When do traits matter? A hierarchical Bayesian model of Brachiopod survival

Peter D Smits
Committee on Evolutionary Biology, University of Chicago

April 30, 2015

1 Introduction

How do differences in taxon traits effect differences in extinction risk? Jablonski [10] hypothesizes that as baseline extinction risk increases, the importance of taxon traits, such as ecology or environmental preference, should decrease in importance. The trait where this pattern should be weakest in terms of magnitude is geographic range [10]. This hypothesis has strong implications for any model of extinction risk: the effect of taxon traits should decrease when baseline extinction intensity increases. This can be stated in a discrete background versus mass extinction framework [10] or a continuous rate variation framework [28].

Geographic range is widely considered the most important taxon trait for estimating differences extinction risk at nearly all times [9–11, 19]. This is logical because if mortality is randomly distributed spatially, a taxon with a large geographic range is less likely to be completely wiped out than a taxon with restricted range. This is a strong prediction based on strong prior evidence.

Miller and Foote [16] demonstrated that, during mass extinctions, taxa which are associated with open ocean environments have a greater extinction risk than those taxa that favor epicontinental seas. During periods of background extinction, however, they found no consistent difference between taxa favoring either environment. Because of this study, the following prediction for survival patterns can be made: as extinction risk increases, taxa associated with open ocean environments should generally increase in extinction risk.

Of additional interest is the possibility of a potential nonlinear relationship between environmental preference and taxon duration. A long standing hypothesis is that generalists or unspecialized taxa will have greater survival than specialists [1, 14, 17, 18, 25] SMITS, IN PREP. A simple expectation for continuous traits is that this nonlinearity would manifest as a downward facing function, with taxa with intermediate trait values having a greater expected durations than taxa

with high or low trait values. By utilizing a continuous measure of environmental preference, it is relatively straight forward to allow for this non-linearity.

I adopt a hierarchical Bayesian survival modelling approach for multiple reasons, one being that it represents a conceptual and statistical unification of the paleontological dynamic and cohort survivorship approaches [1, 2, 21–23, 26, 27]. In this case, origination cohorts are the groups and the mean survival model corresponds to the dynamic survivorship model. By using a Bayesian framework I am best able to quantify the uncertainty inherent in the estimates of the effects of taxon traits on survival especially in cases where the covariates of interest (species traits) are themselves known with error.

Hierarchical modelling, sometimes called "mixed-effects modeling," is a statistical approach which explicitly takes into account the structure of the observed data in order to model both the with and between group structure [5, 6]. In this approach, the units of study (e.g. genera) belong to a single grouping (e.g. origination cohort) that are also considered draws from a probability distribution (e.g. all cohorts, observed and unobserved). The degree of similarity between the groups is then estimated simultaneously as the other parameters of interest (e.g. covariate effects) [6]. The subsequent estimates are then partially pooled together, where parameters from groups with large samples or effects are allowed to be far away from the overall group mean while those for groups with small samples or effects are pulled towards the overall group mean.

This partial pooling is one of the greatest advantages of hierarchical modeling. By letting the groups "support" each other, parameter estimates much better reflect our uncertainty in their estimates. Additionally, this partial pooling helps control for multiple comparisons and spurious results as small effects with little empirical support are drawn towards the overall group mean [5, 6]. This means that while estimates for an individual group may be weakly constrained, all groups contribute to the weighted average that is the overall group mean.

In this analysis, genera are structured as belonging to origination cohorts. All covariate effects (regression coefficients), as well as the intercept term, were allowed to vary by group. The covariance/correlation between covariate effects was also modeled. This hierarchical structure allows inference for how covariates effects may change with respect to each other while simultaneously estimating the effects themselves, correctly propagating our uncertainty. Additionally, instead of relying on potentially biased point estimates of environmental affinity, I adopt a measurement error model approach where environmental affinity is measured as the difference in the taxon's environmental occurrence pattern and the background environmental occurrence pattern. This approach too correctly propagates our uncertainty, which leads to more valid posterior inference.

2 Methods

2.1 Fossil occurrence information

The dataset analyzed here is derived from the a combination of the occurrence information from Miller and Foote [16] and the body size data from Payne et al. [20]. The Miller and Foote [16] dataset is based on the Paleobiology Database (http://www.paleodb.org); see Miller and Foote [16] for a full description of the inclusion criterion.

Sampled occurrences were restricted to those with latitude and longitude coordinates, assignment to either epicontinental or open-ocean environment, and being of a genus present in the body size dataset. Genus duration was calculated as the number of geologic stages from first appearance to last appearance, inclusive. Genera who's last appearance was in a stage preceding a mass extinction were right censored, and genera with a duration of only one stage and were left censored (see below for explanation of censoring). The covariates used to model genus duration were geographic range size (r), environmental preference (v), and body size (m).

Geographic range was calculated using an occupancy approach. First, all occurrences were projected onto an equal-area cylindrical map projection. Each occurrence was then assigned to one of the cells of a 70×34 regular raster grid placed on the map. Each grid cell represents approximately 250,000 km^2. Following this, for each stage, the total number of grid cells occupied is calculated. The number of grid cells that each genus present occurs in was then calculated and made relative by dividing by the total number of possible cells. Finally, mean relative genus occupancy was calculated as the mean of per stage relative occupancy.

Body size data was sourced directly from Payne et al. [20]. Because those measurements are presented with out quantified error, a measurement error model similar to the one for environmental affinity could not be implemented.

Prior to analysis some covariates were transformed in order to improve interpretation. Geographic range size, which can only have between 0 and 1, was logit transformed. Body size, which is defined for all positive real values was natural log transformed. These covariates were then standardized by mean centering and dividing by two times their standard deviation following Gelman and Hill [5].

2.1.1 Uncertainty in environmental preference

The calculation and inclusion of environmental affinity in the subsequent survival model is a statistical procedure that takes into account our uncertainty based on where fossils tend to occur. Because we cannot directly observe if a fossil taxon had occurrences restricted to only a single environmental, instead we

can only get an of affinity with some amount of uncertainty. One advantage of using a Bayesian analytical context is that both parameters and data are considered random samples from some underlying distribution, which means it is possible to model the uncertainty in our covariates of interest [6]. In this case, this is the genus affnity to either epicontiental settings or not. This approach is conceptually similar to Simpson and Harnik [24] but instead of obtaining a single point estimate, an entire posterior distribution is estimated.

The first step is to determine the probability θ at which genus i occurs in an epicontinental settings based on its own pattern of occurrences. Define e_i as the number of occurrences of genus i in an epicontiental sea and o_i as the number of occurrences of genus i not in an epicontinental sea (e.g. open ocean). Because the value of inters is the probability of occurring in an epicontinental environment, given the observed fossil record, I assume that probability follows a beta distribution. We can then define our sampling statement as

$$e_i \sim \text{Binomial}(e_i + o_i, \theta_i). \tag{1}$$

I used a flat prior of θ_i defined as $\theta_i \sim \text{Beta}(1, 1)$. Because the beta distribution is the conjugate prior for the binomial distribution, the posterior is easy to compute in closed form. The posterior probability of θ is then

$$\theta_i \sim \text{Beta}(e_i + 1, o_i + 1) \tag{2}$$

It is extremely important, however, to take into account the overall environmental occurrence probability of all other genera present at the same time as genus i. This is incorporated as an additional probability Θ. Define E_i as the total number of other fossil occurrences (e.g. excepting for genus i) in epicontinental seas during stages where i occurs and O_i as the number of other fossil occurrences not on epicontinental seas. We can then define the sampling statement as

$$E_i \sim \text{Binomial}(E_i + O_i, \Theta_i). \tag{3}$$

Again, I used a flat prior of Θ_i defined as $\Theta_i \sim \text{Beta}(1, 1)$. The posterior of Θ is then simply defined as

$$\Theta_i \sim \text{Beta}(E_i + 1, O_i + 1) \tag{4}$$

I then define the environmental affinity of genus i as $v_i = \theta_i - \Theta_i$. v_i is a value that can range between -1 and 1, where negative values indicate that genus i tends to occur in open ocean environments while positive values indicate that genus i tends to occur in epicontiental environments.

