`agTrend`: An `R` package for estimating trends of aggregated abundance

Devin S. Johnson[1] and Lowell Fritz

National Marine Mammal Laboratory, Alaska Fisheries Science Center
NOAA National Marine Fisheries Service,
Seattle, Washington, U.S.A.
[1]Email: *devin.johnson@noaa.gov*

Summary.

1. We describe an open source `R` package `agTrend` for analyzing regional trends of abundance from sites with uneven sample schedules.

2. The package `agTrend` uses an abundance summary approach to estimate trends. By considering abundance trends in this fashion, rather than a model parameter, we can easily augment missing observations to calculate aggregated abundance and trends.

3. The package uses two hierarchical models to augment missing abundance measurements, while accounting for survey methodology changes and variability due to survey replication. A zero-inflated log-normal distribution is used to model abundance (normalized for methodology changes) and a log-normal distribution to model the observed abundance conditional on the true normalized abundance.

4. The use of `agTrend` is demonstrated with an analysis for regional abundance index trends of Steller sea lions (*Eumitopias jubatus*) in Alaska.

5. The package will be of most use to ecologists and resource managers interested in estimating regional trends of abundance surveys aggregated over several sites when sites have not been surveyed at concurrent times. Hence, regional abundance measurements cannot be directly calculated.

Key words: Abundance, Data augmentation, Hierarchical model, Population growth, Trends

1 Introduction

Estimating trends in growth is a central tenet in management of ecological populations. In many instances, e.g., government agencies, it is often a legal requirement (Hovestadt and Nowicki, 2008). Traditionally, growth trends are defined as the average change in log-abundance over a time period of interest (Humbert et al., 2009). There are many methods for estimating trends. The most common include: simple linear regression of log-abundance (Caughley, 1977), state-space modeling (Holmes and Fagan, 2002; Dennis et al., 2006; Humbert et al., 2009), and Bayesian hierarchical modeling (Sauer and Link, 2002; Link and Sauer, 2002; Ver Hoef and Frost, 2003). While, all of these methods have benefits, they have one downfall that is a major stumbling block for large-scale monitoring programs, trend is estimated from a slope coefficient in the model. Thus, estimating abundance trends of regionally aggregated sites can prove difficult if sites are not surveyed in unison. We propose a methodology and software to overcome this problem by treating a trend as a summary of abundance, rather than a model parameter. To make this method widely available to ecologists we created the add-on package `agTrend` for the `R` statistical environment (R Development Core Team, 2013). The package is available from GitHub (`HTTP://nmml.github.io/agTrend`). There are links and directions for installation on the project website.

In large monitoring programs aggregating site-level abundance into regional abundance can be problematic because sites may not be surveyed in the same years. Thus, site-level abundance cannot simply be "summed up" to form regional abundance observations. Moreover, for long-term studies, survey methods can change prohibiting direct comparison of abundance across years. Hierarchical models can be used to estimate regional-level trends and correct for changing methodology, however, the parameter is interpreted as the average trend which is not the same as the trend of the regional total abundance. Small sites are weighted equally with large sites in assessing the average trend (Ver Hoef and Frost, 2003).

To circumvent this problem, we take an approach initially suggested by Link and Sauer (2002), but garnering little subsequent attention, for estimating a "composite trend" over several sites. Using Bayesian Markov Chain Monte Carlo (MCMC) methods (see Givens and Hoeting 2005) and a hierarchical model to augment missing site data we can summarize the posterior distribution of any function of the regional abundance, including trends, without referring to a model parameter. In addition, Bayesian MCMC methods have many benefits for trend analysis beyond data augmentation including the ability to make statements such as "the probability that the regional population is declining at more that $x\%$ is y" (Wade, 2000).

The augmentation procedure used by `agTrend` is based on two hierarchical processes,

the observation process and the abundance process. The observation model accounts for changes in survey methodology or environmental conditions over the course of the monitoring program that affect the observed abundance, say n_{ij}, for site $i = 1, \ldots, I$, at time $t_j, j = 1, \ldots, t_J$, The abundance process models the *normalized* abundance, N_{ij}. The normalized abundance refers to the abundance that would be observed had the surveys been conducted under what could be termed "ideal" conditions. The normalization allows for proper comparisons of abundance across years. In addition, if n_{ij} is an estimate of N_{ij} (e.g., Johnson et al. 2013), then N_{ij} can be interpreted as a true abundance at site i in year j if it were conducted under ideal conditions. Ver Hoef and Frost (2003) use a similar approach to correct harbor seal (*Phoca vitulina*) surveys to ideal conditions. The regional (aggregated) abundance is defined as $\tilde{N}_j = \sum_i N_{ij}$.

If $\mathbf{N} = (N_{11}, N_{12}, \ldots, N_{IJ})'$ were observed in its entirety, one could directly calculate the regional abundance, $\tilde{\mathbf{N}} = (\tilde{N}_1, \ldots, \tilde{N}_J)'$, and summarize average growth with the function $r(\tilde{\mathbf{N}})$, where $r(\cdot)$ could be the least-squares slope of $\log \tilde{\mathbf{N}}$ over the time period of interest. However, $\mathbf{N}$ is only partially observed, at best. Therefore, we have to account for the uncertainty of the missing observations, changes in survey methodology, and measurement error. The `agTrend` package uses an MCMC procedure to augment the missing components to fully account for each source of uncertainty.

The remainder of the paper is organized as follows. In the following section we describe the specifics of the site-level models used by `agTrend` to augment missing data. We also discuss the posterior inference `agTrend` produces. In the last section, we illustrate the use of `agTrend` with data from a monitoring program for the endangered western Steller sea lion (*Eumitopias jubatus*) in Alaska.

2 Methods

2.1 Abundance augmentation models

For the site augmentation models we have chosen to use a zero-inflated log-normal model rather than a traditional count data model, e.g., Poisson, because it is more flexible with respect to count over- and *under*- dispersion. In the example case-study of Steller sea lion trends in this paper under-dispersion may occur at large rookery sites due to the philopatric mating behavior of sea lions. O'Hara and Kotze (2010) argue against the use of log transformations and normal models, but, the biases of their results faded for abundances beyond 20. Typically, monitoring programs involve abundance larger than 20. Further, the source of bias for small abundances is induced by the "fudge factor" (O'Hara and Kotze, 2010) necessary for transformation of abundances of zero, i.e., $y_{ij} = \log(n_{ij}+c)$, where $c \leq 1$ can be chosen arbitrarily. This problem is handled in `agTrend` through the

use of a zero-inflated (ZI) version of the log-normal distribution.

