Paper 450-2013

Markov Chains and Zeros in My Data: Bayesian Approaches in SAS® that Address Zero-Inflation in Count Data

Matthew Russell, Dept. of Forest Resources, University of Minnesota, St. Paul, MN;
Brian Gray, US Geological Survey, La Crosse, WI

ABSTRACT

There has been an increase in development over the past decade on methods that estimate count data. In recent releases of SAS/STAT software, a number of procedures that perform Bayesian methods have recently been incorporated. A common modeling problem across many disciplines is that of addressing zero-inflation or larger-than-expected proportions of zeros in count data. This challenge is exacerbated when count means and probabilities of structural zeros are heterogeneous. Zero-inflated models are commonly fit using maximum likelihood methods. This paper uses examples from the ecological literature to perform Bayesian analyses on discrete data with zero-inflation. We focus primarily on the PROC MCMC, but also address use of Bayesian methods in the FMM and GENMOD procedures. We fit zero-inflated count models under conditional binomial, Poisson, and negative binomial assumptions, and with and without random intercept effects on structural zero and/or stochastic count components.

INTRODUCTION

SAS® recently introduced methodologies for Bayesian analysis. Some of these methodologies were added to existing procedures (e.g., PHREG, GENMOD, and LIFEREG), while PROC MCMC represents a relatively recent procedure. Researchers from many disciplines have traditionally used the WinBUGS program for performing Bayesian analyses; this tradition is reflected in the breadth of literature devoted to the use of that software, especially in the ecological literature (see McCarthy 2007 and Kéry 2010).

Common in many analyses is assessing the degree to which count data exhibit zero-inflation, defined here as proportions of zeros in excess of that expected under a given count distributional assumption. Hence, count distributions such as the Poisson and negative binomial are useful in describing the stochastic nature of nonnegative integer values. To separate zero and positive counts, a two-stage modeling approach is commonly used for data that display zero-inflation. In this case, one model component estimates the probability of a zero event while a second component estimates one or more parameters of a zero-truncated count distribution. Count regression models can be used to describe various ecological phenomena, may include random effects on model components, and can be estimated using the recently developed SAS procedures that perform Bayesian analysis.

The goal of this study was to evaluate procedures available for fitting zero-inflated models using Bayesian methodologies in SAS/STAT 9.3. We accomplish this by using ecological data. Specific objectives were to (1) evaluate the FMM and MCMC procedures for fitting zero-inflated Poisson and negative binomial models using Markov chain Monte Carlo methods and (2) examine how to apply random effects in Bayesian zero-inflated count models.

DATA

The emerald ash borer (*Agrilus planipennis*) is a nonnative insect introduced to North America in the early 2000's and is destructive to native ash trees. By burrowing underneath the tree's bark, the larvae of the beetle can attack and kill the host ash trees. Recently in 2009, the ash borer was detected in the city of St. Paul, Minnesota, and a quarantine area located in a four-county wide zone was soon implemented to slow the spread of the beetle (Minnesota Department of Natural Resources 2012).

Due to the threat of the beetle on ash trees throughout Minnesota's forests, there is a tremendous interest in locating current populations of ash trees throughout the state. Because of the ash tree's ecological and economic importance, quantifying the presence and abundance of ash populations has direct implications for slowing the possible spread of the ash borer.

Tree measurements from Minnesota's forests were obtained from the USDA Forest Service's Forest Inventory and Analysis database (USDA Forest Service 2012; Table 1). Here, trees were sampled from permanent sample plots 0.40-hectares (1-acre) in size. Measurements on these plots were: plot identification (**`PLOTID`**; a unique identifier for each of the plots found in the state), basal area per hectare (**`BAPH`**; a measure of the density of trees found within the plot), an indicator variable for whether or not one or more ash tree(s) were found in the plot (**`ASH`** = 1 if at least one ash tree was present in the plot; **`ASH`** = 0 if no ash trees were present), a count of the number of ash trees, scaled to represent their number per hectare (**`ASH_TPH`**), and a classification of the type of forest where the plot is located

(**FOREST**). Here, the ash trees that were considered were all species that were coded as a member of the *Fraxinus* genus. Data were collected on 6,306 plots across the state. On 28% of all plots at least one ash tree was observed (Figure 1). A distribution of the observations indicated that a high proportion of zeros was present in the data (Figure 2).