While this approach is noticeably more complicated than previous ones [3, 12, 15, 24] there are some important benefits to both using a continuous measure of affinity as well directly modeling our uncertainty. In order to show case some of these benefits, I performed a simulation analysis of how modal/maximum *a posteriori*(MAP) estimates versus full posterior estimates.

In this simulation, I first defined the "background" epicontinental occurrence θ_b as 0.50 with a small amount of noise. This was represented as a beta distribution

$$\theta_b = \text{Beta}(\alpha = 2500, \beta = 2500). \tag{5}$$

This choice of parameters for the distribution reflects the average number of background occurrences for either epicontinental or open ocean environments per genus.

Using this background occurrence ratio, randomly generated the occurrence patterns 1000 simulated taxa. This was done at multiple sample sizes (1, 2, 3, 4, 5, 10, 25, 50, 100) in order to demonstrate the effects of increasing sample size on the confidence of environmental affinity. For each simulated taxon I calculated the full posterior distribution while assuming a flat Beta prior (Beta$(1, 1)$). Using the full posterior I calculated the MAP probability of occurring in epicontinental environments. The environmental affinity was calculated for each of the simulated taxa using both the full posterior and the MAP estimate. In this toy example, environmental affinity can range between -0.5 and 0.5.

As should be expected, as sample size increases the distribution of MAP estimates converge on the true value (Fig. 1). For taxa with less than 10 occurrences, the MAP estimate is biased towards over estimates of affinity. Note that the mode of the beta distirbution is not defined for situations where there were 0 draws of one of the environmental conditions. Instead, the vertical line is based entirely on the observed occurrences.

In contrast, we can compare the true occurrence probability distribution versus the posterior estimate for a given sample (Fig. 2). When sample sizes are low, posterior estimates are flat and represent a compromise between the likelihood (equivalent to MAP) and the flat prior. Because of this, estimates from small sizes are less likely to be overly biased. This is further emphasized by inspection of the estimates of environmental affinity for the simulated taxa (Fig. 3). Posterior estimates from simulated taxa with small sample size have a much broader distribution that both allows for the extreme observation but still captures the "true" value (0).

By defining environmental preference as the difference in full posterior estimates of occurrence probability, it is possible to include taxa with low sample sizes that are normal discarded [3, 12, 15, 24]. Additionally, 55+% of observed Paleozoic brachiopod genera have less than 10 occurrences which is the sample size range where MAP (or ML) estimates would be most biased. This is preferable to the difference in MAP estimates, especially for taxa with small sample sizes (blue line; Fig. 3).

2.2 Survival model

Genus durations were modeled in a Bayesian parameteric survival analysis framework. Durations were assumed to follow either an exponential or Weibull

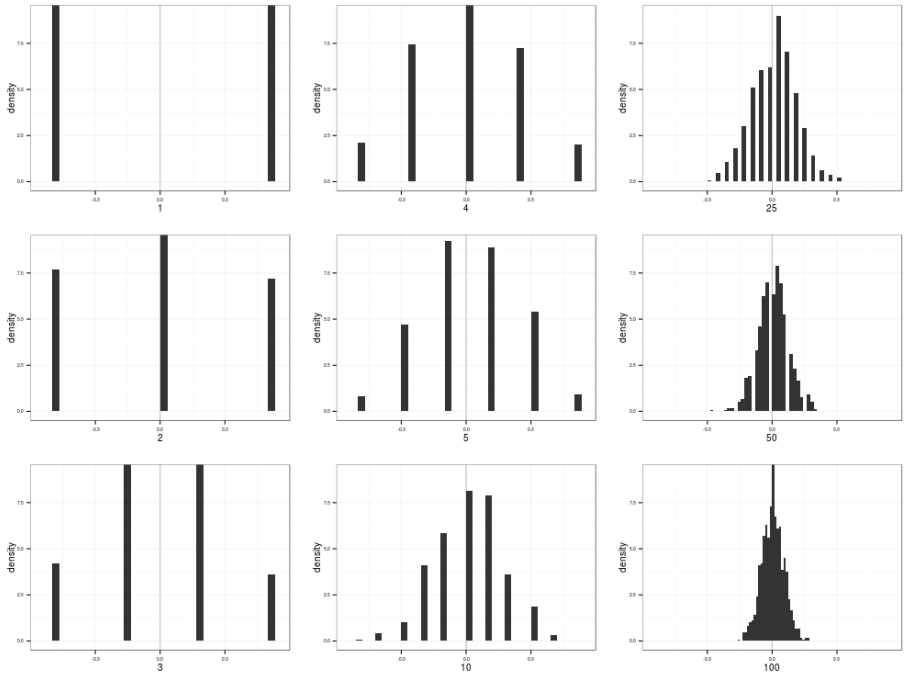

Figure 1: ¡+caption text+¿

distribution. Each of these distributions makes strong assumptions about how duration may effect extinction risk. Use of the exponential distribution assumes that extinction risk is independent of duration. In contrast, use of the Weibull distribution allows for age dependent extinction via the shape parameter α, though only as a monotonic function of duration. Importantly, the Weibull distribution is equivalent to the exponential distribution when $\alpha = 1$. In general, the notation used here follows Gelman and Hill [5], Gelman et al. [6], and STAN MANUAL.

The simplest model of genus duration includes no covariate or structural information. Define y_i as the duration in stages of genus i, where $i = 1, \ldots, n$ and n is the number of observed genera. These two models are them simply defined as

$$\begin{aligned} y_i &\sim \text{Exponential}(\lambda) \\ y_i &\sim \text{Weibull}(\alpha, \sigma). \end{aligned} \tag{6}$$

Note that λ is a "rate" or inverse-scale while σ is a scale parameter, meaning that $\frac{1}{\lambda} = \sigma$.

