncation in this prior is useful from a dimension reduction perspective because it avoids the number of groups increasing with sample size (Murugiah and Sweeting, 2012). Indeed, Casella et al. (2014) justified their use of a strong shrinkage prior by stating that, even when the true number of cluster is large, cluster analysis "will only result in useful inference when the answer contains a relatively small number of clusters". For the same reason, we advocate for modelers interested in dimension reduction to carefully think about the maximum number of groups that is still manageable/interpretable when defining where to truncate the stick-breaking prior, this way limiting the complexity of the solution that is found by the algorithm.

The truncated stick-breaking prior for $\boldsymbol{\theta}$ is constructed indirectly by first defining

$$V_k \sim Beta(1, \gamma)$$

for k=1,…,K-1 and V_K is set to one. The parameters $V_1, \ldots, V_K$ are then used to calculate θ_k, employing the following expressions:

$$\theta_1 = V_1$$

$$\theta_k = V_k \prod_{p=1}^{k-1}(1 - V_p) \text{ for } k>1 \qquad \text{[eqn. 1]}$$

We use the following shorthand to denote this prior:

$$\boldsymbol{\theta} \sim TSB(\gamma)$$

Notice that, according to this prior, the expected proportion of units assigned to cluster k is given by

$$E[\theta_k] = \mathrm{E}[V_k]\prod_{p=1}^{k-1}\left(1-\mathrm{E}\left[V_p\right]\right) = \frac{1}{1+\gamma}\left(1-\frac{1}{1+\gamma}\right)^{k-1} = \frac{\gamma^{k-1}}{(1+\gamma)^k} \text{ for } k<K\text{; and}$$

$$E[\theta_k] = \left(1-\frac{1}{1+\gamma}\right)^{k-1} \text{ for } k=K.$$

The depiction of this equation for $0 < \gamma < 1$ reveals an approximately exponential decay of $E[\theta_k]$ with increasing k and that smaller γ corresponds to faster decay and therefore sparser results (i.e., fewer clusters).

To illustrate how this prior works, say we have a maximum of 6 groups (K=6) and $\boldsymbol{V} = [V_1, \dots, \mathrm{V}_6]$ is equal to $[0.19, 0.33, 0.27, 0.95, 0.47, 1]$. Recall that, by definition, V_6 is set to 1. As illustrated in Table 1, these values for $\boldsymbol{V}$ imply that $\boldsymbol{\theta} = [\theta_1, \dots, \theta_6]$ is equal to $[0.19, 0.27, 0.15, 0.38, 0.01, 0.01]$. Note that θ_5 and θ_6 are very small compared to $\theta_1, \dots, \theta_4$ and that the four first groups account for 99% of all observations. These values suggest the presence of 4 main groups, despite having allowed for up to 6 groups. These results arise because the TSB prior shrinks θ_k to zero for large values of k. Similar to how the components that only explain a small portion of the variation are typically ignored when conducting a Principal Component Analysis (PCA), the remaining groups when using the TSB prior (i.e., groups 5 and 6) are also typically ignored.

(insert Table 1)

An intuitive way of interpreting the expressions for θ_k in Table 1 is to think about a sequential process in which a sampling unit is eventually assigned to cluster k after failing to be

assigned to clusters 1, 2,…, k-1. For example, the expression for θ_3 can be interpreted as the probability that a sampling unit is not assigned to group 1 (equal to 1 - 0.19) times the probability that it is not assigned to group 2 (equal to 1 - 0.33) times the probability that it is assigned to group 3 (equal to 0.27). The name "stick-breaking" originates from the metaphor of sequentially breaking a stick of length 1 into smaller and smaller pieces, as illustrated in Fig. 1.

(insert Fig. 1)

Another benefit of using the TSB prior is that, by weakly identifying the labels of each cluster, it can help to reduce the amount of label switching, a common problem for mixture models which refers to the fact the group labels are unidentified parameters in these models. This problem often leads to poor mixing of MCMC algorithms and generates potentially nonsensical results if posterior distributions of parameters are summarized by their averages (Stephens, 2000).

2. Clustering models

To illustrate the wide applicability of the TSB prior, we describe three probabilistic partition clustering methods that greatly benefit from this prior. All of these clustering models have an observational model in which the response variable y_i, conditioned on the latent cluster membership variable z_i, comes from a distribution that has some parameters indexed by z_i. More explicitly, we assume that:

$$y_i|z_i = k \sim f(\boldsymbol{\beta_k}, \boldsymbol{\phi})$$

where $\boldsymbol{\beta_k}$ is the set of cluster-specific parameters and $\boldsymbol{\phi}$ is a vector containing the remaining parameters that are not cluster specific. We specify *f()* and the priors for $\boldsymbol{\beta_k}$ and $\boldsymbol{\phi}$ in greater detail when describing the individual models used to illustrate the TSB prior. In all models, we assume that the latent cluster membership variable z_i is given by:

$$z_i \sim Cat(\boldsymbol{\theta})$$

and that

$$\boldsymbol{\theta} \sim TSB(\gamma).$$

A commonly used prior for γ is a Gamma distribution (Dunson and Xing, 2009; Manrique-Vallier, 2016; Si and Reiter, 2013). However, we decided to adopt a discrete uniform prior for γ, where this parameter can take any of the following values 0.1, 0.15, 0.2 ,…, 0.95, and 1 with equal probability. This prior was chosen because it ensures $\gamma \leq 1$, it is straight-forward to implement, conforms to our prior belief of equal probability for all possible values of γ, and resulted in good mixing of our MCMC algorithms. Because of the truncation in the stick-breaking prior, $\gamma \leq 1$ ensures that the last group will be smaller than all the other groups (i.e., $E[\theta_k] > E[\theta_K]$ for k=1,…,K-1). All models were fit using Gibbs samplers and detailed information regarding the Full Conditional Distributions (FCDs) used by these algorithms is given in Appendix S1.

Extensive simulations are used to show how the TSB prior can be used within these models to successfully estimate the true number of groups without requiring the fitting of multiple models with different numbers of groups. For all simulated datasets, we vary the true number of groups K and set the parameters within the vector $\boldsymbol{\theta}$ to $1/K$, resulting in clusters of approximately equal size. We estimate the true number of groups by calculating the minimum number of groups that together represent more than 99% of all observations. More specifically,

we assumed that the estimated number of groups $\hat{k}$ is given by $\min_k\left(\sum_{j=1}^{k}\hat{\theta}_j > 0.99\right)$, where $\hat{\theta}_j$ is the posterior mean for group j. Finally, because of the large number of simulations and the large number of parameters within any given model, we assessed convergence by examining trace-plots of the log-likelihood instead of trace-plots of individual parameters.

3. Mixture model applied to movement ecology

3.1. Model description

Hidden Markov models (HMMs) have been extensively used to identify latent behavioral states (e.g., encamped, area restricted search, and transit) based on metrics derived from GPS location data, such as step lengths and turning angles (Morales et al., 2004). The estimation of latent states is valuable to the understanding of animal movement patterns since these states can be used to characterize the function of movements across a landscape when organisms are not directly observable (McClintock et al., 2020; Patterson et al., 2017; Wittemyer et al., 2019). By evaluating behavior-specific movements in relation to environmental covariates, mechanistic drivers of movement and measures of habitat suitability can be discerned from a variety of models, such as resource (Manly et al., 2002) or step-selection functions (Abrahms et al., 2016; Fortin et al., 2005; Wilson et al., 2012; Wittemyer et al., 2019). These latent states can also be used to infer activity budgets, providing a link to an animal's relative energy expenditure (Attias et al., 2018; Christiansen et al., 2013; McClintock et al., 2013; Wilmers et al., 2017; Wilson et al., 2020).

Similar to HMMs, our model also identifies these latent behavioral states, but it does not rely on a Markovian assumption or an underlying mechanistic movement process (e.g., correlated random walk). As a result, our model is better described as a mixture model instead of a HMM.

Furthermore, instead of using the probability density functions typically adopted to model step lengths (e.g., gamma and Weibull) and turning angles (e.g., von Mises and wrapped Cauchy), we discretize these data and use a conditional categorical distribution as the likelihood. While discretizing the data arguably results in the loss of some information content, this approach has the benefit of being able to represent standard and non-standard distributions. This is important because parametric models can be prone to model misspecification (Diana et al., 2020) and it has been shown that relatively minor discrepancies between the data and the standard distributions often adopted within HMMs can lead to the identification of additional but superfluous latent states (Pohle et al. 2017). We believe that the flexibility in representing the distributions of step lengths and turning angles outweighs the relatively minor loss of information, particularly in the context of the large number of observations that arise from these GPS sensors.

Let $y_i^{(1)}$ and $y_i^{(2)}$ denote the step length and turning angle bins, respectively, that observation *i* falls into. We assume that:

$$y_i^{(1)} | z_i = k \sim Cat(\boldsymbol{\phi_{k1}})$$

$$y_i^{(2)} | z_i = k \sim Cat(\boldsymbol{\phi_{k2}})$$

where $z_i = k$ is the latent cluster memberships of observation *i* for data type 1 and 2, respectively. The vectors $\boldsymbol{\phi_{k1}}$ and $\boldsymbol{\phi_{k2}}$ contain the probability that step lengths and turning angles, respectively, fall in each bin given that these observations were assigned to cluster *k*. Finally, our priors are:

$$\boldsymbol{\phi_{k1}} \sim Dirichlet(a)$$

$$\boldsymbol{\phi_{k2}} \sim Dirichlet(a)$$

We assume a relatively sparse prior Dirichlet distribution for $\boldsymbol{\phi_{k1}}$ and $\boldsymbol{\phi_{k2}}$ by setting a to 0.1.