The first model in the hierarchy is the observation model,

$$[n_{ij} \mid N_{ij}, \mathbf{x}_{ij}, \boldsymbol{\gamma}, \sigma_{ij}] = \begin{cases} \mathcal{L}(\mathbf{x}_{ij}'\boldsymbol{\gamma} + \ln N_{ij}, \sigma_{ij}^2) & \text{if } N_{ij} > 0 \\ 0 & \text{if } N_{ij} = 0 \end{cases}, \tag{1}$$

where $\mathcal{L}(\mu, \sigma^2)$ represents a log-normal distribution with location parameter μ and scale parameter σ^2, $\mathbf{x}_{ij}$ are a set of adjustment covariates related to, e.g., environmental conditions or survey methodology, $\boldsymbol{\gamma}$ are the adjustment coefficients, and N_{ij} is the standardized or true abundance. Holmes and Fagan (2002) considered corruption of the data to be random and controlled by σ_{ij}^2; here, we consider the fact that data can become "corrupted" systematically by survey methodology changes over time. In the present version of `agTrend` all σ_{ij}^2 are considered known. If n_{ij} is an estimate of N_{ij} then the estimated standard error of n_{ij} can be given as an argument (e.g., Johnson et al. 2013). Otherwise, `agTrend` sets $\sigma_{ij}^2 = 1.0E-8$ by default, so that $n_{ij} \approx N_{ij} e^{\mathbf{x}_{ij}\boldsymbol{\gamma}}$.

The ZI log-normal model for N_{ij} is given by

$$[N_{ij} \mid \beta_{i0}, \beta_{i1}, \eta_{ij}] = \begin{cases} \mathcal{L}(\beta_{i0} + \beta_{i1}t + \eta_{ij}, \zeta_i^{-1}) & \text{with prob. } p_{ij} \\ 0 & \text{with prob. } 1 - p_{ij} \end{cases}, \tag{2}$$

where β_{i0} and β_{i1} are local linear coefficients, $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iJ})'$ is a random walk of order 2 (RW2; Rue and Held 2005), ζ_i is the local precision parameter, and p_{ij} is the local ZI probability. The RW2 model is used to induce autocorrelation as well as model curvature. The RW2 process is a very smooth time-series approximating a cubic spline (Speckman and Sun, 2003). Finally, the ZI probability is modeled via

$$\text{probit}(p_{ij}) = \theta_{i0} + \theta_{i1}t + \alpha_{ij}, \tag{3}$$

where θ_{i0} and θ_{i1} are local linear trend coefficients and $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iJ})'$ is another RW2 process.

2.2 Bayesian inference

In order to augment the missing N_{ij} and allow for arbitrary trend summaries, we take a Bayesian approach to inference. The MCMC algorithm that `agTrend` uses is summarized in Appendix S1. The MCMC sampler draws a realization from the posterior distribution

$$[\mathbf{N}, \boldsymbol{\varphi} \mid \mathbf{n}, \mathbf{x}] \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \left\{ [n_{ij} \mid N_{ij}, \mathbf{x}_{ij}, \boldsymbol{\gamma}, \sigma_i]^{s(i,j)} [N_{ij} \mid \beta_{i0}, \beta_{i1}, \eta_{ij}] \right\} [\boldsymbol{\varphi}], \tag{4}$$

where $\boldsymbol{\varphi} = (\boldsymbol{\gamma}', \boldsymbol{\theta}', \boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}', \boldsymbol{\zeta})$ is the vector of parameters, $[\boldsymbol{\varphi}]$ is the prior distribution of the parameters, and $s(i, j)$ is an indicator function that equals 1 if site i was surveyed at

time t_j. The parameter prior used by `agTrend` is specified from the following independent priors, where $\mathcal{N}(\mu, \Sigma)$ denotes a normal distribution (of appropriate dimension) with mean μ and variance Σ, and $\mathcal{G}(a, b)$ denotes a gamma distribution with shape a and scale b,

- $[\boldsymbol{\gamma}] = \mathcal{N}(\boldsymbol{\gamma}^{(0)}, \mathbf{Q}_{\gamma}^{-1})$,
- $[\beta_{i0}, \beta_{i1}] = \mathcal{N}(\beta_{i0}^{(0)}, Q_{i0}^{-1}) \times \mathcal{N}(\beta_{i1}^{(0)}, Q_{i1}^{-1})$,
- $[\boldsymbol{\eta}_i] = \mathcal{N}(\mathbf{0}, \tau^{-1}\mathbf{Q}_{RW2}^{-})$ (Note: RW2 process with parameter τ, $\mathbf{Q}_{RW2}$ is a fixed constant matrix, see for details),
- $[\tau] = \mathcal{G}(a_\tau, b_\tau)$,
- $[\zeta_i] = \mathcal{G}(a_\zeta, b_\zeta)$,
- $[\theta_{i0}, \theta_{i1}] = \mathcal{N}(\theta_{i0}^{(0)}, Q_{i0}^{-1}) \times \mathcal{N}(\theta_{i1}^{(0)}, Q_{i1}^{-1})$ (Note: precision parameters not necessarily the same as $\boldsymbol{\beta}_i$ priors),
- $[\boldsymbol{\alpha}_i] = \mathcal{N}(\mathbf{0}, \phi^{-1}\mathbf{Q}_{RW2}^{-})$,
- $[\phi] = \mathcal{G}(a_\phi, b_\phi)$.

We have parameterized the prior distributions using precision rather than variance because this allows the user to set, say $Q_{i0} = 0$, to specify a flat prior distribution for β_{i0}. By default, `agTrend` sets all precision parameters to zero in normal priors for coefficients and, for gamma priors, $a = 0.5$ and $b = 0.00005$.