These simple models can then be expanded to include covariate information as predictors by reparameterizing λ and σ as a regression [13]. Each of the covariates of interest is given its own regression coefficient (e.g. β_{range}) along

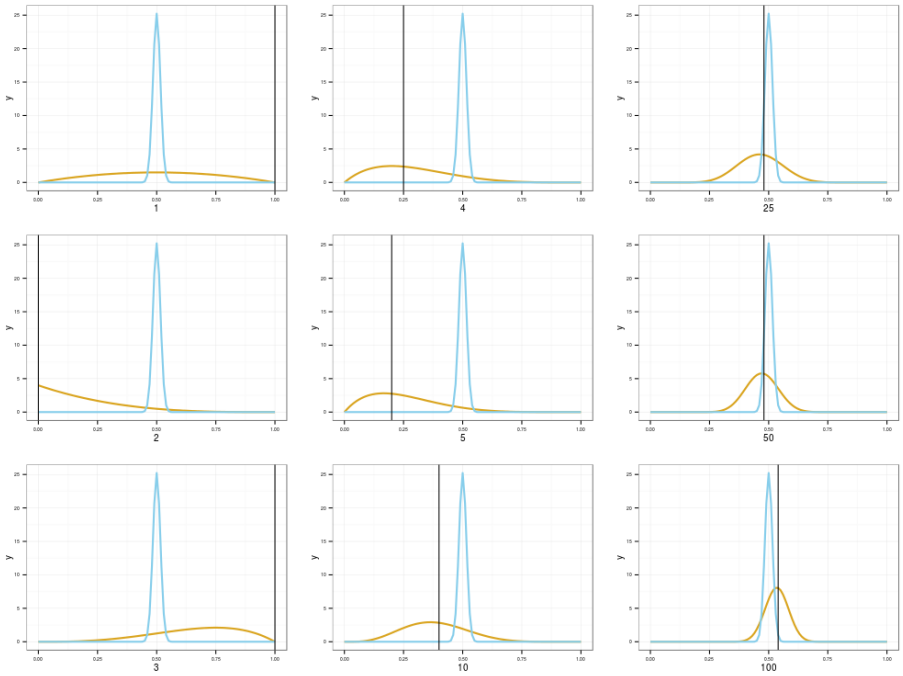

Figure 2: ¡+caption text+¿

with an intercept term β_0. There are some additional complications to the parameterization of σ associated with the inclusion of α as well as interpretability [13]. Both of these cases are written more fully as

$$\begin{aligned} \lambda_i &= \exp(\beta_0 + \beta_r r_i + \beta_v v_i + \beta_{v^2} v_i^2 + \beta_m m_i) \\ \sigma_i &= \exp\left(\frac{-(\beta_0 + \beta_r r_i + \beta_v v_i + \beta_{v^2} v_i^2 + \beta_m m_i)}{\alpha}\right). \end{aligned} \tag{7}$$

The regression equations are exponentiated because both λ and σ are only defined for positive reals. The quadratic term for environmental affinity v is to allow for the possiblity of a nonlinear relationship between environmental affinity and extinction risk.

The models which incorporate both equations 6 and 7 can then be further expanded to allow all of the β coefficients, including β_0, to vary with origination cohort while also modeling their covariance and correlation. This is called a varying-intercepts, varying-slopes model [5]. It is much easier to represent and explain how this is parameterized using matrix notation. First, define $\mathbf{B}$ as $k \times J$ matrix of the k coefficients including the intercept term ($k = 5$) for each of the J cohorts. Second, define $\mathbf{X}$ as a $n \times k$ matrix where each column is one of the covariates of interest. Importantly, $\mathbf{X}$ includes a columns of all 1s which

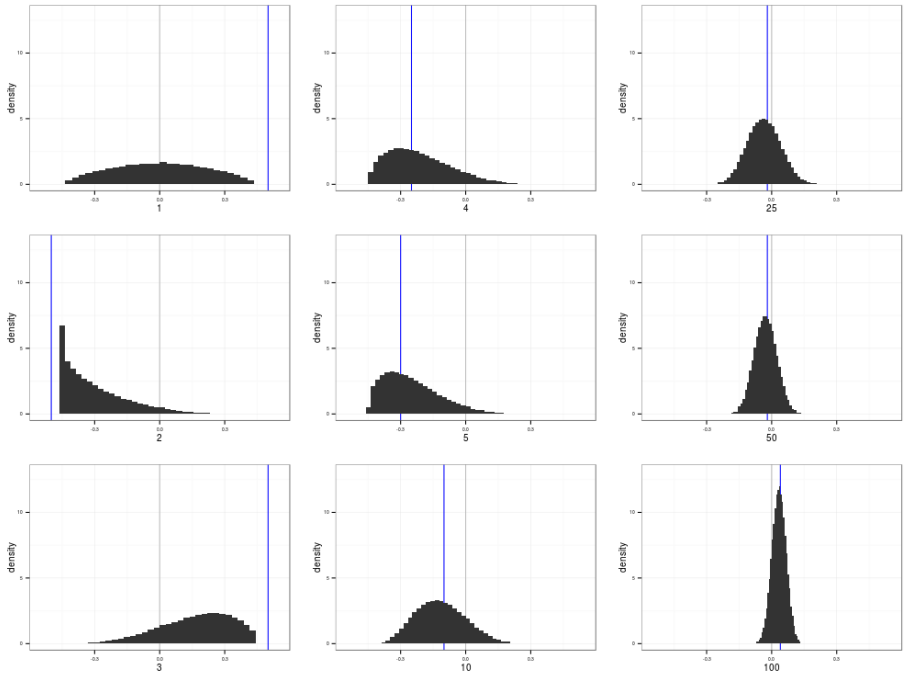

Figure 3: ¡+caption text+¿

correspond to the constant term β_0. Third, define $j[i]$ as the origination cohort of genus i, where $j = 1, \ldots, J$ and J is the total number of observed cohorts.

Using the above hierarchical expansion to the model, we then rewrite λ and σ in matrix notation as

$$\begin{aligned} \lambda_i &= \exp(\mathbf{X}_i \mathbf{B}_{j[i]}) \\ \sigma_i &= \exp\left(\frac{-(\mathbf{X}_i \mathbf{B}_{j[i]})}{\alpha}\right). \end{aligned} \tag{8}$$

At face value, the above parameterization (Eq. 8) is opaque as to how the covariance and correlation between elements of $\mathbf{B}$ are estimated. This becomes more apparent after defining the prior distribution of $\mathbf{B}$. Because $\mathbf{B}$ is a matrix, I used a multivariate normal prior with unknown vector of means μ and covariance matrix $\mathbf{\Sigma}$. This is written as

$$\mathbf{B} \sim \text{MVN}(\mu_{\mathbf{B}}, \Sigma_{\mathbf{B}}). \tag{9}$$

$\mu_{\mathbf{B}}$ is length k vector representing the overall mean of the distributions of β coefficients. $\Sigma_{\mathbf{B}}$ is a $k \times k$ covariance matrix of the β coefficients.

What remains is assigning priors the elements of $\mu_{\mathbf{B}}$ and the covariance matrix $\Sigma_{\mathbf{B}}$. Each of the elements of vector $\mu_{\mathbf{B}}$ were given independent, weakly-informative

normal priors. The prior for $\Sigma_{\mathbf{B}}$ is a bit more complicated. While the conjugate prior distribution for a covariance matrix is the inverse-Wishart [6], because I am using a variant for Hamiltonian Monte Carlo (HMC) called No U-Turn Sampling (NUTS) for posterior estimation as opposed to Gibbs sampling there is not benefit for using a conjugate prior STAN MANUAL. Additionally, the inverse-Wishart distribution strongly constraints the off-diagonal elements of the covariance matrix. Instead, it is better to model the correlation matrix and separate variance terms for each of the k coefficients. This is possible because of the relationship between a covariance and a correlation matrix, defined as

$$\Sigma_{\mathbf{B}} = \text{Diag}(\tau_B)\Omega_{\mathbf{B}}\text{Diag}(\tau_B) \tag{10}$$

where τ_B is a length k vector of variances and $\text{Diag}(\tau_B)$ is a diagonal matrix.

