Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(2) 2010 171-175

A Finite Population Bayes Procedure for Censored Categorical Abundance Data

Mark D. Holland[1], Glen Meeden[1*] and Brian R. Gray[2]
[1]*School of Statistics, University of Minnesota, Minneapolis, MN, USA*
[2]*Upper Midwest Environmental Sciences Center, United States Geological Survey, La Crosse, WI, USA*

Received 27 February 2010; Accepted 13 April 2010

SUMMARY

We propose a Bayes procedure for estimating categorical abundance using data that are observed with error from a random sample from a finite population. The procedure is designed to estimate the proportion of sites in a finite population that belong to each abundance category. Royle and Link (2005) proposed a multinomial mixture model to analyze data of this nature. Holland and Gray (2010) demonstrated that category means would exhibit bias when probabilities of correct category classifications vary among sampling units and this heterogeneity is not modeled. Those authors proposed a modification to the multinomial mixture model that allows correct classification probabilities to vary by sampling unit according to a single normal distribution on a common logit scale. Our proposal allows both correct and incorrect classification probabilities to vary by site and does not require strong assumptions about the nature of the heterogeneity in classification probabilities. We analyze submerged aquatic vegetation data collected by the Long Term Resource Monitoring Program and compare our results to those of Holland and Gray (2010). We also provide simulation results to demonstrate the performance of our proposal and associated credible intervals under several prior distributions.

Keywords: Categorical abundance, Multinomial mixture, Bayes procedure.

1. INTRODUCTION

Categorical abundance measures are commonly employed by ecologists. For example, the abundance of submersed aquatic vegetation (SAV) is monitored in rivers, lakes and estuaries throughout North America using ordered data. This SAV abundance index takes values $y = 0, 1, ..., 5$, with $y = 0$ indicating no observed SAV and $y = 5$ indicating maximum SAV on the sampling instrument (Yin *et al*. 2000). A major concern with the use of these data is that abundance categories are often measured with classification error. In this case, the observed abundance category is less than the true abundance category. These data were previously analyzed under an infinite population assumption (Holland and Gray 2010), but here we propose a method which assumes we are sampling from a finite population.

Royle and Link (2005) proposed a multinomial mixture model for categorical data measured with classification error. This model assumes that the observed abundance category is a realization of a random process that is a mixture of multinomial distributions. Royle and Link (2005) assumed that classification probabilities are constant for all sampling units or that these probabilities can be modeled as a function of covariates. A concern with this approach is that unmodeled variation in probabilities of correct and incorrect classifications may lead to bias in the estimates of those probabilities and, hence, to bias in the estimated probabilities of the latent abundance categories. Holland and Gray (2010) demonstrated that such unmodeled variation yields biased estimates of correct classification and latent abundance class probabilities. The authors proposed a solution that specifies that correct classification probabilities vary

[*]*Corresponding author* : Glen Meeden
E-mail address : glen@stat.umn.edu

according to a logit model with normally distributed random effects, that the magnitude of the variation is common for all correct classification probabilities, and that misclassification probabilities are constant across sites. The complexity of the these multinomial mixture models make assessment of the validity of these assumptions dificult for specific data sets.

The purpose of this paper is to propose a Bayes procedure for estimating the proportion of sites in a population that belong to each possible true abundance class category, without imposing strong assumptions on the nature of the heterogeneity in classification probabilities. We also investigate the sensitivity of the procedure to specification of prior distributions. In Section 2, we introduce a Bayes procedure for analyzing censored categorical data when sampling from a finite population. In Section 3, we present results for estimating abundance of submersed aquatic vegetation (SAV) in the upper Mississippi River using data collected by the Long Term Resource Monitoring Program (LTRMP). In Section 4, we present results from a simulation study to assess the performance of our model by estimating the coverage probabilities of 95% credible intervals and investigate the sensitivity of the procedure to choice of prior distribution hyperparameters.

