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

Nick Grunloh[a], Edward Dick[b], Don Pearson[b], John Field[b], Marc Mangel[a,c]

December 17, 2019

[a] Center for Stock Assessment Research, University of California, Santa Cruz, Mail Stop SOE-2, Santa Cruz, CA 95064, USA.
[b] Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanographic and Atmospheric Administration, 110 McAllister Way, Santa Cruz, CA 95060, USA.
[c] Department of Applied Mathematics and Statistics, Jack Baskin School of Engineering, University of California, Santa Cruz, Mail Stop SOE-2, Santa Cruz, CA 95064, USA.

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 *Sebastes*), 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 introduce unknown bias and 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 Bayesian model averaging approach for inferring spatial pooling strategies across the over-stratified port sampling system. Our modeling approach, along with a computationally robust system of inference and model exploration, allows us to 1) objectively compare alternative models for estimation of species compositions in landed catch, 2) quantify uncertainty in historical landings, and 3) 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.

Introduction

Methods

When model building, most modeling decisions will be dependent on many of the other assumptions of the model. Yet if one desires to emperically test each component of the model, the scientific method would require that each component of the model be tested one-at-a-time against all other modeling choices. This process naturally implies a combinatoric explosion of different tests. For example, if only four modeling options are available in each of three modeling decisions, then $4^3 = 64$ different modeling options must be considered.

For the purpose of fighting against this combinatoric explosion of tests, we take a principled approach of ordering modeling decisions based on the relative importance to prediction. So that the most important features of the model are selected first in a way that is as independent as possible from other model aspects. Furthermore, once a modeling decision is made, it shall be fixed and all subsequent modeling decisions are made conditional on the previously decided assumptions of the model.

In service of this principled approach, first the structure of the data are considered. Based upon the structure of the data a probability model is selected by first testing likelihoods against the data, independent of all other factors. Once a likelihood is selected, priors are considering amoung the parameters of the selected likelihood. Naturally as different likelihoods will introduce different parameters into the model, different choices of prior will be neccessary. Once a probability model is completed by the choice of likelihood and prior, a linear predictor for estimating differences between stratum is constructed in a similar way.

Data

As outlined in Sen (1984) & (1986) the species composition port sampling data are the result of a cluster sampling protocol executed across the many strata of California's commercial fisheries. Each sample is intended to be two fifty-pound clusters selected at random from a stratum. Although port samplers do their best to follow protocol, in reality the port sampling environment does not always allow Sen's original protocol to be followed. The lack of mandatory sampling in California, along with variations in the sampling protocol, may result in only a single cluster being taken, or the size of clusters taken to vary from stratum to stratum based on the particular challenges of sampling each stratum.

Samples are recorded as integer pounds for each observed species, across the landed market categories, gear groups, and port complexes in time (quarters within year). Presently

there are 71 rockfish market categories, although not all market categories are always used. The number of market categories with recorded landings has gone from less than 25 in 1978 to about 55 in 2014, see Figure (1). Landings are grouped into major fishing gear groups (trawl, hook and line, gillnet, fish pot, or other minor categories) and ten major port complexes spanning the California coast, see Figure (2).

The model based methodology proposed here does not rely strongly upon the cluster sampling structure, but rather views each sample as independent and identically distributed (*i.i.d.*) draws from a data generating model, conditional on a parameterization of the stratification system. So long as the parameterization and data generating model are sufficiently robust for handling the behavior of these data, a conditionally *i.i.d.* model of these data will be practically useful 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, a database maintained by the California Cooperative Groundfish Survey (CALCOM, 2018). Just as in Shelton et al. (2012), we aggregate all observed clusters within each unique sample so that the total weight sampled is the sum of pounds in each cluster. Similarly the observed weight for a particular species, in each unique sample, is the sum of all of the observed weights across clusters.

Although model based data analysis has the potential to add significant structure to data, a judicious 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 (see Figures (3 & 4) for a summary of landed weight, the number of landed strata, and the number of samples over the two modeled time periods). For this reason, we refrain from modeling any period where the minimum possible number of effective parameters exceeds the number of samples for the modeled period. Rather than apply models inappropriately, these landings are speciated as the nominal species for their market category. We later demonstrate that due to prioritization in sampling heavily landed, or otherwise commercially relevant categories, this sample size heuristic leaves relatively few landings to be speciated in a statistically uninformed way (i.e. "nominal" speciation). Thus nominal speciation represents a negligible component of the overall expanded landings for most species.

- chunking
  - MCAT
  - 78-82, 83-90, 91-01

Likelihood

For a particular market category, the random variable $Y_{ijklm\nu}$ is the i^{th} observation of the j^{th} species' rounded weight (pounds), in the k^{th} port, caught with the l^{th} gear-group, in the ν^{th} quarter, of year m. The $Y_{ijklm\nu}$ are modeled as $i.i.d.$ observations from some distribution, f, over the whole numbers

$$Y_{ijklm\nu} \overset{i.i.d.}{\sim} f(\theta_{jklm\nu}, \phi). \tag{1}$$

Here $\theta_{ijklm\nu}$ is a linear predictor for inferring the mean weight, and ϕ is a nusiance parameter included (when implied under f) to allow models to more flexibly capture higher moments of the $Y_{ijklm\nu}$. Of particular interest, the residual variance is a key value to quantify.