I used a LKJ prior distribution for $\Omega_{\mathbf{B}}$ as recommended by STAN MANUAL. An LKJ is a single parameter multivariate distribution where values of η greater than 1 concentrate density at the unit correlation matrix, which corresponds to no correlation between the β coefficients. The scale parameter, τ_B, is given a weakly informative half-Cauchy (C^+) prior following Gelman [4].

Given all the above, the exponential distribution based model is then defined, including priors, as

$$\begin{aligned}
y_i &\sim \text{Exponential}(\lambda) \\
\lambda_i &= \exp(\mathbf{X}_i\mathbf{B}_{j[i]}) \\
\mathbf{B} &\sim \text{MVN}(\mu_{\mathbf{B}}, \Sigma_{\mathbf{B}}) \\
\Sigma_{\mathbf{B}} &= \text{Diag}(\tau_B)\Omega_{\mathbf{B}}\text{Diag}(\tau_B) \\
\mu_\kappa &\sim \mathcal{N}(0, 5) \text{ for } \kappa \in 1:k \\
\tau_\kappa &\sim \text{C}^+(1) \text{ for } \kappa \in 1:k \\
\Omega &\sim \text{LKJ}(2).
\end{aligned} \tag{11}$$

The Weibull distribution based model is then also defined as

$$\begin{aligned}
y_i &\sim \text{Weibull}(\alpha, \sigma) \\
\sigma_i &= \exp\left(\frac{-(\mathbf{X}_i\mathbf{B}_{j[i]})}{\alpha}\right) \\
\mathbf{B} &\sim \text{MVN}(\mu_{\mathbf{B}}, \Sigma_{\mathbf{B}}) \\
\Sigma_{\mathbf{B}} &= \text{Diag}(\tau_B)\Omega_{\mathbf{B}}\text{Diag}(\tau_B) \\
\alpha &\sim \text{C}^+(2) \\
\mu_\kappa &\sim \mathcal{N}(0, 5) \text{ for } \kappa \in 1:k \\
\tau_\kappa &\sim \text{C}^+(1) \text{ for } \kappa \in 1:k \\
\Omega &\sim \text{LKJ}(2).
\end{aligned} \tag{12}$$

Note that the above formulations of each model (Eq. 11, 12) does not include how the uncertainty in environmental affinity is included nor how censored observations are included. An explination of including censored observations follows.

2.3 Censored observations

A key aspect of survival analysis is the inclusion of censored, or incompletely observed, data points [8, 13]. The two classes of censored observations encountered in this study were right and left censored observations. Right censored genera are those that did not go extinct during the window of observation, or genera that are still extant. Left censored observations are those taxa that it is only known when it was extinct by. To put another way, this is a taxon that went extinct but the observed duration is an over estimate of the actual duration.

In the context of this study, I considered all genera that had a duration of only one geologic stage to be left censored as we do not have a finer degree of resolution. Conceptually, this is similar to if I was studying, say, survival patterns in rats and an individual had died between the start of the experiment and next time the rats were observed. We know the rat lived no more than day.

The key function for modeling censored observations is the survival function, or $S(t)$. $S(t)$ corresponds to the probability that a genus having existed for t stages will not have gone extinct while $h(t)$ corresponds to the instantaneous extinction rate at taxon age t [13]. For an exponential model, $S(t)$ is defined as

$$S(t) = \exp(-\lambda t), \tag{13}$$

and for the Weibull distribution $S(t)$ is defined as

$$S(t) = \exp\left(-\left(\frac{t}{\sigma}\right)^{\alpha}\right). \tag{14}$$

$S(t)$ is equivalent to the complementary cumulative distribution function, $1-F(t)$ [13].

For right censored observations, instead of calculating the likelihood as normal (Eq. 8) the likelihood of an observation is evaluated using $S(t)$. Conceptually, this approach calculates the likelihood of observing a taxon that existed for at least that long. For left censored data, instead the likelihood is calculated using $1 - S(t)$ which corresponds to the likelihood of observing a taxon that existed no longer than t.

The full likelihood statements incorporating fully observed, right censored, and left censored observations are then

$$\begin{aligned}
\mathcal{L} &\propto \prod_{i \in C} \text{Exponential}(y_i|\lambda) \prod_{j \in R} S(y_j|\lambda) \prod_{k \in L} (1 - S(y_k|\lambda)) \\
\mathcal{L} &\propto \prod_{i \in C} \text{Weibull}(y_i|\alpha, \sigma) \prod_{j \in R} S(y_j|\alpha, \sigma) \prod_{k \in L} (1 - S(y_k|\alpha, \sigma))
\end{aligned} \tag{15}$$

where C is the set of all fully observed taxa, R the set of all right censored taxa, and L the set of all left-censored taxa.

2.4 Parameter estimation

Given the above likelihood and prior statements, the posterior probabilities of all parameters was approximated using a Markov-chain Monte Carlo routine using a variant of Hamiltonian Monte Carlo called the No-U-Turn Sampler [7] as implemented in the probabilistic programming language Stan [**?**]. The estimate of the posterior distribution were approximated from four parallel chains run for 10000 draws split half warm-up and half sampling thinned to every 10 sample for a total of 5000 samples. Chain convergence was assessed via the scale reduction factor $\hat{R}$ where values close to 1 ($\hat{R} < 1.1$) indicate approximate convergence. Convergence means that the chains are approximately stationary and the samples are well mixed [6].

2.5 Model evaluation

Models were evaluated using both a series of multiple posterior predictive checks and an estimate of out-of-sample predictive accuracy.

The motivation behind posterior predictive checks as tools for determining model adequacy is that replicated data sets using the fitted model should be similar to the original data [6]. Systematic differences between the simulations and observed indicate weaknesses of the currently fit model. An example of a technique that is very similar would be inspecting the residuals and Q-Q plots from a linear regression.

The strategy behind posterior predictive checks is to draw simulated values from the joint posterior predictive distribution, $p(y^{rep}|y)$, and then compared to the original observed values [6]. To accomplish this, for each replicate, a single value is drawn from the marginal posterior distributions of each regression coefficient from the final model (Eq. 11, 12). Then, given the covariate information for each of the observations $\mathbf{X}$, a new set of n genus durations are generated giving a single replicated data set y^{rep}. This is repeated 1000 times in order to provide a distribution of possible values that could have been observed given the model.

In order to compare the fitted model to the observed data, various graphical comparisons or test quantities need to be defined. The principal comparison used here is a comparison between non-parameteric approximation of the survival function $S(t)$ as estimated from both the observed data and each of the replicated data sets. The purpose of this comparison is to determine if the model approximates the same survival/extinction pattern as the original data.

