

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

Nick Grunloh, E.J. Dick, Don Pearson, John Field, Marc Mangel

## Abstract

## Introduction

Estimates of landed catch are a key component of many fishery management systems. Stock assessment models (referred to here as assessments) are often conditioned on time series of annual catch, usually under the assumption that catches are known without error. While some assessment models are able to incorporate uncertainty in catch (e.g. Stock Synthesis; Methot and Wetzel, 2013), reliable estimates of catch uncertainty are often unavailable. Without this information, assessment authors often rely on ad-hoc sensitivity analyses which may or may not be incorporated into management advice and/or fail to propagate catch uncertainty into quantities of interest to managers.

Over the past decade, the estimation of catch and associated uncertainty has become a focus for recreational fisheries in the United States (NAS, 2017). Commercial fisheries, on the other hand, are often assumed to have precise estimates of catch by species. This is due in part to the availability of landing receipts (aka fish tickets) which serve as a record of the weight of fish landed into various market categories (sort groups). As noted by Pearson et al. (2008), it is important to recognize that species and market categories are not synonymous. On the U.S. West Coast, for example, it is common for multiple species to be landed within a single market category (CALCOM 2017, PacFIN 2017). This is expected for categories that are clearly designated as mixed-species categories (e.g. “nearshore rockfish”, or species within a particular genus or family). However, in some states, categories that are named after a single species still contain several species, to varying degrees, even after regulations require sorting into a particular category (Pearson et al., 2008). As a result, estimation of landings for a single species based on landing receipt data alone may produce biased estimates of total catch.

Fisherman, dealers, or processors typically decide how to sort species into market categories on a landing receipt. Trained port samplers intercept vessels

offloading catch or during subsequent processing in order to determine the species composition of catch landed in a given market category (Sen 1984, Crone 1995, Tsou et al. 2015). These species composition data are used to partition, or distribute, the weight of landed catch in a market category across species, a process commonly referred to as catch expansion (Pearson and Almany, 1995). To estimate total landings for a single species, the expanded catch is summed across all relevant market categories. Assuming that landing receipt data are a census of the landed catch, uncertainty in total catch for a given species reflects variability in the species composition among port samples. If information is available to estimate sources of bias in the landing receipt data, e.g. underreporting, then it is possible to also incorporate uncertainty in the bias correction factors as well (Bousquet et al. 2010). In this study, we quantify sampling uncertainty for estimates of landed catch, although the modeling framework could be extended to include uncertainty or bias in landing receipt data.

Within market categories, the species composition of landed catch can vary spatially, temporally, by fishing gear, and catch disposition (e.g. fish sold alive or dead). These differences are attributable to many factors, including market preference, fishing behavior, regulatory constraints, and biological/ecological characteristics (e.g. spatial distribution) of the landed species. As a result, estimates of species composition for a given market category are often stratified over time (e.g. quarterly) and across other relevant strata (e.g. ports, gears, catch disposition). Sampling programs often have limited funds, and attempts to reduce bias in species composition estimates through the introduction of additional strata comes at a cost, namely reduced precision (Cochran, 1977; Tomlinson, 1971).

On the U.S. West Coast, port sampling programs allocate effort both spatially and temporally, but many domains of interest (e.g. market category, gear type, catch disposition) remain unsampled or sparsely sampled due to a proliferation of categories over time, logistical constraints, and limited resources (Sen 1986; Crone 1995; Pearson et al. 2008; Tsou et al. 2015). In California, for example, commercial port sampling effort has changed over time and space (Pearson and Almany 1995). For example, regular sampling of California ports north of Point Conception (roughly 34° 27' N. latitude) began as early as 1978, but the more southern ports were rarely sampled prior to 1983. This allocation of effort was largely based on the statewide distribution of landings, diffuse spatial distribution of southern commercial ports, and limitations in funding for port samplers.

When no port samples are collected for landed strata and domains, species composition estimates are ‘borrowed’ from other strata using deterministic algorithms based on expert opinion. These algorithms have unknown bias and precision. In contrast, model-based estimators are increasingly used to estimate quantities of interest for domains with small sample sizes and/or unsampled strata (sometimes referred to as small area estimation; Rao 2003).

Shelton et al. (2012) developed a Bayesian hierarchical statistical framework for

species composition data that pools information among sparsely sampled strata, predicts species compositions for unsampled strata, and can be combined with landing receipts to estimate total landings by species, across market categories and other strata, with associated estimates of uncertainty. Shelton et al. considered hierarchical pooling only among quarters within a single year, comparing generalized linear and hierarchical linear models to trawl data from a single port in California. The authors underscored the need to better understand performance of alternative models, and to overcome issues with computation time, particularly since commercial port sampling data sets often include hundreds of landed strata spanning decades, multiple ports, gear types, and other domains of interest.

Among the U.S. West Coast states, the challenge of estimating landings for sparsely-sampled mixed stock fisheries is perhaps greatest for California. Relative to Oregon and Washington, California has a greater number of commercial ports, market categories, and landed species, with greater species diversity among ports due to the geographical range of the coast. Of particular consequence to the estimation of species compositions is the proliferation of landed market categories over time, particularly during the 1990s (Figure 1). Sampling effort also leveled off in the mid-1990s, with a reduction in effort in the early 2000s. The net result of increased stratification and flat (or reduced) sampling effort over time is a decline in mean sample size per stratum (Figure 1). In this situation, it is critical to understand how efforts to reduce bias (e.g. increasing the number of landed market categories) affect precision of the expanded catch estimates.

Models that take catch uncertainty into account are not new (c.f. Doubleday 1976), but most assessments on the U.S. West Coast assume catch is known without error (PFMC 2018). As a result, catch uncertainty is not propagated into management reference points. However, the implications of catch uncertainty are not limited to stock assessment efforts. In a management context, catch estimates with large (but unknown) uncertainty may cause managers to react to large, high-frequency deviations in estimated catch, and either impose unnecessary restrictions on a fishery, or mistakenly support excessive harvest. This is particularly an issue for prohibited and/or ‘choke’ species, for which information is limited and may be based solely on estimates of discarded catch.

In this study, we evaluate the model-based framework proposed by Shelton et al. (2012) using commercial port sampling data collected in California, U.S.A. We describe species composition data collected by the California Cooperative Groundfish Survey (CCGS, 2017) over the period 1978-1990. We then extend the Shelton et al. framework to address limitations of their approach. Specifically, we evaluate alternative likelihoods to address overdispersion, compare multiple hierarchical structures for pooling information through time, and integrate model predictions across uncertainties in the spatial model structure. Finally, we estimate landed catch by species for both sampled and unsampled strata, and summarize a general framework for quantifying uncertainty including an efficient database design for dissemination of results at any level of aggregation.

