

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

Nick Grunloh<sup>a</sup>, Edward Dick<sup>b</sup>, Don Pearson<sup>b</sup>, John Field<sup>b</sup>, Marc Mangel<sup>a,c</sup>

October 8, 2019

<sup>a</sup> Center for Stock Assessment Research, University of California, Santa Cruz, Mail Stop SOE-2, Santa Cruz, CA 95064, USA.

<sup>b</sup> Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanographic and Atmospheric Administration, 110 McAllister Way, Santa Cruz, CA 95060, USA.

<sup>c</sup> 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

- outline issue
- transition into Motivated from Ole

## Methods

### 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.

- Motivated from Ole, Poisson transition into Overdispersed model
- describe overdispersion concerns
- describe/introduce larger structure of methods section (orders of importance)
  - decide upon a likelihood
  - prior sensitivety
  - linear predictor and random effects

## Likelihood

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

$$Y_{ijklm\nu} \stackrel{i.i.d.}{\sim} f(\theta_{ijklm\nu}, \phi). \quad (1)$$

Here  $\theta_{ijklm\nu}$  is a linear predictor for inferring the mean weight, and  $\phi$  is a nuisance 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 computationally feasible at this time and among these data. See appendix 6.1 for experimentation detail.

- Prior (discussion of sensitivity analysis)
  - basic description
  - variance parameters
  - describe sensitivity analysis
- Linear Predictor
  - basic model
  - describe additions
    - \* Time
    - \* Species Gear?Port?
- describe time chunking
  - 78-82, 83-90, 91-01

## Results

- model comparison across likelihoods
  - MSE, WAIC
  - Posterior predictive spp comp. Violin plots
- Diagnostic plot
- Prior sensitivity
- 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 dirichlet-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 multinomial -> Dirichlet-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 data generating processes to accurately represent higher moments in the data. McCullagh and Nelder (1989, pg. 124) commiserate about the prevalence of overdispersed data in cluster sampling, and 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. 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 attempt to model both the mean and residual variance of the data, with a single parameter for each species. By definition these models do not have additional parameters to model the variance, but rather, residual variances in these models are simply transformations of their mean parameters. Only 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, while producing more accurate measures of center. 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

Among the simplest models for count data are the Poisson and binomial models. Both models are derived under limiting cases for  $\phi$ , and thus under these models  $\phi$  is not an inferred parameter. As a result the Poisson and binomial models rely heavily on their respective data generating processes to accurately represent higher moments in the data. Furthermore, 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  $\phi$  parameter. In these models the the  $\phi$  parameter is 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, while producing more accurate measures of center.

### Poisson Model

$$Y_{ijklm\nu} \stackrel{i.i.d.}{\sim} \text{Poisson}(\theta_{ijklm\nu}) \quad (2)$$

$$\mu_{ijklm\nu} = \textcolor{red}{n}_{ijklm\nu} \exp(\theta_{ijklm\nu}). \quad (3)$$

$$\sigma_{ijklm\nu}^2 = \mu_{ijklm\nu}. \quad (4)$$

### Binomial Model

$$Y_{ijklm\nu} \stackrel{i.i.d.}{\sim} \text{Binomial}(\theta_{ijklm\nu}) \quad (5)$$

$$\mu_{ijklm\nu} = n_{ijklm\nu} \frac{\exp(\theta_{ijklm\nu})}{1 + \exp(\theta_{ijklm\nu})}. \quad (6)$$

$$\sigma_{ijklm\nu}^2 = \mu_{ijklm\nu} \left( 1 - \frac{\mu_{ijklm\nu}}{n_{ijklm\nu}} \right). \quad (7)$$

### Negative Binomial Model

$$Y_{ijklm\nu} \stackrel{i.i.d.}{\sim} \text{NB}(\theta_{ijklm\nu}, \phi) \quad (8)$$

$$\mu_{ijklm\nu} = \textcolor{red}{n}_{ijklm\nu} \exp(\theta_{ijklm\nu}). \quad (9)$$

$$\sigma_{ijklm\nu}^2 = \mu_{ijklm\nu} \left( 1 + \frac{\mu_{ijklm\nu}}{\phi} \right). \quad (10)$$

### Beta-Binomial Model

$$Y_{ijklm\nu} \stackrel{i.i.d.}{\sim} \text{BB}(Y_{ijklm\nu} | \theta_{ijklm\nu}, \phi) \quad (11)$$

$$\mu_{ijklm\nu} = n_{ijklm\nu} \frac{\exp(\theta_{ijklm\nu})}{1 + \exp(\theta_{ijklm\nu})}. \quad (12)$$