VAR NAME	TYPE	LABEL	RANGE
PLOTID	NUM	Plot record ID	1 to 3,606
BAPH	NUM	Basal area ($m^2 ha^{-1}$) of plot	0.2 to 66.2
ASH	NUM	Indicator variable for ash presence/absence	0 or 1
ASH_TPH	NUM	Number of ash trees (count ha^{-1})	0 to 773
FOREST	CLASS	Forest type where the plot is located	One of many (45 forest types total)

Table 1. The Minnesota ash tree dataset.

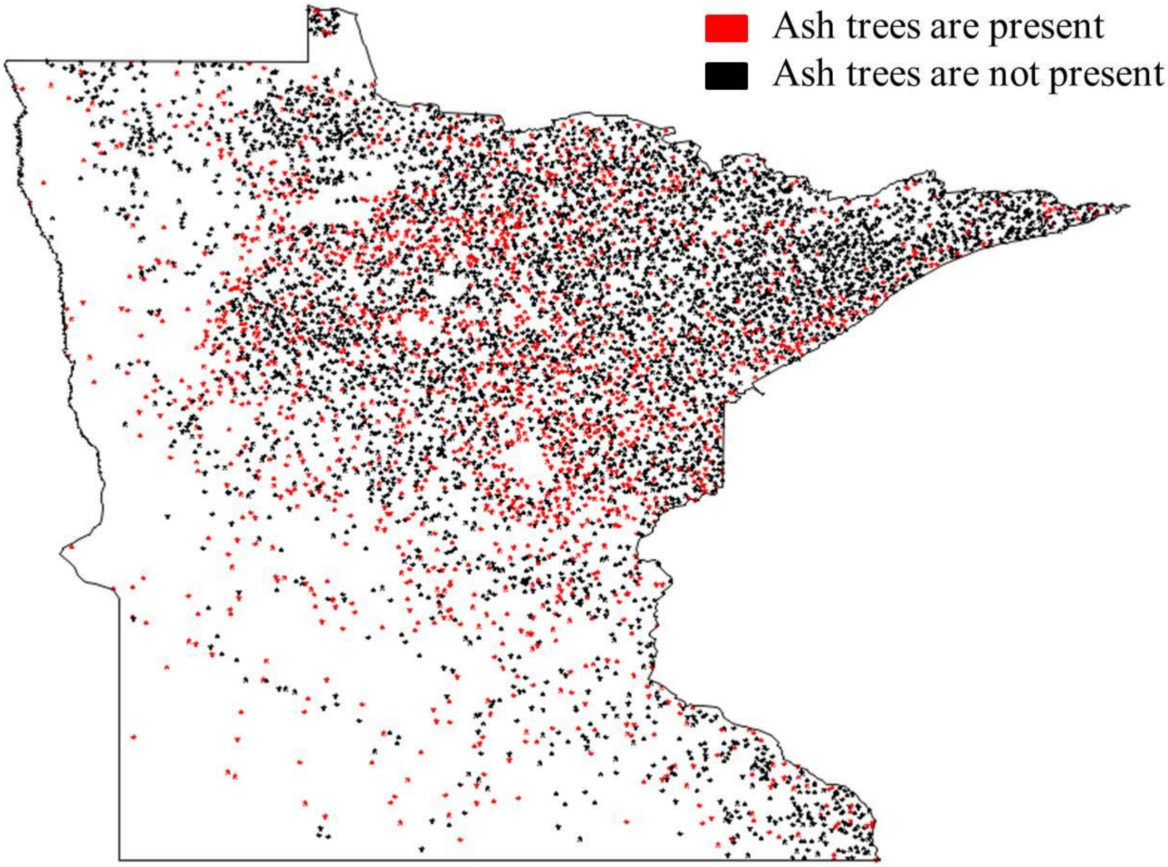

Figure 1. Approximate locations of forest inventory plots in Minnesota measured between 1999 and 2011 indicating the presence/absence of ash trees.

