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 Species 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.

Species dietary and locomotor category assignments were done using the assignments in initial the PBDB which were then reasigned into coarser categories (Table 1). This was done to improve interpretability, increase sample size per category, and make these results comparable to previous studies CITATION.

Fossil occurrences were assigned to 2 My bins ranging through the entire Cenozoic. Taxon duration was measured as the number of bins from the first bin of occurrence to the last bin of occurrence, inclusive.

Species body size estimates were sourced from a large selection of primary literature and compilations, principally the PBDB, PanTHERIA [13], the Neogene Old World Mammal database (Now; `http://www.helsinki.fi/science/now/`), and other large scale data collection efforts [2, 4, 16, 18, 24] CITATION.

This study		PBDB categories
Diet	Carnivore	
	Herbivore	
	Insectivore	
	Omnivore	
Locomotor	Arboreal	
	Ground dwelling	
	Scansorial	

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.

In many cases, species body mass was estimated from anatomical dimensions such as tooth size. These estimates were made using a variety of published regression equations.

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 was 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 [20, 21]. 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 [3, 17].

The number of bioprovinces occupied per time bin was then determined for each species. Additionally, the coefficient of variation (CV) of occupancy was also calculated.

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 the Encyclopedia of Life (`http://eol.org/`), which collects and collates taxonomic information in a single database. This was done programatically using the `taxize` package for R [23].

2

Using the `treebase` package for R [1], I downloaded all phylogenies where at least one of the observed fossil taxon was sampled. This was supplemented with various published phylogenies and taxonomies of fossil mammals [11, 12, 18] MORE CITATIONS.

MRP supertree implemented in the `phytools` package for R [19]

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).

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

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

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. These are sometimes called 'dummy variables'. 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 $\mathbf{X}$, the corresponding β coefficient being 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. Origination cohort is defined as the taxa which all originated during the same temporal bin. 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 [5].

The impact of shared evolutionary history, or phylogeny, was also included as an multivariate normal random effect h. The covariance matrix of h was assumed known, up to a constant υ, from the phylogenetic variance-covariance matrix ($\mathbf{VCV}_{phy}$) [9, 15]. $\mathbf{VCV}_{phy}$ was calculated assuming a Brownian motion model of evolution. The constant υ 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

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

An important part of survival analysis is the inclusion of "censored" observations [10, 14]. These are observations where the failure time has not occurred during the period of interest. 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.

Note that regression coefficients (β-s) are not nececssary directly comparable. Standardizing input variables to have equal standard deviations allows for direct comparison of variables in terms of change in standard deviation [22]. Because there is a mix of continuous and categorical predictors in the above model, standardization is done slighly differently. The expected standard deviation for a binary variable is 0.5, meaning that to make comparisons between binary and continuous variables the continuous inputs must be divided by twice their standard deviation [6]. The above model was fit with both unstandardardized and standardized inputs.

The parameter posteriors were approximated using a Markov-chain Monte Carlo (MCMC) routine implemented in the Stan programming language [25]. Stan implements a Hamiltonian Monte Carlo using a No-U-Turn sampler [8]. Posterior approximation was done using four parallel MCMC chains. Chain convergence was evaluated using the scale reduction factor, $\hat{R}$. 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 [7].

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 assessment 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 [7]. This was done both graphically and numerically.

The model described above was the final model at the end of a continuous model development framework where the sampling and prior distributions were modified to best reflect knowledge of the data, the inclusion of important and necessary covariates, and actually fit the data.

3   Results

θ	mean	sd	2.5%	25%	50%	75%	97.5%	$\hat{R}$
σ	2.19	0.05	2.11	2.16	2.19	2.23	2.28	1.00
α	1.19	0.03	1.14	1.17	1.19	1.21	1.25	1.00

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] B. W. Brook and D. M. J. S. Bowman. The uncertain blitzkrieg of Pleistocene megafauna. *Journal of Biogeography*, 31(4):517–523, Apr. 2004. ISSN 03050270. doi: 10.1046/j.1365-2699.2003.01028.x. URL `http://doi.wiley.com/10.1046/j.1365-2699.2003.01028.x`.