Defining a particular form for f also implies the linear predictor-mean relationship as well as the structure, and scope, of residual variation. For the purposes of accurately modeling not only species composition means, but also higher moments of the data (e.g. variances), it is necessary to recognize model limitations with respect to overdispersed data. The form of f was chosen empirically based upon experimentation among the poisson, binomial, negative binomial, and beta-binomial distributions. Other modeling options were considered for the form of f, but the list above was determined to be the most compuationally feasible at this time and among these data. See appendix 6.1 for modeling details.

As seen in appendix 6.1 the beta-binomial model was selected as the most appropriate and flexible likelihood. With the potential to account for the largest amount of variance, the beta-binomial model may not be the true model, but it certainly offers the most computationally conientient, and accurate, approximate reproduction of the observed data.

Beta-Binomial Model

Inserting the details for the beta-binomial model into equation (1). Now for the i^{th} sample of the j^{th} species' weight, in the k^{th} port, caught with the l^{th} gear-group, in the ν^{th} quarter, of year m we get,

$$y_{ijklm\nu} \overset{i.i.d.}{\sim} BB(\mu_{ijklm\nu},\ \sigma^2_{ijklm\nu}). \tag{2}$$

Above, $\mu_{ijklm\nu}$ is the stratum level mean weight, and $\sigma^2_{ijklm\nu}$ is the stratum level residual variance. $\mu_{ijklm\nu}$ is related to a linear predictor, $\theta_{jklm\nu}$, via the mean function,

$$\mu_{ijklm\nu} = n_{ijklm\nu}\frac{\exp(\theta_{jklm\nu})}{1+\exp(\theta_{jklm\nu})}. \tag{3}$$

Here $n_{ijklm\nu}$ is the observed aggregate cluster size for each sample. Additionally, $\sigma^2_{ijklm\nu}$ is related to $\mu_{ijklm\nu}$ and the overdispersion parameter, ϕ, via the following equation,

$$\sigma^2_{ijklm\nu} = \mu_{ijklm\nu}\left(1 - \frac{\mu_{ijklm\nu}}{n_{ijklm\nu}}\right)\Big(1 + (n_{ijklm\nu} - 1)\ \phi\Big). \tag{4}$$

In the context of the beta-binomial distribution, ϕ is the within-cluster correlation. The situation where $\phi \to 1$ represents identical information content among replicates within a cluster, with maximal overdispersion relative to the binomial distribution. The situation where $\phi \to 0$ represents totally independent information content among replicates within a cluster, and the beta-binomial model approaches the binomial model. ϕ explicitly models average overdispersion across all strata within a market category, while $\mu_{ijklm\nu}$ gives the model flexibility at the stratum level through the linear predictor $\theta_{jklm\nu}$.

The structure of $\theta_{jklm\nu}$ will be developed in section 2.4, but at this point it is worth pointing out that the index notation imples that the linear predictor, $\theta_{jklm\nu}$, is a linear function of species, port, gear-group, year and quarter. For the purpose of achieving an efficient, and theoretcally founded, method of partial pooling amoung these strata, $\theta_{jklm\nu}$ will naturally be composed of some combination of an intercept β_0, fixed effects parameters $\beta^{(\text{fixed})}$, and random effects parameters $\beta^{(\text{random})}$. In the Bayesian parlence, fixed effects amount to parameters with priors that have fixed hyperparameter values, while random effects amount to parameters with priors that have random hyperparameters that themselves are inferred from the data. Admittedly the nomenclature adopted here is philosophically unpure, but I abuse the fixed/random effect terminology for the pragmatic purpose of bridging statistical philosophies in an attempt to more directly communicate the inferencial goals at hand. Of course in the Bayesian paradigm all parameters are random, but still the fixed/random effect terminology is a useful way to consider the different behavior of the parameters included here. For further discussion of fixed and random effects models see (Nelder & McCullagh, 1989) and for a discussion of the Bayesian framing via heirarchical models see (Gelman et al., 2013). Further discussion of prior structure is developed in section 2.3.

Priors

The parameters present in the beta-binomial model are the linear predictor $\theta_{jklm\nu}$ and the overdispersion parameter ϕ. The linear predictor is itself parameterizes with a reference level intercept β_0, some number of fixed effects $\beta^{(\text{fixed})}$, and some number of random effects $\beta^{(\text{random})}$. To complete the Bayesian formulation of the model, priors are expressed, over these parameters, so as to convey any information that the model has external to the data. In this case, priors are expressed in a largely diffuse manner to represent the relative lack of

information external to the data.

Explaination of priors for this model begins with the prior on the intercept,

$$\beta_0 \propto 1. \tag{5}$$

Since the β_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.

In this model fixed effects have a natural interpretation as offsets from the reference level. Each of these offsets are assigned a diffuse normal prior,

$$\beta^{(\text{fixed})} \sim N(0, 32^2). \tag{6}$$

In the null case it is resonable to assume that these offsets may be zero due to a lack of difference from the reference level. However, in the Bayesian setting parameters are random variables, and do not take any one particular value, hence the distributional assumption of normality. Furthermore there is no strong information external to the data in our model to promote the value 0 much more strongly than any other value. The large variance hyperparameter, 32^2, produces a normal distribution that spreads far across the parameter space of fixed effects in this model. It is this large fixed value of the variance hyperparameter which produces behavior similar to classical fixed effect model, and yet even this prior produces an extremly slight amount of posterior shrinkage toward 0, so as to edge on the side of parsimony.