As it stands, the method described above has a built-in sample-size correction, in a sense, for trend inference. That is to say, if n_{ij} is observed for nearly all (i, j), the methodology correction is small, i.e., $\boldsymbol{\gamma} \approx \mathbf{0}$, and the observation variance becomes small, i.e., $\sigma_{ij}^2 \approx 0$, then $Var\{r(\tilde{\mathbf{N}}) \mid \mathbf{n}\} \approx 0$. We term this the *realized* trend because it is a summary of the realized (or nearly so) N_{ij}. It is usually desired to maintain inference over replications of N_{ij} as is the case with traditional regression analysis. Therefore, by default, `agTrend` uses the posterior *predictive* distribution of $\mathbf{N}$ to make trend inference.

The posterior predictive distribution is the Bayesian version of the frequentist notion of sample replication (Gelman et al., 1996). The posterior distribution for abundance is given by

$$[\mathbf{N}^{(p)} \mid \mathbf{n}, \mathbf{x}] \propto \int [\mathbf{N}^{(p)}, \boldsymbol{\varphi} \mid \mathbf{n}, \mathbf{x}]\ [\boldsymbol{\varphi} \mid \mathbf{n}, \mathbf{x}]\ d\boldsymbol{\varphi}, \tag{5}$$

where $N_{ij}^{(p)}$ represents the abundance that would be realized if the abundance process were replicated. A sample from the posterior predictive distribution of every $N_{ij}^{(p)}$ is accomplished within the MCMC at each iteration by using the process model (eqns. 2 and 3; see Appendix S1). For sparsely surveyed populations or populations where $\sigma_{ij}^2 > 0$

and $\boldsymbol{\gamma}$ is uncertain, there will be little or no difference between the realized and the predicted trend inference.

3 Example: Steller sea lion population trends

To illustrate the use of the `agTrend` package, we analyze data from aerial survey monitoring program for western Steller sea lions in Alaska. These data are included with the package, so, this example can be recreated by the reader. In addition, this example, among others, is provided as an `R` demo with the package. Users can type `demo(wdpsNonpups)` after loading the package to run the demo. This will run the entire MCMC augmentation and aggregation, so, it will take a few hours depending on the platform. The example herein does not provide detailed explanation of all the function arguments, but users can type `help(package="agTrend")` in the `R` console window to bring up the manual files for the package.

3.1 Data preparation

The data we are using are the counts of nonpups at rookery and haul-out sites in the western Distinct Population Segment (wDPS; Fritz et al. 2013). These data are included with the package and can be loaded via `data(wdpsNonpups)`. Before we begin the analysis we are going to subset the raw data so we use only those surveys from 1990–2012. Prior to 1990 surveys were sparse, providing only sporadic information. In addition, we also removed those sites that had < 2 nonzero counts over the period. For this example, we are interested in trends within each of the six regions given in the `Region` column.

Prior to 2004, all surveys were conducted by counting animals in photographs taken by hand from an airplane at oblique angles. Starting with the 2004 surveys animals were counted from vertical medium-format photographs. The vertical high resolution photos in the modern surveys tend to produce slightly higher counts (Fritz and Stinchcomb, 2005). Therefore, we add an indicator, `obl`, for surveys using the oblique photo method. The first few rows of the relevant data columns are shown below

```
                site year Region count obl
ADAK/ARGONNE POINT 1992 C ALEU     0   1
ADAK/ARGONNE POINT 1994 C ALEU     0   1
ADAK/ARGONNE POINT 1996 C ALEU   141   1
ADAK/ARGONNE POINT 1998 C ALEU    43   1
ADAK/ARGONNE POINT 2000 C ALEU     8   1
```

3.2 Site-level augmentation models

Now we specify the models used to augment the missing abundance at each site. Not all of the sites have sufficient data to estimate the parameters in the full nonparametric ZI log-normal model. The augmentation function `mcmc.aggregate()` in `agTrend` requires the user to provide a data frame with each row representing a site and two columns giving the models for the trend portion and the ZI portion respectively. In this analysis we used a constant trend, i.e., $\beta_{i1} = 0$, for sites with 5 or fewer nonzero observations. For sites with 6–10 nonzero observations a linear model was used, i.e., $\eta_i = \mathbf{0}$. Finally, for sites with > 10 nonzero observations, the full nonparametric trend was used.
There were not a large number of survey years, so, we elected to use only linear or constant models for the zero-inflation portion, $\boldsymbol{\alpha}_i = \mathbf{0}$. Using the full nonparametric model tended to produce overfitted zero-inflation probabilities. If the number of surveys was 1–5, we set $\theta_{i1} = 0$. If there were not any zero observations at a site a ZI model was not used, i.e., $p_{ij} = 1$ for $j = 1, \ldots, J$. The first few rows of the `wdpsModels` data frame is given below.

```
               site trend zero.infl
ADAK/ARGONNE POINT   lin       lin
 ADAK/CRONE ISLAND   lin       lin
   ADAK/LAKE POINT   RW2      none
             ADUGAK   RW2      none
AFOGNAK/TONKI CAPE   lin       lin
```

Note, the column names need to be as they are shown: the site name is the same as in `wdpsNonpups`, the local trend models are titled "`trend`", and the zero-inflation models are named "`zero.inf`".

3.3 Prior distributions

Although `agTrend` uses default prior distributions if they are left unspecified, it is not always wise to follow that course. If there is little information for some sites, large posterior variance of local augmentation on the log scale can lead to nonsensical results when they are exponentiated. Occasionally, sites can have augmented abundance well beyond reason. There are two mechanisms in `agTrend` to control this from happening. The first method is to specify sensible priors for the $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ parameters. Second, an upper limit can be specified so that abundance samples are not generated beyond the limit.