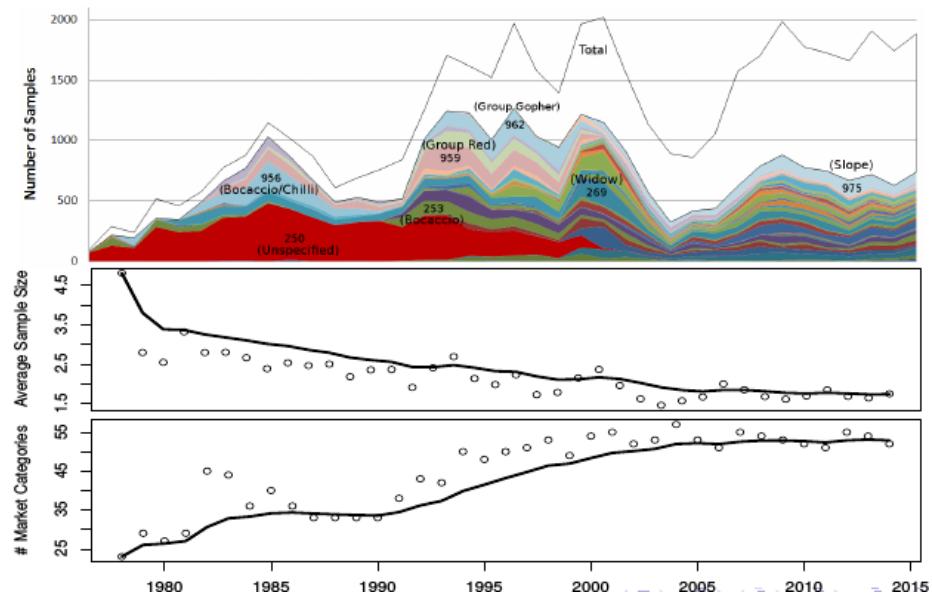


Figure 1: Number of commercial port samples per market category in California, 1978-2014 (upper panel), average sample size per stratum (middle panel), and number of market categories recorded on landing receipts (lower panel). On the lower panels, points indicate observed values, while the black lines represent 9 year moving averages

## Methods

### Data

As outlined in (*cite*) the data in *blank* are the result of a cluster sampling protocol across the many strata, in which each sample is intended to be composed of two fifty pound clusters from the selected stratum. Although port samplers do their best to follow protocol, in reality the port sampling environment does not always allow the exact (*cite*) protocol to be followed. Variations in the sampling protocol may result in only a single cluster being taken clusters taken, or the size of clusters taken to vary, from stratum to stratum based on the particular challenges of sampling each stratum.

The model based methodology proposed here does not rely strongly upon the cluster sampling structure, but rather simply views each sample as independent and identically distributed (*i.i.d.*) draws from some data generating model, conditional on some parameterization of the stratification system. So long as the parameterization and data generating model are sufficiently robust for handling the behavior of these data, the conditionally *i.i.d.* model of these data can be seen as practically useful model for producing predictions about the data generating system.

That said, for the purpose of modeling these data, it is enough to know which clusters were collected as part of which samples, and how big each cluster actually ended up being. This information is readily available from CALCOM (the database maintained by the California Cooperative Groundfish Survey). For the purpose of the models presented here, all of the observed clusters across a unique sample are aggregated so that the total weight sampled is the sum of all of the total pounds sampled in each cluster of each unique sample. Similarly the observed weight for a particular species in each unique sample is the sum of all of the observed weights, for that species, across each cluster of each unique sample.

Although model based data analysis has the potential to add significant structure to data, an ethical application of these methods must always confront the model with enough empirical information to adequately learn about the system. In this setting some market categories and time periods may not be well enough sampled to learn the parameters of the models presented here. For any modeled period where the minimum number of possible parameters exceeds the number of samples for the modeled period, the model is not applied. Rather than apply models inappropriately, these cases are speciated as the nominal species for their market category. We later demonstrate in Figure (*bars*) that due to the prioritization for sampling heavily landed, or otherwise commercially relevant categories, a sample size heuristic such as this, typically only leads to nominal speciation in market categories and time periods where total landings are low. Thus the expanded landings in these periods have a negligible effect on the

overall expanded landings for species in these categories.

- port complexes v. ports? or in intro?
- something about chunking time at regulation time periods
  - introduce idea of modeled period (mcat/time chunk)

## Data Generating Model

For the purposes of accurately modeling not only species composition means, but also higher moments of the data, such as species composition variances, it is necessary to recognize model limitations with respect to over-dispersed data. Among the simplest models for count data are the Poisson and binomial models. Both models are typically specified with a single parameter for modeling all of the moments of the data, and thus they rely heavily on their respective data generating processes to accurately represent higher moments in the data. McCullagh and Nelder (1989, pg. 124) commiserate about the prevalence of over-dispersed data in cluster sampling, and explain ways in which cluster sampling itself may result in over-dispersion.

Extending the Poisson and binomial models to deal with over-dispersion, typically involves adding additional parameters for the purpose of modeling higher moments of the data. The negative binomial (NB) distribution is often used as an over-dispersed extension of the Poisson model, since it can be expressively written as an infinite mixture of Poisson distributions. While the beta-binomial model is typically used to as an over-dispersed extension of the binomial model.

## An Example

To discern between these discrete modeling options we consider a small scale example of the Poisson, binomial, negative binomial, and beta-binomial models fit to the port sampling integer weight data from market category 250, in the Monterey port complex trawl fishery in 1982. (*anywillwork*) This stratum was visited 32 times by port samplers, collecting a total of 59 cluster samples across 55 unique species. Each of the above models are fit to these data.

The Poisson and binomial models attempt to model both the mean and residual variance of the data with a single parameter for each species. By definition these models have residual variances which are tied to the species means. Simply estimating the mean parameters in these cases may not be sufficient to produce models which predict well.

In contrast, the negative binomial and beta-binomial models estimate an additional parameter which can be used to disentangle the mean and residual variance estimates. Thus the negative binomial and beta-binomial models may produce more accurate estimates of the residual variance. Furthermore, in better

modeling the variance, these models may often produce more accurate measures of center as well.

For each of the above mentioned models the predictive species composition distributions are visualized in Figure(*IntervalPlot*) as 95% Highest Density Intervals (HDI) (*citations*), plotted on top of the predictive means for each model and the observed species compositions from the data in Figure(*IntervalPlot*). For brevity we only consider the most prevalent six species in this example (CLPR, BCAC, WDOW, BLGL, ARRA, BANK). Additionally, the MSE, DIC, WAIC, and Bayesian marginal likelihood model probabilities are computed for each model as measures of model fit as seen in Table(*likelihoodtable*).

Table(*likelihoodtable*) show a clear preference for the overdispersed models, with the most overall support for the beta-binomial model. This initial result guides the use of the beta-binomial data generating model for the purposes of building a model to apply at an operational scale.

### Operationalized 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  $\eta^{th}$  quarter, of year  $m$ . As supported by the preliminary results in Figure(*IntervalPlot*) and Table(*likelihoodtable*), the  $y_{ijklm\eta}$  are modeled as observations from a beta-binomial distribution conditional on parameters  $\mu_{jklm\eta}$  and  $\sigma_{jklm\eta}^2$ ,

$$y_{ijklm\eta} \sim BB(\mu_{jklm\eta}, \sigma_{jklm\eta}^2).$$

Where  $\mu_{jklm\eta}$  is the stratum level beta-binomial mean weight and  $\sigma_{jklm\eta}^2$  is the stratum level residual variance.  $\mu_{jklm\eta}$  is related to a linear predictor,  $\theta_{jklm\eta}$ , via the mean function,

$$\mu_{jklm\eta} = n_{ijklm\eta} \frac{\exp(\theta_{jklm\eta})}{1 + \exp(\theta_{jklm\eta})} = n \text{ expit}(\theta_{jklm\eta}) = n \text{ logit}^{-1}(\theta_{jklm\eta}).$$

Here  $n_{ijklm\eta}$  is the observed aggregate cluster size for each sample. Additionally,  $\sigma_{jklm\eta}^2$  is related to  $\mu_{jklm\eta}$  and the overdispersion parameter,  $\rho$ , via the following equation,

$$\sigma_{jklm\eta}^2 = \mu_{jklm\eta} \left(1 - \frac{\mu_{jklm\eta}}{n_{ijklm\eta}}\right) \left(1 + (n_{ijklm\eta} - 1) \rho\right).$$

