Species occurrence as a function of both emergent biological traits and environmental context

Peter D. Smits[1,*]

1. University of Chicago, Chicago, Illinois 60637.

∗ Corresponding author; e-mail: psmits@uchicago.edu.

Manuscript elements:

Keywords:

Manuscript type: Article

Prepared using the suggested LaTeX template for *Am. Nat.*

Introduction

How do species pools change over time as species are recruited or go extinct? When are ecotypes enriched or depleted? How does global and regional environmental context affect the distribution of species ecotypes (e.g. guilds) in a regional species pool?

A regional species pool is the set of species which form communities in a specific region; local communities are subsets of the regional pool. The composition of a regional species pool changes over time due to speciation, migration, extinction. Local scale processes like resource competition only affect the regional species pool if all communities are affected.

Valentine and Bambach how they presented guilds in paleobiology. Bush and Bambach presented an ecocube to describe what how marine invertebrates partition space and resources CITATION. Unique combinations represent what possible ecotypes are observable. The distribution of ecocube occupancy is then normally analyzed using ordination methods and the change in disparity over is estimated CITATION.

Fourth-corner modeling is concerned with explaining either species abundance or presence/absence as a product of species traits, environmental factors, and the interaction between these factors. In modern ecological studies, the matrix being modeled is of species occurrence at localities distributed in region. In this study, the matrix being modeled is of species occurrence in temporal bins across the Cenozoic in North America. These dimensions are all axes of the same three dimensional occurrence matrix: species by locality by time.

One of the greatest challenges with analyzing species occurrence data is the inherent incompleteness of any sample CITATION. In the modern, only presences are certain as an absence can be caused by both the species being truly absent or the species never having been sampled CITATION. For paleontological data in the context of this study, the incomplete preservation of fossil communities combined with the incomplete sampling of what fossils there are means that the true times of origination or extinction may not be observed CITATION.

? found several systematic differences in mammal species durations associated with various species

traits. Omnivorous taxa were found to have, on average, a greater duration than other dietary
28 categories. Additionally, arboreal taxa were found to have a shorter duration than other locomotor categories.

30 An unresolved question from **?** is whether the greater extinction risk faced by arboreal is constant over time or if there was a change in extinction risk at the Paleogene/Neogene boundary.
32 Specifically, the question is whether the extinction risk arboreal taxa increased in the Neogene, driving the loss of arboreal taxa and average extinction risk of arboreal taxa down.

34 There are no observed massive taxonomic turnover events in the North American record, unlike the Neogene record Europe CITATION ALROY OTHERS.

36 The effect of climate on diversity and the diversification process has been the focus of considerable research with many analyses favoring diversification being more biologically-mediated than
38 climate-mediated CITATION. Scale of analysis makes a big difference in interpretation of results, both temporal and geographic. For example when the mammal fossil record analyzed at small
40 temporal and geographic scales a correlation between diversity and climate are observable CITATION CLYDE. However, when the record is analyzed at the scale of the continent and the
42 Cenozoic there is no correlation with diversity and climate CITATION ALROY. This results, however, does not go against the idea that there may be short periods of correlation and that this
44 correlation change or reverse direction over time; instead this result means that there is no single direction of correlation between diversity and climate.

46 There are many global climatic events that may have influenced the distribution of mammal ecotypes regionally, if not globally CITATION. PETM. The Mid-Miocene climactic optimum. The
48 general cooling throughout the Cenozoic and the development of ice-caps in the Neogene. The Oligo-Miocene boundary. The transition from the Paleogene to the Neogene in North America is
50 typically described as the "opening-up" of the landscape as partially forested environments were replaced by savannah and grasslands CITATION.

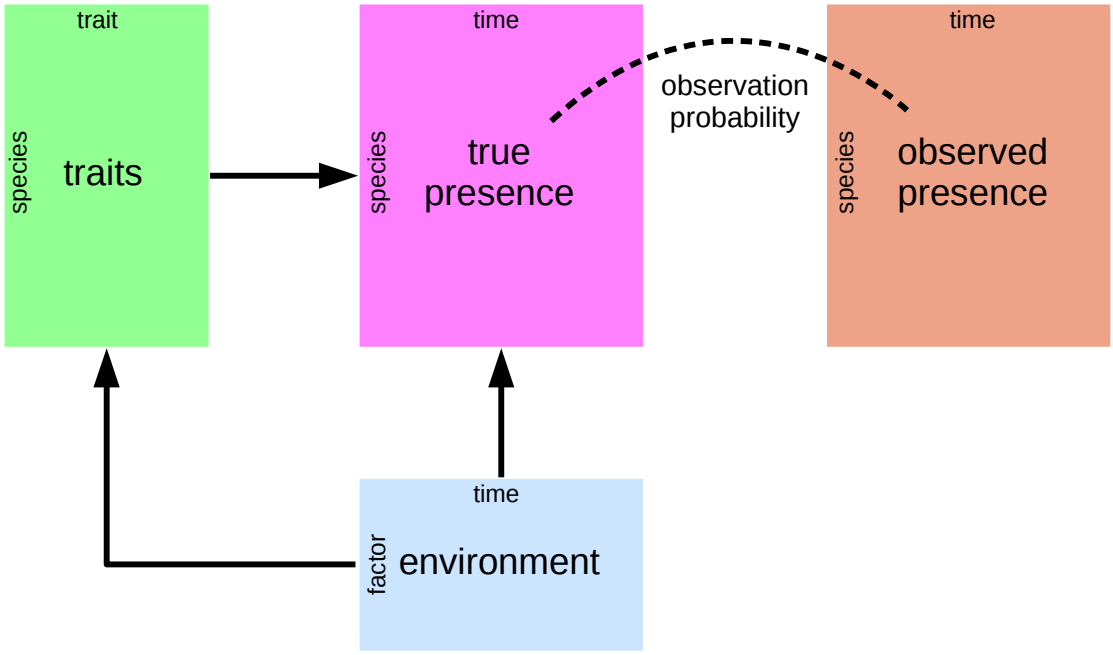

Figure 1: Conceptual diagram of the paleontological fourth corner problem. Figure is based on CITATION

Materials and Methods

Taxon occurrences and species-level information

All fossil occurrence information information was downloaded from the Paleobiology Database. Occurrences (PBDB) were restricted to all Mammalia sampled in North America between the Maastrichtian and Gelasian stages. Taxonomic, stratigraphic, and ecological metadata for each occurrence was included. The raw data is available for download at `http://goo.gl/2slgeU`.

This raw data was then sorted, cleaned, and manipulated programmatically prior to analysis. Species taxonomic assignments given by the PBDB were updated for accuracy and consistency. For example, species classified in the order Artidodactyla were reclassified as Cetartiodactyla. These

re-assignments follow **?** and were based on CITATIONS. Additionally, Taxa who's life habit was classified as either volant (i.e. Chiroptera) or aquatic (e.g. Cetacea) were excluded from this analysis because of both differences in fossilization potential and applicability to the study of terrestrial species pools.

The life habit and dietary categories provided through the PBDB where coarsened to increase per ecotype sample size; this coarsening follows the same procedure as **?**. Additionally, life habit category was further modified to break-up the vague "ground-dwelling" category; re-classifying these species by ankle posture gives more precise information about that species' environmental context. Ground-dwelling taxa were reassigned following **?** by species taxonomic context. Species ecotype is defined as the interaction between life habit and diet categories. Ecotype categories with less than 10 species havig ever been that combination were excluded, yielding a total of 18 of 21 possible ecotypes.

Table 1: Species trait assignments in this study are a coarser version of the information available in the PBDB. Information was coarsened to improve per category sample size and uniformity and followed this table.

This study		PBDB categories
Diet	Carnivore	Carnivore
	Herbivore	Browser, folivore, granivore, grazer, herbivore.
	Insectivore	Insectivore.
	Omnivore	Frugivore, omnivore.
Locomotor	Arboreal	Arboreal.
	Ground dwelling	Fossorial, ground dwelling, semifossorial, saltatorial.
	Scansorial	Scansorial.