3.2. Simulated movement data

We systematically varied the number of clusters K from 2 to 10 and simulated 10 datasets for each setting. We assumed that both step lengths and turning angles were discretized into 15 bins. Each simulated dataset contained 15,000 observations closely following the generative model described above. To ensure that each cluster was sufficiently distinct from the other clusters, we relied on a discretized normal distribution for $\boldsymbol{\phi_{k1}}$ and $\boldsymbol{\phi_{k2}}$. We assumed that the means were evenly distributed across the 15 bins and that the standard deviation was ¼ of the distance between means. For example, for 3 clusters, this discretized normal distribution peaked at the 1st, 8th, and 15th bins, respectively, and the standard deviation was equal to (8-1)/4=1.75. Finally, we set the maximum number of groups to 15 and ran the Gibbs samplers for 10,000 iterations for each simulated dataset.

3.3. Empirical movement data

We rely on GPS telemetry data from 20 individuals of the southern three-banded armadillo (*Tolypeutes matacus*), a species classified as Near Threatened (A2cd) by the IUCN Red List of Threatened Species and highly prioritized for conservation in Brazil (ICMBio, 2014). These data were collected from two sites in the Pantanal wetlands of Brazil using a GPS tracking device with approximately 5-min interval fixes. For each captured individual, age, sex, reproductive status, and body mass were measured. Additional information regarding this system and data can be found in Attias et al. (2020).

Location errors and missing location fixes are widely acknowledged problems with GPS location data (Bjorneraas et al., 2010; Ranacher et al., 2016). To properly analyze these data, we used a number of filtering steps. First, we excluded the data that were collected while armadillos

were in their burrows, usually during the daytime, since no movements occurred during this time. Second, we only retained measurements at 5 ± 1 min intervals to ensure that the derived speed and turning angles were comparable. Speed was calculated as step length divided by the time interval (i.e., the amount of time between successive GPS fixes) and turning angle was calculated as the change in direction between successive steps. Finally, we removed observations for which speed was greater than the 99.9% percentile (equal to 0.71 m/s) to remove biologically implausible movements. After all of these filtering steps, our final data set contained 13,671 observations from 20 individuals. Speed was discretized into bins of equal widths (0.1 m/s) up to 0.6 m/s with the final bin containing all observations > 0.6 m/s, resulting in 7 bins. Turning angle was discretized into 10 evenly spaced bins between $-\pi$ and π.

We set the maximum number of groups to 15 and ran our Gibbs sampler for 20,000 iterations, discarding the first half of the iterations as burn-in. Besides identifying and characterizing each behavioral state, the goal of this analysis was also to gain insights regarding the basic ecology of this poorly known species. To this end, we explore how different factors influence the probability of each behavioral state using a post-hoc generalized linear mixed model (GLMM) with random effects for each individual using the R package 'lme4' (Bates et al., 2015). In addition to the individual level information from the armadillos, we include time of day, ambient temperature, and precipitation as additional predictor variables. Temperature and precipitation are based on daily averages obtained from automatic stations of the Brazilian National Institute of Meteorology (INMET), located in the municipality of each study site. Differently from the clustering process, in which a single model was fit to data from both sexes, we fitted separate GLMM models for females and males. In these models, random effects were included for each individual. It is important to note that uncertainty from the mixture model is

not appropriately propagated to the GLMM parameter estimates. However, we believe that this model can still be useful in helping to interpret the results from the mixture model.

4. Mixture model applied to species occurrence data (i.e., Biogeographic Region model)

4.1. Model description

In this section, we focus on the clustering of locations with similar species composition. These locations are often spatially clustered, resulting in areas that have been variously called biogeographical regions (BR), bioregions, regions of common profile, forest types, or bird conservation regions in the literature. The delineation of these areas is a common task in ecology because it has important implications for both basic and applied scientific questions, such as those in historical biogeography, conservation, and natural resources management (Hill et al., 2017; Kreft and Jetz, 2010; Vilhena and Antonelli, 2015). A review of methods to delineate these areas is provided in Hill et al. (2020).

Let y_{is} denote the number of times that species s (s=1,…,S) was seen in location i (i=1,…,n). We assume that y_{is} arises from a Binomial distribution given by:

$$y_{is}|z_i = k \sim Binomial(n_i, \phi_{ks})$$

where z_i is the latent group membership and n_i is the number of observation opportunities in location i. Notice that z_i influences this Binomial distribution by determining the subscript of the parameter ϕ_{ks}, where k=1,…,K. The parameter ϕ_{ks} represents the presence probability of species s if location i belongs to cluster k. Therefore, the vector $[\phi_{k1}, \dots, \phi_{kS}]$ characterizes cluster k in relation to its species composition. Finally, we adopt the following priors:

$$\phi_{ks} \sim Beta(a, b)$$

We assume that a=b=1, resulting in a uniform prior for ϕ_{ks}.

4.2. Simulated biogeographic data

The true number of groups K was set to 2, 4, 8, 16 and 32 for this model. Ten datasets were generated for each setting and all simulated datasets had 2679 locations and 443 species, similar to the bird data set that was used for one of our case studies. We generated the simulated data closely following the generative model described above. To retrieve the true number of groups, we set the maximum number of groups to 50 and ran the Gibbs samplers for 1,000 – 5,000 iterations for each simulated dataset.

4.3. Empirical biogeographic data

The Alberta Biodiversity Monitoring Institute (ABMI) monitors large-scale responses of biodiversity to environmental change in Alberta, Canada. The program reports on the status and trends of species by establishing species-habitat relationships, determining species' response to various land-use changes, and producing predictive maps. The information on the trend and status of biodiversity, derived from these species-specific results, is then used to support natural resource and land-use decision making in Alberta. While species-specific models are typically created, results are often summarized across species depending on their shared response to natural or human disturbance (e.g., forestry, agriculture) to highlight major results that can be of particular interest in a given region.

The goal of this analysis is to identify the major plant communities in the forested and prairie regions of Alberta, enabling the characterization of biodiversity across large spatial-scales. These results are useful in summarizing the response patterns to disturbances of a large number of species, helping to convey the results (display, interpret, and explain) to land-use managers. We

used presence/absence data on vascular plant species in Alberta collected by the Alberta Biodiversity Monitoring Institute (ABMI). Sites surveyed by ABMI in terrestrial habitat were spaced throughout Alberta using the 20 km National Forest Inventory grid. At each site, a 100m × 100m survey area was established and each survey area was further divided into four 50m × 50m square plots. Because our model lumps these four square plots together, the number of observation opportunities in survey area i (i.e., n_i) is equal to 4 and the number of times species s was seen in this survey area (i.e., y_{is}) is an integer between 0 and 4. All vascular plant species in these plots were identified and their presence/absence in each of these plots was recorded. We focused on data from 2007 to 2018 because of the consistent data collection protocol from this period of time. We also eliminated data from very rare species, defined as species that were present in less than 1% of the sites, resulting in a final dataset with a total of 1,082 sites and 351 species. Details about data collection can be found in ABMI (2014).

The maximum number of groups was set to 50 and the Gibbs sampler was run for 10,000 iterations, discarding the first half of the iterations as burn-in. To enable the visualization of the spatial distribution of the identified clusters, we fit post-hoc Bayesian logistic regressions to the results from the BR model and then use these regression models to create spatial predictions. Predictor variables for these logistic regressions included two climate variables (i.e., mean annual temperature and precipitation) and percentage of land area covered by nine habitat types (i.e., deciduous forest, pine forest, white spruce forest, mixed wood forest, black spruce forest, fens with trees, swamps with trees, open wetland [fen/marsh], and grass/shrub), and five types of anthropogenic landscapes (i.e., harvested forest stands, vegetated strips along linear features [e.g., trails, roads, and railways], crops/pastures, urban industrial/mines, and paved and gravel roads). We considered these variables because they are biologically meaningful and are available

433 throughout the entire study area at the spatial scale of 1 km^2. Similar to our analysis of the
434 mixture model results applied to the movement data, it is important to note that uncertainty from
435 the BR model is not appropriately propagated to the logistic regression parameter estimates.
436 However, we believe that the derived maps based on this logistic regression can still be useful in
437 helping to interpret the results from this mixture model.

438

439 ## 5. Species Archetype models

440 ### 5.1. Model description

441 Species Archetype (SA) models were originally developed by Dunstan et al. (2011) to cluster
442 species that responded similarly to environmental gradients (i.e., species that had similar
443 regression parameters). While the original model followed a relatively standard mixture of
444 regression models approach (Grun and Leisch, 2008), this model was subsequently improved by
445 allowing each species to have a separate intercept (Dunstan et al., 2013), enabling species-
446 specific differences in overall prevalence.

447 SA models have been put forward as a potentially effective strategy to group species,
448 resulting in species archetypes (i.e., groups of species that respond in a similar fashion to the
449 environment) (Dunstan et al., 2011; Dunstan et al., 2013; Hui et al., 2013). Furthermore, SA
450 model results can also simplify conservation management decision by enabling managers to
451 focus on a small set of species archetypes, instead of having to evaluate a multitude of species,
452 each with their own idiosyncratic response to the environment (Dunstan et al., 2011; Dunstan et
453 al., 2013; Hui et al., 2013).