$\rho$  is the within cluster correlation. The situation where  $\rho \rightarrow 1$  represents identical information content among replicates within a cluster, with maximal overdispersion relative to the binomial distribution. The situation where  $\rho \rightarrow 0$

represents totally independent information content among replicates within a cluster, and the beta-binomial model approaches the binomial model.  $\rho$  explicitly models average overdispersion across all stratum, while  $\mu_{jklmn}$  gives the model flexibility at the stratum level through the  $\theta$  linear predictors,

$$\theta_{jklmn} = \beta_0 + \beta_j^{(s)} + \beta_k^{(p)} + \beta_l^{(g)} + \beta_{m\eta}^{(t)}.$$

Firstly,  $\theta$  includes a reference level intercept ( $\beta_0$ ). Secondly,  $\theta$  is factored among the many strata by additive offsets from  $\beta_0$  for each of the species ( $\beta_j^{(s)}$ ), port-complex ( $\beta_k^{(p)}$ ), and gear-group ( $\beta_l^{(g)}$ ) categories. Finally year and quarter parameters are indicated generally here inside the  $\beta_{m\eta}^{(t)}$  term. Several forms for  $\beta_{m\eta}^{(t)}$  are explored each implying a different prior and partial pooling strategies as described in the following section(*Priors*).

## Priors

To complete the Bayesian formulation of this model, priors are expressed in a largely diffuse manner.

$$\begin{aligned}\beta_0 &\propto 1 \\ \{\beta_j^{(s)}, \beta_k^{(p)}, \beta_l^{(g)}\} &\sim N(0, 32^2)\end{aligned}$$

Since the  $\beta_0$  reference level is chosen arbitrarily, with no conception of which values it may take, no restrictions are placed on the value of the intercept. The species ( $\beta_j^{(s)}$ ), port-complex ( $\beta_k^{(p)}$ ), and gear-group ( $\beta_l^{(g)}$ ) offsets are assigned diffuse normal priors. The large fixed values of the prior variance hyperparameters produce behavior similar to classical fixed effect models for species, port-complex, and gear- group parameters.

In returning to the time parameter model,  $\beta_{m\eta}^{(t)}$ , it is useful to consider how the inclusion of predictively superfluous parameters may cause overfitting and weaken model performance through the bias-variance dilemma (Ramasubramanian, K., & Singh, A., 2016). Simply put, the bias-variance dilemma means that model formulation is not simply a bias reduction task, but rather the goal is to formulate models which reduce bias, while jointly minimizing uncertainty. Janyes (2003, pg. 511) describes how the inclusion of estimation bias via the Bayesian methodology may produce better performing estimates, more quickly, than unbiased counterparts. Among the simplest ways to see the principle is in the structure of the MSE performance metric, and how it can be explicitly written to value both estimator bias and variance, as follows.

$$\text{MSE}(\hat{\theta}) = \mathbb{E} \left[ (\hat{\theta} - \theta)^2 \right] = \overbrace{\mathbb{E} \left[ (\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 \right]}^{\text{Var}(\hat{\theta})} + \overbrace{\left( \mathbb{E}(\hat{\theta}) - \theta \right)^2}^{\text{Bias}(\hat{\theta}, \theta)^2}$$

Furthermore a model can minimize bias, without regard for estimation uncertainty, by including one model parameter to be fit to each observation. These parameter estimates are totally unbiased, however such a model is also predictively useless since each estimated parameter is specifically bound to a particular observation, and thus such a model does not generalize.

For modeling  $\beta_{m\eta}^{(t)}$  we consider a spectrum of models which span a wide range of partially pooled models with several different predictive structures as seen below.

(M1)

$$\begin{aligned}\beta_{m\eta}^{(t)} &= \beta_m^{(y)} + \beta_\eta^{(q)} \\ \beta_m^{(y)} &\sim N(0, 32^2) \\ \beta_\eta^{(q)} &\sim N(0, 32^2)\end{aligned}$$

(M1) represents a fixed effects model for additive year and quarter parameters. Here each year and quarter receive totally independent and diffuse priors.

(M2)

$$\begin{aligned}\beta_{m\eta}^{(t)} &= \beta_m^{(y)} + \beta_\eta^{(q)} \\ \beta_m^{(y)} &\sim N(0, v^{(y)}) \\ \beta_\eta^{(q)} &\sim N(0, v^{(q)})\end{aligned}$$

(M2) estimates two hierarchical variance parameters,  $v^{(y)}$  and  $v^{(q)}$ .  $v^{(y)}$  has the effect of partially pooling information among year parameters, while  $v^{(q)}$  partially pools information among quarter parameters. The actual degree of pooling among each of the years and quarters is determined by the data. Depending on the posterior distributions of  $v^{(y)}$  and  $v^{(q)}$ , the  $\beta_m^{(y)}$  and  $\beta_\eta^{(q)}$  may be shrunk back toward the common mean (for small  $v$ ) or allowed to take largely distinct values (in the case of large estimates of the  $v$ ).

(M3)

$$\begin{aligned}\beta_{m\eta}^{(t)} &= \beta_m^{(y)} + \beta_\eta^{(q)} + \beta_{m\eta}^{(y:q)} \\ \beta_m^{(y)} &\sim N(0, v^{(y)}) \\ \beta_\eta^{(q)} &\sim N(0, v^{(q)}) \\ \beta_{m\eta}^{(y:q)} &\sim N(0, v)\end{aligned}$$

(M3) functions similarly as (M2), in that it has hierarchical partial pooling among both the  $\beta_m^{(y)}$  and  $\beta_\eta^{(q)}$  parameters, except that it introduces a two-way interaction term between year and quarter. This interaction term allows estimates for particular quarters to differ from year to year, as opposed to the previous models in which quarters within a year are assumed to be identical from year to year.

Furthermore the  $\beta_{m\eta}^{(y:q)}$  are also modeled hierarchically to introduce a single variance parameter,  $v$ , shared among all of the  $m\eta$  time chunks. Although this interaction term adds many parameters to the model, the shared  $v$  parameter functions to shrink extraneous  $\beta_{m\eta}^{(y:q)}$  estimates back toward the common stratum mean.

(M4)

$$\begin{aligned}\beta_{m\eta}^{(t)} &= \beta_{m\eta}^{(y:q)} \\ \beta_{m\eta}^{(y:q)} &\sim N(0, v)\end{aligned}$$

(M4) simplifies (M3) by excluding year and quarter main effects. This leaves all temporal information in the data to be modeled solely by the quarterly  $\beta_{m\eta}^{(y:q)}$  interaction terms. This model represents more opportunity for partial pooling through time than (M3), as fewer time parameters are introduced. Furthermore all of the  $\beta_{m\eta}^{(y:q)}$  are hierarchically pooled back toward a single common stratum mean via the single shared variance parameter,  $v$ .

(M5)

$$\begin{aligned}\beta_{m\eta}^{(t)} &= \beta_{m\eta}^{(y:q)} \\ \beta_{m\eta}^{(y:q)} &\sim N(0, v_\eta)\end{aligned}$$

(M5) is largely the same as (M4), but it represents slightly less potential partial pooling through its hierarchical prior variances,  $v_\eta$ , on  $\beta_{m\eta}^{(y:q)}$ . Here interaction terms are allowed to partially pool interactions across years, within a common quarter, but since each quarter is assigned a separate variance parameter no pooling is possible between quarters.