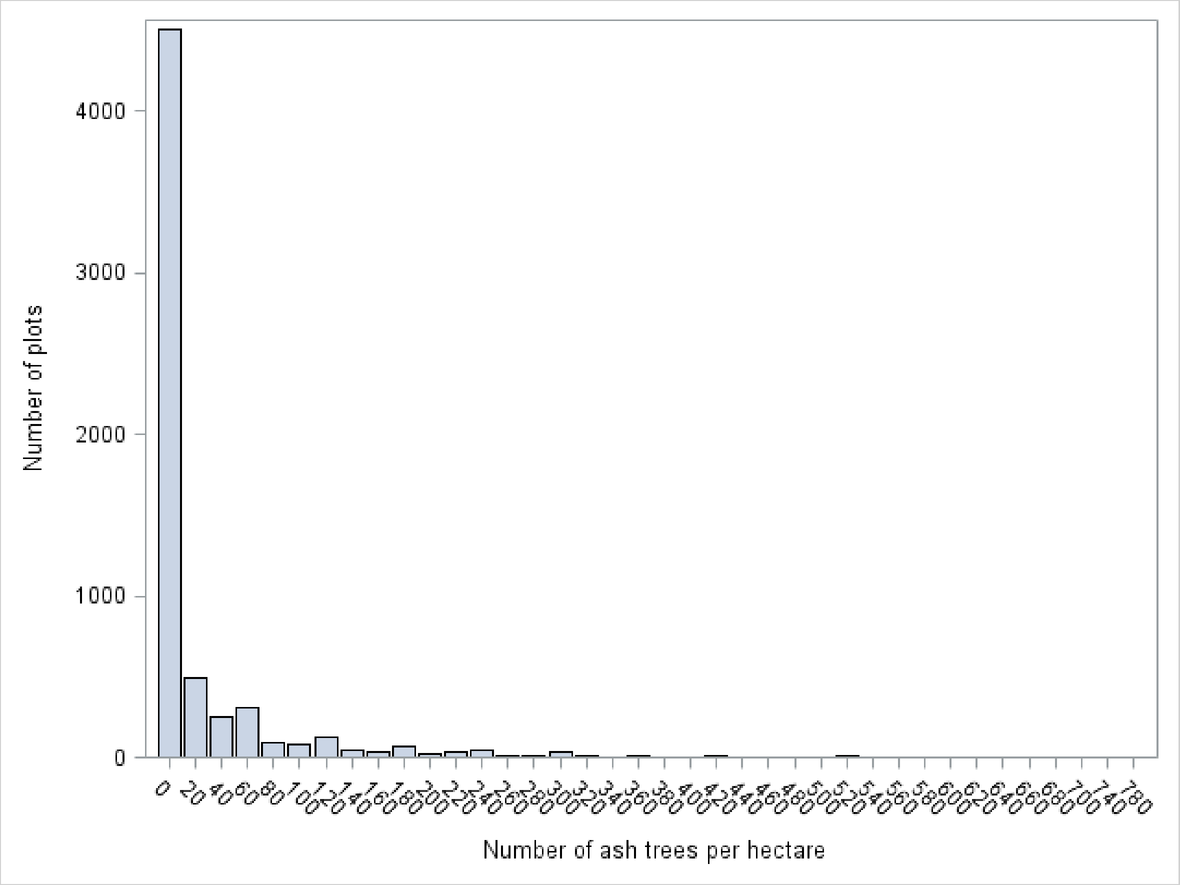

Figure 2. Histogram of number of plots in Minnesota measured between 1999 and 2011 displaying ash tree abundance.

ZERO-INFLATED COUNT MODELS

The Poisson regression model is the benchmark model for count data and can be applied to data whether the response variable is a count or continuous, but becomes restrictive when estimating attributes other than the mean (Winkelmann 2008). Negative binomial regression models are count models that include an overdispersion parameter, making them more flexible than Poisson models. To account for data with a high proportion of zero counts, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models are two types of mixture models that estimate zero and positive counts separately (Welsh et al. 1996; Gray 2005). Hence, comparing the Poisson (P) and negative binomial (NB) families of models becomes an evaluation of the degree of overdispersion in the data, unobserved heterogeneity, and excess zero values (Winkelmann 2008; p. 174).

We might be interested in a zero-inflated model that estimates both the presence and abundance of ash trees in Minnesota. A ZIP probability is estimated by the mass function

$$f_{ZIP}(y) = \begin{cases} \pi + (1-\pi)\mathrm{e}^{-\lambda}, & y = 0 \\ (1-\pi)\dfrac{\lambda^{y}\mathrm{e}^{-\lambda}}{y!}\ , & y = 1, 2, 3, \ldots \end{cases}$$

where y denotes an ash tree count, λ the count mean, and π the probability of zero occurrence. In the case of the Poisson, the variance is equal to its mean.