There is no information in the `wdpsNonpup` data to identify the photo switch effect. There are not any overlapping surveys that used each method at the same site in the

same year. However, there exist other data from another Steller sea lion survey in Southeast Alaska, (loaded via: `data(photoCorrection)`), where this effect was investigated. We used the mean and standard error of the observed photo change effect to create an informative prior for the single γ parameter.

```
ln.photo.ratio <- log(photoCorrection$X2000OBL/photoCorrection$X2000VERT)
gamma<- list(
  gamma.0=mean(ln.photo.ratio),
  Q.gamma=(length(ln.photo.ratio)-1)/var(ln.photo.ratio))
```

Now, for the $\boldsymbol{\beta}_i$, we defined $[\beta_{i0}] = \text{flat}$ and $[\beta_{i1}] = \mathcal{N}(0, 200)$ for those sites where a linear trend was used for augmentation ("lin" and "RW2"). The chosen precision implies that the rate of growth for any site will remain with $\pm 20\%$. This is sensible for a K-selected, long-lived species such as Steller sea lions. Note, we are using the `Matrix` package to take advantage of sparse matrix algebra in the MCMC sampler.

```
List of 2
 $ gamma:List of 2
  ..$ gamma.0: num -0.039
  ..$ Q.gamma: num 8317
 $ beta :List of 2
  ..$ beta.0: num [1:369] 0 0 0 0 0 0 0 0 0 0 ...
  ..$ Q.beta:Formal class 'ddiMatrix' [package "Matrix"] with 4 slots
  .. .. ..@ diag    : chr "N"
  .. .. ..@ Dim     : int [1:2] 369 369
  .. .. ..@ Dimnames:List of 2
  .. .. .. ..$ : NULL
  .. .. .. ..$ : NULL
  .. .. ..@ x       : num [1:369] 0 200 0 200 0 200 0 200 0 200 ...
```

We used the default priors for the remaining parameters.

We also made use of the upper bound capability in `agTrend`. The upper limit for each site was defined to be $3 \times \max(n_{ij})$. We believed it is unlikely that realistic survey count values would exceed this upper limit. This keeps the simulated abundances within reason. The first few rows of the upper bound data frame are given below.

```
               site upper
 ADAK/ARGONNE POINT   423
  ADAK/CRONE ISLAND   180
    ADAK/LAKE POINT  3024
              ADUGAK  1908
 AFOGNAK/TONKI CAPE    48
```

Note, the upper bound column must be labeled "`upper`."

3.4 Augment counts and estimate trends

Now we can begin drawing samples from the posterior predictive trend distribution using the `mcmc.aggregate()` function.

```
fit <- mcmc.aggregate(start=1990, end=2012, data=wdpsNonpups,
                      obs.formula=~obl-1, model.data=wdpsModels,
                      aggregation="Region", abund.name="count",
                      time.name="year", site.name="site",
                      burn=1000, iter=5000, thin=5,
                      prior.list=prior.list, upper=upper,
                      keep.site.param=TRUE, keep.site.abund=TRUE,
                      keep.obs.param=TRUE)
```

Even though we are only interested in trends from 2000–2012, we used a start time of 1990 to retain all of the predicted $N_{ij}^{(p)}$ to examine and use later. In addition, we set all `keep.* = TRUE` to retain the individual site predictions, parameters, and γ samples. In order to calculate regional trends for just 2000–2012, the `updateTrend()` function is used.

```
trend2000 <- updateTrend(x=fit, start=2000, end=2012, type="pred")
```

Although we do not demonstrate it here, there is also a function, `newAggregation()`, that can be called if readers would like to evaluate trends for another aggregation of sites, e.g., we could analyze the trend for the wDPS as a whole. Also, regional abundance forecasting may be accomplished by setting `end` greater than the maximum survey time.

3.5 Steller sea lion trend results

The results show that there is substantial regional variation in count trends for wDPS Steller sea lions (Table 1 and Figure 1). In the far western portion of the population (W ALEU) the population is declining at approximately -7% yr^{-1}, while in some regions in the eastern portion (W GULF and E GULF) the population is increasing at 4% yr^{-1} or better. In addition, populations in the C ALEU and C GULF seem to be relatively stable with some evidence that C ALEU is declining and C GULF is increasing, though nonsignificantly.

The posterior predictive and realized distributions of aggregated sea lion counts are illustrated in Figure 1. In years where only a few sites are surveyed, the realized distribution of counts closely matches the predictive distribution. An extreme example occurs

in 2012 where only sites in W ALEU were surveyed, so, the predictive and realized distributions are the same in the other regions (aside from Monte Carlo differences visible in Figure 1). Examination of Figure 2 allows us to illustrate the augmenting process for two individual sites. The Glacier haul-out site results (Figure 2a–b) illustrate application of the ZI portion in the process model. Even though Glacier is a relatively large haul-out and seems to be growing there is a nontrivial probability of observing $n_{ij} = 0$ throughout the time period. This is due to the fact that a haul-out is populated with juveniles and adults without pups. They are more susceptible to disturbance, natural or anthropocentric. In contrast, the Marmot rookery (Figure 2c) is a large rookery site where a ZI model was not used because it would be extremely unrealistic to observe a count of zero. The curvature of the predictive envelope, modeled by $\boldsymbol{\eta}_i$, indicates a decreasing rate decline from the 1990s into the 2000s.

Acknowledgments

The findings and conclusions in the paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service, NOAA. Reference to trade names does not imply endoresement by the National Marine Fisheries Service, NOAA.

References

Caughley G. 1977. *Analysis of vertebrate populations*. Wiley New York.

Dennis B, Ponciano JM, Lele SR, Taper ML, Staples DF. 2006. Estimating density dependence, process noise, and observation error. *Ecological Monographs* **76**: 323–341.

Fritz L, Sweeney K, Johnson DS, Lynn M, Gilpatrick J. 2013. Aerial and ship-based surveys of Steller sea lions (*Eumetopias jubatus*) conducted in Alaska in June-July 2008 through 2012, and an update on the status and trend of the western Distinct Population Segment in Alaska. Technical Report NMFS-AFSC-In prep, U. S. Department of Commerce, NOAA, Washington, D. C.

Fritz LW, Stinchcomb C. 2005. Aerial, ship, and land-based surveys of steller sea lions (*Eumetopias jubatus*) in the western stock in alaska, june and july 2003 and 2004. Technical Report NMFS-AFSC-153, U. S. Department of Commerce, NOAA, Washington, D. C.

Gelman A, Meng XL, Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**: 733–807.

Givens G, Hoeting JA. 2005. *Computational Statistics*. New York: Wiley.

Holmes EE, Fagan WF. 2002. Population viability analysis for corrupted data sets. *Ecology* **83**: 2379–2386.

Hovestadt T, Nowicki P. 2008. Process and measurement errors of population size: their mutual effects on precision and bias of estimates for demographic parameters. *Biodiversity Conservation* **17**: 3417–3429.

Humbert JY, Mills LS, Horne JS, Dennis B. 2009. A better way to estimate population trends. *Oikos* **118**: 1940–1946.