When considering random effect priors it is useful to consider how overparameterized models may cause overfitting and weaken model performance through the bias-variance dilemma (Ramasubramanian and Singh, 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) describes how the inclusion of even a small amount 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.

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[\ (\hat{\theta} - \theta)^2\ \right] = \overbrace{\mathbb{E}\Big[\ \Big(\hat{\theta} - \mathbb{E}(\hat{\theta})\Big)^2\ \Big]}^{\text{Var}(\hat{\theta})} + \overbrace{\Big(\ \mathbb{E}(\hat{\theta}) - \theta\ \Big)^2}^{\text{Bias}(\hat{\theta},\theta)^2} \tag{7}$$

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.

Random effects typically imply groupings of many parameters that can be pooled in someway. Oftentimes interaction terms appear as random effects due to the combinatoric explosion of the number parameters implied by the interaction of different categories. In section 2.4 a number of random effects models are considered. When framed in the Bayesian context each included random effect parameter recieves the following heirarchical prior structure.

$$\beta^{(\text{random})} \sim N(0, v) \tag{8}$$

$$v \sim IG(1,\ 2 \times 10^3) \tag{9}$$

Similarly to fixed effects, $\beta^{(\text{random})}$ is distributed as a normal distribution with zero mean. However in contrast to the large fixed variance of fixed effect priors, random effects (in the Bayesian context) introduce a heirarchical layer to the prior structure around the prior variance for $\beta^{(\text{random})}$. The parameter v is introduced to model the prior variance for $\beta^{(\text{random})}$. In the posterior, the data is used to determine the extent of shrinkage the $\beta^{(\text{random})}$ should experience toward zero by inferring relatively larger or smaller values for v. As the above model learns the posterior of the hierarchical variance parameter, v, it affects the degree of shrinkage, as well as the effective number of parameters held within the respective hierarchies (Gelman et al., 2013). To achieve this, each variance parameter must itself be assigned a prior to be estimated. Any hierarchical variance parameters, v, included in the model is assigned a diffuse inverse gamma (IG) prior. See appendix 6.2 for further discussion about selecting appropriate priors.

Above the $\beta^{(\text{fixed})}$ and $\beta^{(\text{random})}$ are place holders for a particular fixed or random effect parameter to be included in the model. Each effect to be included in the model is either included as a fixed or random effect. The particular effects to be included in the linear predictor are discussed later in section 2.4.

Finally the overdispersion parameter, ϕ, is assigned a diffuse normal prior on the logit scale, $\text{logit}(\phi) \sim N(0, 2^2)$. 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 ± 3.91), includes almost the entirety of the back transformed unit interval (i.e. 0.5 ± 0.48).

- Prior (discussion of sensativity analysis)
  - hook to appendix to describe sensativeity analysis (prior predictive, other plots)
  - transition to linear predictor

Linear Predictor

As previously mentioned, the linear predictor for this model is composed of a linear combination of an intercept, some number of fixed effects, and some number of random effects. Recall the natural stratification groups outlined by (Sen, 1984) are port-complex, gear-group, year, and quarter. Additionally, the process of port sampling produces the species identification as a natural covariate along side each measured weight. Within a market category, these factors are the primary variables considered in modeling.

Due to the computational difficulty of these data and long run times involved with fitting models with many parameters (particularly models with many heirarchical variances) the linear predictor is constructed in a bottom-up and largely greedy way. That is to say, we grow our eventual linear predictor starting from a linear predictor that only includes an intercept and a species fixed effect. Effects are added incrementally until our eventual completed form is achieved by sequentially including the next best effect, one at a time. At times (particularly when considering random effects) the modeling effort can operate with the oracle of some previously understood correlation structure among categories. The goal is that stepping through possible modeling options for the $\theta_{jklm\nu}$ in this way should produce a simultaneously parsimonious and flexibile model that will preform well both for the observed strata as well as provide a structure for accurately extending prediction into unobserved strata.

A categorical handling the available predictors results in $\theta_{jklm\nu}$ taking the form of a simple sum of the relevant intercept β_0, fixed effect $\beta^{(\text{fixed})}$, and random effects $\beta^{(\text{random})}$ coefficients in a given stratum,

$$\theta_{jklm\eta} = \beta_0 + \sum \beta^{(\text{fixed})} + \sum \beta^{(\text{random})}. \tag{10}$$

Firstly θ is factored among the many strata by the inclusion of fixed effect parameters for each of the species $\beta_j^{(s)}$, port-complexes $\beta_k^{(p)}$, and gear-groups $\beta_l^{(g)}$. Secondly random effects parameters are included for species:port interactions $\beta_{jk}^{(s:p)}$ and year:quarter interactions $\beta_{m\nu}^{(y:q)}$. Experiments about the form of the linear predictor are outlined in appendix 6.3. In particular, attention is given to the inclusion, or removal, of year $\beta_m^{(y)}$ and quarter $\beta_\nu^{(q)}$ fixed effect parameters. As well as which two way interaction terms to include as random effects.

