

Improving Catch Estimation Methods in Sparsely Sampled Mixed-Stock Fisheries.

Nick Grunloh, Edward Dick, Don Pearson, John Field, Marc Mangel

Abstract

Effective management of exploited fish populations, requires accurate estimates of commercial fisheries catches to inform monitoring and assessment efforts. In California, the high degree of heterogeneity in the species composition of many groundfish fisheries, particularly those targeting rockfish (genus *Sebastodes*), leads to challenges in sampling all potential strata, or species, adequately. Limited resources and increasingly complex stratification of the sampling system inevitably leads to gaps in sample data. In the presence of sampling gaps, ad-hoc species composition point estimation is currently obtained according to historically derived “data borrowing” (imputation) protocols which do not allow for uncertainty estimation or forecasting. In order to move from the current ad-hoc “data-borrowing” point estimators, we have constructed Bayesian hierarchical models to estimate species compositions, complete with accurate measures of uncertainty, as well as theoretically sound out-of-sample predictions. Furthermore, we introduce a computational method for discovering consistent “borrowing” strategies across over-stratified data. Our modeling approach, along with a computationally robust system of inference and model exploration, allows us to start to understand the effect of the highly stratified, and sparse, sampling system on the kinds of inference possible, while simultaneously making the most from the available data.

Significance

In order to understand how fish populations respond to fishing, it is critical to obtain accurate estimates of how many fish are removed from the ocean (catch) and to quantify the precision of those estimates. Traditionally, population dynamics models used to measure this response to fishing (“stock assessments”) are conditioned on a time series of annual catches. These catch estimates are often treated as being known without error, despite the fact that they are derived from sampling programs that estimate the proportion of unsampled strata must

be “filled in” through a process sometimes referred to on the U.S. West Coast as “borrowing” (i.e. data imputation). Historically, methods used to “borrow” information among strata have been different species found within multiple sampling strata. Sampling error introduces uncertainty into estimates of the catch, and ad-hoc in nature and driven by expert opinion of local managers (Sen et al. 1984, 1986) (Pearson and Erwin 1997). We seek to improve upon this practice through development of a model-based approach that provides estimates of catch and associated uncertainty, as well as an objective, defensible framework for model selection and data imputation. Although the theoretical basis for a model based estimation of species composition in mixed stock fisheries has been advanced (Shelton et al., 2012), it has not yet been implemented successfully using actual historical or contemporary data.