I also did a graphical examination of the deviance residuals. While normal residuals are defined as $y_i^{rep} - y_i$, deviance residuals are a specific class of residuals derived with non-normal errors in mind. The definition of deviance residuals for a Weibull regression, of which the above models can be considered, is as follows. First define the cumulative hazard function $\Lambda(t)$ for the Weibull

distribution [13]. Given $S(t)$ (Eq. 14), the cumulative hazard function is

$$\Lambda(t) = -log\left(S\left(t\right)\right). \tag{16}$$

Next, define martingale residuals m as

$$m_i = I_i - \Lambda(t_i). \tag{17}$$

I, called the inclusion vector, is vector of length n where $I_i = 1$ means the observation is completely observed and $I_i = 0$ means the observation is censored. Martingale residuals have a mean of 0, range between 1 and $-\infty$, and can be viewed as the difference between the observed number of deaths between 0 and t_i and the expected number of deaths based on the model. However, martingale residuals are asymmetrically distributed, and can not be interpreted in the same manner as standard residuals.

The solution to this is to use deviance residuals, D, which are defined as a function of martingale residuals and takes the form

$$D_i = \text{sign}(m_i)\sqrt{-2[m_i + I_i log(I_i - m_i)]}. \tag{18}$$

Deviance residuals have a mean of 0 and a standard deviation of 1 by definition [13].

The exponential and Weibull models were compared for out-of-sample predictive accuracy using the widely-applicable information criterion (WAIC) [29]. Because the Weibull model reduces to the exponential model when $\alpha = 0$, our interest is not in choosing between these models. Instead comparison of WAIC values is useful for better understanding the effect of model complexity on out-of-sample predictive accuracy. The calculation of WAIC used here corresponds to the "WAIC 2" formulation recommended by Gelman et al. [6].

WAIC can be considered fully Bayesian alternative to the Akaike information criterion, where WAIC acts as an approximation of leave-one-out cross-validation which acts as a measure of out-of-sample predictive accuracy. WAIC is calculated starting with the log pointwise posterior predictive density calculated as

$$\text{lppd} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\Theta^S)\right), \tag{19}$$

where n is sample size, S is the number posterior simulation draws, and Θ represents all of the estimated parameters of the model. This is similar to calculating the likelihood of each observation given the entire posterior. A correction for the effective number of parameters is then added to lppd to adjust for overfitting. The effective number of parameters is calculated, following derivation and recommendations of [6], as

$$p_{\text{WAIC}} = \sum_{i=1}^{n} V_{s=1}^{S}(\log p(y_i|\Theta^S)). \tag{20}$$

where V is the sample posterior variance of the log predictive density for each data point.

Given both equations 19 and 20, WAIC is then calculated

$$\text{WAIC} = \text{lppd} - p_{\text{WAIC}}. \tag{21}$$

When comparing two or more models, lower WAIC values indicate better out-of-sample predictive accuracy. Importantly, WAIC is just one way of comparing models. When combined with posterior predictive checks it is possible to get a more complete understanding of model fit.

3 Results

As stated above, posterior approximations for both the exponential and Weibull models achieved approximate stationarity after 10,000 steps as all parameter estimates have an $\hat{R} < 1.1$ REF TABLES.

Comparisons of the survival functions estimated from 1000 posterior predictive estimated data sets to the estimated survival function of the observed genera demonstrates that both the exponential and Weibull models approximately capture the observed pattern of extinction (Fig. 4). The major difference in fit between the two models is that the Weibull model has a slightly better fit for longer lived taxa than the exponential model.

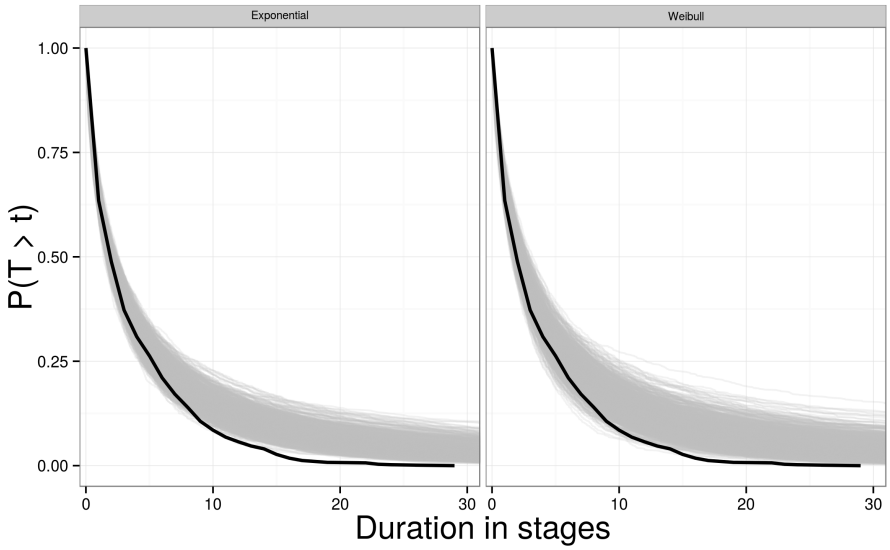

Figure 4: ¡+caption text+¿

Additionally, the Weibull model is expected to have slightly better out-of-sample predictive accuracy when compared to the exponential model VALUES. This is congruent with graphical comparisons of the survival functions (Fig. 4). Because the difference between the WAIC scores is small, both results from the exponential and Weibull models will be analyzed.

Estimates of the overall overall mean covariate effects $\mu_{\mathbf{B}}$ can be considered time-invariant generalizations for brachiopod survival for the Paleozoic SMITS IN PREP (Fig. 5). Consistent with expectations, geographic range size has a negative effect on extinction risk where genera with large ranges having greater durations than genera with small ranges. I also find no time-invariant effect of body size on genus duration.

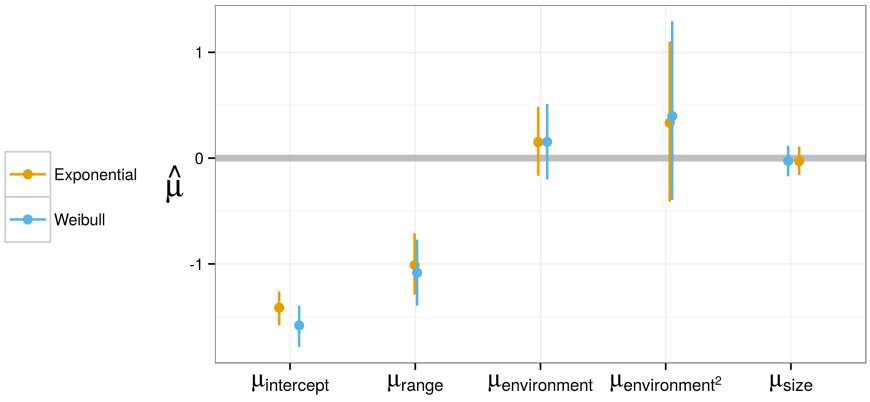

Figure 5: ¡+caption text+¿

Interpretation of the effect of environmental preference v on duration is slightly more involved. Because a quadratic term is the equivalent of an interaction term, both μ_v and μ_{v^2} have to be interpreted together because it is illogical to change values of v and only affect one of the coefficients. To determine the nature of the effect of v on duration I calculated the multiplicative effect of environmental preference on extinction risk.

Given mean estimated extinction risk $\tilde{\sigma}$, we can define the extinction risk multiplier of an observation with environmental preference v_i as

$$\frac{\tilde{\sigma_i}}{\tilde{\sigma}} = \exp\left(\frac{-(\mu_v v_i + \mu_{v^2} v^2)}{\alpha}\right). \tag{22}$$

This exponentiated quadratic function has a y-intercept of log(0) or 1 by definition. Equation 22 can be either upward or downward facing. A downward facing function indicates that genera of intermediate environmental preference

have greater durations than either extreme, and *vice versa* for upward facing functions.