Johnson DS, Ream RR, Towell RG, Williams MT, Guerrero JDL. 2013. Bayesian clustering of animal abundance trends for inference and dimension reduction. *Journal of Agricultural Biological and Environmental Statistics* **In press**.

Link W, Sauer J. 2002. A hierarchical analysis of population change with application to cerulean warblers. *Ecology* **83**: 2832–2840.

O'Hara RB, Kotze DJ. 2010. Do not log-transform count data. *Methods in Ecology and Evolution* **1**: 118–122.

R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rue H, Held L. 2005. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Sauer J, Link W. 2002. Hierarchical modeling of population stability and species group attributes from survey data. *Ecology* **83**: 1743–1751.

Speckman P, Sun D. 2003. Fully bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika* **90**: 289–302.

Ver Hoef JM, Frost KJ. 2003. A Bayesian hierarchical model for monitoring harbor seal changes in Prince William Sound, Alaska. *Environmental and Ecological Statistics* **10**: 201–219.

Wade P. 2000. Bayesian methods in conservation biology. *Conservation Biology* **14**: 1308–1316.

Table 1: Regional trend estimates for 2000–2012. The trend estimates are given in present growth form (i.e., $\lambda = 100(e^r - 1)$). The columns are the posterior median, lower, and upper 95% highest probability density credible intervals of λ.

	Median	Lower CI	Upper CI
W ALEU	-7.23	-9.04	-5.56
C ALEU	-0.56	-1.45	0.43
E ALEU	2.39	0.92	3.94
W GULF	4.01	2.49	5.42
C GULF	0.87	-0.34	2.18
E GULF	4.51	1.63	7.58

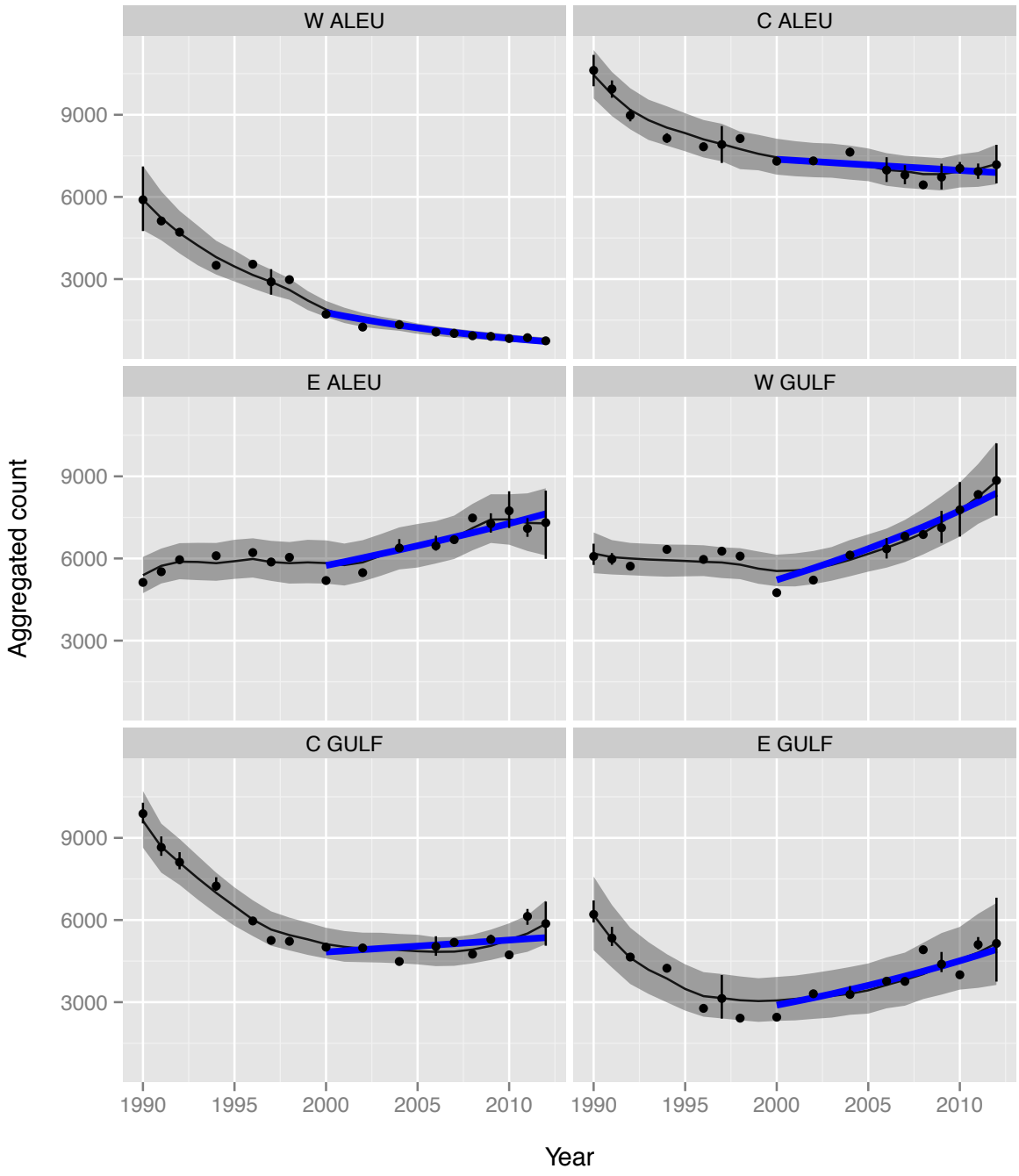

Figure 1: Predictive distribution of aggregated abundance and trends. The grey envelope is the 90% highest probability density credible interval of the posterior predictive counts. The points and error bars represent the observed counts with augmented missing values (realized count distribution). The blue lines are the fitted least-squares predictive trend. The black line is the median of the posterior predictive counts.