The primary difference in the NB compared to the P distribution is the incorporation of an overdispersion parameter α. Analogous to the ZIP model, a ZINB probability is estimated by the mass function

$$f_{ZINB}(y)=\begin{cases}\pi+(1-\pi)\left(\dfrac{1}{1+\mu\alpha}\right)^{1/\alpha} & , \quad y=0\\[2ex](1-\pi)\dfrac{\Gamma(y+1/\alpha)}{\Gamma(y+1)\ \Gamma(1/\alpha)}\left(\dfrac{1}{1+\mu\alpha}\right)^{1/\alpha}\left(\dfrac{\mu\alpha}{1+\mu\alpha}\right)^{y} & , \quad y=1,2,3,\ldots\end{cases}$$

where *y* denotes the ash tree count and μ the count mean. The variance of the NB defined above is $\mu+\alpha\mu^2$. In the case of the Minnesota ash tree data, each were related to a system of linear predictors $\mathbf{X}\boldsymbol{\beta}$, where **X** is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of regression coefficients to be estimated.

A log-link function was used to model variation in the means of ash tree absence and positive counts (abundance), respectively. For comparison, we included a random effects parameter (specified for each forest type) on the intercept term of the linear predictor equation estimating ash tree presence and abundance later using PROC MCMC.

PROC FMM

The FMM procedure fits statistical models when response variables are a mixture of univariate distributions (SAS Institute Inc. 2011). Hence, PROC FMM is ideal for data with an excess number of zeros as in the case of the ash tree data. In addition, Bayesian methods can be implemented in the procedure by specifying the `bayes` statement. The number of Monte Carlo simulations, number of simulations to use as burn-in, and the thinning parameter which aims to reduce autocorrelation between successive Monte Carlo runs can be set with the `nmc=`, `nbi=`, and `thin=`, statements, respectively.

The ash tree data is contained in the dataset `PLOT`. To specify a zero-inflated model in FMM, two model statements are required. The first indicates the dependent (**`ASH_TPH`**) and independent variable (**`BAPH`**) to use in the model, while the second statement adds the zero-inflation component. The mixing component was modeled under a logistic link assumption. We will use the same random seed throughout these procedures by specifying the `seed=` option. A 10,000 Monte Carlo run was following a 2,000-run burn-in:

```
ods graphics on;
proc fmm data=PLOT seed=4572;
model ASH_TPH = BAPH / dist=poisson;
model           +    / dist=constant;
bayes nmc=10000 nbi=2000 thin=3;
run;
ods graphics off;
```

We used the default (Gamerman 1997) sampling algorithm and noninformative prior distributions Normal(0,1000) were placed on equation parameters. As an assessment of model convergence, the trace, autocorrelation and posterior density plots were provided. An example for the coefficient associated with **`BAPH`** (mean of 0.0290) can be seen in Figure 3.