2. BAYES PROCEDURE DEFINITION

Consider a finite population of size N with $y = (y_1, ..., y_N)$ the unknown characteristics of interest. In the ecological context of abundance estimation, y_i represents the latent abundance class at site i (Royle and Link 2005). We assume that each y_i can only take on the $K + 1$ values 0, 1, ..., K. A convenient Bayesian model for the situation where little is known about the population is to assume that

$$y_i|\lambda \overset{iid}{\sim} \text{multinomial}(1, \lambda)\ i = 1, ..., N$$

$$\lambda \sim \text{Dirichlet}\ (\varepsilon, ..., \varepsilon)$$

where $\varepsilon > 0$ is some small number. The joint prior distribution is then

$$p(y) = \int ... \int \prod_{i=1}^{N} p(y_i|\lambda)\ p(\lambda)\, d\lambda$$

$$= \frac{\Gamma((K+1)\,\varepsilon) \prod_{k=0}^{K} \Gamma(c_k(y) + \varepsilon)}{\Gamma(\varepsilon)^{K+1}\, \Gamma((K+1)\,\varepsilon + N)}$$

where for $k = 0, 1, ..., K$ and a fixed y we define $c_k(y)$ to be the number of y_i's which are equal to k.

Let s be a sample of size n and let $y(s)$ denote the units in the sample. Then a standard calculation yields

$$p(y_j, j \notin s|y(s)) = \int ... \int \prod_{j \notin s} p(y_i|\lambda)\ p(\lambda|y(s))\, d\lambda \tag{2.1}$$

Given the sample $y(s)$, it is easy to simulate complete copies of the population using the above formula because

$$p(\lambda \mid y(s)) \sim \text{Dirichlet}(c_0(y(s)) + \varepsilon, c_1(y(s)) + \varepsilon, ..., c_K(y(s)) + \varepsilon)$$

where $c_k(y(s))$ is the number of units in the sample that have k as their value.

Now in our problem we do not observe the y values directly but obtain r independent possibly "censored" observations at each site in the sample. Let

$$x_j = (x_{j1}, ..., x_{jr})$$

denote the r observations at site j. We assume that $y_j \geq \max_i x_{ji}$, $i = 1, ..., r$. Given $y_j = k > 0$, our model for x_j is that the x_{ji}'s are iid multinomial(1, θ_k) where $\theta_k = (\theta_{j0}, \theta_{j1}, ..., \theta_{jk})$. We assume that the prior distribution for θ_k is Dirichlet($\varepsilon_1, ..., \varepsilon_1$) where $\varepsilon_1 > 0$ is some small number. Note when $y_j = 0$ we take $\theta_0 = 1$. Holland and Gray (2010) refer to θ_{jk} as the correct classification probability and $\theta_{j0}, ..., \theta_{j(k-1)}$ as misclassification probabilities at site j.

For a site j with $y_j = k > 0$, one finds by integrating out θ_j that

$$p(x_j \mid y_j = k) = \frac{\Gamma((k+1)\varepsilon_1) \prod_{i=0}^{k} \Gamma(\varepsilon_1 + w_j(i))}{\Gamma(\varepsilon_1)^{k+1} \Gamma((k+1)\,\varepsilon_1 + r)}$$

where $w_j(i)$ is the number of x_{ji}'s which equal i. It is easy to check that the formula remains true when $k = 0$.