Table 2: Posture assignment based on taxonomy

Order	Family	Stance
	Ailuridae	plantigrade
	Allomyidae	plantigrade
	Amphicyonidae	plantigrade
	Amphilemuridae	plantigrade

Continued on next page

Table 2 – continued from previous page

Order	Family	Stance
	Anthracotheriidae	digitigrade
	Antilocapridae	unguligrade
	Apheliscidae	plantigrade
	Aplodontidae	plantigrade
	Apternodontidae	scansorial
	Arctocyonidae	unguligrade
	Barbourofelidae	digitigrade
	Barylambdidae	plantigrade
	Bovidae	unguligrade
	Camelidae	unguligrade
	Canidae	digitigrade
	Cervidae	unguligrade
	Cimolodontidae	scansorial
	Coryphodontidae	plantigrade
	Cricetidae	plantigrade
	Cylindrodontidae	plantigrade
	Cyriacotheriidae	plantigrade
	Dichobunidae	unguligrade
Dinocerata		unguligrade
	Dipodidae	digitigrade
	Elephantidae	digitigrade
	Entelodontidae	unguligrade
	Eomyidae	plantigrade
	Erethizontidae	plantigrade
	Erinaceidae	plantigrade

Continued on next page

Table 2 – continued from previous page

Order	Family	Stance
	Esthonychidae	plantigrade
	Eutypomyidae	plantigrade
	Felidae	digitigrade
	Florentiamyidae	plantigrade
	Gelocidae	unguligrade
	Geolabididae	plantigrade
	Glyptodontidae	plantigrade
	Gomphotheriidae	unguligrade
	Hapalodectidae	plantigrade
	Heteromyidae	digitigrade
	Hyaenidae	digitigrade
	Hyaenodontidae	digitigrade
	Hypertragulidae	unguligrade
	Ischyromyidae	plantigrade
	Jimomyidae	plantigrade
Lagomorpha		digitigrade
	Leptictidae	plantigrade
	Leptochoeridae	unguligrade
	Leptomerycidae	unguligrade
	Mammutidae	unguligrade
	Megalonychidae	plantigrade
	Megatheriidae	plantigrade
	Mephitidae	plantigrade
	Merycoidodontidae	digitigrade
Mesonychia		unguligrade

Continued on next page

Table 2 – continued from previous page

Order	Family	Stance
	Mesonychidae	digitigrade
	Micropternodontidae	plantigrade
	Mixodectidae	plantigrade
	Moschidae	unguligrade
	Muridae	plantigrade
	Mustelidae	plantigrade
	Mylagaulidae	fossorial
	Mylodontidae	plantigrade
	Nimravidae	digitigrade
	Nothrotheriidae	plantigrade
Notoungulata		unguligrade
	Oromerycidae	unguligrade
	Oxyaenidae	digitigrade
	Palaeomerycidae	unguligrade
	Palaeoryctidae	plantigrade
	Pampatheriidae	plantigrade
	Pantolambdidae	plantigrade
	Periptychidae	digitigrade
Perissodactyla		unguligrade
	Phenacodontidae	unguligrade
Primates		plantigrade
	Procyonidae	plantigrade
	Proscalopidae	plantigrade
	Protoceratidae	unguligrade
	Reithroparamyidae	plantigrade

Continued on next page

Table 2 – continued from previous page

Order	Family	Stance
	Sciuravidae	plantigrade
	Sciuridae	plantigrade
	Simimyidae	plantigrade
	Soricidae	plantigrade
	Suidae	digitigrade
	Talpidae	fossorial
	Tayassuidae	unguligrade
	Tenrecidae	plantigrade
	Titanoideidae	plantigrade
	Ursidae	plantigrade
	Viverravidae	plantigrade
	Zapodidae	plantigrade

Species mass information was gathered from multiple different sources where a plurality of the body size estimates are from the PBDB. Body part measurements for many species are also available through the PBDB. Just as with **?**, these measurements and corresponding regression equations were used to get mass estimates for more species. Additional mass estimates and body part measurements were sourced from CITATIONS, the Neogene Old World Database as in **?**. Mass was log-transformed and then mean-centered and rescaled by dividing by two-times its standard deviation; this insures that the magnitude of effects for both continuous and discrete covariates are comparable and follows CITATION.

All fossil occurrences from 64 to 2 million years ago (Mya) were binned into 31 2 million year (My) bins. This temporal length was chosen because it is approximately the resolution of the North American mammal fossil record.

Table 3: Regression equations used in this study for estimating body size. Equations are presented with reference to taxonomic grouping, part name, and reference.

Group	Equation	log(Measurement)	Source
General	$\log(m) = 1.827x + 1.81$	lower m1 area	?
General	$\log(m) = 2.9677x - 5.6712$	mandible length	?
General	$\log(m) = 3.68x - 3.83$	skull length	?
Carnivores	$\log(m) = 2.97x + 1.681$	lower m1 length	?
Insectivores	$\log(m) = 1.628x + 1.726$	lower m1 area	?
Insectivores	$\log(m) = 1.714x + 0.886$	upper M1 area	?
Lagomorph	$\log(m) = 2.671x - 2.671$	lower toothrow area	?
Lagomorph	$\log(m) = 4.468x - 3.002$	lower m1 length	?
Marsupials	$\log(m) = 3.284x + 1.83$	upper M1 length	?
Marsupials	$\log(m) = 1.733x + 1.571$	upper M1 area	?
Rodentia	$\log(m) = 1.767x + 2.172$	lower m1 area	?
Ungulates	$\log(m) = 1.516x + 3.757$	lower m1 area	?
Ungulates	$\log(m) = 3.076x + 2.366$	lower m2 length	?
Ungulates	$\log(m) = 1.518x + 2.792$	lower m2 area	?
Ungulates	$\log(m) = 3.113x - 1.374$	lower toothrow length	?

Environmental and temporal covariates

The group-level covariates in this study are descriptors of species' environmental context, such as global temperature and floral interval.

Global temperature was calculated from Mg/Ca isotope data across the Cenozoic CRAMER CITATION. WHY DID I DO THIS? TOM EZARD HAS USED THIS AND IT IS PROBABLY BETTER THAN THE ZACHOS CURVE. Two aspects of the Mg/Ca-based temperature curve were included in this analysis: mean and range. Both were calculated as the mean of all respective estimates for each 2 My temporal bins. Both mean and range were then rescaled as above: subtract mean, divide by twice the standard deviation.

Floral interval is a holistic descriptor of which taxonomic groups are present and their relative modernity of those taxa for a given temporal interval CITATION. Graham CITATION defines four intervals from the Cretaceous to the Pliocene; only three of these intervals are included in this analysis.

Modelling species occurrence

Two different models were used in this study: a pure-presence model and a birth-death model. Both models at their core are hidden Markov model with an absorbing state. The difference between these two models is if the probability of a species origination and survival are considered equal or different.

Data augmentation

All presence/absence observations are incomplete. The hidden Markov model at the core of this analysis allows for observed absences to be used meaningfully to estimate the number of unobserved species. Of specific concern in this analysis is the unknown "true" size of the dataset; how many species could have actually been observed? While many species have been observed, the natural incompleteness of all observations, especially in the case of paleontological data, there are obviously many species which were never sampled.

Let N by the total number of observed species, M be the upper limit of possible species that could have existed given a model of species presence, and N^* is the all-zero histories where $N^* = M - N$. This approach assumes that $\hat{N} \sim \text{Binomial}(M, \psi)$ where $\hat{N}$ is the estimated "true" number of species and ψ is the probability that any augmented species should actually be "present." Because M is user defined, this approach effectively gives ψ a uniform prior over N to M CITATION. For a more detailed explanation of data augmentation and its application here, please see CITATION, CITATION, and CITATION. For this study, $M = N + \lfloor N/4 \rfloor$.

Data imputation is the process of estimating missing data for partially observed covariates CITATION, this is simple in a Bayesian context because data are also parameters CITATION. Augmented species also have no known mass so a mass estimate must be imputed for each possible species CITATION ROYLE. This procedure assumes that mass values for augmented species are from the same distribution as observed species. The distribution of observed mass values is estimated as part of the model, and new mass values are then generated from this distribution.

Table 4: Observation parameters

Parameter	dimensions	explanation
y	$N \times T$	observed species presence/absence
z	$N \times T$	"true" species presence/absence
p	T	probability of observing a species that is present at time t
m	N	species log mass, rescaled
α_0	1	average log-odds of p
α_1	1	change in average log-odds of p per change mass
r	T	difference from α_0 associated with time t
σ	1	standard deviation of r

This approach is an example of imputing data missing completely at random CITATION. Because log mass values are rescaled as a part of this study, the body mass distribution is already known ($\mathcal{N}(0, 0.5)$); augmented species body mass just simply drawn from this distribution.

In addition to body mass information, the augmented species need an ecotype classification. Because these species are completely unknown, they were all classified as "augmented," an additional grouping indicating their unknown biology. This classification has no biological interpretation.

Observation process

The type of hidden Markov model used in this study has three characteristic probabilities: probability p of observing a species given that it is present, probability ϕ of a species surviving from one time to another, and probability π of a species first appearing CITATION. In this formulation, the probability of a species going extinct is $1 - \pi$. For the pure-presence model $\phi = \pi$, while for the birth-death model $\phi \neq \pi$.

The probability of observing a species that is present p is modeled as a logistic regression was a time-varying intercept and species mass as a covariate. The effect of species mass on p was assumed linear and constant over time and given a prior reflecting a possible positive relationship; these assumptions are reflected in the structure of the model Equation 1. The parameters associated with this part of the model are described in Table 4.