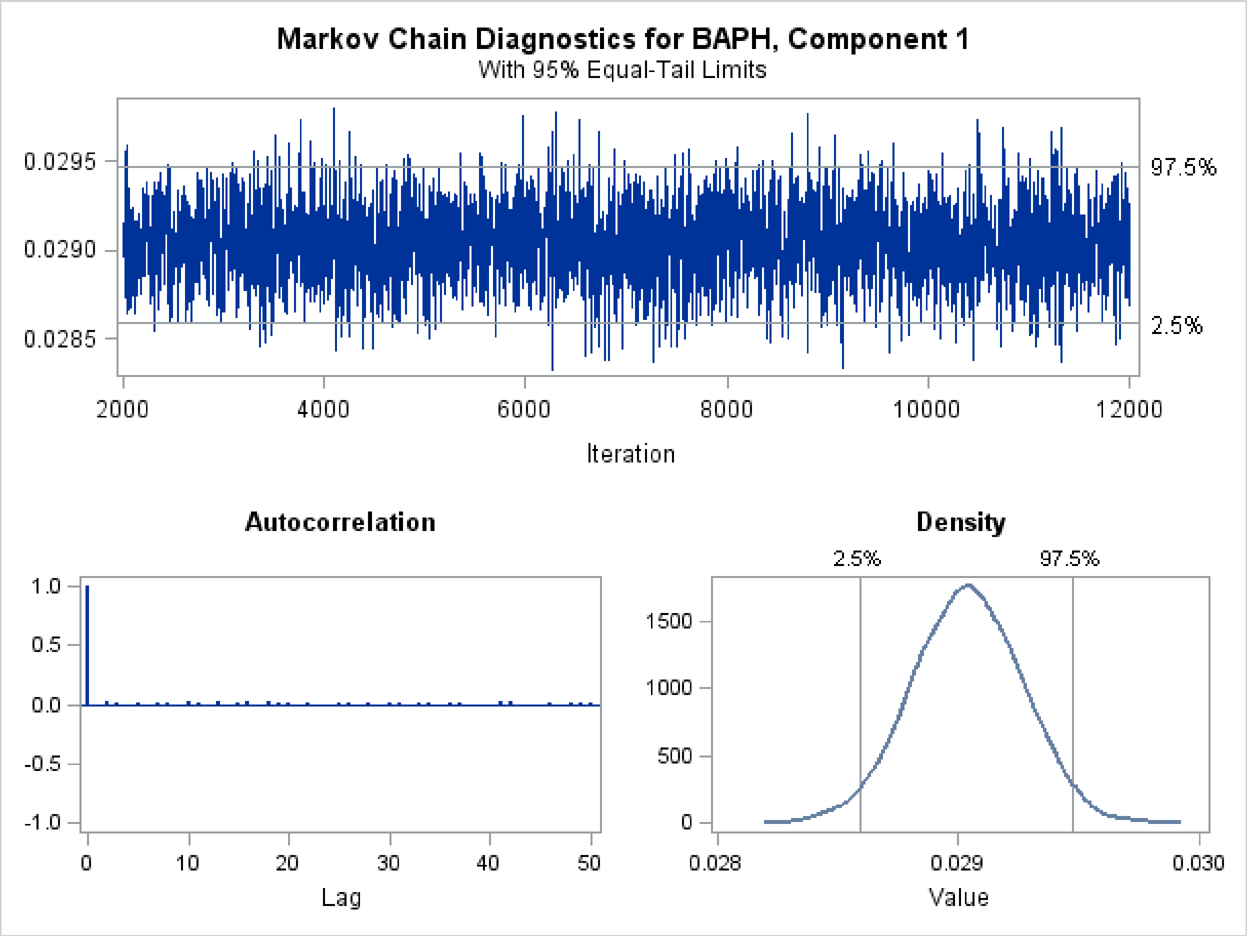

Figure 3. Trace, autocorrelation, and posterior density plots from the FMM output.

Diagnostic plots for the for the `BAPH` parameter indicated good mixing of the Markov Chain, that successive Markov chains were effectively noncorrelated, and that the 95% highest posterior density interval excluded the value zero. The density suggests that ash trees are more abundant with increasing **`BAPH`** of a plot. As a side note, a standard Poisson model could have been fit by disregarding the second `model` statement. The negative binomial distribution is not supported for Bayesian analysis by the FMM procedure.

PROC MCMC

The MCMC procedure is a general purpose Markov chain Monte Carlo procedure designed to fit Bayesian models. PROC MCMC differs substantially from other SAS procedures in that inference is solely Bayesian (SAS Institute Inc. 2011). While some options are common across the MCMC and FMM procedures, coding may be considerably more complex under the former. Key differences are that parameters for coefficients and variance parameters are specified with the `parms` statement, and prior distributions with the `prior` statement. Noninformative priors are here specified as Normal(0,1000). Regression coefficients refer to either the abundance (`pbeta`) or logistic (`lbeta`) variables. Means on the measurement scale are denoted `mu` and `PI`, respectively. The `llike` component specifies that the zero-inflated Poisson model be fit to the response variable **`ASH_TPH`**. The `model` statement specifies the conditional distribution of the data given the parameters.

In PROC MCMC, we might be interested in estimating a similar model predicting the presence/abundance of ash trees:

```
ods graphics on;
proc mcmc data=PLOT seed =4572 nmc=10000 nbi=2000 thin=3 dic propcov=quanew monitor
=(_parms_) outpost=post_ZIP;
parms pbeta0 4.09 pbeta1 0.029 lbeta0 1.34 lbeta1 -0.032;
prior pbeta: ~ normal(0,var=1000);
prior lbeta: ~ normal(0,var=1000);
       link1 = lbeta0 + lbeta1*BAPH;
       PI   = ( 1/(1+exp(-link1)));
       mu=exp(pbeta0 + pbeta1*BAPH);
       llike=log(pi*(ASH_TPH eq 0) + (1-pi)*pdf("poisson", ASH_TPH,mu));
       model general(llike);
run;
ods graphics off;
```

We used the default (N-Metropolis) sampling algorithm and. An example for the coefficient associated with the `pbeta1` can be seen in Figure 4.