If x is the collection of all the x_j's and θ is the collection of all the θ_i's then we can write the joint probability structure for our model as

$$p(\lambda, y, \theta, x) = p(\lambda)p(y \mid \lambda)p(\theta \mid y)p(x \mid \theta)$$
$$= p(\lambda) \prod_{i=1}^{N} p(y_i \mid \lambda)\, p(\theta_i \mid y_i)\, p(x_i \mid \theta_i)$$

Now by integrating over λ and the θ_i's we can find the joint probability function $p(x, y)$. Similarly, given the sample, we can find $p(x(s), y(s))$ by summing over all possible values of the x_i's and the y_i's for units not in the sample. Using the previous work we find that

$$p(x(s), y(s)) = p(y(s))\, p(y(s) \mid x(s))$$
$$= \frac{\Gamma((K+1)\varepsilon) \prod_{k=0}^{K} \Gamma(c_k(y) + \varepsilon)}{\Gamma(\varepsilon)^{K+1}\, \Gamma((K+1)\,\varepsilon + n)} \times$$
$$\prod_{j \in s} \frac{\Gamma((y_j + 1)\varepsilon_1) \prod_{i=0}^{y_j} \Gamma(\varepsilon_1 + w_j(i))}{\Gamma(\varepsilon_1)^{y_j + 1}\, \Gamma((y_j + 1)\,\varepsilon_1 + r)}$$

In principle we could find $p(y(s) \mid x(s))$. Then given $x(s)$ we could generate a value for $y(s)$ from this distribution and then using equation (2.1) generate a complete copy of the population. By repeating this procedure, we can approximate the Bayes estimate of any parameter of interest. More details on the Bayesian approach to finite population sampling can be found in Ghosh and Meeden (1997).

The problem is that we cannot generate independent copies of $y(s)$ from $p(y(s) \mid x(s))$. Instead, we use MCMC to generate a sequence of copies of $y(s)$ which we can treat as essentially independent drawns from this distribution. To do so, we select a starting value for $y(s)$ using the discrete uniform distribution over all possible values at each site. Then we simulate a long chain to obtain a single simulated value of $y(s)$. To obtain the next value of $y(s)$, we generate a new starting value and again simulate a long chain. We will treat this sequence of $y(s)$ values as independent draws from the distribution of interest.

Next we describe how the chain is generated for a fixed $x(s)$. Let $y(s)$ be the present state of the chain which is consistent with $x(s)$. To get the proposal we pick a site at random from the sample and then select a new value at random from its possible values, except that the current value may not be selected. This makes it clear that we should only select from the units whose value is known not to be K. Let $\tilde{y}(s)$ denote the new proposal state. Suppose at the selected site k_1 is the current value and k_2 is the proposed value. Then we move to the new state with probability given by

$$\text{ratio} = \frac{p(x(s), \tilde{y}(s))}{p(x(s), y(s))}$$
$$= \frac{(c_{k_2}(y(s)) + \varepsilon)\, \Gamma((k_2 + 1)\,\varepsilon_1)\, \Gamma((k_1 + 1)\,\varepsilon_1 + r)}{(c_{k_1}(y(s)) + \varepsilon)\, \Gamma((k_1 + 1)\,\varepsilon_1)\, \Gamma((k_2 + 1)\,\varepsilon_2 + r)}$$
$$\times\ \Gamma(\varepsilon_1)^{k_1 - k_2}$$

Note that the ratio does not depend on the $w_j(i)$'s. The only information used from the x_j's is the lower bound for each site in the sample.

3. ABUNDANCE ESTIMATES OF SUBMERSED AQUATIC VEGETATION

We estimate the proportion of sites in each abundance class category for the SAV species wild celery (*Vallisneria americana*). Holland and Gray (2010) analyzed this data set using a multinomial mixture model with heterogeneous correct classification probabilities. The data were collected by the LTRMP in 1999 from the open water portion of Navigation Pool 13 (located near Clinton, Iowa) of the Upper Mississippi River (Yin *et al*. 2000). A simple random sample of size n = 210 of sites was selected from the finite population of N = 14,471 possible sampling locations. At each selected site, r = 6 surveys were taken. Sampling plots approximated a square doughnut (Thompson, 2002, p. 280), with surveys located systematically within the plot. The short distance between survey locations (~3 m) suggests the possibility of nontrivial among-site variation in θ_k (Holland and Gray 2010). This data set has maximum unknown abundance index value K = 5. However, sites with observed values of x_{ji} = 4 or 5 are very rare, so we collapse $x_{ji} \geq 3$ to x_{ji} = 3. The resulting maximum abundance index value is K = 3.