Table 5: Parameters for the model of presence in the pure-presence model

Parameter	dimensions	explanation
z	$N \times T$	"true" species presence/absence
θ	$N \times T - 1$	probability of $z = 1$
a	$T - 1 \times D$	ecotype-varying intercept; mean value of log-odds of θ
m	N	species log mass, rescaled
b_1	1	effect of species mass on log-odds of θ
b_2	1	effect of species mass, squared, on log-odds of θ
U	$T \times D$	matrix of group-level covariates
γ	$U \times D$	matrix of group-level regression coefficients
Σ	$D \times D$	covariance matrix of a
Ω	$D \times D$	correlation matrix of a
τ	D	vector of standard deviations for each ecotype a_d

$$
\begin{aligned}
y_{i,t} &\sim \text{Bernoulli}(p_{i,t} z_{i,t}) \\
p_{i,t} &= \text{logit}^{-1}(\alpha_0 + \alpha_1 m_i + r_t) \\
r_t &\sim \mathcal{N}(0, \sigma)
\end{aligned}
\tag{1}
$$

Pure-presence process

For the pure-presence model there is only a single probability dealing with the presence of a species θ. This probability was modeled as multi-level logistic regression with both species-level and group-level covariates CITATION. The parameters associated with pure-presence model are presented in Table 5 and the full sampling statement in Equation 2.

The species-level of the model (Eq. 2) is a logistic regression with varying-intercept that varies by ecotype. Additionally, species mass was included as a covariate associated with two regression coefficients allowing a quadratic relationship with log-odds of occurrence. This assumption is based on the known distribution of mammal body masses where species with intermediate mass values are more common than either small or large bodied species. These assumptions are also reflected in the choice of priors for these regression coefficients.

The values of each ecotype's intercept are themselves modeled as regressions using the group-level covariates associated with environmental context. Each of these regressions has an associated

variance of possible values of each ecotype's intercept CITATION. In addition, the covariances between ecotype intercepts, given this group-level regression, are modeled CITATION.

Following recommendations from CITATION, CITATION, CITATION all parameters not modeled elsewhere were given weakly informative priors. Weakly informative means that priors do not necessarily encode actual prior information but instead help regularize or weakly constrain posterior estimates. These priors have a concentrated probability density around and near zero; this has the effect of tempering our estimates and help prevent overfitting the model to the data CITATION.

$$
\begin{aligned}
y_{i,t} &\sim \text{Bernoulli}(p_{i,t} z_{i,t}) & \alpha_0 &\sim \mathcal{N}(0,1) \\
p_{i,t} &= \text{logit}^{-1}(\alpha_0 + \alpha_1 m_i + r_t) & \alpha_1 &\sim \mathcal{N}(1,1) \\
r_t &\sim \mathcal{N}(0,\sigma) & \sigma &\sim \mathcal{N}^{+}(1) \\
z_{i,t} &\sim \text{Bernoulli}(\theta_{i,t}) & b_1 &\sim \mathcal{N}(0,1) \\
\theta_{i,t} &= \text{logit}^{-1}(a_{t,j[i]} + b_1 m_i + b_2 m_i^2) & b_2 &\sim \mathcal{N}(-1,1) \\
a &\sim \text{MVN}(u\gamma, \Sigma) & \gamma &\sim \mathcal{N}(0,1) \\
\Sigma &= \text{diag}(\tau)\Omega\text{diag}(\tau) & \tau &\sim \mathcal{N}^{+}(1) \\
& & \Omega &\sim \text{LKJ}(2)
\end{aligned}
\tag{2}
$$

Birth-death process

In the birth-death model, $\phi \neq \pi$ and so each of these probabilities are modeled separately but in a similar manner to how θ is modeled in the pure-presence model (Eq. 2). The parameters associated with the birth-death presence model are presented in Table 6 and the full sampling statement, including observation (Eq. 1), is described in Equation 3.

Similar to the pure-presence model, both ϕ and π are modeled as logistic regressions with varying-intercept and one covariate associated with two parameters. The possible relationships between mass and both ϕ and π are reflected in the parameterization of the model and choice of

Table 6: Parameters for the model of presence in the pure-presence model

Parameter	dimensions	explanation
z	$N \times T$	"true" species presence/absence
ϕ	$N \times T - 1$	probability of $z_{\_,t} = 1 \vert z_{\_,t-1} = 0$
π	$N \times T - 1$	probability of $z_{\_,t} = 1 \vert z_{\_,t-1} = 1$
a^{ϕ}	$T - 1 \times D$	ecotype-varying intercept; mean value of log-odds of θ
a^{π}	$T - 1 \times D$	ecotype-varying intercept; mean value of log-odds of θ
m	N	species log mass, rescaled
b_1^{ϕ}	1	effect of species mass on log-odds of ϕ
b_1^{π}	1	effect of species mass on log-odds of π
b_2^{ϕ}	1	effect of species mass, squared, on log-odds of ϕ
b_2^{π}	1	effect of species mass, squared, on log-odds of π
U	$T \times D$	matrix of group-level covariates
γ^{ϕ}	$U \times D$	matrix of group-level regression coefficients
γ^{π}	$U \times D$	matrix of group-level regression coefficients
Σ^{ϕ}	$D \times D$	covariance matrix of a^{ϕ}
Σ^{π}	$D \times D$	covariance matrix of a^{π}
Ω^{ϕ}	$D \times D$	correlation matrix of a^{ϕ}
Ω^{π}	$D \times D$	correlation matrix of a^{π}
τ^{ϕ}	D	vector of standard deviations for each ecotype a_d^{ϕ}
τ^{π}	D	vector of standard deviations for each ecotype a_d^{π}

priors (Eq. 3).

The intercepts of ϕ and π both vary by species ecotype and those values are themselves the product of group-level regression using environmental factors as covariates (Eq. 3); this is identical to the

pure presence model (Eq. 2).

$$
\begin{aligned}
y_{i,t} &\sim \text{Bernoulli}(p_{i,t} z_{i,t}) & \Sigma^{\phi} &= \text{diag}(\tau^{\phi}) \Omega^{\phi} \text{diag}(\tau^{\phi}) \\
p_{i,t} &= \text{logit}^{-1}(\alpha_0 + \alpha_1 m_i + r_t) & \Sigma^{\pi} &= \text{diag}(\tau^{\pi}) \Omega^{\pi} \text{diag}(\tau^{\pi}) \\
r_t &\sim \mathcal{N}(0, \sigma) & \rho &\sim \text{U}(0, 1) \\
\alpha_0 &\sim \mathcal{N}(0, 1) & b_1^{\phi} &\sim \mathcal{N}(0, 1) \\
\alpha_1 &\sim \mathcal{N}(1, 1) & b_1^{\pi} &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{N}^{+}(1) & b_2^{\phi} &\sim \mathcal{N}(-1, 1) \\
z_{i,1} &\sim \text{Bernoulli}(\rho) & b_2^{\pi} &\sim \mathcal{N}(-1, 1) \\
z_{i,t} &\sim \text{Bernoulli}\left(z_{i,t-1} \pi + \sum_{x=1}^{t} (1 - z_{i,x}) \phi_t \right) & \gamma^{\phi} &\sim \mathcal{N}(0, 1) \\
\phi_{i,t} &= \text{logit}^{-1}(a^{\phi}_{t,j[i]} + b_1^{\phi} m_i + b_2^{\phi} m_i^2) & \gamma^{\pi} &\sim \mathcal{N}(0, 1) \\
\pi_{i,t} &= \text{logit}^{-1}(a^{\pi}_{t,j[i]} + b_1^{\pi} m_i + b_2^{\pi} m_i^2) & \tau^{\phi} &\sim \mathcal{N}^{+}(1) \\
a^{\phi} &\sim \text{MVN}(u \gamma^{\phi}, \Sigma^{\phi}) & \tau^{\pi} &\sim \mathcal{N}^{+}(1) \\
a^{\pi} &\sim \text{MVN}(u \gamma^{\pi}, \Sigma^{\pi}) & \Omega^{\phi} &\sim \text{LKJ}(2) \\
& & \Omega^{\pi} &\sim \text{LKJ}(2)
\end{aligned}
\tag{3}
$$

Posterior inference and model adequacy

Programs that implement joint posterior inference for the above models (Eqs. 2, 3) were implemented in the probabilistic programming language Stan CITATION. The models used here both feature latent discrete parameters in the large matrix z (Tables 4, 5, 6; Eqs. 1, 2, 3). All methods for posterior inference implemented in Stan are derivative based which causes complications for actually implementing the above models because integers do not have derivatives. Instead of implementing a latent discrete parameterization, the posterior probabilities of all possible states of the latent parameters z were estimated (i.e. marginalized).

Species durations at minimum range-through from the FAD to the LAD, but the incompleteness of

all observations means that the actual time of origination or extinction is unknown. The
182 marginalization approach used here means that the probabilities all possible histories for a species
are calculated, from the end members of the species having existed for the entire study interval and
184 the species having only existed between the directly observed FAD and LAD to all possible
intermediaries CITATION.