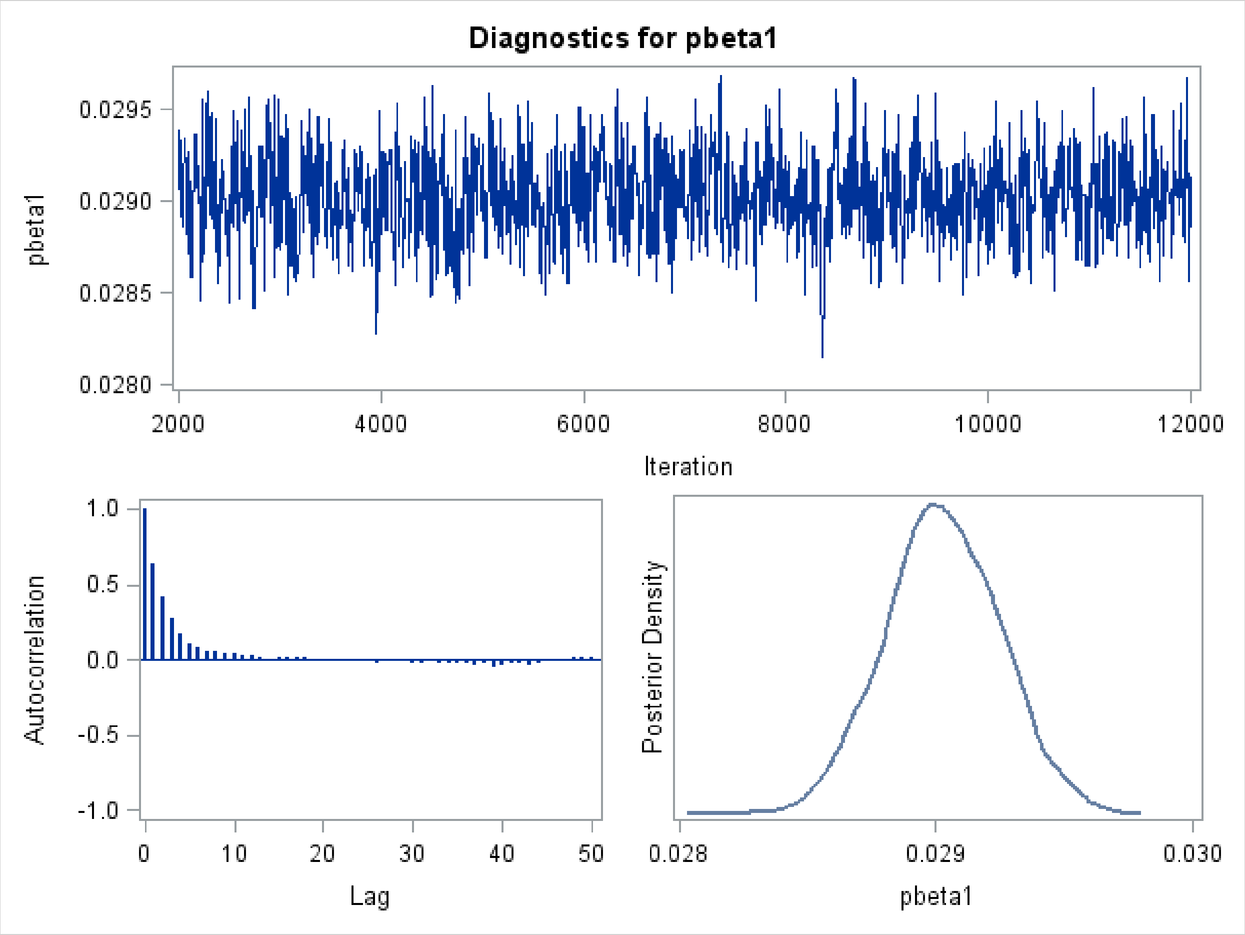

Figure 4. Trace, autocorrelation, and posterior density plots from the MCMC output.

Trace, autocorrelation, and posterior plots for `pbeta1` (i.e., the coefficient representing BAPH; mean of 0.2896) appear acceptable (Figure 4). Posterior estimates from the Markov chains are available in the output dataset `post_ZIP`.