We obtained parameter estimates under two prior distributions. First, we set $\varepsilon = \varepsilon_1 = 1$. In this case, we simulated 500 complete copies of the population y vector, using 500,000 steps in the MCMC chain to generate each simulated value of $y(s)$. Next, we set $\varepsilon = \varepsilon_1 = 0.5$. Using this second prior specification, we simulated 500 complete copies of the population y, using 750,000 iterations for each MCMC chain. We

simulated each chain for more steps under the second prior distribution because the chains required more iterations to explore the entire state space. Bayes estimates of the proportion of sites that belong to each possible abundance category were obtained using the mean of the proportions from each simulated copy of the population y vector. We obtained approximate 95% credible intervals for each parameter using the 2.5 and 97.5 percentile of the simulated proportions of sites in each category. The two priors yielded point estimates of the proportion of sites in each abundance category that were comparable at intermediate k values (k = 1, 2), but which differed at extrema (Table 3.1). The $\varepsilon = \varepsilon_1 = 0.5$ prior yielded a larger estimate for the proportion of sites with $k = 0$ and a smaller estimate of the proportion of sites with $k = 3$. Precision appeared poorer under the $\varepsilon = \varepsilon_1 = 0.5$ assumption.

Holland and Gray (2010) analyzed this data set under an infinite population assumption with heterogeneous correct classification probabilities and estimated that $\hat{\lambda}$ = (0.68, 0.18, 0.05, 0.08). Those authors also estimated λ under the constant classification probability assumption proposed by Royle and Link (2005). This method yielded $\hat{\lambda}$ = (0.73, 0.13, 0.07, 0.07). The estimated λ vector under both prior distributions using the current proposal is similar to the previous estimates. The 95% credible interval for each λ_i, i = 0, ..., 3, obtained under the finite population assumption contains the corresponding estimate from the multinomial mixture model using both fixed and heterogeneous classification probabilities.

Table 3.1. Bayes estimates of population proportion of sites in each abundance category under two different prior distributions. Standard deviations of Bayes estimates from each simulated copy of population are given in parentheses. Approximate 95% credible intervals are provided in brackets.

	$k = 0$	$k = 1$	$k = 2$	$k = 3$
$\varepsilon = \varepsilon_1 = 1$	0.704	0.161	0.054	0.081
	(0.034)	(0.029)	(0.024)	(0.022)
	[0.63, 0.77]	[0.11, 0.22]	[0.01, 0.11]	[0.05, 0.13]
$\varepsilon = \varepsilon_1 = 0.5$	0.714	0.160	0.056	0.069
	(0.032)	(0.027)	(0.020)	(0.020)
	[0.65, 0.77]	[0.11, 0.22]	[0.02,0.10]	[0.04, 0.11]

4. SIMULATION RESULTS

We investigated the influence of the choice of prior hyperparameters, ε and ε_1 using a simulation study. We simulated a population of N = 10,000 sites with λ = (0.5, 0.25, 0.15, 0.1).