	Time Bin							
	1	2	3	4	5	6	7	8
Observed	0	0	0	1	0	1	1	0
Certain	?	?	?	1	1	1	1	?
Potential	0	0	0	1	1	1	1	0
Potential	0	0	1	1	1	1	1	0
Potential	0	1	1	1	1	1	1	0
Potential	1	1	1	1	1	1	1	0
Potential	0	0	0	1	1	1	1	1
Potential	0	0	1	1	1	1	1	1
Potential	0	1	1	1	1	1	1	1
Potential	1	1	1	1	1	1	1	1

Figure 2: Conceptual figure of all possible occurrence histories for an observed species. The first row represents the observed presence/absence pattern for a single species at eight time points. The second row corresponds to the known aspects of the "true" occurrence history of that species. The remaining rows correspond to all possible occurrence histories that are consistent with the observed data. The process of parameter marginalization described in the text

186 The combined size of the dataset and large number of parameters in both models (Eqs. 2, 3),
specifically the total number of latent parameters that are the matrix z, means that stochastic
188 approximate posterior inference is computationally very slow even using HMC. Instead, an
approximate Bayesian approach was used: variational inference. A recently developed automatic
190 variational inference algorithm called "automatic differention variational inference" (ADVI) is

implemented in Stan and was used here CITATION. ADVI assumes that the posterior is Gaussian but still yields a true Bayesian posterior; this assumption is similar to quadratic approximation of the likelihood function used in maximum likelihood inference CITATION. The principal limitation of assuming the joint posterior is Gaussian is that the true topology of the log-posterior isn't represented; this is a particular burden for scale parameters which are bound to be positive (e.g. standard deviation).

After fitting both models (Eqs. 2, 3) using ADVI, model adequacy and quality of fit was assessed using a series of posterior predictive checks CITATION CITATION. Because all Bayesian models are inherently generative, simulations of new data sets is "free" CITATION. By simulating many theoretical data sets using the observed covariate information the congruence between predictions made by the model and the observed empirical data can be assessed. By combining multiple posterior predictive tests of congruence between empirical and simulated values of interest, the holistic adequacy of the model can be analyzed CITATION.

An example posterior predictive check used in this study was comparing the observed average number of observations per species to a distribution of simulated averages; if the empirically observed value sits in the middle of the distribution than the model is adequate in reproducing the observed number of occurrences per species.

Posterior simulations for time series are start with the values at t = 1 and then just simulating forward.

Results

Comparing the fits of the pure-presence and birth-death models

Comparison of the posterior predictive performance of the pure-presence and birth-death models reveals a striking difference in quality of the models' fits to the data (Fig. **??**). The birth-death model is clearly able to reproduce the observed average number of occurrence, in contrast to the

pure-birth model which greatly underestimates the ovserved average number of occurrences. The interpretation of these results is that the results of the birth-death model are more representative of the data than the pure-presence model, though further inspection of the posterior parameter estimates if generally called for CITATION.

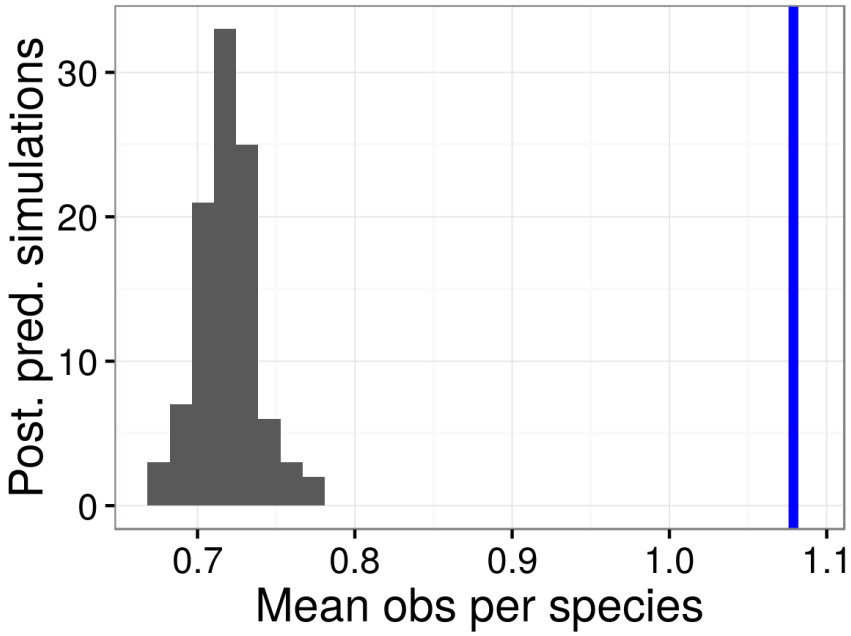

Figure 3: Comparison of the average observed number of occurrences per species (blue line) to the average number of occurrences from 100 posterior predictive datasets using the posterior estimates from the pure-presence model.

Analysis of diversity

Acknowledgements

I would like to thank K. Angielzcyk, M. Foote, P. D. Polly, and R. Ree for helpful discussion and advice. This entire study would would not have been possible without the Herculean effort of the many contributors to the Paleobiology Database. In particular, I would like to thank J. Alroy and M. Uhen for curating most of the mammal occurrences recorded in the PBDB. This is Paleobiology Database publication XXX.

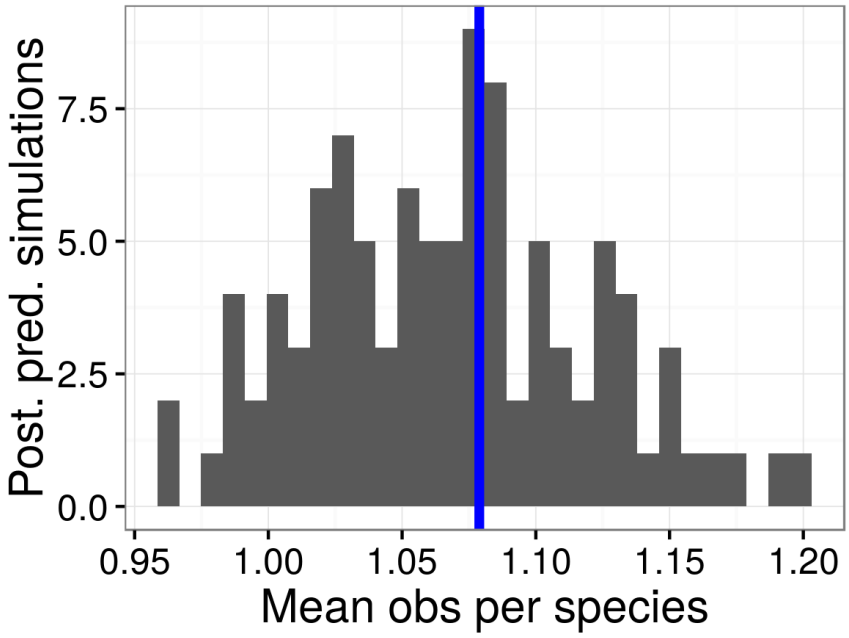

Figure 4: Comparison of the average observed number of occurrences per species (blue line) to the average number of occurrences from 100 posterior predictive datasets using the posterior estimate from the birth-death model.

