FishBase Nutrient Model Framework

AARON MACNEIL

June 18, 2021

1 Fishbase Nutrient Analysis Tool

Research has demonstrated the potential importance of fish as a critical source of micronutrients for many people, particularly within middle and low income countries (Vaitla et al. 2018, Hicks et al. 2019). Yet measured nutrient values are relatively scarce for fish, with typically few species represented from only a few countries. To overcome this data limitation, we developed a Bayesian hierarchical model that includes phylogenetic information (reflecting the interrelatedness of fish species) as well as trait-based information (reflecting key aspects of fish diet, thermal regime, and energetic demand) to predict concentrations of seven key nutrients (calcium, iron, omega-3, protein, selenium, vitamin A, and Zinc) for the world's marine and inland fish species.

It is important to recognize that the predictions generated by our statistical model represent a set of extreme out of sample predictions - using information from less than 10% of fish species to predict the nutrient content for the remaining 90% plus species. Yet these predictions also represent the best available information about what the nutrient content of the world's fishes might be. As such, this codebase is a work in progress that we expect to be constantly updated as new data or new covariate information becomes available. Recent fieldwork from our team sampling tropical fishes in Seychelles has shown reasonable out of sample predictive ability from our original model published in Hicks et al. 2019. However we also expect that the model will provide bad estimates for some species and for some locations.

Therefore we ask that you, dear user, let us know how the model is performing against your own observations, and we hope that you will be willing to contribute new data to this project using a clear sampling and analysis protocol, publishing your new data, and alerting our InFoods collaborators to the new data. We also welcome new model structure ideas to be shared with our model developer. These contributions will help improve the nutrient predictions available in FishBase and, ultimately, the quality of fish-derived nutrient data guiding food policy around the world.

2 Codebase

A GitHub Repo has been put together to host the codebase for the Nutrient Analysis Tool.

2.1 Repo Contents

The NutrientFishbase repo includes a few key files:

NutrientFishbase/model/FishBase_Nutrient_Models.py: Python code for estimating model parameters from observed nutrient data (from species in **NutrientFishbase/data/all_nutrients_active.csv** and traits from **NutrientFishbase/data/all_traits_active.csv**).

NutrientFishbase/model/FishBase_Nutrient_Predictions.py: Python code for using nutrient model posteriors to predict nutrient content for unobserved species (from **NutrientFishbase/data/all_traits_for_predictions.csv**), based on phylogeny and traits.

2.2 Repo Use

The models include several python package dependencies, including Pandas and PyMC3.

To run the models, simply download the **FishBase_Nutrient_Models.py** file and run it in python

python run FishBase_Nutrient_Models

which will grab the required files from GitHub and generate a range of plots and files for each nutrient. Generated plots for each nutrient (X) include:

X_LooPit.jpg: a three panel figure including a plot of the observed data (Yi) with their posterior predictive means, a plot of the leave-one-out probability integral transform (LOO-PIT) for the data against a uniform distribution, and a plot of the LOO-PIT expected cumulative density function (ECDF) and an expected uniform CDF, all of which look for ways in which the model is failing to fit the observed and expected data. An outline of these plots, and the source for their code, can be found here.

X_ObsPred.jpg: a two-panel plot of the within-model observed (red) vs predicted (blue) values and their 95% highest posterior density intervals, and a plot of the distribution of the within-model observed and predicted values. These provide some measure of model fit and show how the models fail, generally at the highest end, suggesting additional covariates are needed to predict

rare, high concentration nutrient values.

X__PriorPC.jpg: a plot of the prior predictive distribution and the observed data.

X__Trace.jpg: a large figure depicting both the posterior distribution and trace for each model parameter.

X__results.csv: a flat file containing the trace for each paremeter as an individual column.

X__Summary.csv: a flat file with summary statistics for each parameter including the posterior mean, standard deviation (sd), lower 94% highest posterior density interval (hdi_3%), upper posterior density interval (hdi_97%), mean Monte Carlo standard error (mcse_mean), standard deviation Monte Carlo standard error (mcse_sd), effective sample size mean (ess_mean), effective sample size standard deviation (ess_sd), effective sample size central tendency (ess_bulk), effective sample size distribution tail (ess_tail), and convergence ratio (r_hat).

To generate predictions, simply download the **FishBase__Nutrient__Predictions.py** file (after you've run the models) and run it in python, **python run FishBase__Nutrient__Predictions**, which will grab the **X__results.csv** files and will use covariates to generate posterior predictive values for all the species listed in **all__traits__for__predictions.csv** and generate two files:

Species__Nutrient__Predictions.csv: a flat file with summary statistics for each nutrient for each species, including the scientific name of the species (species), the FishBase species code (spec_code), a highest posterior predictive density value (X_mu), a lower 95% highest posterior predictive density interval (X_l95), a lower 50% highest posterior predictive density interval (X_l50), an upper 50% highest posterior predictive density interval (X_h95), and an upper 95% highest posterior predictive density interval (X_h95).