These tests result in the following parsimonious form for the linear predictor,

$$\theta_{jklm\nu} = \beta_0 + \beta_j^{(s)} + \beta_k^{(p)} + \beta_l^{(g)} + \beta_{m\nu}^{(y:q)} + \beta_{jk}^{(s:p)}. \tag{11}$$

- chunking
  - MCAT
  - 78-82, 83-90, 91-01

Species Composition Prediction

Bayesian inference of the above model gives access to the full posterior distribution of all of the parameters of the model, given the data. 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; any required moments are simply obtained by computing the desired empirical moments from the posterior samples. Additionally, the fact that the parameters are full distributions means that any functions of those parameters are themselves random variables with the function representing a transformation of those parameters.

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

$$p(y^*_{jklm\nu}|\underline{y}) = \iint \mathrm{BB}\Big(y^*_{jklm\nu}|\theta_{jklm\nu}, \phi\Big) P\Big(\theta_{jklm\nu}, \phi|\underline{y}\Big) d\theta_{jklm\nu} d\phi.$$

Here BB is the data generating beta-binomial distribution for a predictive observation and $P(\theta_{jklm\nu}, \phi_{jklm\nu}|y)$ is the posterior distribution of the parameters given the observed data. $\theta_{jklm\nu}$ and ϕ are integrated numerically via Monte Carlo integration to produce samples from the posterior predictive distribution, $p(y^*_{jklm\nu}|y)$, for sampled weights in the $jklm\nu^{th}$ stratum.

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

$$\pi^*_{jklm\nu} = \frac{Y^*_{jklm\nu}}{\sum_j Y^*_{jklm\nu}} \quad \underline{Y}^*_{klm\nu} \neq \underline{0}.$$

For a particular market category, $\pi^*_{jklm\nu}$ is predicted proportion of species j in the k^{th} port, caught with the l^{th} gear, in the ν^{th} quarter, of year m.

Expansion of Landed Catch to Species

For a particular market category, speciated landings simply amounts to the multiplication of the known total landings ($\lambda_{klm\nu}$), reported on landing receipts in the $klm\nu^{th}$ stratum, with the posterior predictive species composition, $\pi^*_{jklm\nu}$, as follows

$$\lambda^*_{jklm\nu} = \lambda_{klm\nu}\pi^*_{jklm\nu}.$$

$\lambda^*_{jklm\nu}$ is then the posterior predictive landings for species j in the $klm\nu^{th}$ stratum of a particular market category. Recall that since $\pi^*_{jklm\nu}$ is a random variable, then so is $\lambda^*_{jklm\nu}$. Computing the variance of $\lambda^*_{jklm\nu}$ simply amounts to computing the variance of random draws from the $\lambda^*_{jklm\nu}$ distribution. Furthermore, any level of aggregation of $\lambda^*_{jklm\nu}$ is easily obtained by summing $\lambda^*_{jklm\nu}$ draws across the desired indices. For example to obtain the distributions of yearly catch of Bocaccio in a particular market category (i.e. aggregated across ports, gears, and quarters) one simply fixes j to Bocaccio, and computes the following transformation of $\lambda^*_{jklm\nu}$,

$$\lambda^*_{j\cdot\cdot m\cdot} = \sum_k \sum_l \sum_\nu \lambda^*_{jklm\nu}.$$

Distribution summaries such as quantiles, means, or variances may be computed by computing those metrics from the random draws of the resulting $\lambda^*_{j\cdot\cdot m\cdot}$ distributions.

Results

- model comparison across likelihoods
  - MSE, WAIC
  - Posterior predictive spp comp. Violin plots
- Diagnostic plot
- Prior sensativity
- linear predictor diagnostic plots and tell story
- time series plot and tell story

Discussion

Likelihood

Admittedly the structure of these data as rounded pounds do not immediately cry out for the above counting distributions. Rather one might consider modeling these data as censored observations of a normal response, or possibly one might be interested in modeling the multivariate structure as a multinomial (Shelton et al. (2012) shows that this is implied under the poisson response) or dirichilet-multinomial response. I grant that indeed such models would be lovely to explore, however from the pragmatic perspective these models are very difficult to fit in this setting.

- poisson -> NB
- binomial -> BB
- poisson -> multinomial (Ole)
- overdispersed multimonial -> Dirichelette-Multinomial
- motivated by calcom, not comparison
- Poisson (Ole) v. Beta-binomial (overdispersed model)
  - Poisson/Binomial/NB/BB Comparison
  - Violin plots
- Paragraph about model selection techniques
  - MSE, WAIC
  - Diagnostic
- String together longest time series plots possible
  - WDOW, BCAC, CHILI, CNRY

Figures

Appendix

Appendix A: Likelihood Experiments

For the purposes of accurately modeling not only species composition means, but also higher moments of the data (e.g. variances), it is necessary to recognize model limitations with respect to overdispersed 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 assumed data generating processes to accurately represent higher moments in the data.

One might expect the sampling approach in this setting to produce data which are theoretically similar to a multivariate multinomial model. Shelton (2012), describes how to fit a version of that model via the multinomial–Poisson transformation on these data. The multinomial–Poisson transformation gives the result that fitting regression parameters via conditionally *i.i.d.* Poisson models produces the same parameter inference as fitting the joint multinomial model (?, ?). Thus the Poisson model presented below produces the same inference as joint multinomial model.