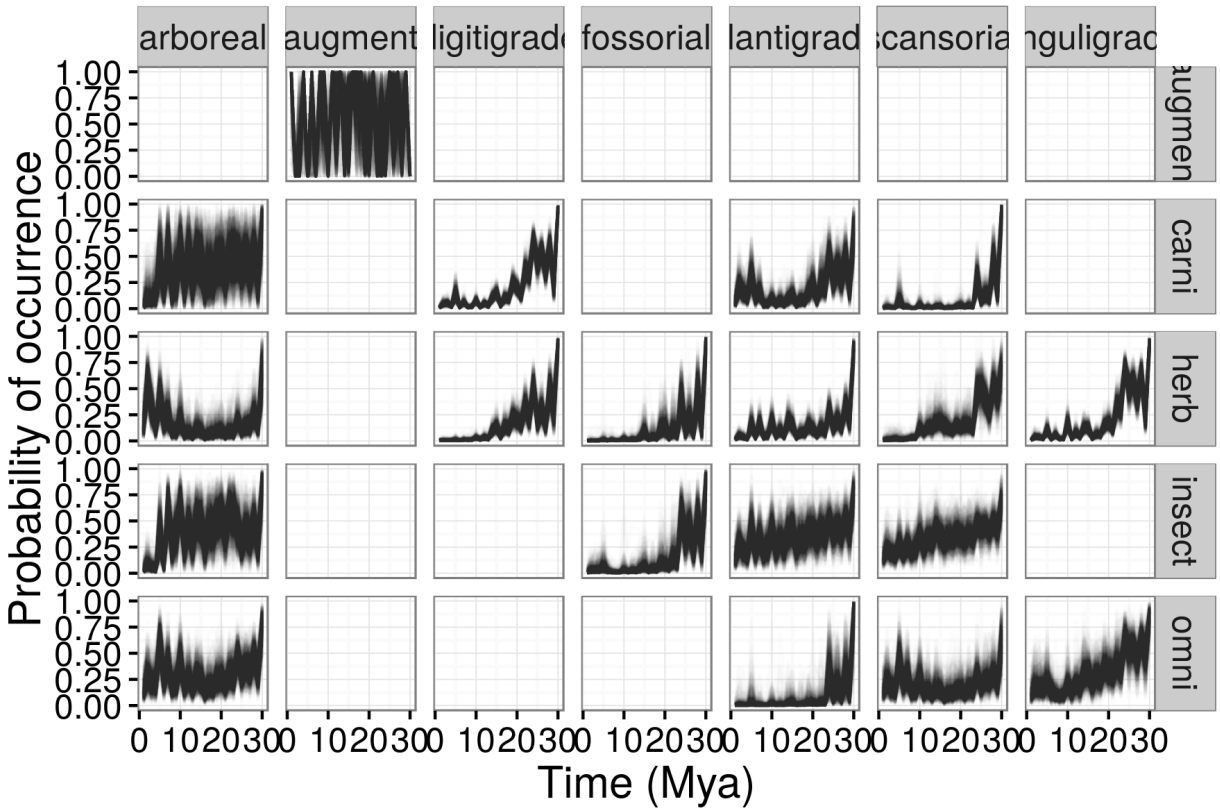

Figure 5: Probability of a mammal ecotype occurring over time as estimated from the pure-presence model. Each panel depicts 100 random samples from the model's posterior. The columns are by locomotor category and rows by dietary category; their intersections are the observed and analyzed ecotypes. Panels with no lines are ecotypes not observed in the dataset.

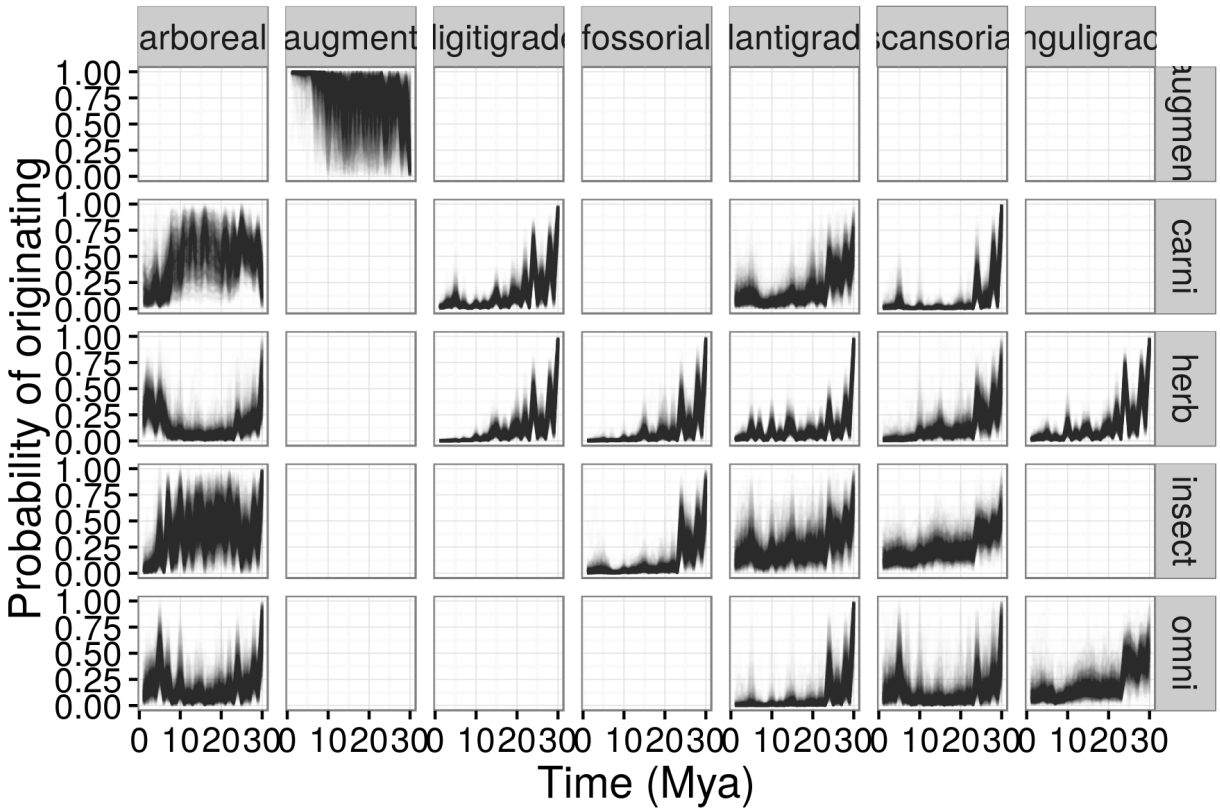

Figure 6: Probability of a mammal ecotype origination probabliities at each time point as estimated from the birth-death model. Each panel depicts 100 random samples from the model's posterior. The columns are by locomotor category and rows by dietary category; their intersections are the observed and analyzed ecotypes. Panels with no lines are ecotypes not observed in the dataset.

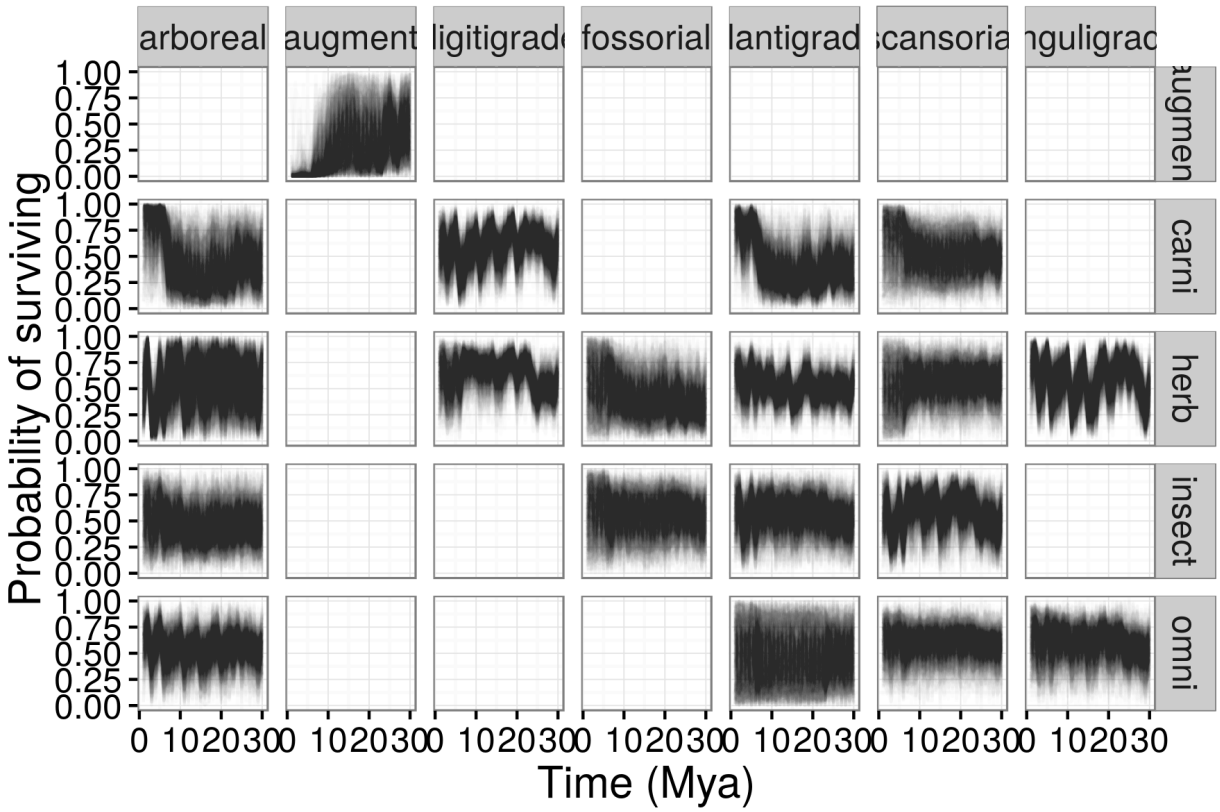

Figure 7: Probability of a mammal ecotype survival probabilities at each time point as estimated from the birth-death model. Each panel depicts 100 random samples from the model's posterior. The columns are by locomotor category and rows by dietary category; their intersections are the observed and analyzed ecotypes. Panels with no lines are ecotypes not observed in the dataset.