454 Let y_{is} denote the presence (=1) or absence (=0) of species s (s=1,…,S) in location i
455 (i=1,…,n). Because this is a binary variable, we assume the following Bernoulli distribution:

$$y_{is}|z_s = k \sim Bernoulli\left(\Phi(\alpha_s + \boldsymbol{x_i^T}\boldsymbol{\beta}_k)\right)$$

where z_s is the latent cluster membership of species *s*, α_s is a species-specific intercept, $\boldsymbol{x_i^T}$ is a vector of location-specific covariates, $\Phi()$ is the standard normal cumulative distribution function, and $\boldsymbol{\beta}_k$ is a vector containing the regression slopes for cluster *k*, where k=1,…,K. Notice that the latent variable z_s influences this Bernoulli distribution by determining the subscript of the vector $\boldsymbol{\beta}_k$. In other words, species that belong to the same cluster *k* have the same slope parameters $\boldsymbol{\beta}_k$, essentially having identical responses to covariates. We adopt a probit link (instead of the more common logit link) because it enables the straight-forward fitting of the model using the data augmentation scheme described in Albert and Chib (1993). More specifically, we assume the existence of another set of latent variables ω_{is} such that:

$$y_{is} = 1 \text{ if } \omega_{is} > 0$$

$$y_{is} = 0 \text{ otherwise}$$

where

$$\omega_{is} \sim N(\alpha_s + \boldsymbol{x_i^T}\boldsymbol{\beta}_k, 1)$$

For the remaining parameters, we adopt the following priors, given by:

$$\alpha_s \sim N(0,10)$$

$$\boldsymbol{\beta}_k \sim N(\mathbf{0}, \mathbf{I})$$

where $\mathbf{I}$ is the identity matrix.

5.2. Simulated data

Similar to the BR model, the true number of groups *K* was set to 2, 4, 8, 16 and 32 for this model. Ten datasets were generated for each setting and, similar to the empirical data, all simulated datasets had 2679 locations and 443 species. We generated the simulated data closely following the generative model described above. Furthermore, we simulated the slope

parameters $\boldsymbol{\beta}_k$ from standard normal distributions. However, to simulate the species-specific intercepts, we assumed that $\alpha_s \sim N(0, 0.4^2)$. The standard deviation for α_s was set to 0.4 to avoid creating data in which certain species are almost always present or almost always absent. This is important because it would be difficult for our models to assign these species to their correct groups given that this assignment depends on the accurate estimation of the slope parameters $\boldsymbol{\beta}_k$. We also assumed that six uncorrelated covariates were available, which were generated from standard normal distributions. The maximum number of groups was set to 50 and the Gibbs samplers were run for 1,000 – 5,000 iterations for each simulated dataset.

5.3. Empirical data

The Breeding Bird Survey (BBS) is a long-term program that monitors the status and trend of bird populations in North America. In brief, data are collected annually in June by trained participants along randomly established roadside routes approximately 39 km long with stops 0.8 km apart. At each stop, a 3-min point count is conducted (Pardieck et al., 2017).

The BBS actually records count data (rather than presence/absence) per stop in each route. However, these counts may include the same individual observed multiple times and bird detection may vary by species and environmental conditions (e.g., weather or traffic noise). To avoid some of the issues with the count data and to be able to illustrate the use of the SA model described previously, we convert these count data into presence/absence of each species in each route. Furthermore, we subset the BBS data for the year of 2015 and eliminate data from very rare species, defined here as species that were present in less than 10 routes. In total, the final dataset used for analysis contained information on 443 species and 2679 routes, spread throughout Canada and the United States.

To understand how bird species are affected by climatic variables, we gathered average temperature and precipitation for North America from 1970-2000 for the month of June from WorldClim (www.worldclim.org/version2), with a spatial resolution of 5 arc-minutes (10 km grid). Because each species archetype can potentially have non-linear associations with precipitation and temperature, we relied on B-splines to capture the association between these environmental variables and species presence. More specifically, B-spline basis functions were included in the model for temperature and precipitation, where knots were *a priori* set to 10%, 20%, …, 90% percentiles of the corresponding environmental variables. Additional information regarding different types of splines and basis functions can be found in Wood (2017) and similar functional clustering ideas can be found in Dunson (2010). By identifying the niche breadth in relation to temperature and precipitation of the different species groups, this analysis is able to identify which of these groups are more likely to be impacted by changes in precipitation, changes in temperature, or changes in both variables. We set the maximum number of groups to 50 and ran our Gibbs sampler for 10,000 iterations, discarding the first half of the iterations as burn-in.

6. Software

To enable readers to reproduce the results described in this article and to use the different models highlighted here for their own data, we have created an R package called EcoCluster that enables straight-forward fitting of the BR model and the SA model, both with the TSB prior. We have also created an R package called bayesmove (Cullen et al., in review) that enables the straight-forward fitting of the mixture model with the TSB prior used for the movement data. These packages can be readily downloaded from our public GitHub account

(https://github.com/drvalle1/EcoCluster) and CRAN (https://CRAN.R-project.org/package=bayesmove). The vignettes accompanying these R packages illustrate how to fit these models and interpret their results.

Results

1. Simulated data results

We find that the proposed models with the TSB prior are able to successfully recover the true number of groups for all models, with a slight decrease in performance for the BR model when the true number of groups is equal to 32 (top panels in Fig. 2). Importantly, we find that the data contain substantial information on γ (the parameter that governs the TSB prior), with posterior means for γ that are relatively small for sparse settings (i.e., when only few groups exist) versus closer to 1 when many groups exist (bottom panels in Fig. 2). We also find that all of the proposed algorithms were able to accurately retrieve the parameter values used to simulate the data (i.e., $\boldsymbol{\phi_{k1}}$ and $\boldsymbol{\phi_{k2}}$ for the mixture model applied to the movement data; ϕ_{ks} for the BR model; and $\alpha_s, \boldsymbol{\beta}_k$ for the SA model; data not shown).

(insert Fig. 2)

The standard approach of fitting models with different number of clusters and then selecting the optimal number of clusters is much more computationally expensive. For example, using the real datasets, we found that the time required to fit the mixture model with the TSB prior applied

to movement data and a maximum of 15 groups was equal to 17% (25 min. vs. 149 min.) of the time need to fit multiple mixture models, one for each number of clusters (2 to 15). Similarly, fitting the BR model once with the TSB prior and a maximum of 50 groups corresponds to approximately 6% (10 min. vs. 189 min.) of the total time required to vary the number of clusters from 2 to 50 and fit individual BR models for each setting. Finally, fitting the SA model once with the TSB prior and a maximum of 50 groups took 4% (40 min. vs. 990 min.) of the time required to run multiple SA models, one for each pre-specified number of clusters (2 to 50).

2. Empirical results for the mixture model applied to movement data

Our model identified two behavioral states (out of a maximum of 15 possible states) that together comprise 99% of all observations. The first state is comprised mostly of slower and more tortuous movements (hereafter "foraging" state, Figs. 3a and 3b) while the second state includes faster and more directed movements (hereafter "transit" state, Figs. 3c and 3d). When exploring these results, we find that, while the daily number of observations assigned to the foraging state is very similar between males and females (Fig. 3e), males tend to have a higher number of observations assigned to the transit state (Fig. 3f). To determine how covariates influence these behavioral states, we fit a post-hoc generalized linear mixed model (GLMM), where the binary response variable was equal to 1 for the transit state and 0 otherwise. The larger proportion of the transit state for males in comparison to females is evident by the much larger intercept for males when compared to females (Table 2). Furthermore, we find a quadratic relationship between the probability of exhibiting the transit state with time of night (see Fig. 3g). Finally, we do not find a strong influence of precipitation, temperature, or region, on the proportion of the transit state (Table 2). These results suggest that armadillos from both regions

570 behave similarly and that precipitation/temperature have no measurable effect on the proportion
571 of the transit state, despite the fact that decreased daily temperatures have been associated with
572 an overall lower duration of activity period (Attias et al., 2018).
573
574 (insert Fig. 3)
575 (insert Table 2)
576
577 ## 3. Empirical results for the BR model applied to the vascular plant data from Alberta
578 The BR model identified 7 (out of 50) major groups, representing 99.5% of all locations.
579 This analysis resulted in substantial dimension reduction given that, instead of having to
580 separately examine the results for 351 species, the BR model enables us to focus just on the
581 results from these 7 groups. To simplify the interpretation and enable the spatial visualization of
582 the patterns identified by the BR model, we fitted a post-hoc logistic regression to the results
583 from this model. Predictions from these regression models reveal striking spatial patterns (Fig.
584 4). For example, group three had a strong association with temperature and precipitation, with
585 most species in this group being relatively rare species that are mainly restricted to the colder
586 Rocky Mountains, Upper Foothills, and Canadian shield natural regions (e.g., Engelmann Spruce
587 *Picea engelmannii* and Rocky Mountain alpine fir *Abies bifolia*). On the other hand, group two
588 had a positive association with most of the other remaining upland forest types (i.e., deciduous,
589 white spruce, mixed wood, and harvested stands). Groups 4 and 5 were mostly restricted to
590 lowland forest types (black spruce and fens with trees). Interestingly, group one was strongly
591 associated with the proportion of cultivated land (e.g., crop and pasture), agreeing with the fact
592 that many of the species that dominate this group are either introduced or cultivated species (e.g.,

593 Canola *Brassica napus* and Barley *Hordeum vulgare*). Group seven, on the other hand, was
594 strongly associated with highly anthropogenic landscapes, with substantially increased presence
595 probabilities in regions with higher urban/industrial/mines areas and associated vegetated strips
596 along railways, roads and trails (soft-linear). The characteristic species in this group include
597 White Sweet Clover *Melilotus albus*, Yellow Sweet Clover *M. officinalis*, and Scentless
598 Chamomile *Tripleurospermum inodorum.*
599
600 (insert Fig. 4)
601
602 ## 4. Empirical results for the SA model applied to the Breeding Bird survey data
603 All of the 50 species groups in the SA model had species in them but 95% of all the species
604 were contained in the first 40 of these groups. As expected, several groups were strongly
605 associated with temperature and/or precipitation, typically exhibiting unimodal relationships
606 between average prevalence and these environmental variables. An example of the results for 4
607 species groups are shown in Fig. 5. The results for all the other species groups are available in
608 Appendix S2.
609
610 (insert Fig. 5)
611
612 The line graphs in Fig. 5 illustrate how all the species within a species archetype respond
613 in a similar fashion to the environment. The heat maps of the predicted average prevalence for
614 different combinations of temperature and precipitation provide a depiction of the environmental
615 space occupied by these species groups (i.e., the realized niche, Fig. 5). These figures illustrate

that some species archetypes are relatively insensitive to precipitation but very sensitive to temperature (e.g., species archetype 31), some are relatively insensitive to temperature but very sensitive to precipitation (e.g., species archetype 5), while finally some groups are sensitive to both temperature and precipitation (e.g., species archetype 21). These results can potentially be useful to highlight which sets of species are more likely to be impacted by different facets of climate change (e.g., Tingley et al., 2012), enabling the prioritization of these species for conservation purposes.