Due to the theoretical restrictitions of these simple models, and the reality the port sampling environment it is possible (or likely) that the actual data in CALCOM deviate substanatively from the platonic ideals implyed by the one parameter counting distributions. In this appendix we emperically question the appropriatness of the assumptions of various data generating processes, and explore models that loosen these assumptions by accounting for the overdispersion present in the CALCOM data. It is also interesting to note that McCullagh and Nelder (1989) commiserate about the prevalence of overdispersed data in cluster sampling designs. They explain ways in which cluster sampling itself may result in overdispersion.

Extending the Poisson and binomial models to deal with overdispersion, typically involves adding additional parameters for the purpose of modeling higher moments of the data (in this settling we call that parameter ϕ). The negative binomial (NB) distribution is often used as an overdispersed extension of the Poisson model, since it can be expressly written as an infinite mixture of Poisson distributions. The beta-binomial model is used as an overdispersed extension of the binomial model.

The Poisson and binomial models are both derived under limiting cases for ϕ, and thus under these models ϕ is not an infered parameter. As a result, these models are only capable

of capturing a relatively rigid scope of response behavior. In contrast, the negative binomial and beta-binomial models use the data to estimate the ϕ parameter. In these models the ϕ parameter is used to disentangle the mean and residual variance estimates. Thus the negative binomial and beta-binomial models may produce more accurate measures of center, by more accurately estimating the residual variance.

We develop an example on a subset of data to evaluate statistical support for overdispersed models.

Poisson Model

$$Y_{ij} \overset{i.i.d.}{\sim} Poisson(\theta_j) \tag{12}$$

$$\mu_j = \exp(\theta_j). \tag{13}$$

$$\sigma_j^2 = \mu_j. \tag{14}$$

Binomial Model

$$Y_{ij} \overset{i.i.d.}{\sim} Binomial(\theta_j) \tag{15}$$

$$\mu_{ij} = n_{ij} \frac{\exp(\theta_j)}{1 + \exp(\theta_j)}. \tag{16}$$

$$\sigma_{ij}^2 = \mu_j \Big(1 - \frac{\mu_j}{n_{ij}}\Big). \tag{17}$$

Negative Binomial Model

$$Y_{ij} \overset{i.i.d.}{\sim} NB(\theta_j, \phi) \tag{18}$$

$$\mu_j = \exp(\theta_j). \tag{19}$$

$$\sigma_j^2 = \mu_j \Big(1 + \frac{\mu_j}{\phi}\Big). \tag{20}$$

Beta-Binomial Model

$$Y_{ij} \overset{i.i.d.}{\sim} BB(Y_{ij}|\theta_j, \phi) \tag{21}$$

$$\mu_{ij} = n_{ij} \frac{\exp(\theta_j)}{1 + \exp(\theta_j)}. \tag{22}$$

$$\sigma_{ij}^2 = \mu_{ij} \Big(1 - \frac{\mu_{ij}}{n_{ij}}\Big) \Big(1 + (n_{ij} - 1)\ \phi\Big). \tag{23}$$

We develop an example on a subset of data to evaluate statistical support for overdispersed models, see appendix 6.1, which we have subsequently used for the purposes of applying at an operational scale

- Likelihood
  - describe overdispersion concerns
  - outline model descriptions
  - poisson, NB, beta, beta-binomial
  - describe experiment

Appendix B: Prior Experiments

‘

Appendix C: Linear Predictor Experiments

Appendix D: Complete Model

$$Y_{ijklm\nu} \overset{i.i.d.}{\sim} BB(Y_{ijklm\nu} | \theta_{jklm\nu}, \phi) \tag{24}$$

$$\theta_{jklm\nu} = \beta_0 + \beta_j^{(s)} + \beta_k^{(p)} + \beta_l^{(g)} + \beta_{m\nu}^{(y:q)} + \beta_{jk}^{(s:p)} \tag{25}$$

$$\Big\{\beta_j^{(s)}, \beta_k^{(p)}, \beta_l^{(g)}\Big\} \sim N(0, 32^2) \tag{26}$$

$$\beta_{m\nu}^{(y:q)} \sim N(0, v^{(y:q)}) \qquad \beta_{jk}^{(s:p)} \sim N(0, v^{(s:p)}) \tag{27}$$

$$v^{(y:q)} \sim IG(1,\ 2 \times 10^3) \quad v^{(s:p)} \sim IG(1,\ 2 \times 10^3) \tag{28}$$