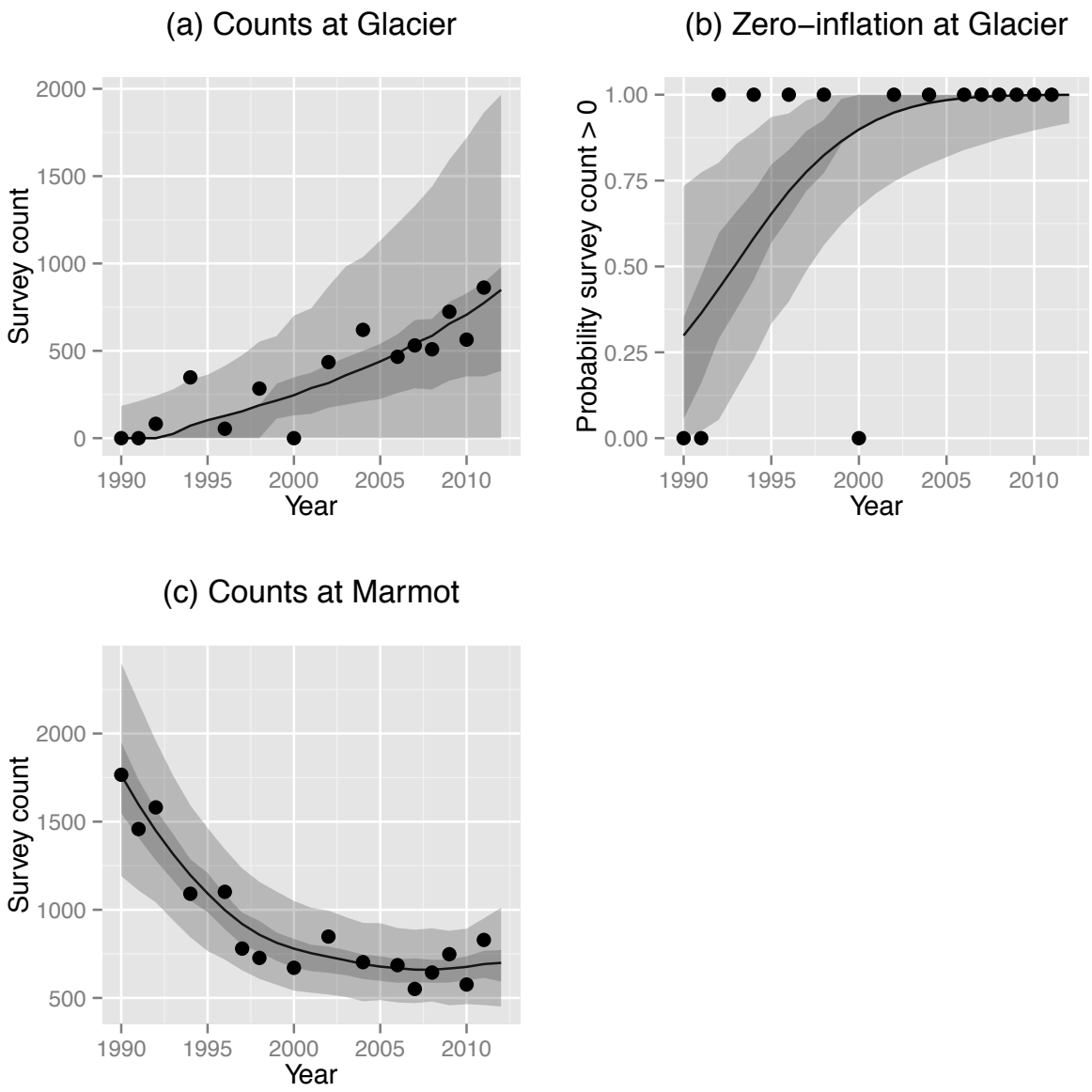

Figure 2: Predicted survey count at the Glacier haul-out and Marmot rookery. The dark and light grey envelopes are 50% and 90% highest probability density credible intervals, respectively. For plots (a) and (c) the points are the observed counts, n_{ij}, while for plot (b), the points are indicators of positive counts. The black line is the median posterior predictive counts for (a) and (c), while in (b) it is the median p_{ij}.

S1: MCMC details for the `R` package `agTrend`

Supplementary information for "`agTrend`: An `R` package for estimating trends of aggregated abundance"

Devin S. Johnson[1] and Lowell Fritz

National Marine Mammal Laboratory, Alaska Fisheries Science Center
NOAA National Marine Fisheries Service,
Seattle, Washington, U.S.A.
[1]Email: *devin.johnson@noaa.gov*

S1.1 Data augmentation models

The more mathematical details of this appendix require that we reparameterize the data augmentation models in vector form. The MCMC code in `agTrend` uses vectorized versions of the model to speed computation. We add the following notation to that given in the main portion of the paper. The dimensions are relatively starightforward to assertain and well as the entries for the design matrices, thus, we avoid giving the details here, but Johnson et al. (2013) provides a similar model for which a more detailed description is given.

The process model is defined by two submodels, (1) the positive count model and (2) the zero-inflation model. To construct this model in vector form, we break the realized abundance, N_{ij} into a function of two random variables q_{ij}, an indicator that $N_{ij} > 0$, and z_{ij}, the log count if $q_{ij} = 1$. The normalized abundance is given by

$$N_{ij} = q_{ij} \exp\{z_{ij}\}. \tag{1}$$

To model q_{ij} we use the probit method of Albert and Chib (1993). So, define the vector $\tilde{\mathbf{q}}_i$, where

$$[\tilde{\mathbf{q}}_i] = \mathcal{N}(\mathbf{T}\boldsymbol{\theta}_i + \boldsymbol{\alpha}_i, \mathbf{I}),$$

and

- $\mathbf{T}$ is a $J \times 2$ matrix with ones in the first column and sequential times from $t_1, \ldots, t_J$,
- $\boldsymbol{\theta}_i = (\theta_{i0}, \theta_{i1})'$

- $[\boldsymbol{\alpha}_i] = [(\alpha_{i1}, \ldots, \alpha_{iJ})'] = \mathcal{N}(\mathbf{0}, \phi_i^{-1}\mathbf{Q}_{RW2}^{-1})$, where $\mathbf{Q}_{RW2}$ is the precision weights matrix for a random walk of order 2 (see Rue and Held 2005).