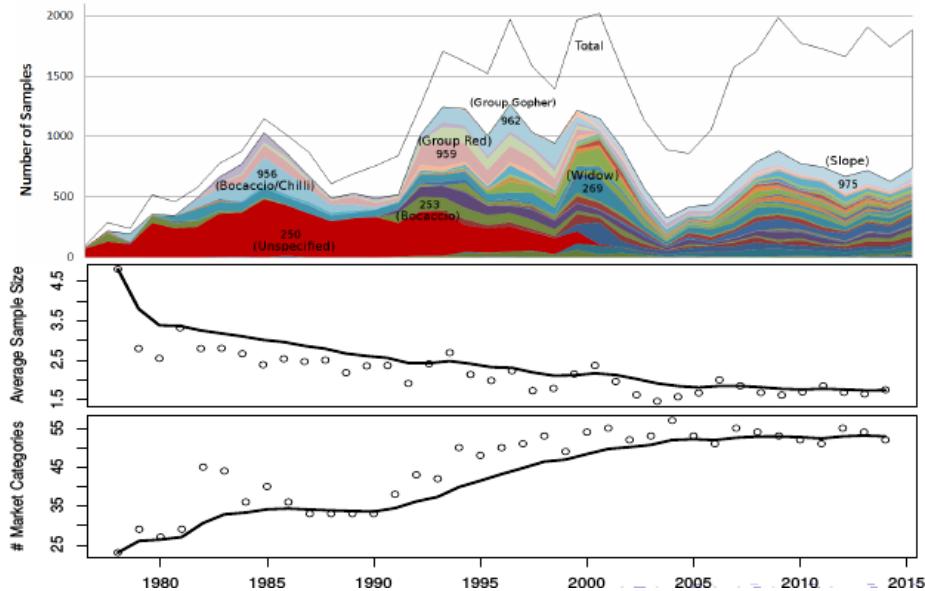


Figure 1: Spase Data

The difficulties associated with the existing ad-hoc approach are magnified by an increase in the number of sampling strata over time, specifically the number of “market categories,” into which fishermen and dealers sort their catch (Figure 1, Bottom). The increase in the number of market categories (sampling strata) has not been matched by increases in sampling effort, resulting in a decline in the average number of samples per stratum (Figure 1, Middle). In other words, data are becoming more sparse, increasing our uncertainty in estimates of catch. Since the data are also stratified over a number of ports, fishing gear types, years, and quarters, inference is not possible without some sort of stratum pooling. Rather than rely so heavily on the previous, ad-hoc pooling rules which change based on the availability of samples, we hope to standardize any necessary pooling through an exhaustive search of the space (possible configurations) of pooled

models. Pooling (and partial pooling) among strata is achieved using Bayesian hierarchical statistical models and model averaging (Gelman et al., 2014).

Methods

Data, Sampling, & Prep

Model

For a particular market category, $y_{ijklm\eta}$ is the i^{th} sample of the j^{th} species' weight, in the k^{th} port, caught with the l^{th} gear, in the η^{th} quarter, of year m . The $y_{ijklm\eta}$ are modeled as observations from a Beta-Binomial distribution (BB) conditional on parameters θ and ρ .

$$y_{ijklm\eta} \sim BB(y_{ijklm\eta} | \theta, \rho).$$

Given observed overdispersion relative to the Poisson and Binomial distributions, the Beta-Binomial model makes use of a correlation parameter, ρ , to better model uncertainties, while maintaining a flexible model on stratum means through the linear predictor. The linear predictor parameters, θ , are then factored as follows among the many strata,

$$\theta_{ijklm\eta} = \beta_0 + \beta_j^{(s)} + \beta_k^{(p)} + \beta_l^{(g)} + \beta_{m\eta}^{(y:q)}.$$

As a Bayesian model, we specify any information external to the dataset, through our priors on the parameters, $p(\theta)$. Our priors are largely diffuse normals, representing relatively little prior information, producing behavior similar to classical fixed effect models on species ($\beta_j^{(s)}$), port ($\beta_k^{(p)}$), and gear ($\beta_l^{(g)}$) parameters. Our priors on time parameters ($\beta_{m\eta}^{(y:q)}$) are modeled similarly to a classical random effects model, which uses the data to estimate a shared variance among all year-quarter interaction terms. Such a hierarchical prior thru time, imposes the prior information that data thru time share some degree of similarity, however the exact degree of similarity is not specified, rather the degree of similarity among time parameters is itself a parameter to be estimated from the data. In recent years, inference on these models has become faster and easier to compute through the use of computational Laplace approximations (Rue et al., 2009); we compute inferences on the above model in R (R Core Team, 2015) using the R-INLA package (Rue et al., 2013).

Species Composition

Applying the bayesian predictive framework to the above model gives the following expressions for predicted weight in each stratum,

$$p(y_{jklm\eta}^*|y) = \iint \text{BB}\left(y_{jklm\eta}^*|\theta_{jklm\eta}, \rho\right) P\left(\theta_{jklm\eta}, \rho|y\right) d\theta_{jklm\eta} d\rho.$$

$p(y_{jklm\eta}^*|y)$ is computed via monte carlo integration and represents the model's full predictive distribution for the j^{th} species' weight, in the k^{th} port, caught with the l^{th} gear, in the η^{th} quarter, of year m . Considering the same type of prediction across all species in a particular stratum allows for the calculation of predictive species compositions. The following joint transformation of the species' predictive weights result in predictive species compositions,

$$\pi_{jklm\eta}^* = \frac{y_{jklm\eta}^*}{\sum_j y_{jklm\eta}^*} \quad y_{jklm\eta}^* \neq 0.$$

Because the y^* are random variables, and π^* is nothing more than a transformation of the y^* , π^* is too a random variable. Furthermore once inference is complete, we can easily sample these distributions and compute any desired moments from these samples.