Discussion

Determining the number of clusters is a long-standing challenge for a range of clustering algorithms. The standard approach to deal with this problem for model-based clustering consists of fitting models with different number of groups and selecting the optimal number of groups using indices such as AIC or BIC, an approach that can be very computationally intensive and that has been reported to often overestimate the true number of groups. Here we show how Bayesian clustering models, when used in conjunction with sparsity inducing priors such as the TSB prior described here, can determine the number of clusters without requiring the fitting of multiple models.

To illustrate how a wide range of Bayesian clustering models can benefit from sparsity-inducing priors, we show with simulated data how the truncated-stick breaking (TSB) prior can successfully estimate the true number of groups for three types of clustering models (i.e., two mixture models, one applied to movement data and the other applied to species occurrence data, and a SA model which clusters species according to how they respond to the environment).

Nevertheless, we believe that the ability to identify the existing clusters is likely to depend on several factors, including the type of model, how distinctive the clusters are from one another, the size of each group, and the amount of available data. For example, additional simulations in which groups were allowed to vary in size revealed that the BR model did not perform as well as the other models in this setting (Appendix S3). A closer examination of the BR model results revealed that this model had a challenging time correctly assigning some of the plots to the rare groups (i.e., groups assigned to less than 10 plots) because these groups were rare and therefore much harder to characterize. Additional research needs to be conducted to better characterize the circumstances in which the TSB prior is likely to work well and when it is likely to fail.

We also show that the standard approach of varying the number of groups and fitting multiple models is much more computationally expensive. Some might argue that using AIC or BIC based on fitting multiple models is only computationally problematic if the algorithms used to fit these models are slow (e.g., MCMC algorithms). Our experience has been that several of the alternative clustering models that rely on optimization (e.g., SAM and HMM in the "ecomix" and "momentuHMM" R packages, respectively) instead of MCMC algorithms are also relatively slow because they often require multiple model fits for a given number of groups due to the multimodality of the likelihood surface. This is further exacerbated if different numbers of groups need to be tested and a bootstrapping approach is required to estimate parameter uncertainty (e.g., as in SAM within the "ecomix" package). As a result, despite the intuition that optimization algorithm will always be faster than MCMC algorithms, in practice this is not always true because of the multiple model fits that are required by these optimization-based methods.

660 We demonstrate how these models can unveil important environmental management and
661 ecological insights. In the mixture model applied to the movement data from the three-banded
662 armadillos, we identified two latent behavioral states which were labeled foraging and transit.
663 Additionally, we found that males tend to exhibit a greater proportion of time in the transit state
664 than females and that the proportion of this state peaks midway through the night. These sexual
665 differences regarding the transit state are likely related to the species' socio-biology, as the
666 increased transit state of the promiscuous males should increase their chances of encountering
667 receptive females. Indeed, males have been recently shown to have larger home ranges than
668 females (Attias et al., 2020) and, according to our results, this difference is unlikely to be related
669 to the acquisition of energetic resources by the larger males, as there were no noticeable
670 differences in the amount of foraging state between sexes (Fig. 3e).
671 The BR model enabled substantial dimension reduction by summarizing the results from 351
672 species into 7 major groups. Similar to forest types
673 (https://data.fs.usda.gov/geodata/rastergateway/forest_type/) and Bird Conservation regions
674 (https://nabci-us.org/resources/bird-conservation-regions/), these results can be used for
675 conservation and management purposes. For example, our results have identified a plant
676 community that is heavily influenced by anthropogenic disturbance. By mapping the spatial
677 distribution of this group, our analysis can enable the spatial prioritization of restoration and
678 invasive species elimination initiatives. Furthermore, the monitoring of this group is likely to be
679 critical in identifying the main drivers of environmental change in the region and developing
680 effective mitigation strategies.
681 In relation to the SA model applied to the 2015 survey data on North American breeding
682 birds, we have identified species groups that respond similarly to temperature and precipitation.

683 This enables the identification of sets of species that are likely to be more impacted by changes
684 in precipitation, by changes in temperature, or by both. Interestingly, differently from the other
685 two applications, the SA model still identified the existence of 50 groups, which was the
686 maximum number of groups allowed by our analysis. These results suggest that there are
687 probably more groups than what we have allowed for in this analysis. We believe that this might
688 be due to the flexibility of the environmental response curves and the relatively rigid structure of
689 SA models, which require species to have the same set of slope parameters. As a result,
690 relatively minor changes in how these species respond to their environment, particularly when
691 there are a lot of observations for any given species, can foster the creation of many small groups
692 instead of few large groups. Future research could devise a different formulation for the SA
693 model so that species can be grouped together even if they differ slightly in how they respond to
694 the environment. These results are also important to highlight that, despite the use of a sparsity-
695 inducing prior, the model might still reveal that a sparse solution (i.e., a few clusters) is not
696 supported by the data. In this situation, the modeler has to decide to either use the results as they
697 are, because a larger number of groups would be unmanageable, or re-run the analysis with a
698 larger number of groups.

699 An important limitation in our analysis of the armadillo movement data and the plant
700 occurrence data from Alberta is that we relied on post-hoc regression models to better interpret
701 our mixture model results. The problem with this two-stage approach is that it does not properly
702 propagate the uncertainty associated with the mixture model results, potentially leading to over-
703 confident inference and predictions. While this might not be too troublesome for exploratory
704 studies like ours, this is an important problem for more confirmatory analyses. There are
705 relatively few methods that have been developed that avoid these post-hoc analyses (see review

706 in Hill et al. (2020)). Nevertheless, the few existing single-stage methods require multiple
707 models to be fit to determine both the optimal number of groups and the optimal set of
708 covariates. Properly propagating the uncertainty associated with all these decisions is an area of
709 active research even for these single-stage models.

710 It is important to note that, because our primary goal was to show the versatility of the TSB
711 prior, we have not provided a more in-depth comparison of the three example models to other
712 commonly used models. Such a model comparison could be useful future research. Furthermore,
713 we have focused on ecological clustering applications, but the TSB prior is likely to be useful for
714 a much broader range of applications (e.g., for use of the Dirichlet process for genetic clustering,
715 see Huelsenbeck and Andolfatto (2007) and references therein). Also, we have focused on
716 models where the primary interest is on the identified latent structure (i.e., the identified clusters)
717 because we believe that this is the type of dimension-reduction result that ecologists find more
718 revealing and insightful. Indeed, many of the ecological applications involving the Dirichlet
719 process and its extensions rely on the identified clusters to draw insights regarding, for example,
720 animal movement and migration patterns (Diana et al., 2020; Valle et al., 2017), temporal
721 dynamics of seal pup rookeries (Johnson et al., 2013), and spatial distribution of bird
722 communities (Valle et al., 2018). However, we acknowledge that the Dirichlet process has been
723 used for a much wider range of applications, some of which are not focused on identifying
724 clusters. For example, in ecology, the Dirichlet process has been used for density estimation
725 (Dorazio et al., 2008), to develop spatial models of the expected number of birds (Rodriguez and
726 Dunson, 2011), and to generate a more parsimonious description of the covariance matrix
727 between species in joint species distribution models (Taylor-Rodriguez et al., 2017).

It is important to note that our approach does not apply to all clustering methods. For instance, many clustering approaches are algorithmic and do not rely on an underlying statistical model, precluding the use of our approach. Even among clustering approaches based on statistical models, adopting a prior will only make sense for models fitted within a Bayesian framework. Finally, attempts to fit models with the TSB prior using packages such as JAGS (Plummer, 2003) and Stan (Stan Development Team, 2020) may result in label switching and convergence problems (e.g., Sollmann et al., 2020). The reason for this is that we have observed that a critical step for our customized MCMC algorithms to perform well is to order the identified clusters (from largest to smallest) during the burn-in phase. While this ordering does not change the likelihood (cluster labels are unidentified in standard mixture models), it does influence the TSB prior. Mixture models often have multimodal posteriors/likelihood functions (Scrucca et al., 2016; Stephens, 2000) and the ordering of clusters helps the model with the TSB prior find the highest peak, this way reducing label switching and convergence issues. Developing approaches for ordering clusters within packages, such as JAGS or Stan, is an important area for future research.