Now, defining q_{ij} to be the indicator that $\tilde{q}_{ij} > 0$ implies the probability, p_{ij}, that $q_{ij} = 1$ is described by the probit regression model,

$$\text{probit}(p_{ij}) = \theta_{i0} + \theta_{i1}t_j + \alpha_{ij}.$$

For all times, $t_1, \ldots, t_J$, we define the positive portion of abundance via the vector eqution

$$\mathbf{z}_i = \mathbf{T}\boldsymbol{\beta}_i + \boldsymbol{\eta}_i + \boldsymbol{\delta}_i,$$

where

- $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1})'$,
- $[\boldsymbol{\eta}_i] = [(\eta_{i1}, \ldots, \eta_{iJ})'] = \mathcal{N}(\mathbf{0}, \tau_i^{-1}\mathbf{Q}_{RW2}^{-1})$
- $[\boldsymbol{\delta}_i] = \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\boldsymbol{\delta}_i}^{-1})$, where $\mathbf{Q}_{\boldsymbol{\delta}_i} = \zeta_i\mathbf{I}$.

It does not matter that we observe $n_{ij} = 0$ which implies $N_{ij} = 0$ for some i and j, because $q_{ij} = 0$ for those observations. Thus, z_{ij} is unidentafiable and the posterior distribution will only depend only on the model structure itself and the posterior distribution of the parameters. So, it will not affect the parameter inference.

The vectorized version of the observation model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{M}\mathbf{z} + \boldsymbol{\epsilon},$$

where

- $\mathbf{y}$ is the vector of $\log n_{ij}$ such that $n_{ij} > 0$ ($\mathbf{y}$ will be shorter than $\mathbf{n}$ if some $n_{ij} = 0$),
- $\mathbf{z} = (\mathbf{z}_1', \ldots, \mathbf{z}_I')'$,
- $\mathbf{M}$ is a matrix that selects the appropriate z_{ij} to match y_{ij}
- $[\boldsymbol{\epsilon}] = \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\boldsymbol{\epsilon}}^{-1})$,
- $\mathbf{Q}_{\boldsymbol{\epsilon}}$ is a diagonal matrix with the appropriate σ_{ij}^{-2} at the entry corresponding to y_{ij}.

S1.2 MCMC posterior sampling

The full conditional disctributions used for MCMC sampling depend on the prior distributions selected. The `agTrend` package uses the following prior distributions, where $\mathcal{N}$ and $\mathcal{G}$ denote a normal and gamma distribution respectively.

- $[\boldsymbol{\gamma}] = \mathcal{N}(\boldsymbol{\gamma}_0, \mathbf{Q}_{\boldsymbol{\gamma}})$
- $[\boldsymbol{\beta}_i] = \mathcal{N}(\boldsymbol{\beta}_i^{(0)}, \mathbf{Q}_{\boldsymbol{\beta}_i}^{-1})$
- $[\tau_i] = \mathcal{G}(a_\tau, b_\tau)$
- $[\zeta_i] = \mathcal{G}(a_\zeta, b_\zeta)$
- $[\boldsymbol{\theta}_i] = \mathcal{N}(\boldsymbol{\theta}_i^{(0)}, \mathbf{Q}_{\boldsymbol{\theta}_i}^{-1})$
- $[\phi_i] = \mathcal{G}(a_\phi, b_\phi)$

S1.2.1 $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_i$, $\boldsymbol{\eta}_i$, and $\mathbf{z}_i$ updates

All of the full conditional distributions for MCMC updates of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_i$, $\boldsymbol{\eta}_i$, and $\mathbf{z}_i$ are multivariate normal distributions of the form $\mathcal{N}(\mathbf{D}^{-1}\mathbf{d}, \mathbf{D}^{-1})$. The $\mathbf{D}$ and $\mathbf{d}$ values for each parameter are given in the following table.

Parameter	$\mathbf{D}$	$\mathbf{d}$
$\boldsymbol{\gamma}$	$\mathbf{X}'\mathbf{Q}_{\boldsymbol{\epsilon}}\mathbf{X} + \mathbf{Q}_{\boldsymbol{\gamma}}$	$\mathbf{X}'\mathbf{Q}_{\boldsymbol{\epsilon}}(\mathbf{y} - \mathbf{M}\mathbf{z}) + \mathbf{Q}_{\boldsymbol{\gamma}}\boldsymbol{\gamma}_0$
$\boldsymbol{\beta}_i$	$\mathbf{T}'\mathbf{Q}_{\boldsymbol{\delta}_i}\mathbf{T} + \mathbf{Q}_{\boldsymbol{\beta}_i}$	$\mathbf{T}'\mathbf{Q}_{\boldsymbol{\delta}_i}(\mathbf{z}_i - \boldsymbol{\eta}_i) + \mathbf{Q}_{\boldsymbol{\beta}_i}\boldsymbol{\beta}_i^{(0)}$
$\boldsymbol{\eta}_i$	$\mathbf{Q}_{\boldsymbol{\delta}_i} + \tau_i\mathbf{Q}_{RW2}$	$\mathbf{Q}_{\boldsymbol{\delta}_i}(\mathbf{z}_i - \mathbf{T}\boldsymbol{\beta}_i)$
$\mathbf{z}_i$	$\mathbf{M}_i'\mathbf{Q}_{\boldsymbol{\epsilon}}\mathbf{M}_i + \mathbf{Q}_{\boldsymbol{\delta}_i}$	$\mathbf{M}_i'\mathbf{Q}_{\boldsymbol{\epsilon}}(\mathbf{y} - \mathbf{X}\boldsymbol{\gamma})_i + \mathbf{Q}_{\boldsymbol{\delta}_i}(\mathbf{T}\boldsymbol{\beta}_i + \boldsymbol{\eta}_i)$

The matrix $\mathbf{M}_i$ is the submatrix of $\mathbf{M}$ created by extracting those rows associated with the ith site. Also, $(\mathbf{y} - \mathbf{X}\boldsymbol{\gamma})_i$ denotes those etries of $\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}$ associated with the ith site. Recall that $\mathbf{z}_i$ is updated for all j, even in the case $n_{ij} = 0$.