(M6)

$$\begin{aligned}\beta_{m\eta}^{(t)} &= \beta_{m\eta}^{(y:q)} \\ \beta_{m\eta}^{(y:q)} &\sim N(0, v_m)\end{aligned}$$

(M6) follows the same idea as (M5), however here interaction terms are allowed to partially pool interactions within a common year, across the quarters of that year, but not between years. (M6) often involves fitting slightly more parameters than (M5) because, at least in this setting, it is typical to model more than four years of data at once.

Hierarchical variance parameters are estimated from the data. As the above models learn the posteriors of the hierarchical variance parameters, it affects the degree of shrinkage as well as the effective number of parameters held within the respective hierarchies (Gelman, 2014). To achieve this, each variance parameter must itself be assigned a prior to be estimated. For all of the hierarchical variance parameters included in the above models  $v$  is assigned a diffuse  $v \sim IG(1, 2 \times 10^3)$  prior.

Finally the overdispersion parameter,  $\rho$ , is assigned a diffuse normal prior on the logit scale,  $\text{logit}(\rho) \sim N(0, 2^2)$ . The  $N(0, 2^2)$  prior is indeed a symmetric, and far reaching, prior when back transformed to the unit interval. To notice this, it is helpful to realize that the central 95% interval for a  $N(0, 2^2)$  (i.e.  $0 \pm 3.91$ ), includes almost the entirety of the back transformed unit interval (i.e.  $0.5 \pm 0.48$ ).

Table(*priortable*) shows a clear preference for model (M4). It is worth mentioning that among all of the models considered here, (M4) offers the largest potential for hierarchical partial pooling among the time parameters. For the duration of the methods we build upon this preliminary result.

## Species Composition Prediction

Estimating model (M4) in a Bayesian way gives access to the full posterior distribution of all of the parameters of the model. It is useful to emphasize that in the Bayesian setting, these parameters have full distributions, and they are typically handled as a large number of samples from the joint posterior distribution of the parameters. Once the posterior sampling is complete, this simplifies parameter mean and variance estimation; required moments are simply obtained by computing the desired moments from the posterior samples. Additionally, the fact that the parameters are full distributions means that any functions which contain, or are derived from, parameters are themselves random variables with the function representing a random variable transformation of those parameters.

To obtain predicted species compositions from this model, first consider the posterior predictive distribution of sampled weight for a particular stratum.

$$p(y_{jklm\eta}^*|y) = \iint \text{BB}\left(y_{jklm\eta}^*|\mu_{jklm\eta}, \sigma_{jklm\eta}^2\right) P\left(\mu_{jklm\eta}, \sigma_{jklm\eta}^2|y\right) d\mu_{jklm\eta} d\sigma_{jklm\eta}^2.$$

Here BB is the data generating beta-binomial distribution for a predictive observation and  $P(\mu_{jklm\eta}, \sigma_{jklm\eta}^2|y)$  is the posterior distribution of the parameters given the observed data. Integration of the parameters,  $\mu_{jklm\eta}$  and  $\sigma_{jklm\eta}^2$ , is done by Monte Carlo integration to obtain samples from the predictive distribution,  $p(y_{jklm\eta}^*|y)$ , for sampled weights in the  $jklm\eta^{th}$  stratum.

Obtaining predictive species compositions from predictive weights amounts to computing the following transformation,

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

For a particular market category,  $\pi_{jklm\eta}^*$  is the models representation of the observation level species composition for species  $j$  in the  $k^{th}$  port, caught with the  $l^{th}$  gear, in the  $\eta^{th}$  quarter, of year  $m$ .

## Model Exploration & Averaging

Presently, stratum with diminishingly small sample sizes are managed by an ad-hoc “data borrowing” protocol, as outlined in Pearson and Erwin (1997). The protocol for “data borrowing” calls for pooling only when forced to fill holes brought about by unsampled strata. Naturally, such a pooling protocol introduces bias to fill in unsampled strata, however due to the mathematically unstructured way in which this bias is introduced, it is hard to quantitatively justify these “data borrowing” rules.

Model (M4) avoids temporal ad-hoc “borrowing” protocols used in Pearson and Erwin (1997) by making use of its hierarchical structure to fill temporal holes with a posterior predictive distribution for unseen time periods within the modeled period. This hierarchical structure uses the data to estimate the degree of pooling through time, rather than ad-hoc “data borrowing”.

Despite the benefits of modeling these data as Bayesian hierarchical models, port sampling data still remains sparse. Given the degree of sparsity in these data it is certainly possible that models which consider an additional degree of data pooling between port complexes may offer predictive benefits. In exploring strategies for pooling data across space it is necessary to formalize the port complex pooling scheme in a way which provides a mathematically understandable and scalable structure to build upon.

Given the categorical nature of port complex variables, the typical hierarchical regularization priors among port complexes are not appropriate. Rather, we

frame port complex pooling as a model uncertainty problem, in which we consider some degree of port complex pooling, but the exact degree of pooling, and the particular partitioning of the pooled port complexes are not known.

Port complex pooling is achieved by repeatedly fitting model (M4) with different partitionings of the port complex variables within a particular market category and modeling time period. This model exploration exercise explores the possible ways to produce groupings of the existing port complexes so as to discover predictively useful partitionings of the port complexes. Insisting that the port complex groupings be partitions of the available port complexes provides a well defined mathematical structure for exploring the space of pooled port complexes.

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

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

In the case of California the set of items to be partitioned is the set of port complexes in California, of which there are  $K = 10$ , implying a grand total of  $B_{10} = 115975$  ways of partitioning the port complexes in California in each market category and modeled time period. The brute force model selection strategy of computing all 115975 of these partitionings strategies is computationally infeasible. However, not all pooling schemes represent biologically relevant models. For example, perhaps it is reasonable to pool only among adjacent ports (i.e. no discontinuities between port complex pooling in space), or perhaps it is reasonable to assert that biologically similar regions can only extend across a small number of ports.

Here only adjacent port poolings are considered, such that the maximum size of a port complex grouping is three port complexes. These are the only two constraints that are enforced on port complex partitions here, although many other constraints may, in theory, be chosen. These constraints were chosen so as to mirror the currently accepted protocols in Pearson and Erwin (1997) within the context of this framework. When these two simple constraints are applied, the number of models to explore in each modeled period is reduced to a much more manageable 274 models.

An exhaustive search of the models in the biologically constrained subspace of  $B_{10}$ , allows for a concrete comparison of the relative predictive accuracy of each partitioning. Additionally the partitioned models provide a set of candidate models for use in Bayesian Model Averaging (BMA) (Hoeting et al., 1999). BMA, as applied here, essentially allows the model exploration strategy to average across all relevant port complex partitionings and adds robustness to the final species composition estimates.