We may be interested in including random intercept terms on the abundance and logistic components when fitting a ZIP model. If we wish to allow the intercept term to vary for each forest type (**`FOREST`**) coded in the ash tree data, we can name the random effects on the abundance and logistic components `delta` and `chi`, respectively, and make use of two random statements. We will specify noninformative priors for these random terms using an inverse gamma distribution.

```
ods graphics on;
proc mcmc data=PLOT seed =4572 nmc=10000 nbi=2000 thin=3 dic propcov=quanew monitor
=(_parms_ delta chi) outpost=post_ZIP_RE;
parms pbeta0 4.09 pbeta1 0.029 lbeta0 1.34 lbeta1 -0.032 delta_s2 1 chi_s2 1;
prior pbeta: ~ normal(0,var=1000);
prior lbeta: ~ normal(0,var=1000);
prior delta_s2: ~ igamma(0.1,s=0.01);
prior chi_s2: ~ igamma(0.1,s=0.01);

random delta~normal(0,var=delta_s2) subject=FOREST ;
random chi~normal(0,var=chi_s2) subject=FOREST ;

       link1 = lbeta0 + chi + lbeta1*BAPH;
       PI   = ( 1/(1+exp(-link1)));
       mu=exp(pbeta0 + delta + pbeta1*BAPH);
       llike=log(pi*(ASH_TPH eq 0) + (1-pi)*pdf("poisson", ASH_TPH,mu));
       model general(llike);
run;
ods graphics off;
```

An example for the coefficient associated with the pbeta1 coefficient (mean of 0.0384) can be seen in Figure 5.

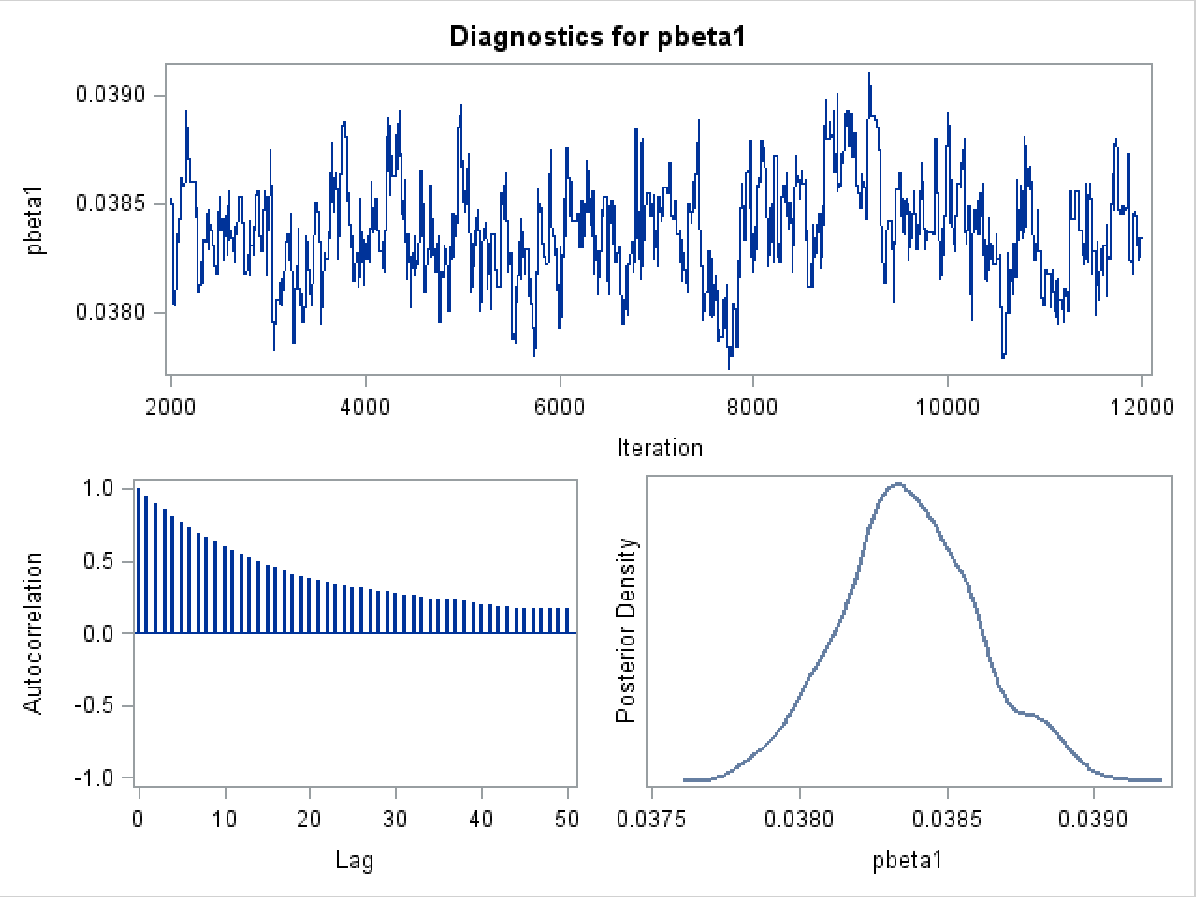

Figure 5. Trace, autocorrelation, and posterior density plots from the MCMC output fit with random intercept effects.