[3] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

[4] M. Freudenthal and E. Martín-suárez. Estimating body mass of fossil rodents. *Scripta Geologica*, 145:1–130, 2013.

[5] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[6] A. Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, pages 2865–2873, 2008. doi: 10.1002/sim.

[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall, Boca Raton, FL, 3 edition, 2013.

[8] M. D. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *arXiv*, 1111(4246), 2011. URL `http://arxiv.org/abs/1111.4246`.

[9] E. A. Housworth, P. Martins, and M. Lynch. The Phylogenetic Mixed Model. *The American Naturalist*, 163(1):84–96, 2004.

[10] J. G. Ibrahim, M.-H. Chen, and D. Sinha. *Bayesian Survival Analysis*. Springer, New York, 2001.

[11] C. M. Janis, K. M. Scott, and L. L. Jacobs. *Evolution of Tertiary mammals of North America. Vol. 1. Terrestrial carnivores, ungulates, and ungulatelike mammals*. Cambridge University Press, Cambridge, 1998.

[12] C. M. Janis, G. F. Gunnell, and M. D. Uhen. *Evolution of Tertiary mammals of North America. Vol. 2. Small mammals, xenarthrans, and marine mammals*. Cambridge University Press, Cambridge, 2008.

[13] K. E. Jones, J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. D. L. Orme, K. Safi, W. Sechrest, E. Boakes, C. Carbone, C. Connolly, M. J. Cutts, J. K. Foster, R. Grenyer, M. Habib, C. A. Plaster, S. A. Price, E. A. Rigby, J. Rist, A. Teacher, O. R. P. Bininda-Emonds, J. L. Gittleman, G. M. Mace, and A. Purvis. PanTHERIA : a species-level database of life history , ecology , and geography of extant and recently extinct mammals. *Ecology*, 90(9):2648, 2009.

[14] D. G. Kleinbaum and M. Klein. *Survival analysis: a self-learning text*. Springer, New York, NY, 2 edition, 2005.

[15] M. Lynch. Methods for the analysis of comparative data in evolutionary biology. *Evolution*, 45(5):1065–1080, 1991.

[16] R. T. McKenna. *Potential for Speciation in Mammals Following Vast , Late Miocene Volcanic Interruptions in the Pacific Northwest*. Masters, Portland State University, 2011.

[17] 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/`.

[18] P. Raia, F. Carotenuto, F. Passaro, D. Fulgione, and M. Fortelius. Ecological specialization in fossil mammals explains Cope's rule. *The American naturalist*, 179(3):328–37, Mar. 2012. ISSN 1537-5323. doi: 10.1086/664081. URL `http://www.ncbi.nlm.nih.gov/pubmed/22322221`.

[19] L. J. Revell. phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223, 2012.

[20] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–23, Jan. 2008. ISSN 1091-6490. doi: 10.1073/pnas.0706851105. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2234100&tool=pmcentrez&rendertype=abstract`.

[21] M. Rosvall, D. Axelsson, and C. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(14):13–24, 2009. URL `http://www.springerlink.com/index/H8193132U6432363.pdf`.

[22] H. Schielzeth. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1(2):103–113, Feb. 2010. ISSN 2041210X. doi: 10.1111/j.2041-210X.2010.00012.x. URL `http://doi.wiley.com/10.1111/j.2041-210X.2010.00012.x`.

[23] Scott Chamberlain and Eduard Szocs. taxize - taxonomic search and retrieval in r. *F1000Research*, 2013. URL `http://f1000research.com/articles/2-191/v2`.

[24] F. A. Smith, J. H. Brown, J. P. Haskell, S. K. Lyons, J. Alroy, E. L. Charnov, T. Dayan, B. J. Enquist, S. K. M. Ernest, E. A. Hadly, K. E. Jones, D. M. Kaufman, P. A. Marquet,

B. A. Maurer, K. J. Niklas, W. P. Porter, B. Tiffney, and M. R. Willig. Similarity of Mammalian Body Size across the Taxonomic Hierarchy and across Space and Time. *The American Naturalist*, 163(5):672–691, 2004.

[25] Stan Development Team. Stan: A c++ library for probability and sampling, version 2.5.0, 2014. URL `http://mc-stan.org/`.