For the  $\iota^{th}$  model in a set of candidate models  $M$ , then the BMA weight for  $M_\iota$  follows directly from Bayes Theorem as,

$$\omega_\iota = \Pr(M_\iota|y) = \frac{p(y|M_\iota)p(M_\iota)}{\sum_\iota p(y|M_\iota)p(M_\iota)}.$$

Where  $\omega_\iota$  is the posterior probability that model  $\iota$  is the true data generating model of the data, conditional on the subspace of candidate models and the observed data.  $\omega_\iota$  is then straightforwardly used to average posteriors across all of the models, as

$$\bar{p}(\theta|y) = \sum_\iota \omega_\iota p(\theta|y, M_\iota).$$

## Expansion

$$\lambda_{jklm\eta}^* = \lambda_{klm\eta} \pi_{jklm\eta}^*$$

## Results

### Data

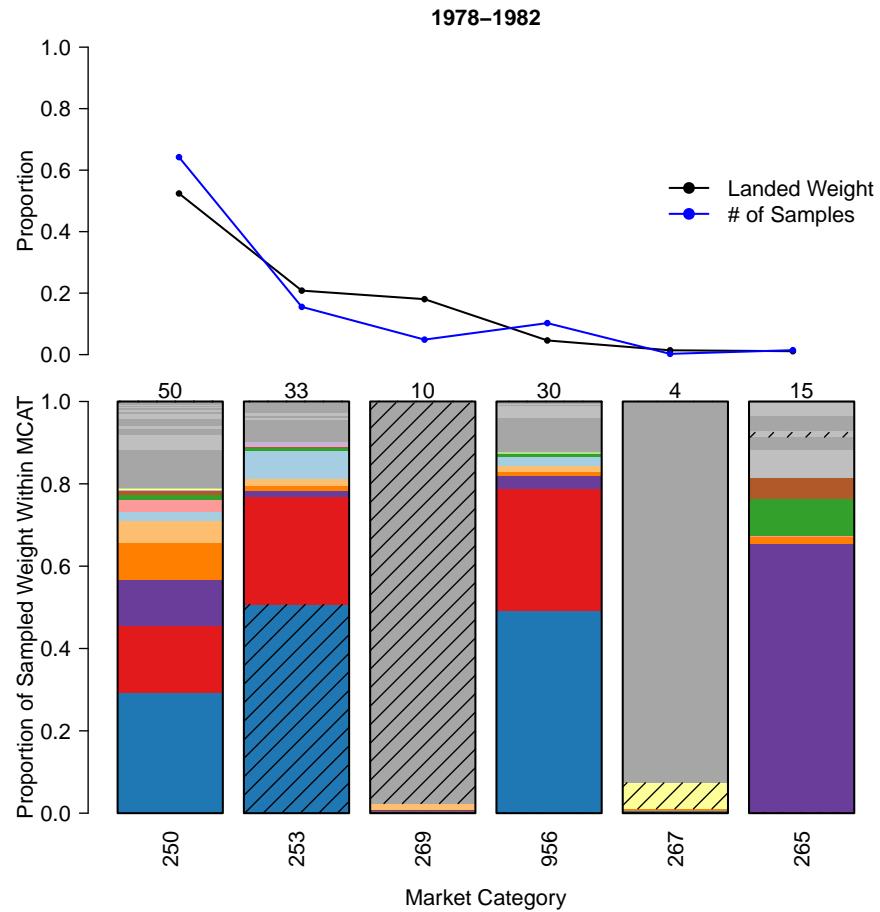
Figure (*bars*) shows the relative sampling effort and landings across rockfish market categories accounting for the top 99% percent of landings for two modeled periods. In the top panel of each modeled period the black line represented the proportion of landings in each market category, while the blue line represents the proportion of samples, by number, taken in each market category. In the bottom panel of each modeled period, colors represent the proportion of sampled weight of twelve commercially relevant species. Alternating dark and light grey panels fill in all other species. On the top of each bar the number of observed species in each market category is listed. Hatched regions indicated the nominal species of a market category where applicable.

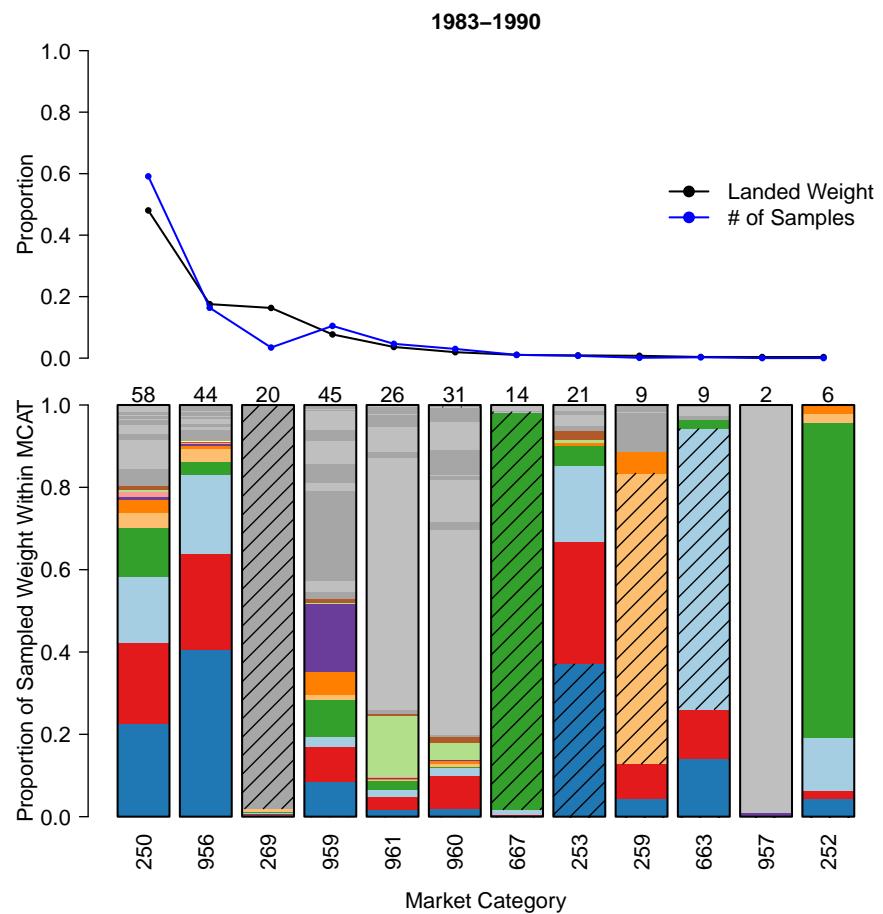
Notice that sampling effort both tracks landings as well as the number of species in each market category. These two factors are among the most important factors to prioritize in sampling for the sake of applying our model's results to a large proportion of landings (*What proportion of landings is the model expanding?*).

Furthermore due to this pattern in port sampling, when sample sizes become too small to apply our model, the expanded landings tend to represent a negligible proportion of total landings.

The lower panels of Figure (*bar*) visually demonstrates just how many different species are landed into commercially relevant market categories. Although market categories often carry names that label them with a nominal species, Figure

(bar) makes it abundantly clear that these names can mislead our thinking about the purity, and consistency, of these categories through time. To drive this point home, consider market category 267. The nominal label for market category 267 is Brown, while Brown rockfish only amounts to 6.3% of the sampled weight in 1978-1982. In 1978-1982 market category 267 might be better named Widow as Widow rockfish amounts to 92.6% of sampled weight in this time period, however considering the most recent time period (2000-2015), in Appendix Figure (panel), market category 267 samples are composed of 99.6% Brown and 0% Widow.





	BANK
	BCAC
	ARRA
	BLGL
	BLCK
	CLPR
	YTRK
	CNRY
	BLUR
	VRML
	BRWN
	RDBD

## Data Generating Model

### Example

Table(likelihood) shows model performance metrics for the toy example considered from market category 250 for the Monterey Trawl fishery in 1982. Here we consider Mean Squared Error (MSE; computed on the species composition scale), deviance information criterion (DIC), widely applicable information criterion (WAIC), and marginal Bayesian model probabilities ( $pr(M|y)$ ). These measures span a wide range of model selection philosophies and yet here they all consistently agree in ranking the likelihood models. Both of the overdispersion models considered here (NB and BB) outperform the more standard Poisson and binomial count models, with the most support being for the BB model and the Poisson model showing the least support.