The estimated mean effect of environmental preference as a multiplier of mean extinction risk can then be simply visualized (Fig. 6). This figure depicts 1000 posterior predictive estimates of Eq. 22 across all possible values of v. The number indicates the posterior probability that the function is downward facing, with generalists having lower extinction risk/greater duration than either type of specialist. Note that the inflection point/optimum of Fig. 6 is at approximately 0, something that is observable given the estimate of μ_v (Fig. 5).

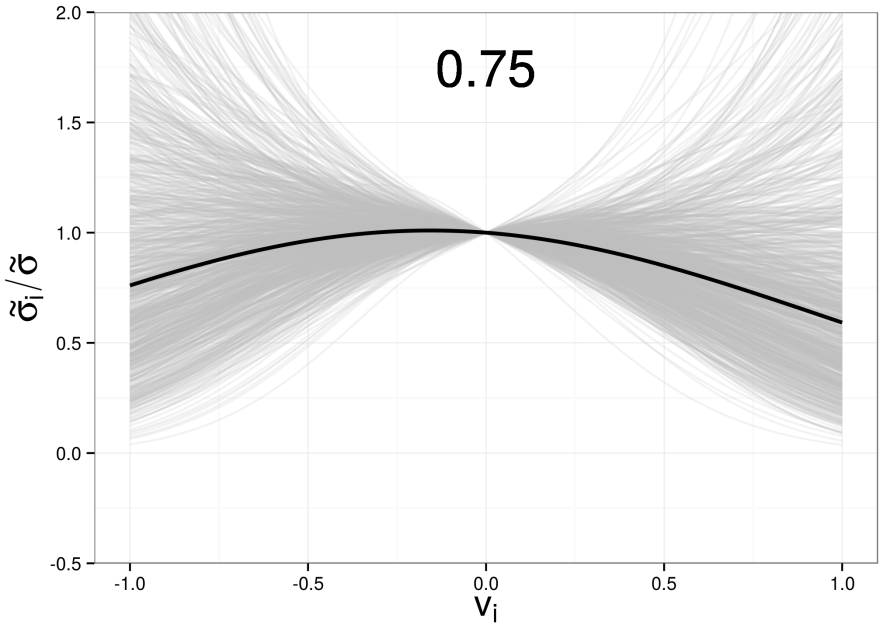

Figure 6: ¡+caption text+¿

The matrix Σ describing the covariance between the different coefficients describes how these coefficients might vary together across the origination cohorts. Similar to how this was modeled (Eq. 11, 12), Σ can be decomposed into a vector of standard deviations τ and a correlation matrix Ω.

The estimates of the standard deviation of between cohort coefficient estimates τ_B vary greatly (Fig. 7). Coefficients with greater values of τ have greater between cohort variation. The covariate effect with the greatest between origination cohort variation is β_{v^2} with β_r being the second largest. Both β_v and β_m have little between cohort variation, as both have less variation than mean extinction risk β_0. However the amount between cohort variation in estimates of β_v means that it is possible that the function describing the effect of environment can be

upward facing for some cohorts (Eq. 22), meaning that environmental generalists being shorted lived than specialists.

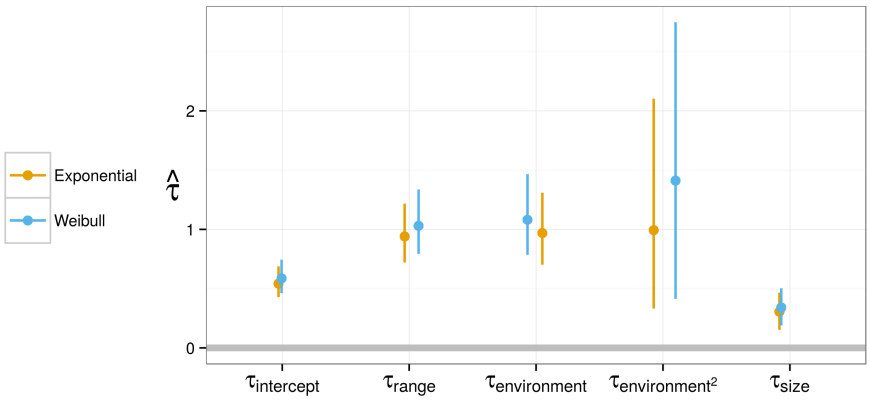

Figure 7: ¡+caption text+¿

The correlation matrix Ω shows a few important correlations between the coefficients (Fig. 8). The correlation terms describe the relationship between the coefficients and how their estimates may vary together across cohorts. Of particular note are the correlations between the intercept term β_0, or expected extinction risk, and the biological covariates (Fig. 8 first column/last row). Keep in mind that when β_0 is low, extinction risk is low; and conversely, when β_0 is high, then extinction risk is high.

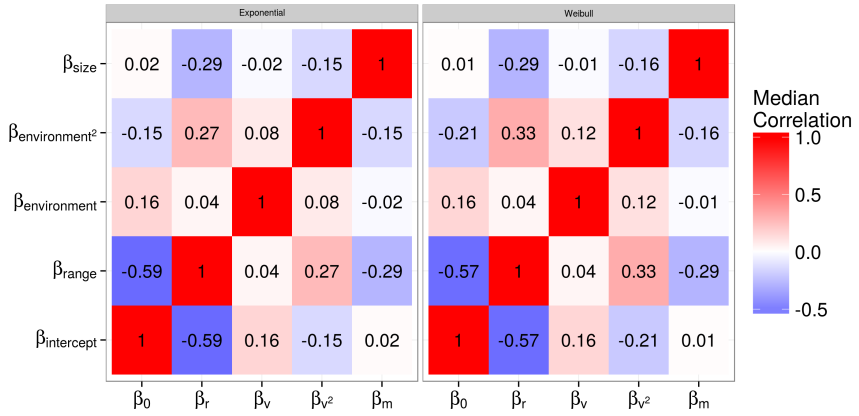

Figure 8: ¡+caption text+¿

The marginal posterior estimates for the correlations between estimates of the β_0 and the effects of the biological covariates indicate that the relationships between expected extinction risk and either geographic range β_r is of note (Fig. 9). The negative correlation between β_0 and β_r implies that as extinction risk increases, the effect/importance of geographic range on genus duration increases.

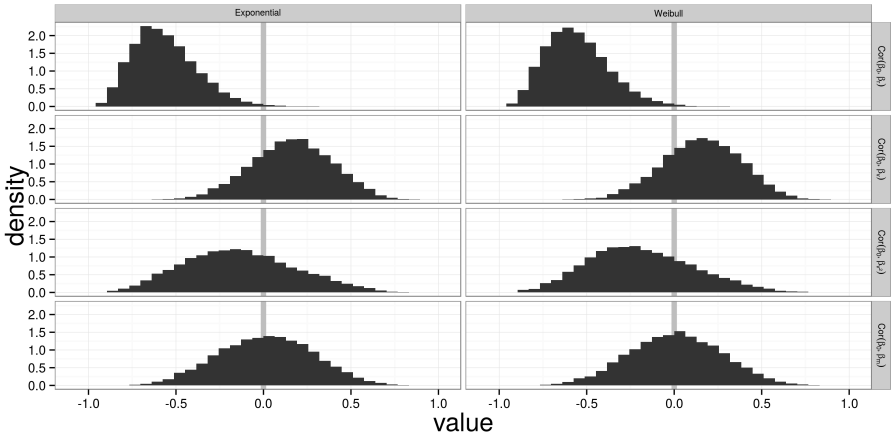

Figure 9: ¡+caption text+¿

In addition to just analyzing the covariance matrix between the coefficient estimates, it is important to also observe the individual origination cohort level estimates.