$$\text{logit}(\phi) \sim N(0, 2^2) \tag{29}$$

Figures

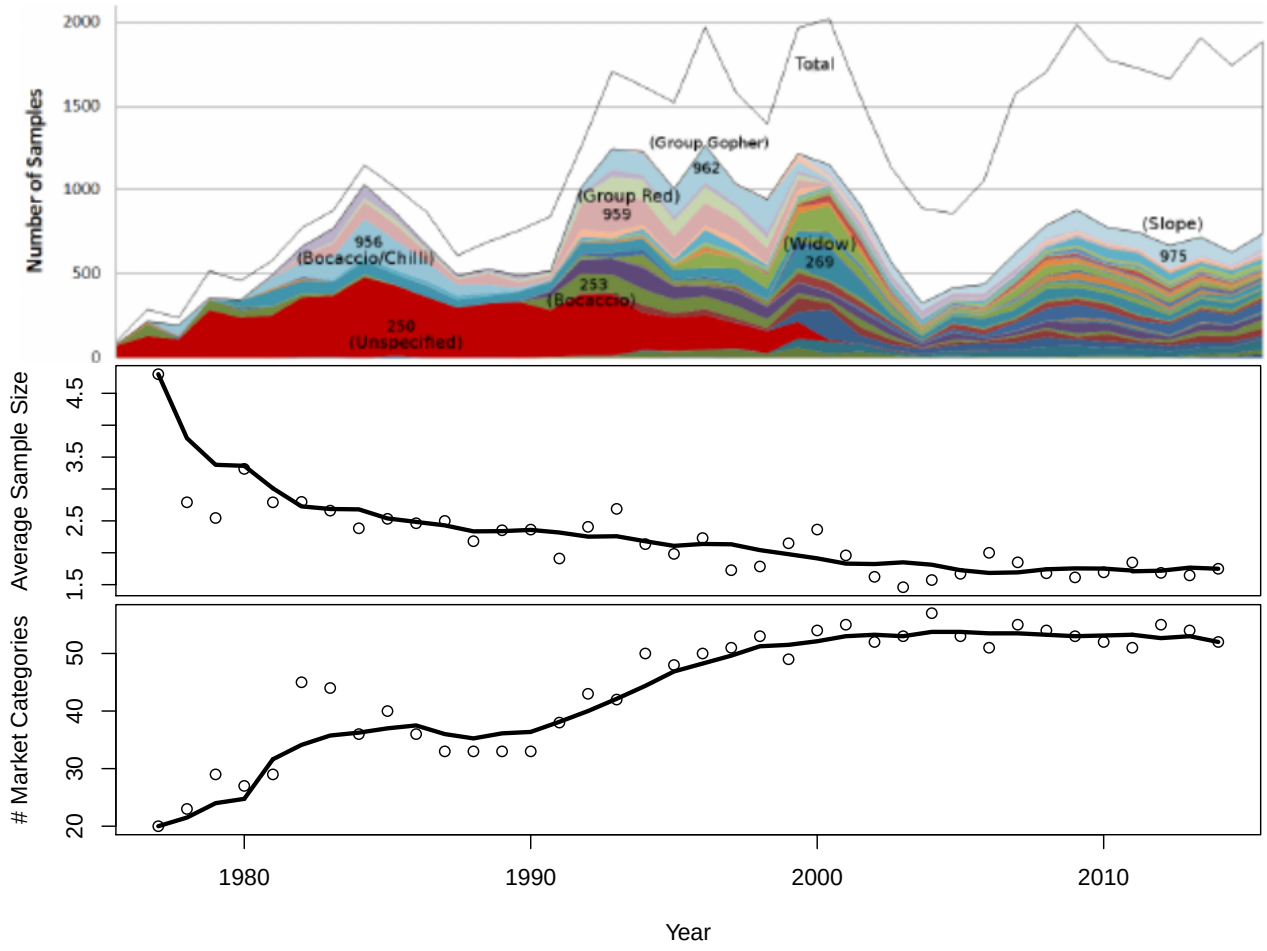

Figure 1: Number of commercial port samples per market category in California, 1978-2014 (upper panel), average sample size per stratum (middle panel), and number of market categories recorded on landing receipts (lower panel). On the lower panels, points indicate observed values, while the black lines represent 8 year moving averages