Despite the limitations described above, it is likely that clustering approaches will greatly benefit from sparsity-inducing approaches like the TSB prior in the same way that a wide range of regression models has benefitted from sparsity-inducing approaches (e.g., regularization penalties or strong priors) to improve predictions and identify the most important predictor variables (Hooten and Hobbs, 2015). Several extensions to this prior, already developed in Bayesian nonparametrics, may be profitably exploited in the future for clustering applications in ecology and environmental science. For example, the Pitman-Yor process (also known as the two-parameter Poisson Dirichlet process; Pitman, 1995; Pitman and Yor, 1997) is a

751 generalization of the Dirichlet Process that offers more flexible clustering rates and cluster size
752 tail behaviors. Alternatively, a hierarchical Dirichlet process (Teh et al., 2006) could be used to
753 capture nested clusters. Finally, a truncated stick-breaking prior together with a kernel based
754 approach (Reich and Fuentes, 2007) could be used so that spatially proximate sites are more
755 likely to cluster together, or a probit stick-breaking process (Rodriguez and Dunson, 2011) could
756 be adopted to ensure that sites with similar environmental conditions and/or species with similar
757 trait values are more likely to be clustered together.
758 Despite the fact that the stick-breaking prior has a relatively long-history in statistics,
759 relatively few ecological modelers have used this prior, probably due to the general lack of
760 awareness among quantitative ecologists and environmental scientists regarding how this prior
761 can help a wide range of applied clustering problems. This is particularly surprising given the
762 prominent role of clustering methods in ecology and related disciplines (Legendre and Legendre,
763 2012). With this article, we hope to help remedy this situation by better characterizing the
764 benefits of this approach while at the same time providing R packages that implement these
765 methods, enabling the straight-forward fitting of these models by quantitative scientists.
766

767 **Acknowledgements**

768 This work was partly supported by the US Department of Agriculture National Institute of
769 Food and Agriculture McIntire–Stennis project 1005163 and by the US National Science
770 Foundation awards 1458034 and 2040819 to DV. The funders had no role in study design, data
771 collection and analysis, decision to publish, or preparation of the manuscript. NA thanks the
772 Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES, number 1575316) and

773 Fundacao de Apoio ao Desenvolvimento do Ensino, Ciencia e Tecnologia do Estado de Mato
774 Grosso do Sul (FUNDECT, process 23/200.715/2013) for the scholarships granted, and Teak
775 Resources Co., Embrapa Pantanal, and Fazenda Santa Tereza for logistic support during the
776 fieldwork.
777

778 **Author's contributions**

779 DV conceived the ideas, designed the methodology, analyzed the data, and wrote the first draft
780 of the article. NA provided the armadillo movement data. NA and JC helped to interpret the
781 results from the mixture model applied to this movement data. YJ obtained the covariate data for
782 the SA model. BB created the EcoCluster R package while JC created the bayesmove R package.
783 EA provided the ABMI dataset and helped interpret the corresponding results. All authors
784 contributed critically to the manuscript and gave final approval for publication.
785

786 **Literature cited**

787 ABMI, 2014. Terrestrial field data collection protocols (abridged version). Alberta Biodiversity
788 Monitoring Institute, Alberta, Canada.
789 Abrahms, B. et al., 2016. Lessons from integrating behavior and resource selection: activity-
790 specific responses of African wild dogs to roads. Animal Conservation, 19(3): 247-255.
791 Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data.
792 Journal of the American Statistical Association, 88(422): 669-679.

793 Attias, N., Gurarie, E., Fagan, W.F., Mourao, G., 2020. Ecology and social biology of the
794 southern three-banded armadillo. Journal of Mammalogy: 1-14.
795 Attias, N., Oliveira-Santos, L.G.R., Fagan, W.F., Mourao, G., 2018. Effects of air temperature on
796 habitat selection and activity patterns of two tropical imperfect homeotherms. Animal
797 Behavior, 140: 129-140.
798 Auerbach, D.A. et al., 2016. Towards catchment classification in data-scarce regions.
799 Ecohydrology, 9: 1235-1247.
800 Azeria, E., Sanmartin, I., As, S., Carlson, A., Burgues, N., 2007. Biogeographical patterns of the
801 East African coastal forest vertebrate fauna. Biodiversity and Conservation, 16: 883-912.
802 Azeria, E.T. et al., 2009. Using null model analysis of species co-occurrences to deconstruct
803 biodiversity patterns and select indicator species. Diversity and Distributions, 15: 958-
804 971.
805 Baskerville, E.B. et al., 2011. Spatial guilds in the Serengeti food web revealed by a Bayesian
806 group model. PLOS Comp Biol, 7(12): e1002321.
807 Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using
808 lme4. J Stat Softw, 67(1): 1-48.
809 Berkhin, P., 2006. A survey of clustering data mining techniques. In: Kogan, J., Nicholas, C.,
810 Teboulle, M. (Eds.), Grouping multidimensional data. Springer, Berlin, Heidelberg.
811 Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the
812 Integated Completed Likelihood. IEEE Transactions on Pattern Analysis and Machine
813 Intelligence, 22(7): 719-725.

814 Bjorneraas, K., van Moorter, B., Rolandsen, C.M., Herfindal, I., 2010. Screening Global
815 Positioning System Location Data for Errors Using Animal Movement Characteristics.
816 The Journal of Wildlife Management, 74(6): 1361-1366.
817 Blackwell, D., MacQueen, J.B., 1973. Ferguson distributions via Polya urn schemes. Annals of
818 Statistics, 1: 353-355.
819 Bouveyron, C., Brunet-Saumard, C., 2014. Model-based clustering of high-dimensional data: a
820 review. Computational Statistics and Data Analysis, 71: 52-78.
821 Casella, G., Moreno, E., Giron, F.J., 2014. Cluster analysis, model selection, and prior
822 distributions on models. Bayesian Analysis, 9(3): 613-658.
823 Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. NbClust: An R package for
824 determining the relevant number of clusters in a data set. J Stat Softw, 61(6).
825 Christiansen, F., Rasmussen, M.H., Lusseau, D., 2013. Inferring activity budgets in wild animals
826 to estimate the consequences of disturbances. Behavioral Ecology, 24(6): 1415-1425.
827 Crane, H., 2016. The ubiquitous Ewens sampling formula. Statistical Science, 31(1): 1-19.
828 Daudin, J.J., Picard, F., Robin, S., 2008. A mixture model for random graphs. Statistics and
829 Computing, 18: 173-183.
830 Depraetere, N., Vandebroek, M., 2014. Order selection in finite mixtures of linear regression:
831 literature review and a simulation study. Statistical Papers, 55: 871-911.
832 Diana, A., Matechou, E., Griffin, J., Johnston, A., 2020. A hierarchical dependent Dirichlet
833 process prior for modelling bird migration patterns in the UK. Annals of Applied
834 Statistics, 14(1): 473-493.
835 Dorazio, R.M. et al., 2008. Modeling unobserved sources of heterogeneity in animal abundance
836 using a Dirichlet process prior. Biometrics, 64: 635-644.

837 Dunson, D.B., 2010. Nonparametric Bayes applications to biostatistics. In: Hjort, N.L., Holmes,
838 C., Muller, P., Walker, S.G. (Eds.), Bayesian nonparametrics. Cambridge University
839 Press, Cambridge, UK.
840 Dunson, D.B., Xing, C., 2009. Nonparametric Bayes modeling of multivariate categorical data.
841 Journal of the American Statistical Association, 104(487): 1042-1051.
842 Dunstan, P.K., Foster, S.D., Darnell, R., 2011. Model based grouping of species across
843 environmental gradients. Ecol Modell, 222: 955-963.
844 Dunstan, P.K., Foster, S.D., Hui, F.K.C., Warton, D.I., 2013. Finite mixture of regression
845 modeling for high-dimensional count and biomass data in Ecology. Journal of
846 Agricultural, Biological, and Environmental Statistics, 18(3): 357-375.
847 Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theoretical Population
848 Biology, 3: 87-112.
849 Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics,
850 1(2): 209-230.
851 Fortin, D. et al., 2005. Wolves influence elk movements: behavior shapes a trophic cascade in
852 Yellowstone National Park. Ecology, 86: 1320-1330.
853 Foster, S.D., Hill, N.A., Lyons, M., 2017. Ecological grouping of survey sites when sampling
854 artefacts are present. J R Stat Soc Ser C Appl Stat.
855 Fraley, C., Raftery, A.E., 2007. Model-based methods of classification: using the mclust
856 software in chemometrics. J Stat Softw, 18(6).
857 Grun, B., Leisch, F., 2008. Finite mixtures of generalized linear regression models. In: Shalabh,
858 Heumann, C. (Eds.), Recent advances in linear models and related areas. Springer,
859 Heidelberg, Germany.

860 Hahn, P.R., Carvalho, C.M., 2015. Decoupling shrinkage and selection in Bayesian linear
861 models: a posterior summary perspective. Journal of the American Statistical
862 Association, 110(509): 435-448.
863 Hill, N. et al., 2020. Determining marine bioregions: a comparison of quantitative approaches.
864 Methods in Ecology and Evolution, 11: 1258-1272.
865 Hill, N.A. et al., 2017. Model-based mapping of assemblages for ecology and conservation
866 management: a case study of demersal fish on the Kerguelen Plateau. Diversity and
867 Distributions, 23: 1216-1230.
868 Hooten, M.B., Hobbs, N.T., 2015. A guide to Bayesian model selection for ecologists.
869 Ecological Monographs, 85(1): 3-28.
870 Huelsenbeck, J.P., Andolfatto, P., 2007. Inference of Population Structure Under a Dirichlet
871 Process Model. Genetics, 175: 1787-1802.
872 Hui, F.K.C., Warton, D.I., 2015. Order selection in finite mixture models: complete or observed
873 likelihood information criteria? Biometrika, 102(3).
874 Hui, F.K.C., Warton, D.I., Foster, S.D., Dunstan, P.K., 2013. To mix or not to mix: comparing
875 the predictive performance of mixture models vs. separate species distribution models.
876 Ecology, 94(9): 1913-1919.
877 ICMBio, 2014. Plano de ação nacional para a conservação do tatu-bola.
878 Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. Journal of
879 the American Statistical Association, 96(453): 161-173.
880 Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Computing Surveys,
881 31(3): 264-323.