In comparison to the overall mean extinction risk $\mu_{intercept}$, cohort level estimates β_0 show some amount of variation through time as expected by estimates of $\tau_{intercept}$ (Fig. 10). A similar, if slightly greater, amount of variation is also observable in cohort estimates of the effect of geographic range β_0 (Fig. 11). Again, smaller values of β_0 correspond to lower expected extinction risk. Similarly, smaller values of β_0 correspond to greater decrease in extinction risk with increasing geographic range

How environmental effect varies between cohorts can be observed by using the cohort specific estimates of the coefficients $\mathbf{B}_j$. Following the exact same procedure for generating figure 7, but substituting $\beta_{v;j}$ and $\beta_{v^2;j}$ for μ_v and μ_{v^2}, the cohort specific effect of environmental preference as a multiplier of mean extinction risk can be calculated. This was done only for the Weibull model, though the observed pattern should be similar for the exponential model.

As expected based on the estimates of τ_v and τ_{v^2}, there is greater variation in the peakedness of the function than there is variation between upward and downward facing functions (Fig. 12. Only 9 of the 31 cohorts have less than 50% posterior probability that generalists are shorter lived than specialists, though

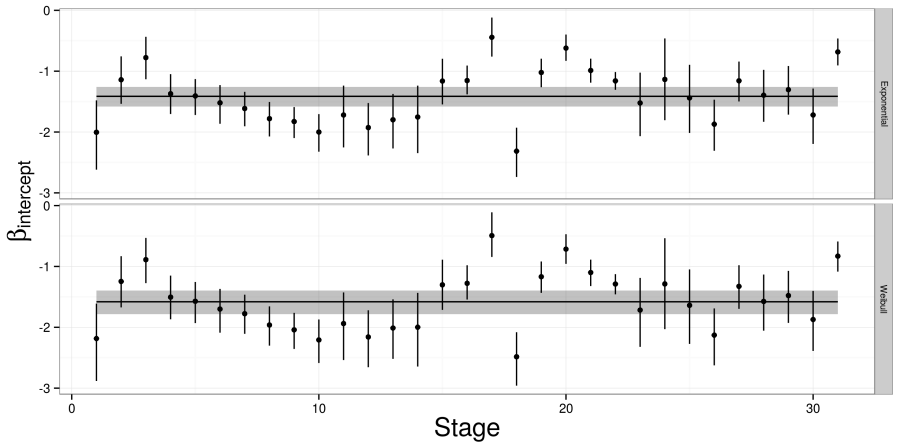

Figure 10: ¡+caption text+¿

2 of those cases have approximately a 50% probability of being either upward or downward facing. This is congruent with the 0.70+ posterior probability that μ_{v^2} is positive/the relationship is downward facing, which corresponds to approximately 8 out of 31 cohorts.

Additionally, a quite striking pattern emerges when the inflection point of the function is either far away from the y-intercept (x = 0, y = 1) or when there is little evidence non-linearity (Fig. 12). Cohort 21 and 20, for example, have almost linear relationships between environmental preference and duration multiplier. This type of relationship occurs when β_{v^2} approaches 0, flattening the non-linear curvature of the relationship.

References

[1] T. K. Baumiller. Survivorship analysis of Paleozoic Crinoidea: effect of filter morphology on evolutionary rates. *Paleobiology*, 19(3):304–321, 1993.

[2] M. Foote. Survivorship analysis of Cambrian and Ordovician Trilobites. *Paleobiology*, 14(3):258–271, 1988.

[3] M. Foote. Substrate affinity and diversity dynamics of Paleozoic marine animals. *Paleobiology*, 32(3):345–366, Sept. 2006. ISSN 0094-8373. doi: 10.1666/05062.1. URL `http://www.bioone.org/doi/abs/10.1666/05062.1`.

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

[5] A. Gelman and J. Hill. *Data Analysis using Regression and Multi-*

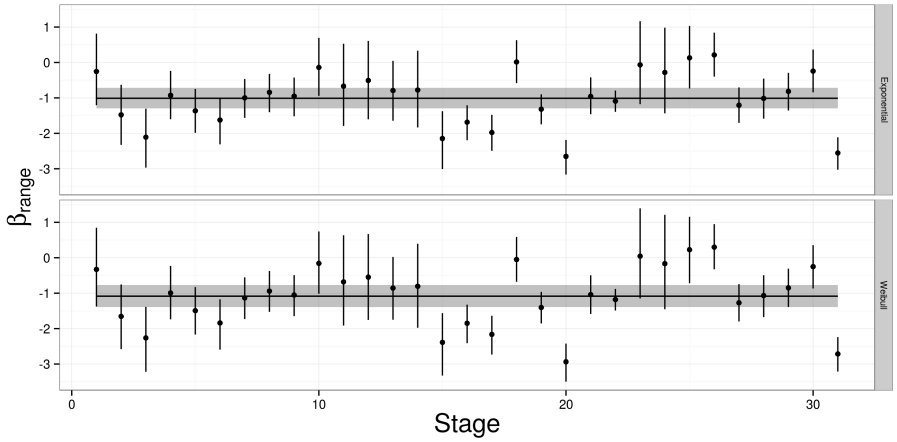

Figure 11: ¡+caption text+¿

level/Hierarchical Models. Cambridge University Press, New York, NY, 2007.

[6] 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.

[7] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

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

[9] D. Jablonski. Background and mass extincitons: the alternation of macroevolutionary regimes. *Science*, 231(4734):129–133, 1986.

[10] D. Jablonski. Heritability at the species level: analysis of geographic ranges of cretaceous mollusks. *Science*, 238(4825):360–363, Oct. 1987. ISSN 0036-8075. doi: 10.1126/science.238.4825.360. URL `http://www.ncbi.nlm.nih.gov/pubmed/17837117`.

[11] D. Jablonski and K. Roy. Geographical range and speciation in fossil and living molluscs. *Proceedings. Biological sciences / The Royal Society*, 270 (1513):401–6, Feb. 2003. ISSN 0962-8452. doi: 10.1098/rspb.2002.2243. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1691247&tool=pmcentrez&rendert`

[12] W. Kiessling and M. Aberhan. Environmental determinants of marine benthic biodiversity dynamics through Triassic Jurassic time. *Paleobiology*, 33(3):414–434, 2007.

[13] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2nd edition, 2003.

[14] L. H. Liow. A test of Simpson's "rule of the survival of the relatively unspecialized" using fossil crinoids. *The American naturalist*, 164 (4):431–43, Oct. 2004. ISSN 1537-5323. doi: 10.1086/423673. URL http://www.ncbi.nlm.nih.gov/pubmed/15459876.