	Poisson	Binomial	NB	BB
MSE	0.06412	0.06264	0.05171	0.04479
DIC	1342.27	1571.46	345.89	340.86
WAIC	1421.61	1665.41	345.09	341.66
$pr(M y)$	$\approx 0$	$\approx 0$	$\approx 10^{-7}$	$\approx 1$

Figure(likelihoods) shows observed species compositions plotted against Poisson, Binomial, NB, and BB predictive intervals. Here 95% highest density intervals (HDI) are shown to visualize the predictive accuracy of each model against the data.

The large spread of the observed species compositions seen in Figure(likelihoods) visually demonstrate the degree of overdispersion present in port sampling data. The Poisson and binomial models disregard this overdispersion to prioritize fitting the data mean. The NB and BB models explicitly model overdispersion in the data, albeit via differing mechanisms, and as such they predict a larger subset of the data.

The split beta-binomial intervals seen in Figure(likelihoods) are the consequence of confining a large amount of residual variability to the unit interval. The beta-binomial is the only model considered here, which estimates such a large degree of variability and thus it is the only model that produces predictive species composition distributions of the sort. Figure(violin) shows the beta-binomial predictive distributions as a violin plot demonstrating how the beta-binomial model arranges predictive density over the unit interval. The predictive intervals in Figure(likelihoods) are the smallest possible regions on each of these densities so that each intervals contain 95% probability. Species composition is bounded on  $[0, 1]$ , thus in the presence of large variability predictive density may aggregate around the bounds. For the example of STRK, notice that although the predictive HDI is split, the vast majority of density lies directly atop the data.

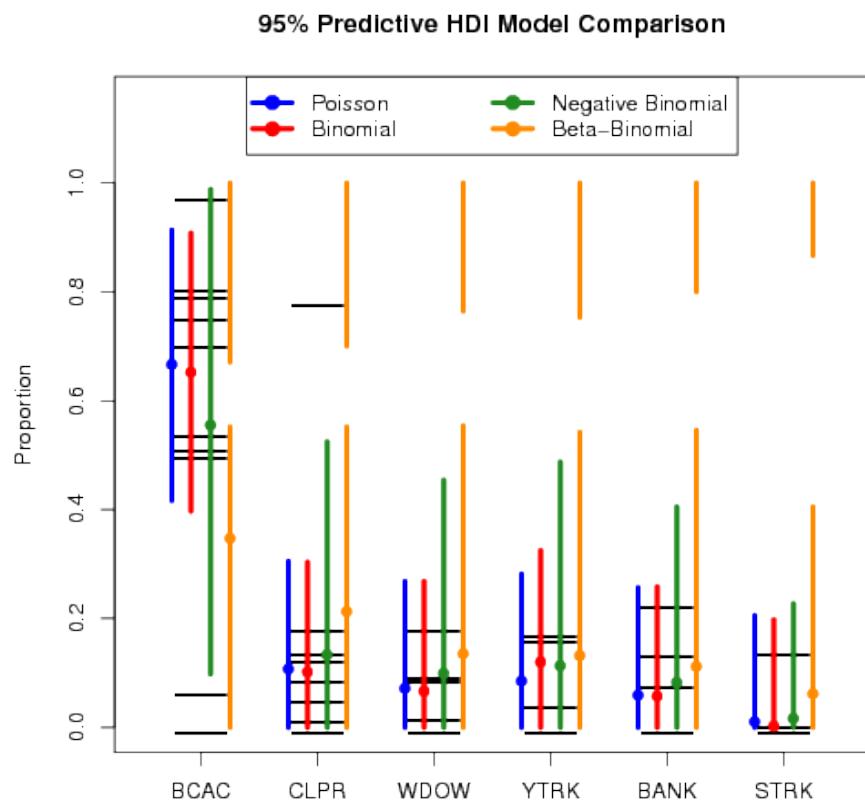


Figure 2: Interval Plot

### Beta–Binomial Posterior Predictive Species Compositions

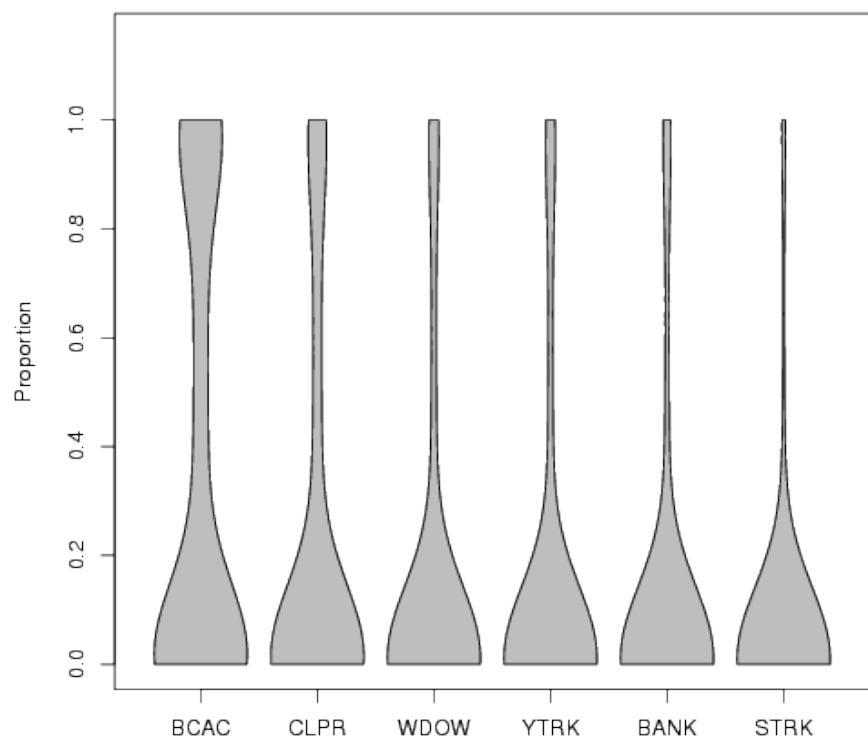


Figure 3: Violin Plot

## Predictor and Prior Selection

	M1	M2	M3	M4	M5	M6
MSE	0.127245	0.127042	0.126801	0.122373	0.127236	0.126573
DIC	39790.26	39491.64	39244.91	37231.70	39407.02	39406.41
WAIC	39745.58	39446.51	39192.25	37182.93	39354.11	39353.49
$pr(M y)$	$\approx 0$	$\approx 0$	$\approx 0$	$\approx 1$	$\approx 0$	$\approx 0$

Table(*Predictor*) displays the relative support for the model structure on the  $\beta_{m\eta}^{(t)}$  time parameters. From M1 to M4 the models represent a spectrum of models with an increasing potential of shrinkage among time parameters. Models M5 and M6 represent models which build in complexity, from M4, via the inclusion of multiple hierarchical variance parameters among the interaction terms.

Across all of the time models, model M4 displays consistent support over all other candidate models considered here. Model M4 represents a model with maximal potential for pooling through time, while still maintain the ability to model differences in seasonality from year to year.

As a final check of the model structure and the implied prior information the prior predictive is considered. The prior predictive distribution summarizes the information that is intrinsic to the model structure itself, in the absence of data. The prior predictive of modeled weight is considered over a 100 pound cluster size, which is consistent with aggregating the two nominal 50 pound cluster samples described by Sen (1984) in the original sampling protocol.

As seen in Figure(*prior Predictive*) the prior predictive of (M4) is both symmetric and quite diffuse over the 100 pound domain. The U shape of the distribution is a side effect of the diffusion of the selected prior. As data are added to the model, the indecisive U shape collapses toward the data in the posterior.