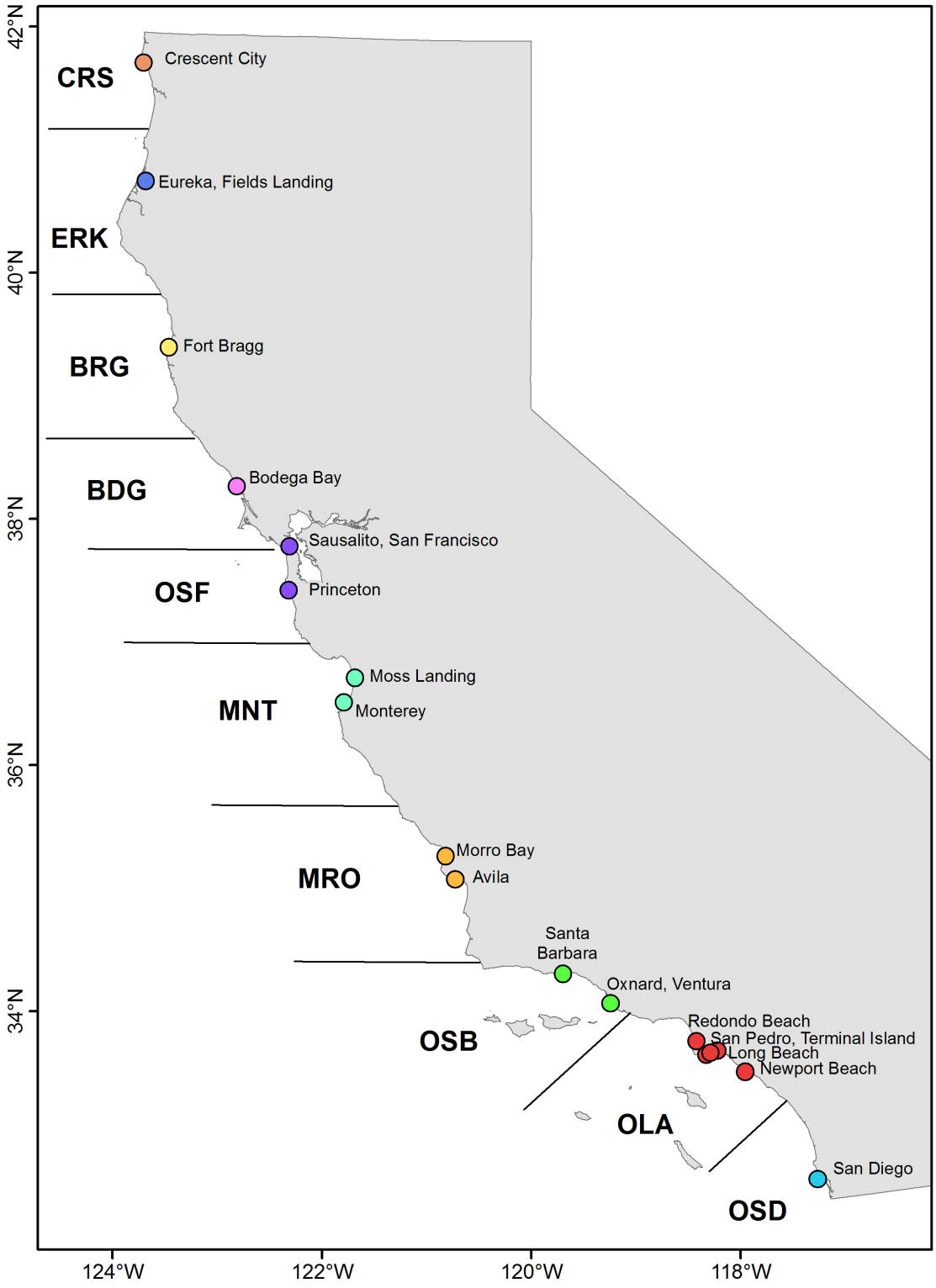

Figure 2: Map showing the ports in California that account for at least 95% of landings. Separating lines show how ports have been aggregated into port complexes.

				1978 - 1982			1983 - 1990		
Category	Market Category	Description	Nominal Species or Group	Tons	# Strata	# Samples	Tons	# Strata	# Samples
Multi-species	250	Rockfish, unspecified	UNSPECIFIED ROCKFISH	36539.3	524	1021	55332	1048	2933
	262	Thornyheads	THORNYHEADS, UNSPECIFIED	8512.2	202	237	27929	406	392
	956	Rockfish, group bocaccio/chili	UNSPECIFIED ROCKFISH	3213.7	47	127	20227	655	870
	957	Rockfish, group bolina	UNSPECIFIED SHELF ROCKFISH	27.6	27	0	417	426	1
	958	Rockfish, group deepwater reds	ROCKFISH GROUP 3	16.3	1	0	19	10	0
	959	Rockfish, group red	UNSPECIFIED ROCKFISH	225.1	41	9	8883	843	501
	960	Rockfish, group small	UNSPECIFIED ROCKFISH	1.8	6	2	2223	439	118
	961	Rockfish, group rosefish	ROCKFISH GROUP 6	162.1	13	12	4179	377	327
	962	Rockfish, group gopher	UNSPECIFIED ROCKFISH	0.0	0	0	314	225	2
	963	Rockfish, large red	UNSPECIFIED ROCKFISH	0.0	0	0	0	0	0
"Single-species"*	245	Rockfish, cowcod	COWCOD	10.9	38	1	273	294	31
	246	Rockfish, copper (whitebelly)[1]	COPPER ROCKFISH	6.8	18	0	6	93	0
	247	Rockfish, canary	CANARY ROCKFISH	0.4	1	0	62	51	0
	248	Rockfish, yelloweye	YELLOWEYE ROCKFISH	0.0	0	0	0	0	0
	249	Rockfish, vermilion	VERMILION ROCKFISH	4.8	21	0	1	34	0
	251	Rockfish, black-and-yellow	BLACK AND YELLOW ROCKFISH	0.2	1	0	0	5	0
	252	Rockfish, black	BLACK ROCKFISH	197.4	104	0	403	194	1
	253	Rockfish, bocaccio	BOCACCIO	14512.9	184	224	1029	79	44
	254	Rockfish, chilipepper	CHILIPEPPER ROCKFISH	68.7	48	0	90	102	2
	255	Rockfish, greenspotted	GREENSPOTTED ROCKFISH	10.6	6	0	4	7	0
	256	Rockfish, starry	STARRY ROCKFISH	2.6	10	0	2	7	0
	257	Rockfish, darkblotched	DARKBLOTCHED ROCKFISH	0.0	0	0	0	0	0
	258	Rockfish, China	CHINA ROCKFISH	78.2	68	1	48	147	4
	259	Rockfish, yellowtail	YELLOWTAIL ROCKFISH	287.5	116	0	868	223	11
	263	Rockfish, gopher	GOPHER ROCKFISH	232.4	95	0	35	51	0
	264	Rockfish, pinkrose	PINKROSE ROCKFISH	0.0	0	0	0	0	0
	265	Rockfish, yelloweye[2]	YELLOWEYE ROCKFISH	774.8	175	27	108	99	0
	267	Rockfish, brown	BROWN ROCKFISH	981.3	246	9	186	111	3
	268	Rockfish, rosy	ROSY ROCKFISH	0.8	3	1	7	14	0
	269	Rockfish, widow	WIDOW ROCKFISH	12575.6	75	132	18802	374	497

1. Market category 246 is no longer used since whitebelly rockfish is now considered copper rockfish.
2. Market category 265 was redefined from red rockfish to yelloweye rockfish in 1981 by CDFW.