[15] A. I. Miller and S. R. Connolly. Substrate affinities of higher taxa and the Ordovician Radiation. *Paleobiology*, 27(4):768–778, Dec. 2001. ISSN 0094-8373. doi: 10.1666/0094-8373(2001)027¡0768:SAOHTA¿2.0.CO;2. URL http://www.bioone.org/doi/abs/10.1666/0094-8373%282001%29027%3C0768%3ASAOHTA%3E2.0.CO%3E

[16] A. I. Miller and M. Foote. Epicontinental seas versus open-ocean settings: the kinetics of mass extinction and origination. *Science*, 326(5956): 1106–9, Nov. 2009. ISSN 1095-9203. doi: 10.1126/science.1180061. URL http://www.ncbi.nlm.nih.gov/pubmed/19965428.

[17] S. Nürnberg and M. Aberhan. Habitat breadth and geographic range predict diversity dynamics in marine Mesozoic bivalves. *Paleobiology*, 39(3):360–372, Apr. 2013. ISSN 0094-8373. doi: 10.1666/12047. URL http://www.bioone.org/doi/abs/10.1666/12047.

[18] S. Nürnberg and M. Aberhan. Interdependence of specialization and biodiversity in Phanerozoic marine invertebrates. *Nature communications*, 6:6602, Jan. 2015. ISSN 2041-1723. doi: 10.1038/ncomms7602. URL http://www.ncbi.nlm.nih.gov/pubmed/25779979.

[19] J. L. Payne and S. Finnegan. The effect of geographic range on extinction risk during background and mass extinction. *Proceedings of the National Academy of Sciences*, 104:10506–11, June 2007. ISSN 0027-8424. doi: 10.1073/pnas.0701257104. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1890565&tool=pmcentrez&rendert

[20] J. L. Payne, N. A. Heim, M. L. Knope, and C. R. Mcclain. Metabolic dominance of bivalves predates brachiopod diversity decline by more than 150 million years. *Proceedings of the Royal Society B*, 281:20133122, 2014.

[21] D. M. Raup. Taxonomic survivorship curves and Van Valen's Law. *Paleobiology*, 1(1):82–96, Jan. 1975. ISSN 0036-8075. doi: 10.1126/science.49.1254.50. URL http://www.ncbi.nlm.nih.gov/pubmed/17777225.

[22] D. M. Raup. Cohort Analysis of generic survivorship. *Paleobiology*, 4(1): 1–15, 1978.

[23] C. Simpson. *Levels of selection and large-scale morphological trends*. PhD thesis, University of Chicago, 2006.

[24] C. Simpson and P. G. Harnik. Assessing the role of abundance in marine bivalve extinction over the post-Paleozoic. *Paleobiology*, 35(4):631–

647, Dec. 2009. ISSN 0094-8373. doi: 10.1666/0094-8373-35.4.631. URL http://www.bioone.org/doi/abs/10.1666/0094-8373-35.4.631.

[25] G. G. Simpson. *Tempo and Mode in Evolution*. Columbia University Press, New York, 1944.

[26] L. Van Valen. A new evolutionary law. *Evolutionary Theory*, 1:1–30, 1973. URL http://ci.nii.ac.jp/naid/10011264287/.

[27] L. Van Valen. Taxonomic survivorship curves. *Evolutionary Theory*, 4: 129–142, 1979.

[28] S. C. Wang. On the continuity of background and mass extinction. *Paleobiology*, 29(4):455–467, Dec. 2003. ISSN 0094-8373. doi: 10.1666/0094-8373(2003)029¡0455:OTCOBA¿2.0.CO;2. URL http://www.bioone.org/doi/abs/10.1666/0094-8373%282003%29029%3C0455%3AOTCOBA%3E2.0.CO%3E

[29] S. Watanabe. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

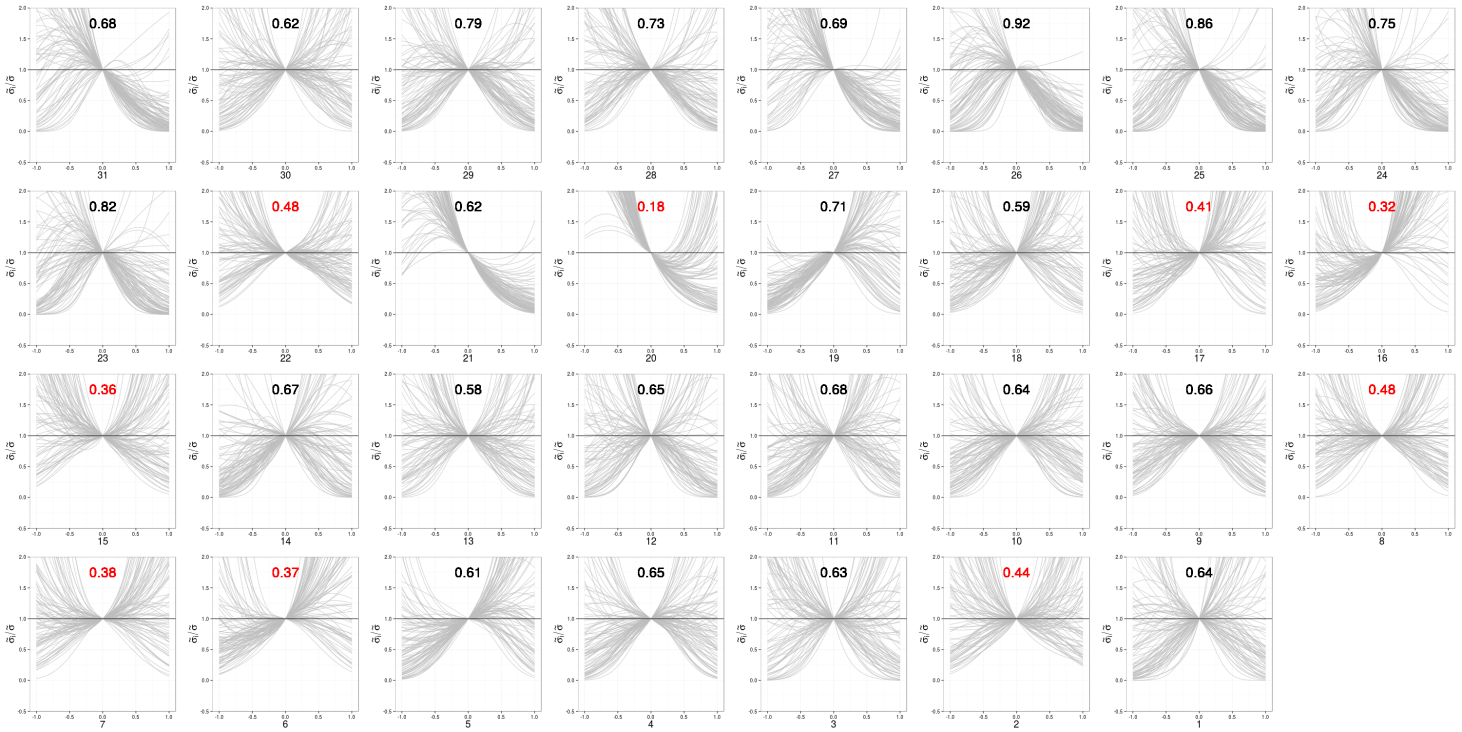

Figure 12: ¡+caption text+¿