882 Jameel, Y., Valle, D., Kay, P., 2020. Spatial variation in the detection rates of frequently studied
883 pharmaceuticals in Asian, European and North American rivers. Science of the Total
884 Environment.
885 Johnson, D.S., Ream, R.R., Towell, R.G., Williams, M.T., Guerrero, D.L., 2013. Bayesian
886 clustering of animal abundance trends for inference and dimension reduction. Journal of
887 Agricultural, Biological, and Environmental Statistics, 18(3): 299-313.
888 Kreft, H., Jetz, W., 2010. A framework for delineating biogeographical regions based on species
889 distributions. Journal of Biogeography, 37: 2029-2053.
890 Legendre, P., Legendre, L., 2012. Numerical ecology. Elsevier Science, Amsterdam.
891 Lyons, M.B., Foster, S.D., Keith, D.A., 2017. Simultaneous vegetation classification and
892 mapping at large spatial scales. Journal of Biogeography: 1-12.
893 MacArthur, R.H., 1957. On the relative abundance of bird species. Proc Natl Acad Sci USA, 43:
894 293-295.
895 MacEachern, S.N., Muller, P., 1998. Estimating Mixture of Dirichlet Process Models. Journal of
896 Computational and Graphical Statistics, 7(2): 223-238.
897 Manly, B.F.J., McDonald, L.L., Thomas, D.L., McDonald, T.L., Erickson, W.P., 2002. Resource
898 selection by animals: statistical design and analysis for field studies. Kluwer Academic
899 Publishers, The Netherlands.
900 Manrique-Vallier, D., 2016. Bayesian population size estimation using Dirichlet process
901 mixtures. Biometrics, 72: 1246-1254.
902 McClintock, B.T. et al., 2020. Uncovering ecological state dynamcis with hidden Markov
903 models. Ecol. Lett.

McClintock, B.T., Russell, D.J., Matthiopoulos, J., King, R., 2013. Combining individual animal movement and ancillary biotelemetry data to investigate population-level activity budgets. Ecology, 94(4): 838-849.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley Series in Probability and Statistics. Wiley-Interscience.

Morales, J.M., Haydon, D.T., Frair, J., Holsinger, K.E., Fryxell, J.M., 2004. Extracting more out of relocation data: building movement models as mixtures of random walks. Ecology, 85: 2436-2445.

Murugiah, S., Sweeting, T., 2012. Selecting the precision parameter prior in Dirichlet process mixture models. Journal of Statistical Planning and Inference, 142: 1947-1959.

Papaspiliopoulos, O., Roberts, G.O., 2008. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika, 95(1): 169-186.

Pardieck, K.L., Ziolkowski, D.J., Lutmerding, M., Campbell, K., Hudson, M.A.R., 2017. North American Breeding Bird Survey Dataset 1966-2016, version 2016.0, U. S. Geological Survey, Patuxent Wildlife Research Center.

Park, T., Casella, G., 2008. The Bayesian Lasso. Journal of the American Statistical Association, 103(482): 681-686.

Patterson, T.A. et al., 2017. Statistical modelling of individual animal movement: an overview of key methods and a discussion of practical challenges. AStA Advances in Statistical Analysis, 101(4): 399-438.

Pitman, J., 1995. Exchangeable and partially exchangeable random partitions. Probability Theory and Related Fields, 102: 145-158.

926 Pitman, J., Yor, M., 1997. The two-parameter Poisson-Dirichlet distribution derived from a
927 stable subordinator. The Annals of Probability, 25: 855-900.
928 Plummer, M., 2003. JAGS: A program for analysis of Bayesian graphical models using GIbbs
929 sampling.
930 Pohle, J., Langrock, R., van Beest, F.M., Schmidt, N.M., 2017. Selecting the number of states in
931 hidden Markov models: pragmatic solutions illustrated using animal movement. Journal
932 of Agricultural, Biological, and Environmental Statistics, 22(3): 270-293.
933 Ramette, A., 2007. Multivariate analyses in microbial ecology. FEMS Microbiology Ecology,
934 62: 142-160.
935 Ranacher, P., Brunauer, R., Trutschnig, W., van der Spek, S., Reich, S., 2016. Why GPS makes
936 distances bigger than they are. Int J Geogr Inf Sci, 30(2): 316-333.
937 Reich, B.J., Fuentes, M., 2007. A multivariate semiparametric Bayesian spatial modeling
938 framework for hurricane surface wind fields. The Annals of Applied Statistics, 1(1): 249-
939 264.
940 Rodriguez, A., Dunson, D.B., 2011. Nonparametric Bayesian models through probit stick-
941 breaking processes. Bayesian Analysis, 6(1): 145-178.
942 Saxena, A. et al., 2017. A review of clustering techniques and developments. Neurocomputing,
943 267: 664-681.
944 Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. mclust5: clustering, classification and
945 density estimation using Gaussian finite mixture models. R Journal, 8(1): 289-317.
946 Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica, 4: 639-650.

947 Si, Y., Reiter, J.P., 2013. Nonparametric Bayesian multiple imputation for incomplete
948 categorical variables in large-scale assessment surveys. Journal of Educational and
949 Behavioral Statistics, 38(5): 499-521.
950 Sollmann, R. et al., 2020. A Bayesian Dirichlet process community occupancy model to estimate
951 community structure and species similarity. Ecol Appl.
952 Stan Development Team, 2020. Stan Modeling Language Users Guide and Reference Manual.
953 Stephens, M., 2000. Dealing with label switching in mixture models. J R Stat Soc Series B,
954 62(4): 795-809.
955 Taylor-Rodriguez, D., Kaufeld, K., Schliep, E.M., Clark, J.S., Gelfand, A.E., 2017. Joint species
956 distribution modeling: dimension reduction using Dirichlet processes. Bayesian Analysis,
957 12(4).
958 Teh, Y.W., 2011. Dirichlet Process, Encyclopedia of Machine Learning. Springer, pp. 280-287.
959 Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. Journal
960 of the American Statistical Association, 101: 1566-1581.
961 Ter Braak, C.J.F., Hoijtink, H., Akkermans, W., Verdonschot, P.F.M., 2003. Bayesian model-
962 based cluster analysis for predicting macrofaunal communities. Ecol Modell, 160: 235-
963 248.
964 Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via
965 the gap statistic. J R Stat Soc Series B, 63(2): 411-423.
966 Tingley, M.W., Koo, M.S., Moritz, C., Rush, A.C., Beissinger, S.R., 2012. The push and pull of
967 climate change causes heterogeneous shifts in avian elevational ranges. Global Change
968 Biology, 18: 3729-3290.

969 Valle, D., Albuquerque, P., Zhao, Q., Barberan, A., Fletcher Jr., R.J., 2018. Extending the Latent
970 Dirichlet Allocation model to presence/absence data: a case study on North American
971 breeding birds and biogeographic shifts expected from climate change. Global Change
972 Biology. DOI:10.1111/gcb.14412
973 Valle, D. et al., 2017. Individual movement strategies revealed through novel clustering of
974 emergent movement patterns. Sci Rep, 7. DOI:10.1038/srep44052
975 Vilhena, D.A., Antonelli, A., 2015. A network approach for identifying and delimiting
976 biogeographical regions. Nat Commun, 6. DOI:10.1038/ncomms7848
977 Wilmers, C.C., Isbell, L.A., Suraci, J.P., Williams, T.M., 2017. Energetics-informed behavioral
978 states reveal the drive to kill in African leopards. Ecosphere, 8(6): e01850.
979 Wilson, R.P. et al., 2020. Estimates for energy expenditure in free-living animals using
980 acceleration proxies: a reappraisal. Journal of Animal Ecology, 89(1): 161-172.
981 Wilson, R.R., Gilbert-Norton, L., Gese, E.M., 2012. Beyond use versus availability: behavior-
982 explicit resource selection. Wildlife Biology, 18(4): 424-430.
983 Wittemyer, G., Northrup, J.M., Bastille-Rousseau, G., 2019. Behavioral valuation of landscapes
984 using movement data. Philos Trans R Soc Lond B Biol Sci, 374(1781).
985 Wood, S.N., 2017. Generalized Additive Models: an introduction with R. Texts in Statistical
986 Science Series. CRC Press, Boca Raton, FL, 476 pp.
987
988

Tables

Table 1. Example of the calculations involved in the truncated stick-breaking (TSB) prior assuming a maximum number of 6 cluster.

Cluster	V_k	θ_k
1	0.19	0.19
2	0.33	$0.33(1-0.19)=0.27$
3	0.27	$0.27(1-0.33)(1-0.19)=0.15$
4	0.95	$0.95(1-0.27)(1-0.33)(1-0.19)=0.38$
5	0.47	$0.47(1-0.95)(1-0.27)(1-0.33)(1-0.19)=0.01$
6	1 (by definition)	$1(1-0.47)(1-0.95)(1-0.27)(1-0.33)(1-0.19)=0.01$

994 Table 2. GLMM regression coefficients. Statistically significant ($p<0.05$) coefficients are
995 highlighted in bold.

	Female		Male	
Parameters	Estimate	Pr(>|z|)*	Estimate	Pr(>|z|)*
Intercept	**-0.47**	**0.000**	-0.03	0.803
Time	-0.23	0.083	0.00	0.969
$Time^2$	**-0.48**	**0.019**	**-0.49**	**0.030**
Temperature	0.04	0.246	0.01	0.795
Precipitation	0.00	0.843	0.06	0.096
Region	0.25	0.063	0.03	0.872

996 * Note that p-values should be interpreted with care because a) the GLMM model assumes
997 temporal independence within each individual, which is unlikely to be a valid assumption given
998 that these data were collected every 5 minutes; and b) uncertainty from the first-stage mixture
999 model is not taken into account.

