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. This standardization means that a regression coefficient associated with this covariate describes the change in extinction probability per change in standard deviation of geographic range, and that coefficients associated with similarly standardized covariates will be directly interpretable.

Temperature data used as covariates in this analysis are based on Magnesium/Calcium isotope ratios sourced from Cramer et al CITATION. These are considered 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).

Our data source, Cramer et al., provides temperature estimates for every 0.1 My from 0 to 63 Mya; we binned these estimates every 1 My to match with how we binned all fossil occurrences. The mean of these values was used as the temperature estimate for that 1 My interval (Fig. 1). Temperature was transformed and standardized the in the same manner as geographic range.

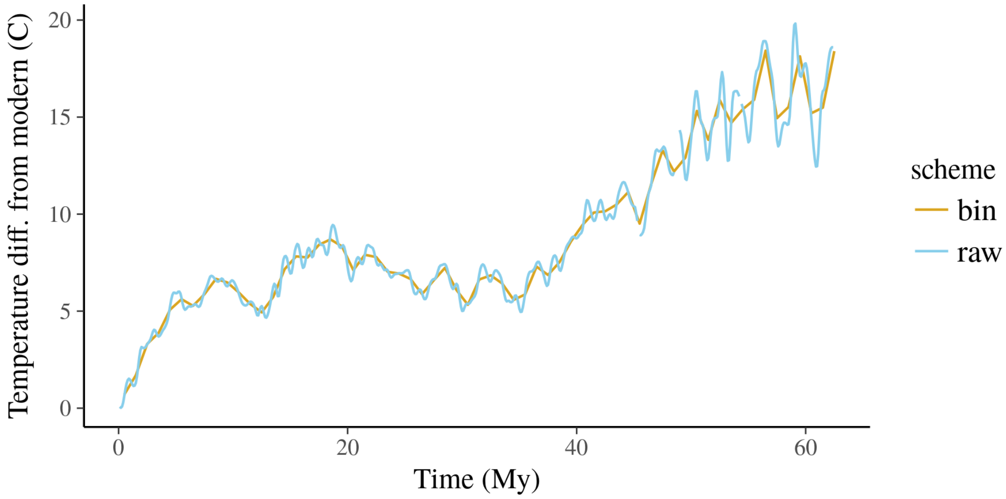

Figure 1: Comparison of initial temperature estimates from Cramer et al. CITATION (goldenrod) versus the binned values used in this analysis (blue). The initial values are for every 0.1 My while our bins are defined for every 1 My.

2 Model Specifications

In survival analysis, the hazard funciton describes the instanenous rate of extinction of a species. Species occurrence are recorded for all k 1 My intervals. For the discrete time intervals $T = 1, \cdots, k$, extinction is defined as $T = t$. The discrete-time hazard function is then written

$$\lambda(t|X) = P(T = t|T \geq t, X), \quad t = 1, \cdots, k \tag{1}$$

where X is a matrix of covariates. This statement is the conditional probability of a species going extinct by the end of the t-th interval given that it survied up until t.

The corresponding survival function, describing the probability of a species going extinct after time t is expressed

$$S(t|X) = P(T > t|X) = \prod i = 1t(1 - \lambda(i|X). \tag{2}$$

Convinently, the hazard function (Eq. 1) is easily reparameterized as logistic regression. This form of logistic regression model defined where $\lambda(t|X) = h(\Theta)$ where $h(.)$ is an logit inverse-link function and Θ is the probability of a taxon going extinction during interval t. Θ is then parameterized as any regression model.

In this case, we opted for a hierarchical/mixed-effects model with multiple non-nested varying intercepts and slopes 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. Secondly, 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

$$\begin{aligned}
t_i &\sim \text{Bernoulli}(\Theta) \\
\Theta &= logit^{-1}(\beta_{w[i],f[i]} + \beta_{w[i],f[i]} r_i + \beta_{w[i],f[i]} d_i + \gamma_1 c_{w[i]} + \gamma_2 l_{w[i]} + a_{d[i],f[i]}) \\
\beta_{w,f} &\sim MVN(\alpha_f, \Sigma) \\
\alpha_f &\sim MVN(0, \Sigma) \\
\gamma &\sim N(0, \sigma) \\
a_{w,f} &\sim MVN(\delta_f, \Sigma) \\
\delta_f &\sim MVN(0, \Sigma)
\end{aligned} \tag{3}$$

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.

The basic core of the model described above 3 can be written in R formula syntax as

$$\begin{aligned}\text{event} \sim \text{range} + \text{range\_diff} + \text{temp} + \text{temp\_lag} + \\ (1 + \text{range} + \text{range\_diff}|\text{mybin/phylum}) + \\ (1|\text{age/phylum}).\end{aligned} \tag{4}$$

This formula is then used with a call to `stan_glmer` as Bernoulli family model which, along with the relevant data and prior choices, yields parameter estimates.

3 Model Parameter Estimation

We implemented our model using the `rstanarm` 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) historical covariates with time-varying intercepts and slopes, 2) no historical coviarates with time-varying intercepts and slopes, 3) historical covariates without time-varying intercepts and slopes, and 4) no historical covariates and no time-varying intercepts and slopes (except age effect).

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 characteristic 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.

4.1 Prediction given class imbalance

When the classes are not approximately balanced, the base probability of 1 versus 0 is below 0.5. When the intercept term of the logistic regression is sufficiently negative (-2 or less), the only way to obtain a linear predictor value of 0.5 or greater is if the covariate effects are large or there the covariates are at extreme values. The 0.5 default cut-off point for binary prediction is much to high. A method for determining the new "optimal" cutpoint involves calculating the ROC curve for the data and finding the point along that curve which maximizes both true positive rate and the false positive rate; the combination with the greatest sum corresponds to the new optimal cutpoint CITATION. Class membership is then predicted given this new cutpoint. This process is done for every posterior draw, as each realization from the posterior predictive distribution has its own optimal cutpoint.

4.2 Cross-validation

Cross-validation is a procedure for estimating the expected out-of-sample predictive error of a model CITATION. By dividing the data into sub-components (folds), fitting a model to some of the folds folds, and then predicting the response values for the held-out fold. For our dataset, expected out-of-sample AUC was estimated using five-fold cross-validation. For time-series data, the folds are approximately equal segments of time. The model is fit to the first fold and the posterior estimates are used to predict the states of the observations in the second fold, then the model is fit to the first and second fold and the posterior states are used to estimate the states from the third fold, and so on with increasingly large numbers of folds used for fitting a model to predict the states from the subsequent fold. With 63 time points, each of the five folds represents approximately 13 time points. Keep in mind, however, that each time point corresponds to many (100-1000) individual observations.

For each posterior draw of the cross-validation predictions, the optimal cutpoint was calculated as described above. Given these class assignments based on estimated optimal cutpoint, the AUC value of the out-of-sample estmates was calculated. AUC values are calculated for the entire fold and for each of the individual time points represented within that fold.