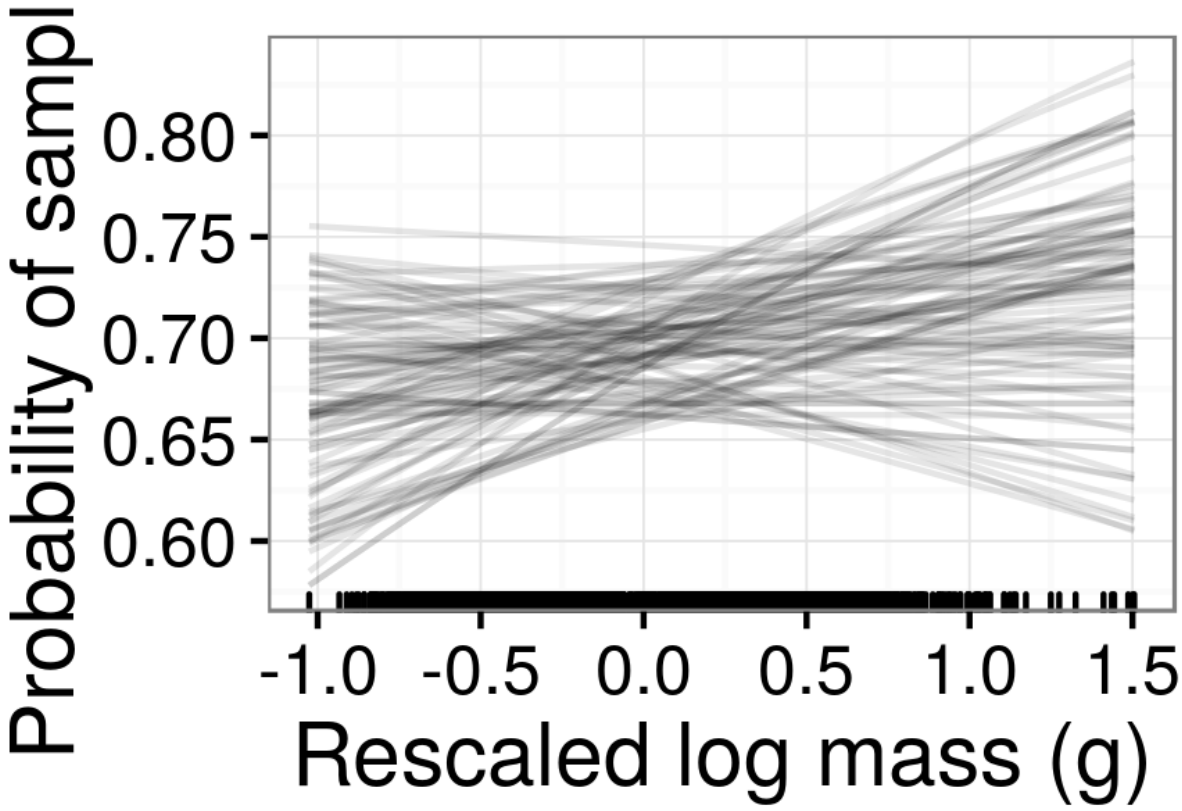

Figure 8: Estimates of the effect of species mass on probability of observing a present species (p). Mass has been log-transformed, centered, and rescaled; this means that a mass of 0 corresponds to the mean of log-mass of all observed species and that mass is in standard deviation units. Estimates are from the pure-presence model.

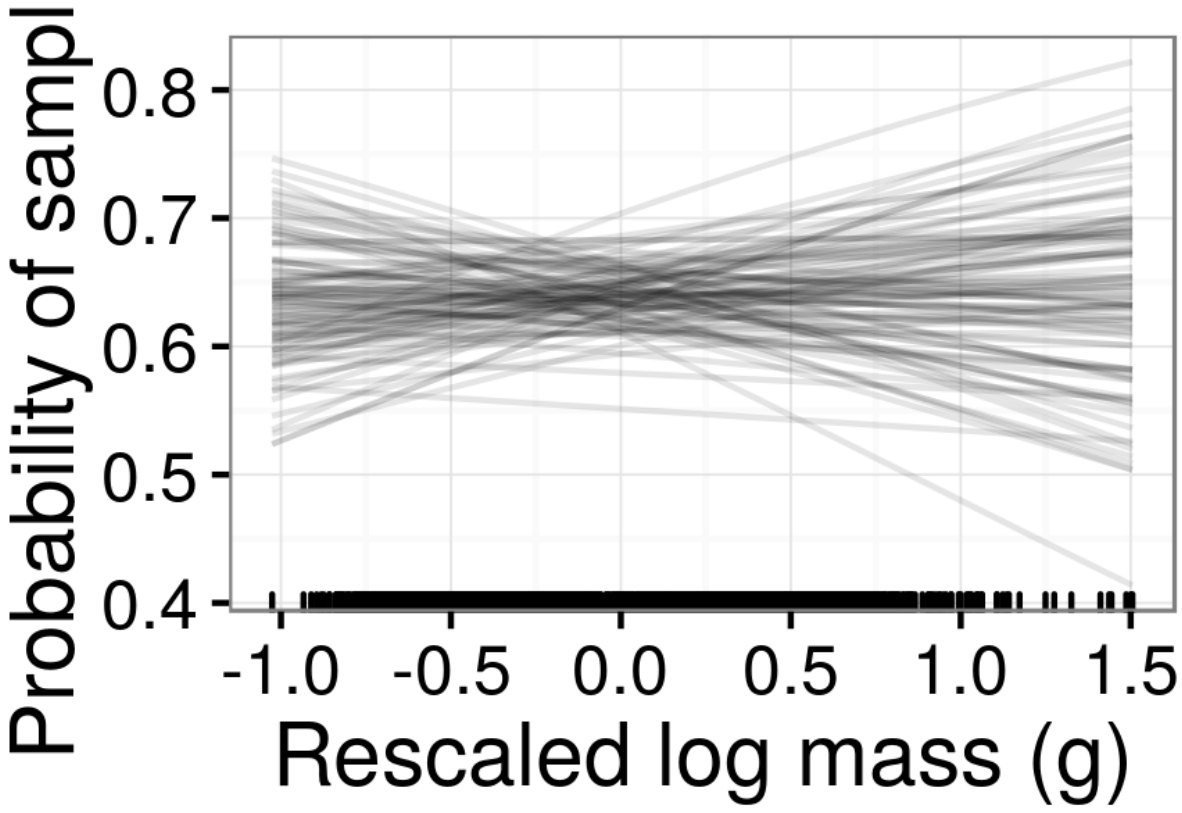

Figure 9: Estimates of the effect of species mass on probability of observing a present species (p). Mass has been log-transformed, centered, and rescaled; this means that a mass of 0 corresponds to the mean of log-mass of all observed species and that mass is in standard deviation units. Estimates are from the birth-death model.

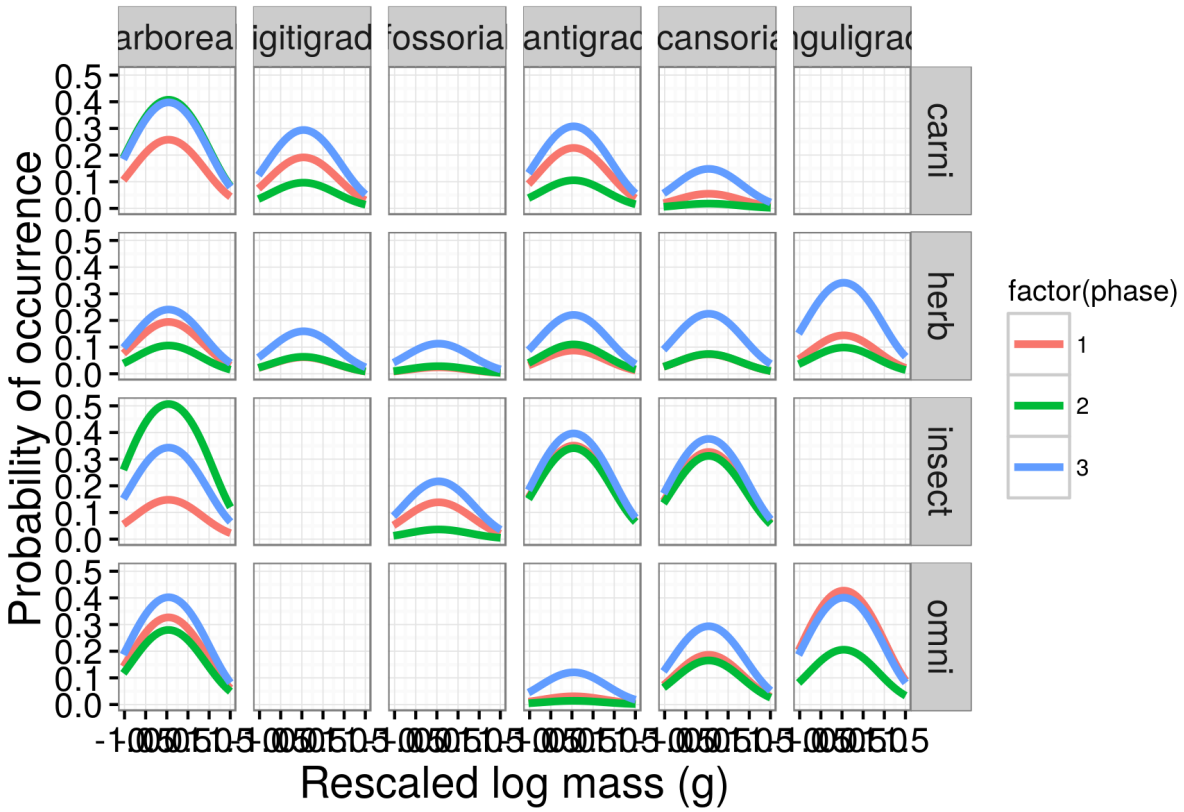

Figure 10: Mean estimate of the effect of species mass on the probability of a species occurrence for each of the three plant phases. The effect of mass is considered constant over time and that the only aspect of the model that changes with plant phase is the intercept of the relationship between mass and occurrence. The three plant phases are indicated by the color of the line. Mass has been log-transformed, centered, and rescaled; this means that a mass of 0 corresponds to the mean of log-mass of all observed species and that mass is in standard deviation units.