1000

1001 **Figure legends**

1002 Fig. 1. Visual representation of the stick-breaking metaphor. From top to bottom, one starts with
1003 a stick of length 1, breaking it into two sticks of length 0.19 and 0.81 (black and red). This latter
1004 piece (red) is then broken again into two sticks of length 0.27 and 0.54 (red and green). This
1005 latter piece (green) is then broken into two sticks of length 0.15 and 0.39 (green and blue). This
1006 process is reiterated multiple times until the maximum number of groups is reached.
1007
1008 Fig. 2. The use of the truncated stick-breaking prior in the different model enables the successful
1009 uncovering of the true number of groups (top panels) and these simulated data contain
1010 considerable information regarding the hyper-parameter of the TSB prior γ (bottom panels). Left
1011 to right panels show the estimated number of groups (top panels) and the estimated γ (bottom
1012 panels) for the mixture model applied to the movement data, the BR model and SA model,
1013 respectively, based on ten simulated datasets for each value of the true number of groups. We
1014 assumed a maximum of 15 groups for the mixture model applied to the movement data, and a
1015 maximum of 50 groups for the BR and SA models. Notice that the x-coordinate of each point
1016 was shifted slightly (i.e., jittered) on the top panels to enable the visualization of overlapping
1017 circles.
1018
1019 Fig. 3. Results for the mixture model applied to movement data. Panels a-d depict the estimated
1020 distributions for speed and turning angle for the two behavioral states identified by the mixture
1021 model. The first state, henceforth “foraging”, is characterized by slower and more tortuous
1022 movements (panels a and b) while the second state, henceforth “transit”, is characterized by
1023 faster and more directed movements (panels c and d). Comparisons are made between females

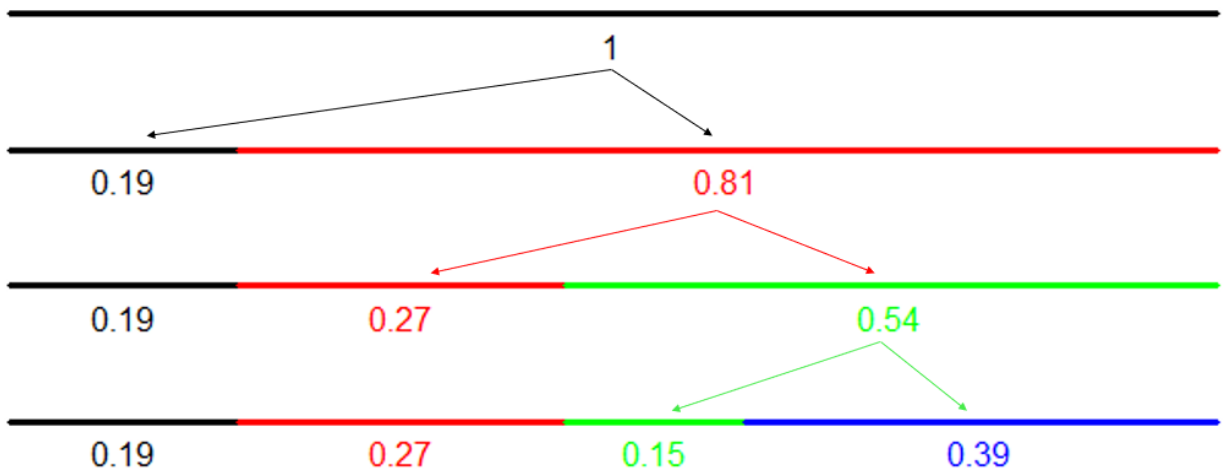
and males regarding the daily number of observations assigned to the foraging state (panel e), and the daily number of observations assigned to the transit state (panel f), and the estimated proportion of transit observations as a function of time of night (panel g).

Fig. 4. Spatial distribution of the groups identified by the Biogeographic Region model. Each panel depicts the predicted presence probability of each group, where cyan to purple indicate probabilities ranging from 0 to 1, respectively. Group numbers are given in the lower left corner of each map.

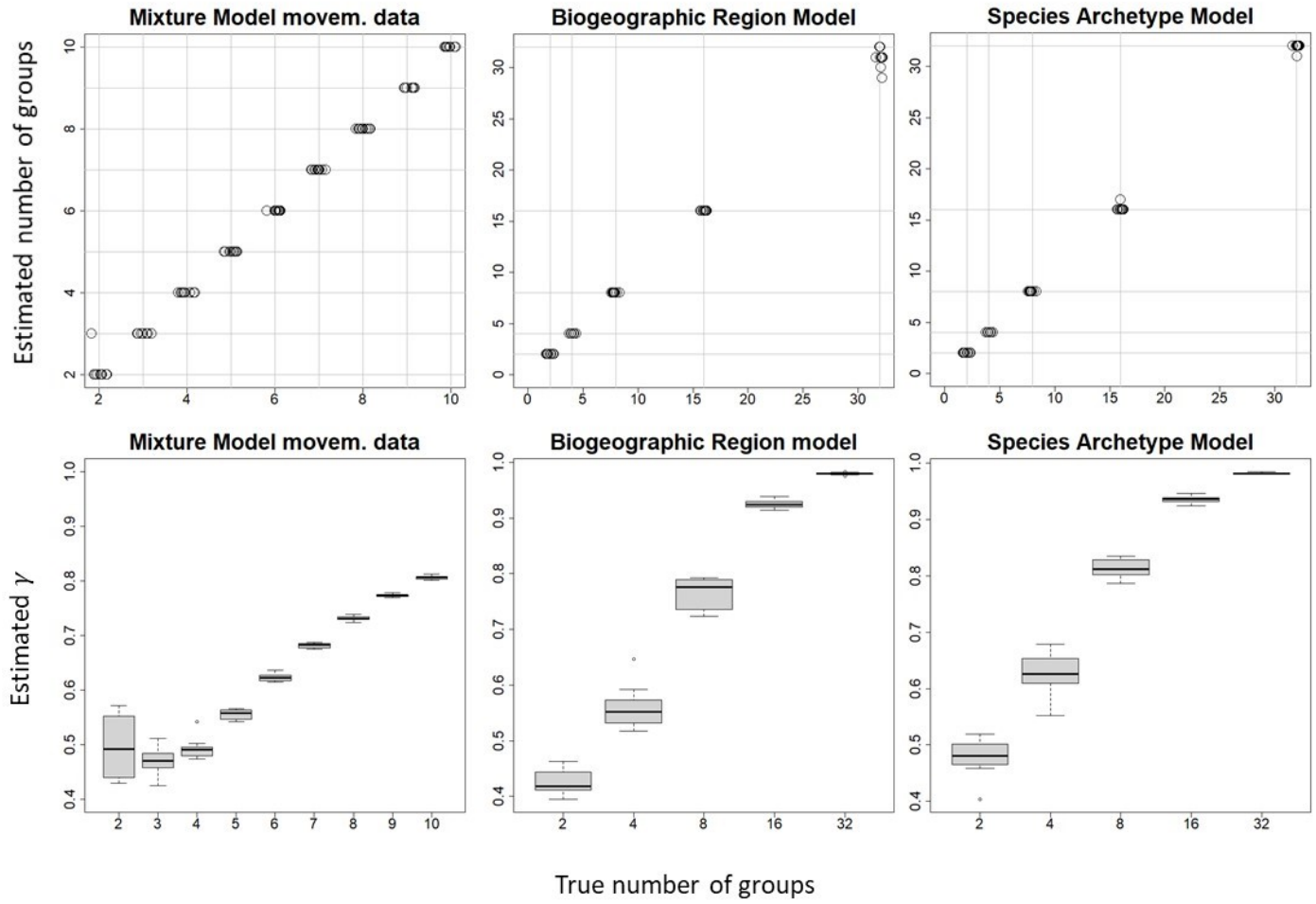
Figure 5. Association between occurrence probability (prevalence) and precipitation and temperature for a subset of the species archetypes identified by the SA model. Each panel displays the results for a particular species archetype (numbers in the top left corner correspond to the species archetype identifier). Individual lines in the line graphs depict the estimated associations for each species within that archetype. For the precipitation line graphs, temperature was set to its mean value. Similarly, for the temperature line graphs, precipitation was set to its mean value. Heat maps show the environmental space of each species archetype by displaying the average prevalence for each temperature and precipitation combination. In all panels, the ranges for precipitation and temperature were limited to the 2.5 and 97.5 percentiles from the original data. Cyan to purple indicate probabilities ranging from 0 to 1.