Species__Obs__predictions.jpg: a plot of the distribution of the predicted nutrients (histogram) against the range (dashed vertical lines) and median (solid vertical line) of the observed data.

2.3 Bayesian Model Covariates

Fish consume nutrients in relation to key aspects of their diet, energetic demand, and thermal regime, in ways that are reflected by their individual species traits. Recognizing this, our statistical models represent these dimensions using traits sourced directly from FishBase. Specifically these include:

1. *Feeding pathway (FP)*: indicates whether nutrients are sourced through a pelagic or benthic food web.

2. *Trophic level (TL)*: represents the number of feeding linkages between primary producers and a given species.

3. *Environment (EN)*: refers to the aquatic regime; one of marine, freshwater, brackish, or mixed (more than one environment).

4. *Water column (WC)*: refers to typical position in the water column; one of pelagic, demersal, reef-associated, bathypelagic, or benthopelagic, each of which has distinct pathways for nutrient input and cycling.

5. *Maximum length (Lmax)*: refers to how long a species is expected to grow, and scales directly with key attributes relating to home range size and metabolism.

6. *Age at maturity (Amat)*: reflects the time at which resources are allocated to reproduction.

7. *Body shape (BS)*: reflects how fish feed and move through their environment; one of flat, elongate (or eel-like), fusiform, or having short-deep bodies.

8. *Geographic zone (GZ)*: represents the thermal regime typical of each species; one of tropical, subtropical, temperate, and polar/deep.

While fish traits are directly linked to where and what fish eat, these characteristics are known to be correlated among related species, resulting in phylogenetically-predictable nutrient content (Vaitla et al. 2018). Therefore, we included phylogenetic relatedness within the correlation structure of our statistical model (see *Model structure* below), using a recently-developed phylogenetic tree for all marine fishes (Rabosky et al. 2013).

Lastly, samples of fish tissue in our nutrients database included nuisance parameters (things that influence sample collection but are not of direct interest), including the tissue type (muscle, whole, whole/parts, unknown; FO) and preparation (wet, dry, unknown; PR).

3 Bayesian Model Structure

The model that underlies our nutrient predictions is a modification of that presented in Hicks et al. 2019, where we removed a couple of covariates, depth and K (growth rate), that had the potential to induce spurious correlation in our posterior effect sizes, given their potential for collider bias in our asserted directed acyclic graph.

As an alternative to the GP phylogenetic covariance model (used in Vaitla et al. 2018 for example) we capitalized on the hierarchical nesting of phylogeny (sensu Thorson 2020), whereby species belong to a given genus, genera to specific families, and families to specific orders. This implies that species-level intercepts in the observed data come from a population related by genus group membership, genera represent samples from families, and families are samples from their parent orders, which can be represented in a hierarchical phylogenetic model that includes a global (overall) mean (0) at the top of a series of a non-centred, hierarchical relationships:

$$
\begin{aligned}
\gamma_0 &\sim N(0,1) \\
\sigma_{ord} &\sim Exp(1) \\
\beta_{oz} &\sim N(0,1) \\
\beta_{ord} =& \gamma_0 + \sigma_{ord}\beta_{oz} \\
\sigma_{fam} &\sim Exp(1) \\
\beta_{fz} &\sim N(0,1) \\
\beta_{fam} &= \beta_{ord} + \sigma_{fam}\beta_{fz} \\
\sigma_{gen} &\sim Exp(1) \\
\beta_{0,gz} &\sim N(0,1) \\
\beta_{0,gen} &= \beta_{0,fam} + \sigma_{gen}\beta_{0,gz} \\
\mu_i &= \beta_{0,gen} + \beta_x X \\
\beta_i &\sim N(\mu_i, \sigma)
\end{aligned}
$$

In both phylogenetic models the set of species level trait covariates was the same

$$\beta_x X = \beta_1 GZ + \beta_2 TL + \beta_4 FP + \beta_5 L_m ax + \beta_6 BS + \beta_8 A_m at + \beta_9 WC$$

Leading to an observation-scale model

$$\mu_{obs} = \mu_i + \gamma_1 FO + \gamma_2 PR$$

While Hicks et al. 2019 used a mix of Normal, Gamma, and Noncentral-t distributions for the data likelihood, we chose to model nutrients (except protein) on the log scale, and used either a Normal (selenium, omega-3)

$$\gamma_{obs} \sim N(\mu_{obs}, \sigma_{obs})$$

or Noncentral-t distribution (protein, zinc, calcium, iron, vitamin A)

$$\gamma_{obs} \; Nt(\mu_{obs}, \sigma_{obs,\tau})$$

Given regularizing priors

$$\begin{aligned} \beta_x, \gamma_x &\sim N(0, 1) \\ \sigma_{obs} &\sim Exp(1) \\ \tau &\sim U(0, 20) \end{aligned}$$

We ran the three models on each of the seven nutrients, using the Python package PyMC3. Models were run with four separately-initiated chains for 5,000 iterations using a No-U-Turn sampler (NUTS).