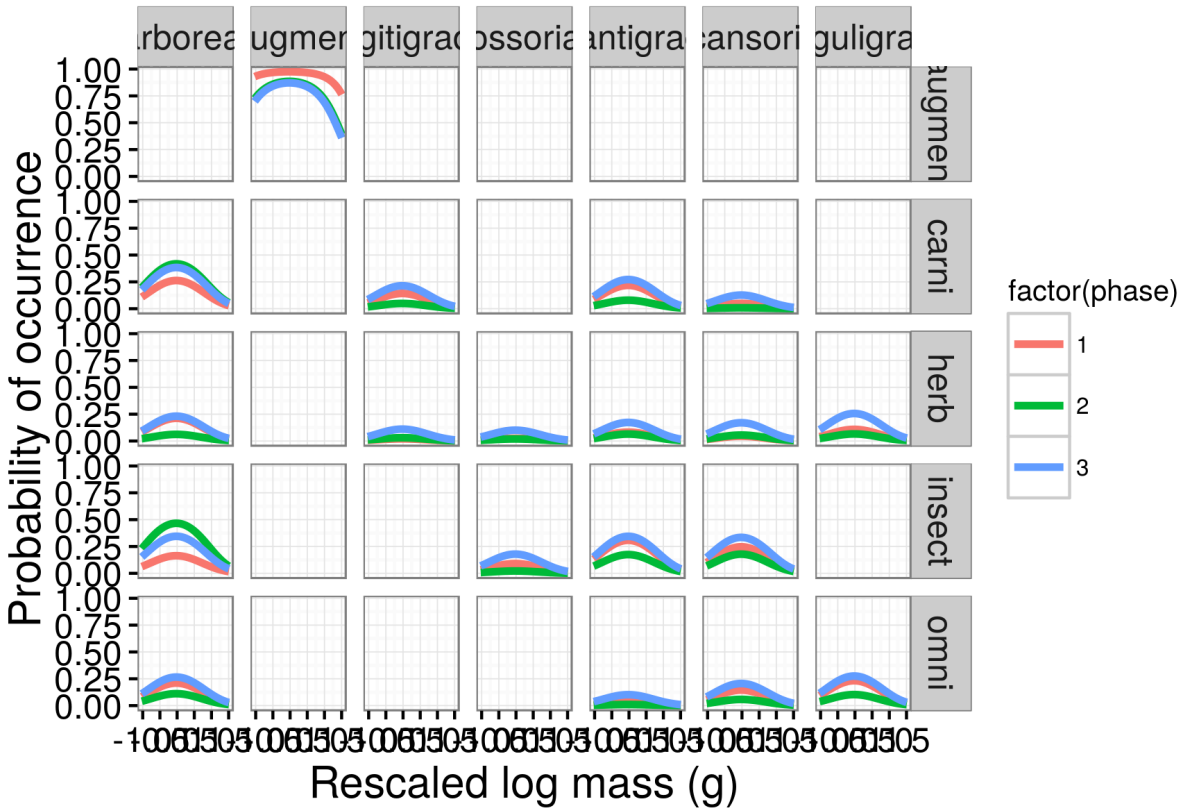

Figure 11: Mean estimate of the effect of species mass on the probability of a species originating for each of the three plant phases. The effect of mass is considered constant over time and that the only aspect of the model that changes with plant phase is the intercept of the relationship between mass and origination. The three plant phases are indicated by the color of the line. Mass has been log-transformed, centered, and rescaled; this means that a mass of 0 corresponds to the mean of log-mass of all observed species and that mass is in standard deviation units.

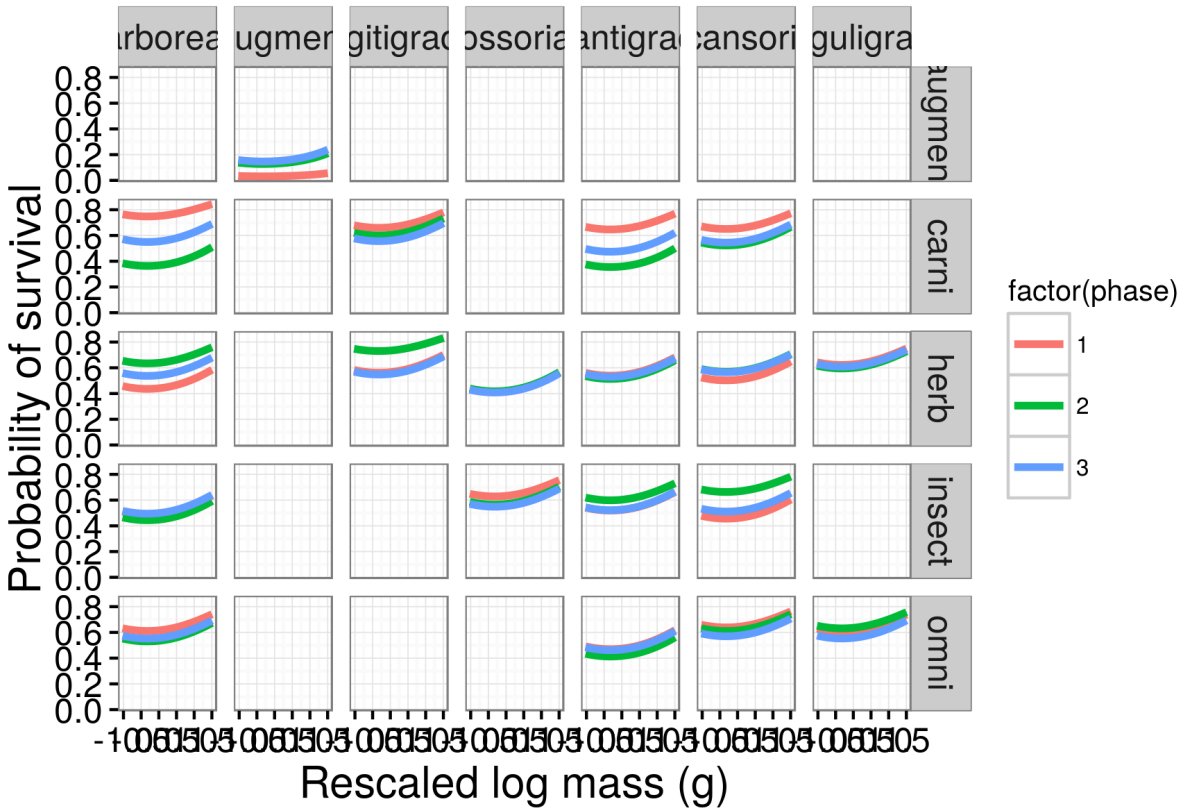

Figure 12: Mean estimate of the effect of species mass on the probability of a species survival for each of the three plant phases. The effect of mass is considered constant over time and that the only aspect of the model that changes with plant phase is the intercept of the relationship between mass and survival. The three plant phases are indicated by the color of the line. Mass has been log-transformed, centered, and rescaled; this means that a mass of 0 corresponds to the mean of log-mass of all observed species and that mass is in standard deviation units.