Figures

Fig. 1



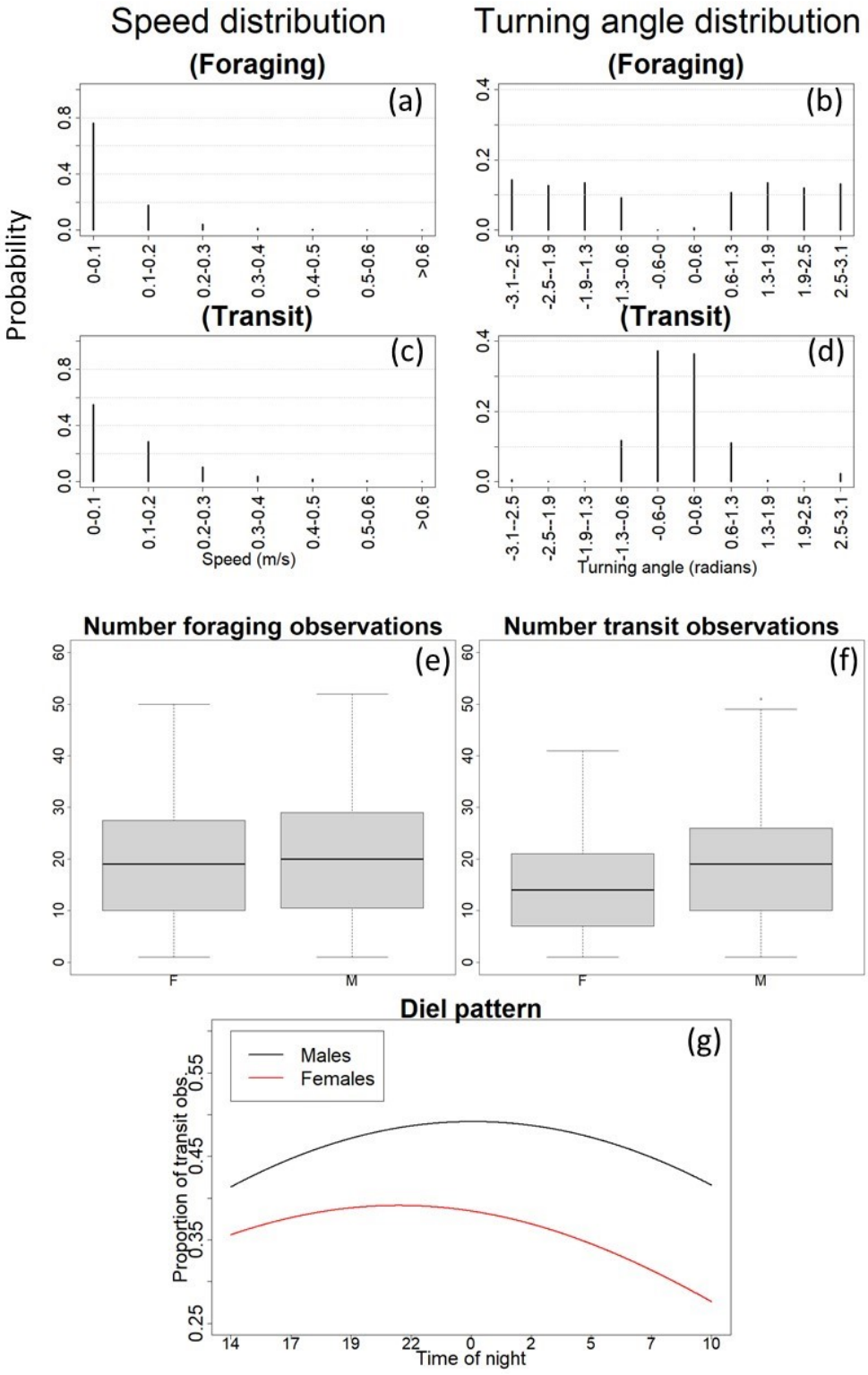
1048 Fig. 2



1049

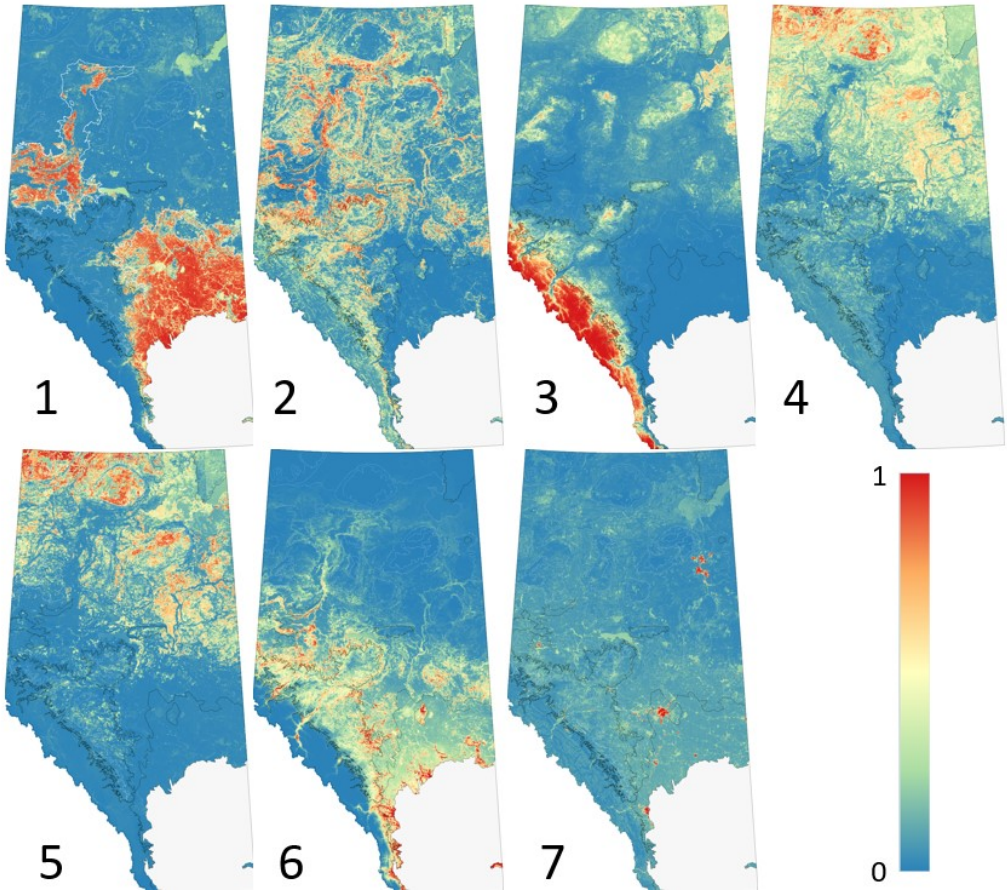
1050

1051 Fig. 3



1052

1053 Fig. 4

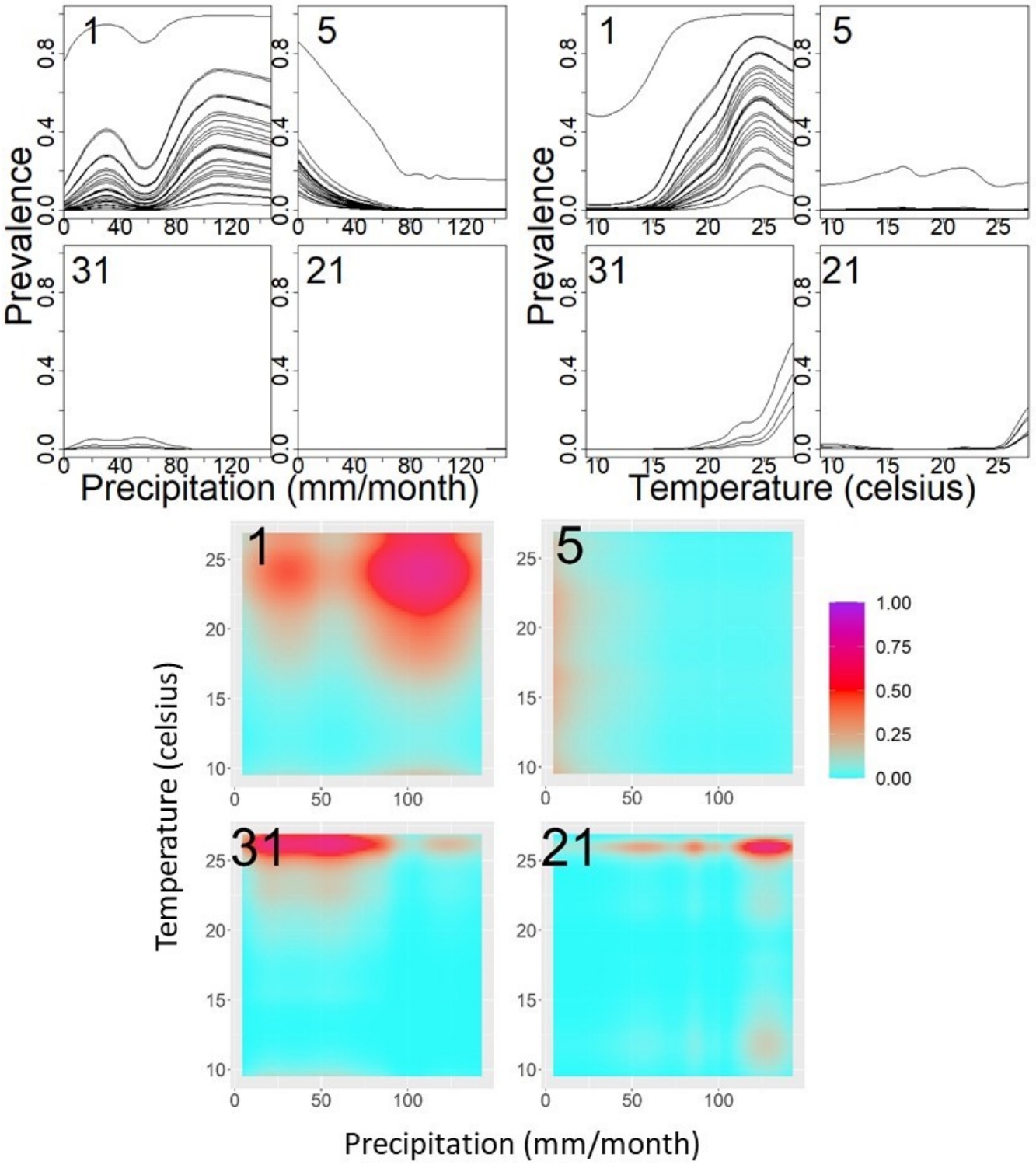


1054

1055

1056

1057 Fig. 5



1058

1059