S1.2.2 τ_i and ζ_i updates

All τ_i and ζ_i updates are are of the form $\mathcal{G}(c, d)$. The following table gives the values for c and d for each parameter

Parameter	c	d
τ_i	$a_\tau + (J-2)/2$	$b_\tau + \boldsymbol{\eta}_i' \mathbf{Q}_{RW2} \boldsymbol{\eta}_i / 2$
ζ_i	$a_\zeta + J/2$	$b_\zeta + (\mathbf{z} - \mathbf{X}_z \boldsymbol{\beta}_i - \boldsymbol{\eta}_i)'(\mathbf{z} - \mathbf{X}_z \boldsymbol{\beta}_i - \boldsymbol{\eta}_i)/2$

S1.2.3 $\tilde{q}_{ij}$, $\boldsymbol{\theta}_i$, and $\boldsymbol{\alpha}_i$ updates

Updating the components of the zero-inflation portion of the model is slightly more challeging than the abundance portion due to the fact $\mathbf{q}_i = (q_{i1}, \ldots, q_{iJ})'$ is partially observed. Therefore, we break sampling of $\tilde{q}_{ij}$ into three parts.

case 1: $n_{ij} = 0$ **observed**

This observation implies that $q_{ij} = 0$ and $\tilde{q}_{ij} < 0$. Thus the full conditional distribution for sampling is $[\tilde{q}_{ij} \mid \ldots] = \mathcal{N}_{(-\infty,0)}(\theta_{i0} + \theta_{i1} t_j + \alpha_{ij}, 1)$, where $\mathcal{N}_{(l,u)}$ refers to a truncated normal distribution bounded below by l and above by u.

case 2: $n_{ij} > 0$ **observed**

This observation implies that $q_{ij} = 1$ and $\tilde{q}_{ij} > 0$. Thus the full conditional distribution for sampling is $[\tilde{q}_{ij} \mid \ldots] = \mathcal{N}_{(0,\infty)}(\theta_{i0} + \theta_{i1} t_j + \alpha_{ij}, 1)$

case 3: n_{ij} **not observed**

In this instance there is no data to condition on in the full conditional, so we follow the model definition to update. The full conditional distribution is $[\tilde{q}_{ij} \mid \ldots] = \mathcal{N}(\theta_{i0} + \theta_{i1} t_j + \alpha_{ij}, 1)$.

For the current valueof $\tilde{\mathbf{q}}_i$ the zero-inflation parameters $\boldsymbol{\theta}_i$, $\boldsymbol{\alpha}_i$, and ϕ_i are updated analogous to $\boldsymbol{\beta}_i$, $\boldsymbol{\eta}_i$, and τ_i. The $\mathcal{N}(\mathbf{D}^{-1}\mathbf{d}, \mathbf{D}^{-1})$ updates for $\boldsymbol{\theta}_i$ and $\boldsymbol{\alpha}_i$ are given in the following table.

Parameter	$\mathbf{D}$	$\mathbf{d}$
$\boldsymbol{\theta}_i$	$\mathbf{T}'\mathbf{T} + \mathbf{Q}_{\boldsymbol{\theta}_i}$	$\mathbf{T}'(\tilde{\mathbf{q}}_i - \boldsymbol{\alpha}_i) + \mathbf{Q}_{\boldsymbol{\theta}_i} \boldsymbol{\theta}_i^{(0)}$
$\boldsymbol{\alpha}_i$	$\mathbf{I} + \phi_i \mathbf{Q}_{RW2}$	$(\tilde{\mathbf{q}}_i - \mathbf{T}\boldsymbol{\theta}_i)$

The precision parameter is updated via $[\phi_i \mid \ldots] = \mathcal{G}(a_\phi + (J-2)/2, b_\phi + \boldsymbol{\alpha}_i' \mathbf{Q}_{RW2} \boldsymbol{\alpha}_i / 2)$, analogous to τ_i.

S1.3 Posterior augmentation and trends

To seemlesly aggregate abundance and summarize trends `agTrend` samples from the posterior distribution of the missing N_{ij}. Using the samples drawn in the previous sections, drawing samples for unobserved N_{ij} are strightforward. Given q_{ij} and z_{ij} we can simply cacluate

$$N_{ij} = q_{ij} e^{z_{ij}}. \tag{2}$$

This will give us the *realized* abundance distribution for all i and j. However, recall that if nearly all sites are observed in years where surveys take place, $\sigma^2_{ij} \approx 0$, and $\boldsymbol{\gamma} \approx \mathbf{0}$, then $Var[r(\mathbf{N})] \approx 0$. Thus, we may want to incorporate uncertainty due to replicated survey effort. Thus, we need to incorporate the uncertainty of survey process via the posterior predictive distribution of abundance.

The posterior predictive abundance is sampled via $N^{(p)}_{ij} = q^{(p)}_{ij} \exp\{z^{(p)}_{ij}\}$ where

- $q^{(p)}_{ij}$ is the indicator that $\tilde{q}^{(p)}_{ij} > 0$,
- $[\tilde{\mathbf{q}}^{(p)}_i \mid \dots] = \mathcal{N}(\mathbf{T}\boldsymbol{\theta}_i + \boldsymbol{\alpha}_i, \mathbf{I})$, and
- $[\tilde{\mathbf{z}}^{(p)}_i \mid \dots] = \mathcal{N}(\mathbf{T}\boldsymbol{\beta}_i + \boldsymbol{\eta}_i, \zeta_i^{-1}\mathbf{I})$.

Upon examination of the q_{ij} and z_{ij} updates, the reader can see that these are the same samples for times and sites with missing n_{ij}. The only difference is that we are sampling $N^{(p)}_{ij}$ for all sites and times regardless of whether n_{ij} was observed. Thus, replicating the survey effort. If sampling is sparse, the inference will be similar.

After augmentation/prediction, the abundances are aggregated from sites in to regional abundance $\tilde{N}_j = \sum_i N^{(p)}_{ij}$. From this point the log abundance is calculated, $\tilde{\mathbf{z}} = \log \tilde{\mathbf{N}}$. Finally, the trend is calculated as

$$r = r(\tilde{\mathbf{N}}) = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\tilde{\mathbf{z}}, \tag{3}$$

i.e., the least-squares summary of average change in log abundance. After the MCMC algorithm is complete, trend inference is calculated as the median r and 95% HPD CI is also calculated.

References

Albert J, Chib S. 1993. Bayesian-analysis of binary and polychotomous reponse data. *Journal of the American Statistical Association* **88**: 669–679.

Johnson DS, Ream RR, Towell RG, Williams MT, Guerrero JDL. 2013. Bayesian clustering of animal abundance trends for inference and dimension reduction. *Journal of Agricultural Biological and Environmental Statistics* **In press**.

Rue H, Held L. 2005. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.