1 Data Specifications

Microfossil occurrence data was downloaded from the Neptune Database `http://www.nsb-mfn-berlin.de/nannotax`. Occurrence information was downloaded for calcareous nannofossils, diatoms, dinoflaggellants, foraminifera, and radiolarians from the entire globe and having occurred between 120 and 0 Mya. This selection of species was then culled to just those species which have their first occurrence at most 63 My; this avoids those taxa which survived the K/Pg boundary, those taxa arrising just after the K/Pg boundary, and lines up the occurrences with the temperature time-series (discussed below).

Data was binned into 1 My bins based on the estimated age of the fossil occurrence. The estimated ages of each occurrence is a product of core-specific age-models and are overly precise. The hope is that binning the data overcomes the heterogenity in age-models between the cores.

The survival or extinction of a taxon is determined for each taxons' occurrences. For every occurrence except the last, the taxon has survived which is indicated by a 0. The last occurrence of the taxon is considered the bin in which the taxon has gone extinct. This protocol means that we are reading the fossil record "as written," a practice that is potentially dangerous CITATIONS but is common with marine microfossil data CITATIONS.

A taxon's geographic range during a 1 My bin was calculated as the maximum great circle distance on an ellipsoid between any two occurrences, also called a geodesic. Geographic range was measured in kilometers.

Geographic range was then log-plus-one transformed and then standardized by zero-centering the data and then dividing by its standard deviation so that geographic range had mean 0 and standard deviation 1.

Temperature data is based on estimates from Cramer et al CITATION. These estimates are based on Magnesian/Calcium isotope ratios and are a more accurate estimate of global temperature than the standard Oxygen isotope-based estimates because Mg/Ca based estimates are not effected by ice-volume and fresh-water input (e.g. metioric water). Cramer et al. provide temperature estimates for every 0.1 My from 0 to 63 Mya; we binned these estimates every 1 My in accordance with the way we binned all fossil occurrences.

2 Model Specifications

In survival analysis, the hazard funciton describes the instanenous rate of extinction at time t given that the species has lived up to that point. For discrete-time survival data, this is expressed as $\lambda(t|X) = P(T = t|T \geq t, X)$ for $t = 1, \cdots N$ where t is the status of a species at a time point, T is the time of extinction of the species, and $i = 1, 2, \cdots, N$ with N being the total number of observations.

In survival analysis the hazard function describes the in stannous rate of extinction for a species given matrix of covariates X, $\lambda(t|X)$. We parameterized the hazard function as a logistic regression/generalized linear model defined $\lambda(t|X) = h(\Theta)$ where $h(.)$ is an inverse logit link function and Θ is the probability of a taxon going extinction during interval t.

Θ is then parameterized as a hierarchical model with multiple non-nested varying intercepts CITATION. First, we considered two aspects of geographic range as continuous covariates: geographic range r during interval t, and the difference d between the geographic range at $t - 1$ and t. We also considered the interaction between geographic range during intervals t and $t - 1$. At the first observation for a taxon, the diff of geographic range is 0.

Second, we considered two aspects of global temperature: mean temperature during interval t, and the lag of mean temperature (i.e. mean temperature during interval $t - 1$.)

The logistic regression can thus be expressed as

$$t_i = logit^{-1}(\beta_{when,fossil.group} + \beta_{when,fossil.group}range_i + \beta_{when,fossil.group}rangediff_i + \beta_{t_i}$$
$$\beta_{when,fossil.group} \sim MVN(\alpha_{fossil.group,\Sigma})$$
$$\alpha_{fossil.group} \sim MVN(0, \Sigma)$$
$$\gamma \sim N(0, \sigma)$$
$$a_{when,fossil.group} \sim MVN(\delta_{fossil.group}, \Sigma)$$
$$\delta_{fossil.group} \sim MVN(0, \Sigma)$$

$$(1)$$

with i indexing the observation and bracket subscripts referencing the class of the ith observation e.g. $j[i]$ is the age of the ith observation. $a_{j[i]}$ is taxon age in bins at time of observation where $j = 1, 2, \cdots, J$ and J is the maximum observed age of any taxon. b_k is the time of observation where $k = 1, 2, \cdots, K$

and K is the number of time bins observed. Finally, c_f is the taxonomic group of an observation where $f = 1, 2, \cdots, F$) and F is the total number of taxonomic groups analyzed.

3   Model Parameter Estimation

We implemented our model using the **r**stanarm package for the R programming language CITATION. This package provides an interface to the Stan probabilistic programming language for writing hierarchical/mixed-effects models in native R. Posterior estimates were obtained through Hamiltonian Monte Carlo, using 2000 steps divided equally between warm-up and sampling. In order to prevent divergent transitions the adapt delta value was increased to 0.99; all other HMC/NUTS sampling parameters were kept at the defaults for rstanarm version XX CITATION.

Posterior convergence was determined using the following general and HMC specific diagnostic criteria: scale reduction factor ($\hat{R}$; target < 1.1), effective sample size (eff; target value eff/steps < 0.0001), number of samples that saturated the maximum trajectory length for avoiding infinite loops (treedepth; target value 0), sample divergence, and the energy Bayesian Fraction of Mission Information (E-BFMI; target value > 0.2). For futher explanation of these diagnostic criteria, see the Stan Manual cITATION.

4   Model Selection and Adequacy

We considered four variants of this model: 1) only age-based varying intercepts, 2) age and time of observation varying intercepts, 3) age and taxonomic group varying intercepts, and 4) all three considered varying intercepts. These models were compared using three measures of model performance: WAIC and LOOIC which are estimates of a model's out-of-sample performance, and ROC which is a measure of a model's predictive performance.

To determine how many non-nested varying-intercept terms were possible to include without overly biasing out models' parameter estimates, the four variant models described above were compared using both WAIC and LOOIC

which are estimates of a model's expected out-of-sample performance; these measures are fully Bayesian, taking into account the entire estimated joint posterior. CITATION. WAIC and LOOIC are interpreted similarly to AIC, with lower values indicating greater expected out-of-sample performance. Additionally, the Bayesian nature of WAIC and LOOIC mean that they are calculated with standard errors allowing for straight forward comparisons. We selected the model with lowest WAIC and LOOIC and all models within one standard error of its WAIC/LOOIC estimate.

Given these selected models, we did a further comparison between models including the change in geographic range from $t-1$ to t, and its interaction with geographic range at t, in the model to determine how much our predictions of extinction risk are improved by the inclusion of "the past."

The model adequacy was measured using the area under the reciever operating characteristic curve (AUC). This measure is commonly used in classification problems like this one as it has the desireable characterisic of comparing the model's true positive rate with its false positive rate, as opposed to only true positive rate measured by accuracy CITATION. This value ranges between 0.5 and 1, with 0.5 indicating no improvement in performance from random and 1 indicating perfect performance. AUC was calculated for the dataset as a whole, and for each of time bins.