## Model Exploration & Averaging

Figure(*modelselection*) shows the results of port complex model selection for the modeled period from 1978 to 1982 in market category 250. Along the top,  $\omega$  BMA weights for the top 10 models are displayed (each column is a distinct model). The following ten rows indicate the ten port complexes in California, and the colored cells indicate how port complexes are partitioned in each model.

Considering Figure(*modelselection*), the best partitioned model gives distinct parameters to CRS and ERK, while pooling BRG/BDG, OSF/MNT/MRO, and OSB/OLA/OSD. This model uses five parameters to model the ten ports complexes in California. Given the set of candidate models explained above, the BMA procedure weights this model at approximately 32% of the truth. Notice that the only difference among the top four models is in how the port complexes

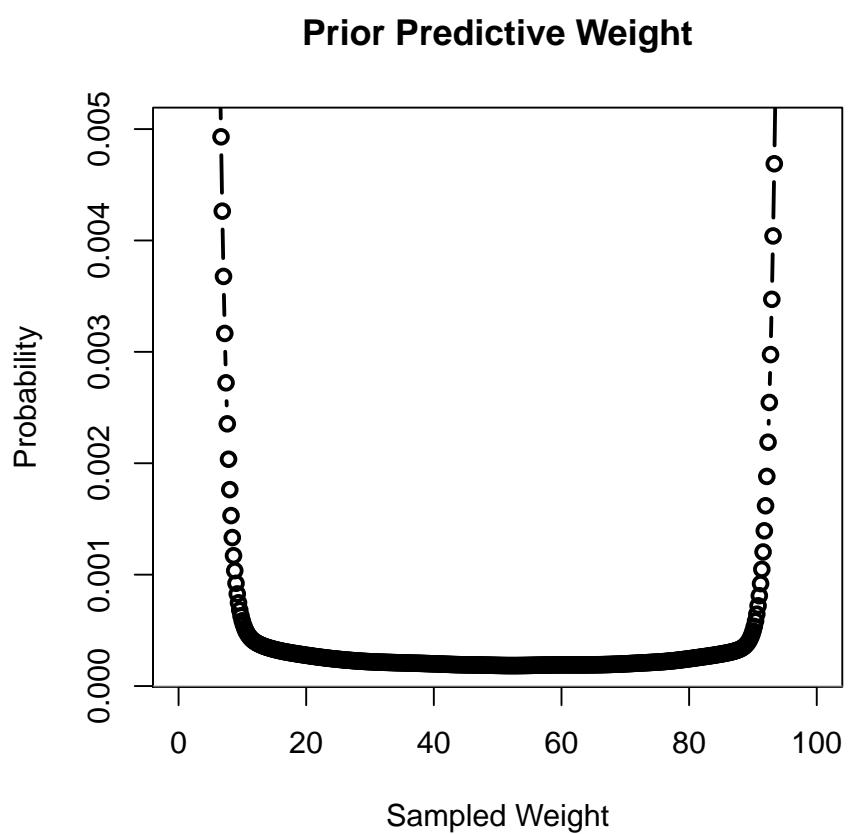


Figure 4: Prior Prediction

MCAT 250										
$\omega$	0.32	0.14	0.13	0.12	0.02	0.02	0.02	0.02	0.02	0.02
CRS										
ERK										
BRG										
BDG										
OSF										
MNT										
MRO										
OSB										
OLA										
OSD										

Figure 5: 1978 to 1982

south of point conception are handled. In fact, when the seven northernly port complexes are fixed in the described partitioning, the top four models represent all of the possible partitionings of the southern three port complexes.

In this modeled period it is known that no species composition sampling (landings?) was done south of point conception, thus it is not surprising that these models perform similarly. When no data is present, parameters simply represent place holders for out of sample prediction. Since the port complexes south of point conception are not informed by data, the predictions are identical in these categories. Since the first model makes identical predictions to the following three, and does so using the fewest parameters, it is correctly identified as the most parsimonious explanation among these data.

When considering how the top four model partitionings share identical structure in the seven northernly port complexes, while exhaustively spanning the candidate partitions south of point conception, it is simple to see that BMA assign's approximately 71% marginal probability to the northernly model structure.

The results shown here only represent a single market category across the time period 1978-1982. Similar results for other market categories and time periods are provided in the appendix Figure(*colortables*).

- zero landings?

## Prediction

Repeatedly fitting model (M4) across port complex partitionings and applying the BMA procedure, ultimately provides access to posterior predictive distributions of the species compositions ( $\pi_{jklm\eta}^*$ ) within a market category and time interval

modeled period. A straight forward way to evaluate the performance of the model in each modeled period is to compare the predictions of the model in each modeled period with the actual observations of species compositions from port samplers.

Species composition posterior predictive distributions are considered via HDI at three levels containing 68%, 95%, and 99% of posterior predictive probability. Table(*Prediction*) shows the proportion of observed species compositions which existed within the HDI across all strata, of each prediction level, in each modeled period.

- Should any of these be excluded?

## 78-82

	68%	95%	99%
250	67.1%	96.1%	98.7%
253	67.3%	96.3%	98.9%
262	67.4%	93.8%	95.3%
265	69.6%	96.0%	97.8%
269	68.2%	88.8%	90.2%
270	68.6%	93.6%	96.7%
956	68.3%	96.7%	99.2%
959	68.5%	96.3%	98.1%
961	69.3%	93.2%	95.3%
AVG	68.3%	94.5%	96.7%

## 83-90

	68%	95%	99%
245	60.8%	94.9%	97.7%
250	68.1%	96.0%	99.0%
253	69.3%	97.1%	98.9%
259	83.8%	91.9%	92.9%
262	68.5%	95.1%	95.9%
269	68.6%	94.2%	94.7%
270	67.9%	94.2%	96.7%
663	68.1%	94.1%	96.3%
667	69.4%	92.5%	93.5%
956	67.5%	96.2%	99.0%
959	67.4%	96.4%	99.0%
960	68.0%	96.1%	98.6%
961	68.6%	94.6%	97.8%

	68%	95%	99%
AVG	68.9%	94.9%	96.9%

Table(*Prediction*) largely shows that the observed proportion of predicted samples aligns appropriately with the predictions made by the model. Considering the average performance across market categories at each prediction level, it appears that prediction is mostly appropriate with the possible exception of the 99% prediction level. The 99% prediction level appears to slightly under-predict on average, indicating that predictive distributions are slightly lighter in the far tails than the data.

## Nuisance Parameters

### 78-82

$\rho$

MCAT	Mean	Median	Posterior SD
250	0.55	0.55	0.004
253	0.39	0.39	0.001
262	0.35	0.35	0.008
265	0.64	0.64	0.002
269	0.52	0.52	0.019
270	0.53	0.54	0.020
956	0.35	0.35	0.007
959	0.47	0.47	0.070
961	0.55	0.55	0.004

$v$

MCAT	Mean	Median	Posterior SD
250	12915.85	18523.12	8699.87
253	22747.87	23063.76	1535.53
262	20254.41	20506.36	2581.87
265	15846.22	16694.98	7601.15
269	20135.05	19975.15	4667.11
270	19931.96	19955.13	6033.35
956	19659.11	19795.60	1227.99
959	19159.69	13375.80	19256.94
961	18631.44	19498.31	7970.44

### 83-90

$\rho$

MCAT	Mean	Median	Posterior SD
245	0.65	0.65	0.014
250	0.51	0.51	0.002
253	0.47	0.47	0.010
259	0.75	0.75	0.009
262	0.41	0.41	0.001
269	0.57	0.57	0.046
270	0.74	0.75	0.027
663	0.51	0.51	0.001
667	0.49	0.49	0.022
956	0.43	0.43	0.003
959	0.55	0.55	0.004
960	0.45	0.45	0.004
961	0.59	0.59	0.001

- high  $\rho$  in 259 and over prediction in 259?
- lower rho might predict more appropriately?