Figure 3: Landed weight (metric tons), number of landed strata (year, quarter, port complex, and gear group), and number of species composition samples by market category and time period. Market categories created after 1990 are not listed (e.g. 678, 679, 964, and 971-976). * "Single-species" market categories are nominal (in name only); landings in these categories often include a mixture of species.

				1978 - 1982			1983 - 1990		
Category	Market Category	Description	Nominal Species or Group	Tons	# Strata	# Samples	Tons	# Strata	# Samples
	270	Rockfish, splitnose	SPLITNOSE ROCKFISH	458.7	93	32	3	7	16
	271	Rockfish, Pacific ocean perch	PACIFIC OCEAN PERCH	175.9	65	0	72	60	0
	650	Rockfish, rougheye	ROUGHEYE ROCKFISH	0.0	0	0	0	0	0
	651	Rockfish, olive	OLIVE ROCKFISH	1.1	7	0	4	32	0
	652	Rockfish, grass	GRASS ROCKFISH	0.1	4	0	0	4	0
	653	Rockfish, pink	PINK ROCKFISH	0.1	1	0	0	4	0
	654	Rockfish, greenstriped	GREENSTRIPED ROCKFISH	0.0	0	0	0	0	0
	655	Rockfish, copper	COPPER ROCKFISH	0.4	9	0	43	77	0
	656	Rockfish, blackspotted	BLACKSPOTTED ROCKFISH	0.0	0	0	0	0	0
	657	Rockfish, flag	FLAG ROCKFISH	0.5	4	0	0	0	0
	658	Rockfish, treefish	TREEFISH	0.0	1	0	0	1	0
	659	Rockfish, kelp	KELP ROCKFISH	0.0	2	0	0	4	0
	660	Rockfish, honeycomb	HONEYCOMB ROCKFISH	0.0	1	0	0	0	0
	661	Rockfish, greenblotched	GREENBLTCHED ROCKFISH	0.1	1	0	0	1	0
	662	Rockfish, bronzespotted	BRONZESPOTTED ROCKFISH	0.0	0	0	0	0	0
	663	Rockfish, bank[3]	BANK ROCKFISH	0.0	1	0	432	54	15
	664	Rockfish, rosethorn	ROSETHORN ROCKFISH	0.0	0	0	0	0	0
	665	Rockfish, blue	BLUE ROCKFISH	176.8	117	0	129	194	2
	666	Rockfish, squarespot	SQUARESPOT ROCKFISH	0.0	0	0	0	0	0
	667	Rockfish, blackgill	BLACKGILL ROCKFISH	9.0	3	1	1213	206	128
	668	Rockfish, stripetail	STRIPETAIL ROCKFISH	0.0	0	0	0	0	0
	669	Rockfish, speckled	SPECKLED ROCKFISH	0.2	2	0	0	2	0
	670	Rockfish, swordspine	SWORDSPINE ROCKFISH	0.0	0	0	0	0	0
	671	Rockfish, calico	CALICO ROCKFISH	0.0	0	0	0	0	0
	672	Rockfish, shortbelly	SHORTBELLY ROCKFISH	2.5	2	0	52	11	1
	673	Rockfish, chameleon	CHAMELEON ROCKFISH	0.0	0	0	0	1	0
	674	Rockfish, aurora	AURORA ROCKFISH	0.0	0	0	0	0	0
	675	Rockfish, redbanded	REDBANDED ROCKFISH	0.0	0	0	1	1	0
	676	Rockfish, Mexican	MEXICAN ROCKFISH	0.0	0	0	0	0	0
	677	Rockfish, shortraker	SHORTRAKER ROCKFISH	0.0	0	0	0	0	0
	970	Rockfish, quillback	QUILLBACK ROCKFISH	0.0	0	0	0	0	0

3. Bank rockfish are sometimes referred to as "red widow" rockfish.

Figure 4: (Continued) Landed weight (metric tons), number of landed strata (year, port complex, and gear group), and number of species composition samples by market category and time period. Market categories created after 1990 are not listed (e.g. 678, 679, 964, and 971-976). * "Single-species" market categories are nominal (in name only); landings in these categories often include a mixture of species.

References

CALCOM. (2018). *California Cooperative Groundfish Survey Database: CDFG.* PSMFC, Belmont, CA; NMFS, Santa Cruz, CA. Retrieved from http://calcomfish.ucsc.edu

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC. Retrieved 2019-01-24, from https://www.taylorfrancis.com/books/9781439898208 doi: 10.1201/b16018

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press. (Google-Books-ID: UjsgAwAAQBAJ)

Nelder, J., & McCullagh, P. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall/CRC.

Sen, A. R. (1984). Sampling commercial rockfish landings in California.

Sen, A. R. (1986). Methodological Problems in Sampling Commercial Rockfish Landings. *Fishery Bulletin*, *84*(2).

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.