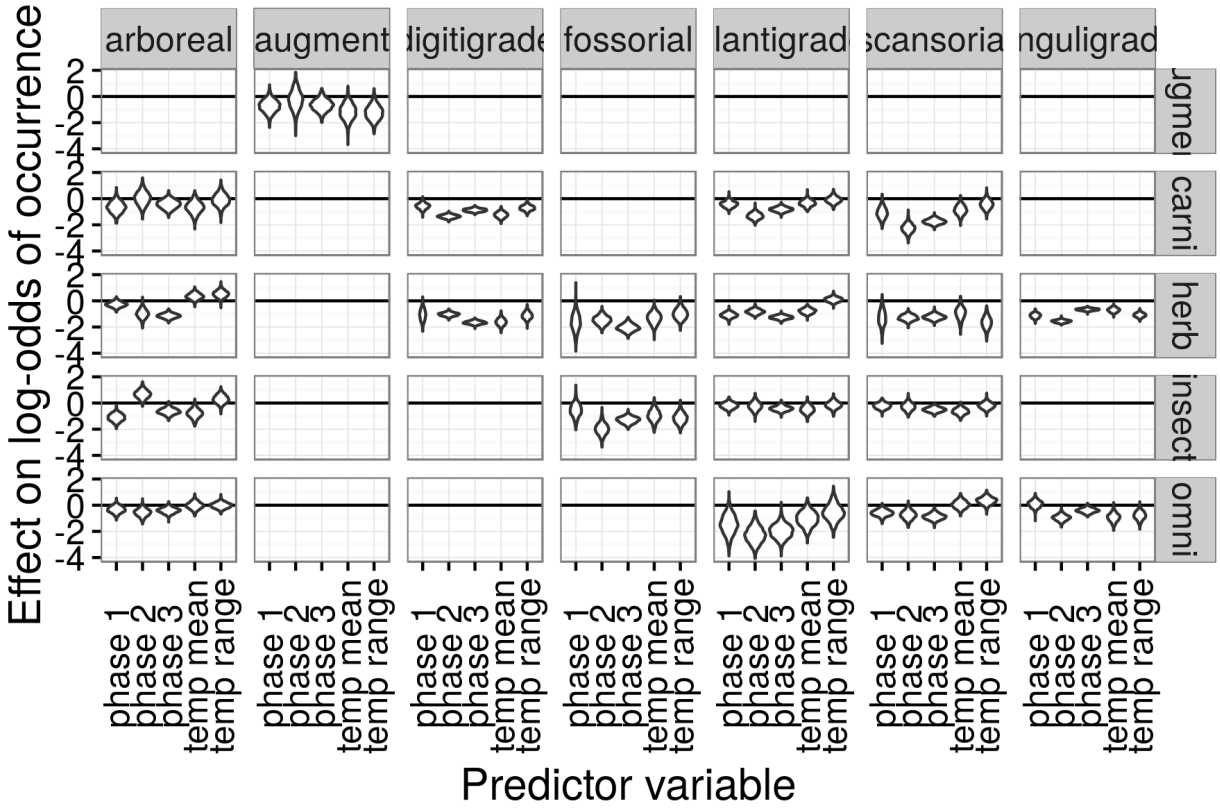

Figure 13: Estimated effects of the group-level covariates describing environmental context on log-odds of species occurrence. These estimates are from the pure-presence model.

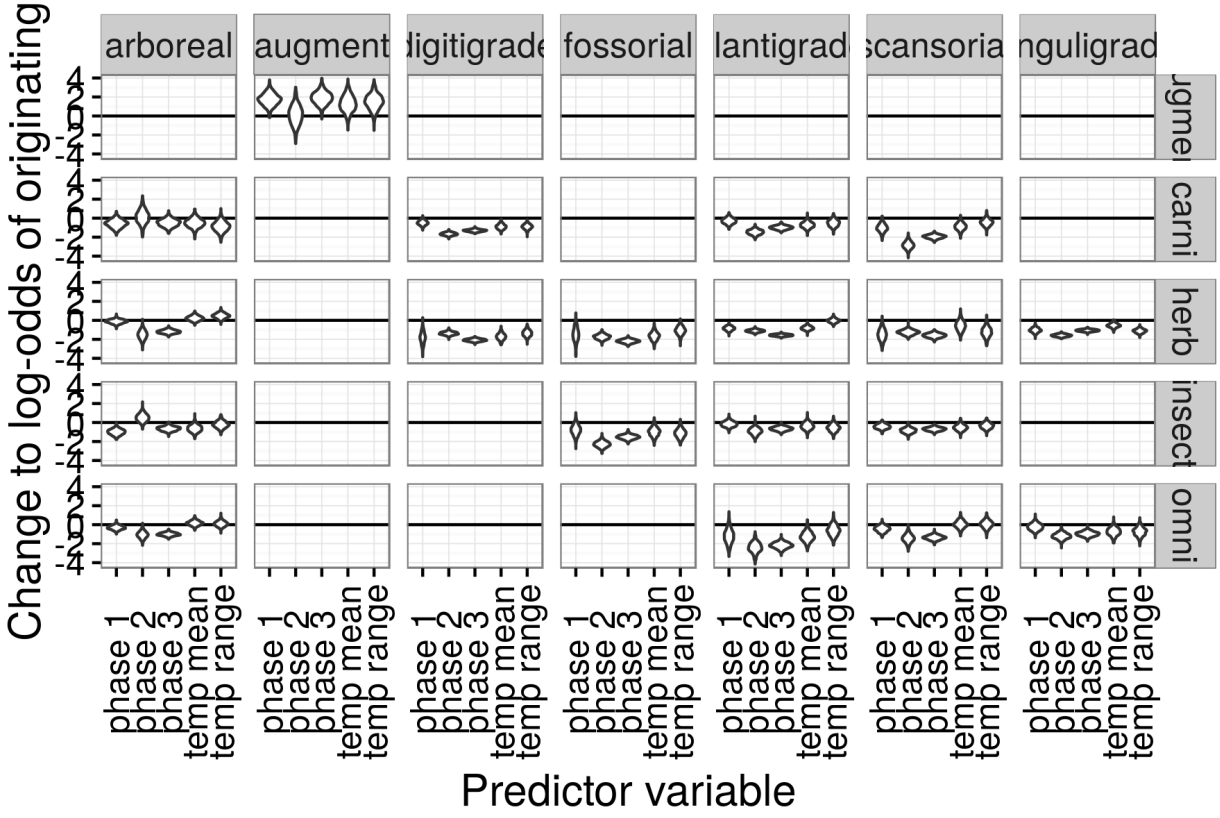

Figure 14: Estimated effects of the group-level covariates describing environmental context on log-odds of species origination. These estimates are from the birth-death model.

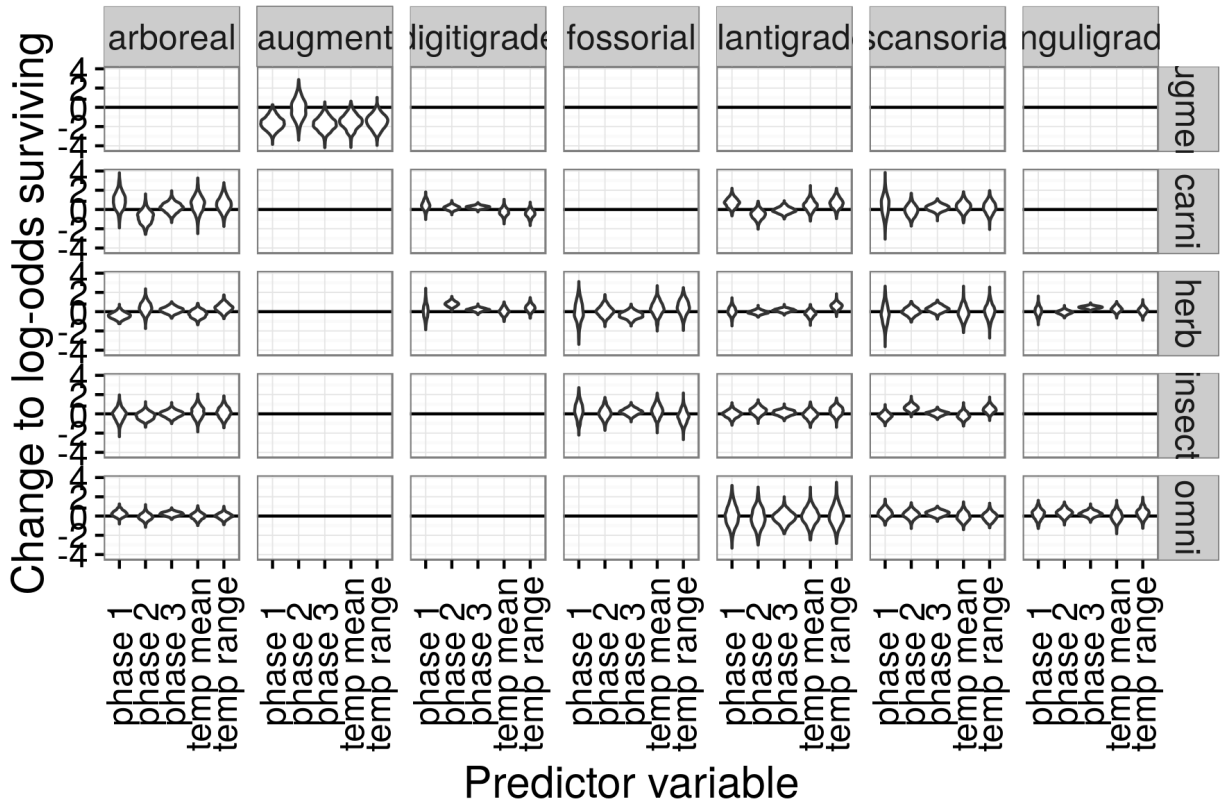

Figure 15: Estimated effects of the group-level covariates describing environmental context on log-odds of species survlval. These estimates are from the birth-death model.