$v$

MCAT	Mean	Median	Posterior SD
245	20211.82	20204.95	1276.83
250	236.03	192.53	134.67
253	20455.18	20140.50	1521.72
259	20246.14	20186.61	898.99
262	20445.49	20348.56	343.70
269	34386.49	25951.03	24030.32
270	20253.34	19908.07	9269.02
663	19563.87	19624.09	331.04
667	20089.55	20078.27	2723.34
956	20581.67	20664.71	913.92
959	19242.41	18707.09	5076.03
960	20059.66	20012.80	1703.89
961	20127.69	20141.04	580.80

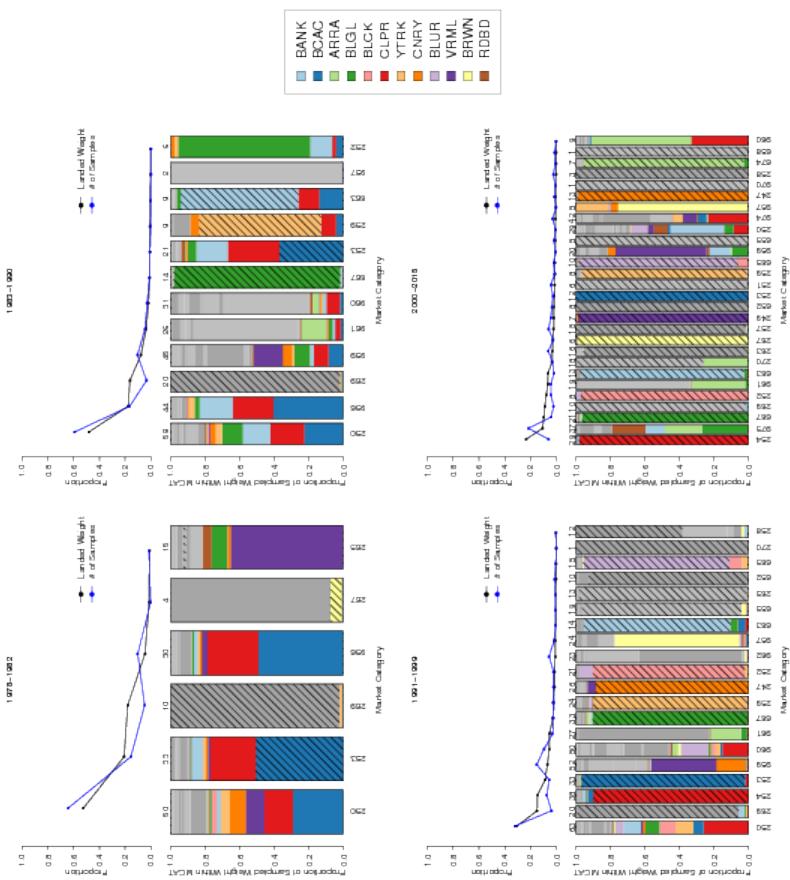
- Degree of smoothing (hierarchical parameters, rho)
- Report mean/variance of example
- Add  $v$  and  $\rho$  tables across all runs

## Discussion

- Variance Estimates; Justify Model based statistics
- Biase/Variance trade off; hierarchical modeling
- Prediction: tails lighter due to inla?

- higher degree Laplace approximation will better capture higher moments of posterior and thus may better capture information in the tails
- high  $\rho$  in 259 and over prediction in 259?
- lower rho might predict more appropriately?
- Model averaging
  - Interpret/Speculate
- Looking Forward
  - forecasting/hindcasting
    - \* simple
    - \* time series models
  - more computation faster
    - \* broader model exploration
    - \* broader spatial expansion
  - Hierarchical Spatial Models
  - DP Models

## Appendix



**78 to 82**

MCAT 250						MCAT 253					
$\omega$	0.32	0.14	0.13	0.12	0.02	$\omega$	0.14	0.14	0.14	0.10	0.06
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					
MCAT 265						MCAT 269					
$\omega$	0.02	0.02	0.01	0.01	0.01	$\omega$	0.19	0.14	0.14	0.13	0.07
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					
MCAT 956						MCAT 262					
$\omega$	0.10	0.08	0.07	0.07	0.06	$\omega$	0.30	0.21	0.06	0.05	0.04
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					

MCAT 270						MCAT 959					
$\omega$	0.08	0.06	0.04	0.03	0.03	$\omega$	0.49	0.49	0.00	0.00	0.00
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					

MCAT 961					
$\omega$	0.05	0.05	0.05	0.04	0.04
CRS					
ERK					
BRG					
BDG					
OSF					
MNT					
MRO					
OSB					
OLA					
OSD					

### 83 to 90

MCAT 250						MCAT 253					
$\omega$	0.73	0.25	0.00	0.00	0.00	$\omega$	0.11	0.09	0.06	0.05	0.05
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					

MCAT 259						MCAT 269					
$\omega$	0.02	0.02	0.02	0.01	0.01	$\omega$	0.64	0.12	0.07	0.06	0.04
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					
MCAT 663						MCAT 667					
$\omega$	0.49	0.49	0.00	0.00	0.00	$\omega$	0.07	0.03	0.03	0.03	0.03
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					
MCAT 956						MCAT 959					
$\omega$	0.26	0.21	0.19	0.11	0.10	$\omega$	0.36	0.22	0.18	0.15	0.02
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					

MCAT 960						MCAT 961					
$\omega$	0.19	0.15	0.10	0.09	0.05	$\omega$	0.17	0.12	0.10	0.08	0.08
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					
MCAT 245						MCAT 262					
$\omega$	0.05	0.05	0.05	0.05	0.04	$\omega$	0.47	0.29	0.07	0.06	0.02
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					
MCAT 270						MCAT 663					
$\omega$	0.04	0.03	0.03	0.03	0.03	$\omega$	0.49	0.49	0.00	0.00	0.00
CRS						CRS					
ERK						ERK					
BRG						BRG					
BDG						BDG					
OSF						OSF					
MNT						MNT					
MRO						MRO					
OSB						OSB					
OLA						OLA					
OSD						OSD					

## References

- [1] CCGS (California Cooperative Groundfish Survey). (2017). Pacific States Marine Fisheries Commission, 350 Harbor Boulevard, Belmont, California, 94002. URL: 128.114.3.187.
- [2] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.
- [3] NAS (National Academies of Sciences, Engineering, and Medicine). 2017. Review of the Marine Recreational Information Program. Washington, DC: The National Academies Press.
- [4] Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382-401.
- [5] Jaynes, E. T. (2003). Probability theory: The logic of science. Cambridge university press.
- [6] Ramasubramanian, K., & Singh, A. (2016). Machine Learning Using R. Apress.
- [7] McCullagh P. & Nelder, J.A. (1989). Generalized Linear Models, 2nd ed. London: Chapman and Hall.
- [8] 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.
- [9] 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/>.
- [10] 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/>.
- [11] 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.
- [12] Sen, A.R. (1984). Sampling commercial rockfish landings in California. NOAA Tech Memo. NOAA-TM-NMFS-SWFSC-45.
- [13] Sen AR. (1986). Methodological problems in sampling commercial rockfish landings. *Fish Bull.* 84: 409-421 .
- [14] 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.