We chose λ such that the abundance class probabilities decreased as abundance increased in an effort to approximate the pattern we observed in the SAV data. For $j = 1, ... N$, if $y_j = k > 0$ we generated $\theta_j = (\theta_{j0}, \theta_{j1}, ..., \theta_{jk}) \sim$ Dirichlet($\varepsilon_{gen}, ..., \varepsilon_{gen}$), where $\varepsilon_{gen} = 0.5$. Next, we simulated $r = 6$ independent observations $x_{ji} \sim$ multinomial(1, θ_k), $i = 1, ..., r$. If $y_j = 0$, then $x_{ji} = 0$, $\forall\, i = 1, ..., r$. We took 100 samples from the population, each of which consisted of $n = 50$ sites. For each sample, we computed the Bayes estimate of the population proportion of sites that fell into each category, $k = 0, ..., K$. The mean of the Bayes estimates and the simulated coverage probabilities of approximate 95% credible intervals for each parameter are reported in Table 4.1. We also supply the mean of the standard deviation of the estimates from each simulated copy of the population for a single data set, and the standard deviation of the Bayes estimates across simulated data sets. Estimates were obtained under two popular "objective" prior distributions specifed by $\varepsilon = \varepsilon_1 = 1$ and $\varepsilon = \varepsilon_1 = 0.5$. For each sample, Bayes estimates were computed using 500 simulated values of $y(s)$. A MCMC chain was simulated for 50,000 iterations to generate each value of $y(s)$.

The simulation results in Table 4.1 demonstrate that when the assumptions of our model hold, the

Table 4.1. Simulation results for Bayes estimates of population proportion of sites in each abundance category under several different prior distributions.

		$k = 0$	$k = 1$	$k = 2$	$k = 3$
$\varepsilon = \varepsilon_1 = 1$	Mean Estimate	0.500	0.246	0.155	0.099
	Mean SD	0.073	0.071	0.062	0.047
	SD of Estimates	0.071	0.067	0.052	0.039
	Coverage prob.	0.930	0.970	1.000	0.960
$\varepsilon = \varepsilon_1 = 0.5$	Mean Estimate	0.529	0.239	0.154	0.078
	Mean SD	0.073	0.069	0.059	0.040
	SD of Estimates	0.078	0.068	0.058	0.040
	Coverage prob.	0.920	0.950	0.910	0.930

proposed method for finite population estimation yields nearly unbiased estimates of λ under the prior distribution specified by $\varepsilon = \varepsilon_1 = 1$. When $\varepsilon = \varepsilon_1 = 0.5$, the estimate of proportion of sites in abundance category $k = 0$ was biased high by a modest 7%, and the estimate of the proportion of sites in abundance category $k = 3$ was biased low by 22%. Using the standard deviation of the proportion of sites in each abundance class category approximates the standard deviation of the estimates across repeated simulated samples from the population well. The simulated coverage probability of 95% credible intervals also appears to be close to nominal coverage probability.

5. DISCUSSION

The model proposed by Holland and Gray (2010) requires the potentially restrictive assumptions that the probabilities of misclassification are constant at all sites, and that the correct classification probabilities vary according to the normal distribution on the logit scale. In contrast, the method we propose allows both correct and incorrect classification probabilities (θ_k) to vary by site, and we specify a vague prior for the distribution of the θ_k vectors. The method of Holland and Gray (2010) allows estimation of correct and incorrect classification probabilities, while our method removes them from the estimation problem through integration.

The proposed method yields estimates of λ that are comparable to those of Holland and Gray (2010) for the SAV data set considered here and requires fewer assumptions about the form of the heterogeneity in classification probabilities. Furthermore, we saw in a simulation example that the proposed method provides reasonable estimates and credible intervals with close to nominal coverage probabilities under several prior distriubtions. Thus, we recommend the current method for general application when the λ vector represents the sole parameter of interest.

REFERENCES

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.

Holland, M.D. and Gray, B.R. (2010). Multinomial mixture model with heterogeneous classification probabilities. *Environ. Ecol. Stat.*, DOI 10.1007/s10651-009-0131-2.

Royle, J.A. and Link, W.A. (2005). A general class of multinomial mixture models for anuran calling survey data. *Ecology*, **86(9)**, 2505-2512.

Thompson, S.K. (2002). *Sampling*. 2nd ed., Wiley, New York.

Yin, Y., Winkelman, J.S. and Langrehr, H.A. (2000). Long term resource monitoring procedures: Aquatic vegetation monitoring. U.S. Geological Survey, Upper Midwest Environmental Sciences Center, La Crosse, WI. LTRMP 95-P002-7.