

Influence of ecological traits on mammal extinction risk

Peter D Smits

Committee on Evolutionary Biology, University of Chicago

Abstract

1 Introduction

2 Methods

2.1 Fossil occurrence information

Fossil occurrence information was downloaded from the Paleobiology Database (PBDB; <http://paleodb.org/>). Occurrence, taxonomic, stratigraphic, and biological information was downloaded for all North American mammals. This data set was filtered so that only occurrences identified to the species level, excluding all “sp.”-s. All aquatic and volant taxa were also excluded. Additionally, all occurrences without latitude and longitude information were excluded.

Using the dietary and locomotor category assignments from the PBDB were then reassigned into coarser categories. This was done to improve interpretability, increase sample size per category, and make these results comparable to previous studies CITATION.

DIET assignment

LOCOMOTION assignment

2.2 Bioprovince occupancy

For each 2 My time bin, a bipartite biogeographic network was created between species occurrences and spatial units. Spatial units were defined as 2x2 latitude–longitude grid cells.

In these bipartite networks, taxa can only be linked to localities and *vice versa*. Taxa are not linked to each other, nor are localities linked.

Emergent bioprovinces within the biogeographic occurrence network were identified using the map equation CITATION. A bioprovince is a set of species–locality connections that are more interconnected within the group than without. This was done for each bin’s biogeographic network using the `igraph` package for R [2, 3].

The number of bioprovinces occupied per time bin was then determined for each species.

2.3 Semi-formal supertree

Because there exists no phylogenetic hypothesis of all Cenozoic fossils mammals from North America, it was necessary to construct a semi-formal supertree. This was done by combining taxonomic information for all the observed species and many published phylogenies.

The taxonomic information from the PBDB served as the basis for additional revision. The taxonomy of many species was updated using information from the Encyclopedia of Life (<http://eol.org/>), which collects and coallates this information into a single database. This was done programatically using the `taxize` package for R [5].

Using the `treebase` package for R [1], I downloaded all phylogenies where at least one of the observed fossil taxon was sampled.

RAIA CITATION tree

MRP supertree implemented in the `phytools` package for R [4]

2.4 Survival model

I implemented a fully Bayesian model of taxon durations. For the sampling distribution or likelihood, I assumed a parametric survival model where the observations followed a Weibull distribution (Eq. 1) with shape α and scale σ defined as in a regression model with \mathbf{X} being a matrix of predictor variables (Eq. 2).

$$\begin{aligned} p(y_i|\alpha, \sigma) &= \text{Weibull}(y_i|\alpha, \sigma) \\ &= \frac{\alpha}{\sigma} \left(\frac{y}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{y}{\sigma}\right)^\alpha\right) \end{aligned} \tag{1}$$

$$\sigma = \exp\left(\frac{-(h_i + r_{j|i} + \sum \beta^T \mathbf{X}_i)}{\alpha}\right) \tag{2}$$

These predictors are as follows. Log of mean occupancy and log body size (g) were used as continuous predictors. For modeling discrete predictors in a regression model, the vector of

states is transformed into a $n \times (k - 1)$ matrix where each column is a series of 1's and 0's indicating the observed's category, k being the number of categories. Only $k - 1$ columns are necessary as the intercept takes on the remaining value. This was done for dietary and locomotor category. In total, this is 5 binary predictors. Finally, a 1 was included in the predictor matrix \mathbf{X} and the corresponding β coefficient represents the intercept.

All β coefficients were given a different, diffuse informative normally distributed prior. These priors were chosen because it is expected that the effect size of each variable on duration will be small, as is generally the case with binary predictors. In all cases, posterior inference was not effected by changes to this choice of prior.