Model Exploration & Averaging

The straight-forward spatial model implied by the categorical port complex variables do not adequately resolve in-sample prediction at the observed sample sizes. Presently these deminishingly small within stratum sample sizes are managed by an ad-hoc "borrowing" protocol outlined by Pearson and Erwin (1997). We aim to formalize this "borrowing" idea via an exhaustive search of spatially pooled models, combined with the formalized process of Bayesian Model Averaging (BMA) to appropriately integrate port-complex pooling model uncertainty into species composition estimates (Hoeting et al. , 1999).

The space of possible pooled models is well defined in terms of the size of the set of items to be partitioned, K , as described by the Bell numbers (B_K),

$$B_K = \sum_{\hat{k}=0}^K \frac{1}{\hat{k}!} \left(\sum_{j=0}^{\hat{k}} (-1)^{\hat{k}-j} \binom{\hat{k}}{j} j^K \right).$$

The most straight-forward solution in the presence of this type of model uncertainty is to compute all B_K possible pooling schemes. However, maybe not all pooling schemes represent biologically relevant models. For example, perhaps it is reasonable to pool only among adjacent ports, or to assert that biologically similar regions can only extend across a small number of ports (if so, how many?).

Each of these hypotheses are easily represented as subsets of the total model space, B_K , as seen in Figure (2). An exhaustive search of the models in these subspaces, and a comparison of the relative predictive accuracy of each model, provides concrete quantitative support for, or against, each of these hypotheses. Through this technique of exhaustive search and measuring relative predictive accuracy, we are able to understand the system to a greater degree than before possible. Furthermore such an exhaustive search of these model spaces allows for even more accurate estimates of species composition, and uncertainty, through the use of Bayesian Model Averaging (BMA) among the candidate models. Bayesian model averaging allows us to account for model uncertainty around these difficult modeling decisions, while combining the respective predictive capabilities of each model of a given subset of model space (Hoeting et al., 1999). Once all of the models of a given model space are computed, combining them to account for model uncertainty, through BMA, requires trivial computation time, but adds substantial robustness to our predictions.

BMA

Bayesian model averaging is straight forward. For the μ^{th} model, of model space \mathbb{M} , a straight forward implementation of Bayes theorem gives,

$$Pr(\mathbb{M}_\mu|y) = \frac{p(y|\mathbb{M}_\mu)p(\mathbb{M}_\mu)}{\sum_\mu p(y|\mathbb{M}_\mu)p(\mathbb{M}_\mu)} = \omega_\mu$$

Where ω_μ is the posterior probability that model μ is the true data generating model of the data, conditional on the subspace of candidate models and the observed data. ω_μ is then straightforwardly used to average together the posteriors of all of the candidate models, as follows $\bar{p}(\theta|y) = \sum_\mu \omega_\mu p(\theta|y, \mathbb{M}_\mu)$.

References

- [1] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.
- [2] Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382-401.
- [3] Pearson, D.E., and Erwin, B. (1997). Documentation of California's commercial market sampling data entry and expansion programs. NOAA Tech Memo. NOAA-TM-NMFS-SWFSC-240.
- [4] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- [5] Rue H., Martino S., Lindgren F., Simpson D., Riebler A. (2013). R-INLA: Approximate Bayesian Inference using Integrated Nested Laplace Approximations. Trondheim, Norway. URL <http://www.r-inla.org/>.

- [6] Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2), 319-392.
- [7] Sen, A.R. (1984). Sampling commercial rockfish landings in California. NOAA Tech Memo. NOAA-TM-NMFS-SWFSC-45.
- [8] Sen AR. (1986). Methodological problems in sampling commercial rockfish landings. *Fish Bull.* 84: 409-421 .
- [9] Shelton, A. O., Dick, E. J., Pearson, D. E., Ralston, S., & Mangel, M. (2012). Estimating species composition and quantifying uncertainty in multispecies fisheries: hierarchical Bayesian models for stratified sampling protocols with missing data. *Canadian Journal of Fisheries and Aquatic Sciences*, 69(2), 231-246.