The impact of origination cohort (j) was included as a random effect r in the parameterization of σ . The most recent temporal bin, 0 - 2 Mya, was excluded, leaving 32 different cohorts. Each cohort was considered an exchangeable sample of a shared general "cohort effect." The individual cohort effects were estimated in a hierarchical framework where the between cohort variance constrained the individual cohort effect estimate. This is done by giving r a normally distributed prior centered at 0 with scale τ which is then estimated from the data. τ is given a diffuse half-Cauchy hyperprior following GELMAN CITATION.

The impact of shared evolutionary history, or phylogeny, was also included as an individual random effect h , which is given a multivariate normal prior centered at 0. The covariance matrix of prior distribution of h is assumed known from the phylogenetic variance-covariance matrix (\mathbf{VCV}_{phy}) up to a constant v . This constant was given a diffuse half-Cauchy prior.

The shape parameter α was assumed constant and was given a diffuse half-Cauchy prior. This is standard practice in survival analysis.

Below is a summary of the priors used for each estimated parameter

$$\begin{aligned}\tau &\sim \text{half Cauchy}(2.5) \\ r &\sim \text{Normal}(0, \tau) \\ v &\sim \text{half Cauchy}(2.5) \\ h &\sim \text{MultiNormal}(0, v \times \mathbf{VCV}_{phy}) \\ \beta &\sim \text{Normal}(0, 10) \\ \alpha &\sim \text{half Cauchy}(2.5)\end{aligned}$$

An important part of survival analysis is the inclusion of "censored" observations. These are observations where the failure time was not directly observed. In this way both duration and event information are modeled simultaneously, as opposed to just duration or event as in linear and logistic regression, respectively.

The most common censored observation is right censored, where the point of extinction had not yet been observed in the period of study. In this case, this means taxa that are still extant. For each right censored observation, the log probability is incremented by the complementary cumulative density function evaluated at the observed duration.

Left censored observations, on the other hand, correspond to observations that went extinct any time between 0 and some known point. In this study, taxa occurring in only a single time bin were left censored. Because of the minimum resolution of the record, we cannot observe if these taxa went extinct in less than that single bin or not. For each left censored observation, the log probability is incremented by the cumulative density function evaluated at the observed duration.

The joint and marginal posteriors of all parameters were approximated using a Markov-chain Monte Carlo (MCMC) routine implemented in the Stan programming language [6]. Stan implements a Hamiltonian Monte Carlo using a No-U-Turn sampler CITATION.

Posterior approximation was done using four parallel MCMC chains. Chain convergence was evaluated using the scale reduction factor, \hat{R} CITATION. Values of \hat{R} close to 1, or less than or equal to 1.1, indicate approximate convergence. Convergence means that the chains are approximately stationary and the samples are well mixed CITATION.

Each chain was evaluated for 2000 steps, with the first 1000 used as warm-up and the last 1000 as samples from the posterior. In total, this yields 4000 samples from the posterior for all estimated parameters.

2.4.1 Posterior predictive checks

The most basic assesement of model fit is that simulated data generated using the fitted model should be similar to the observed. This is the idea behind posterior predictive checks. Using the predictors from each of the observed durations, and randomly drawn parameter estimates from their marginal posteriors, a simulated data set y^{rep} was generated. This process repeated 1000 times and the distribution of y^{rep} was compared with the observed y . This was done both graphically and numerically.

3 Results

4 Discussion

References

- [1] C. Boettiger and D. Temple Lang. *treebase: An R package for discovery, access and manipulation of online phylogenies*, 2014. URL <http://CRAN.R-project.org/package=treebase>. R package version 0.0-7.1.
- [2] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL <http://igraph.org>.

- [3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL <http://www.R-project.org/>.
- [4] L. J. Revell. phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223, 2012.
- [5] Scott Chamberlain and Eduard Szocs. taxize - taxonomic search and retrieval in r. *F1000Research*, 2013. URL <http://f1000research.com/articles/2-191/v2>.
- [6] Stan Development Team. Stan: A c++ library for probability and sampling, version 2.5.0, 2014. URL <http://mc-stan.org/>.