In this case, the trace, autocorrelation, and posterior plots for pbeta1 indicate poor model convergence. To remedy this, we could consider running more Markov chains to attempt to achieve convergence and/or specifying more samples to be used as burn-in. Posterior estimates for fixed and random-effects terms are available in the output dataset post_ZIP_RE.

Observed variance: mean ratio for **ASH_TPH** was 210.8, indicating substantial unmodeled extra Poisson variation in the data. We may address this concern by fitting a conditional ZINB model to the data. To fit a ZINB model to the ash tree data, we will make use of the FCMP procedure to specify that we wish to use the negative binomial distribution. Code to initiate the MCMC appears similar to what was specified for the ZIP model:

```
proc fcmp outlib=sasuser.funcs.test;
function poismean_nb(mean, size);
  return(size/(mean+size));
  endsub;
run;

options cmplib=sasuser.funcs;
run;

ods graphics on;
proc mcmc data=PLOT seed =4572 nmc=10000 nbi=2000 thin=3 dic propcov=quanew monitor
=(_parms_) outpost=post_zinb;
parms pbeta0 4.09 pbeta1 0.029 lbeta0 1.34 lbeta1 -0.032 alpha 1.75;
prior pbeta: ~ normal(0,var=1000);
prior lbeta: ~ normal(0,var=1000);
prior alpha ~ igamma(.01, scale=0.01);

       alpha2=round(alpha+1, 1);

       link1 = lbeta0  + lbeta1*BAPH;
       PI  = ( 1/(1+exp(-link1)));
       mu=exp(pbeta0 + pbeta1*BAPH );
       llike=log(pi*(TPH_ASH eq 0) + (1-pi)*pdf("poisson", ASH_TPH,mu));
       model ASH_TPH~negbin(alpha2, poismean_nb(mu, alpha2));
run;
ods graphics off;
```

We can similarly fit a random ZINB model to the ash tree data, allowing the intercept terms to vary according to each subject=**FOREST**:

```
ods graphics on;
proc mcmc data=PLOT seed =4572 nmc=10000 nbi=2000 thin=3 dic propcov=quanew monitor
=(_parms_) outpost=post_zinb_RE;
parms pbeta0 4.09 pbeta1 0.029 lbeta0 1.34 lbeta1 -0.032 alpha 1.75 delta_s2 1 chi_s2
1;
prior pbeta: ~ normal(0,var=1000);
prior lbeta: ~ normal(0,var=1000);
prior alpha ~ igamma(.01, scale=0.01);
prior delta_s2: ~ igamma(0.1,s=0.01);
prior chi_s2: ~ igamma(0.1,s=0.01);

random delta~normal(0,var=delta_s2) subject=FOREST;
random chi~normal(0,var=chi_s2) subject=FOREST;

       alpha2=round(alpha+1, 1);

       link1 = lbeta0 + chi + lbeta1*BAPH ;
       PI  = ( 1/(1+exp(-link1)));
       mu=exp(pbeta0 + delta + pbeta1*BAPH);
       llike=log(pi*( ASH_TPH eq 0) + (1-pi)*pdf("poisson", ASH_TPH,mu));
       model ASH_TPH~negbin(alpha2, poismean_nb(mu, alpha2));
run;
ods graphics off;
```

Relative model fit was evaluated using the deviance information (DIC). The DIC values indicate the ZINB model with random effects was the best-fitting model (Table 2).

	ZIP			ZINB		
	FMM	MCMC	MCMC w/ random effects	FMM	MCMC	MCMC w/ random effects
DIC	188,962	188,828	122,458	[a]	52,261	44,544

[a] Negative binomial distribution not supported for Bayesian analysis in FMM procedure.

Table 2. Performance of zero-inflated Poisson (ZIP) and negative binomial (ZINB) models as measured by deviance information criterion (DIC) for various Bayesian procedures in SAS that address zero-inflation.

CONCLUSIONS

Fitting zero-inflated count models to the Minnesota ash tree data using Bayesian methods in SAS/STAT 9.3 was straightforward. By incorporating the `bayes` statement into PROC FMM, users can readily compare ZIP models fit with both Bayesian and frequentist methods. The generalized MCMC procedure allows users the opportunity to specify any variety of Bayesian model they wish. PROC MCMC permits the fitting of ZINB models both with and without random-effects terms.

REFERENCES AND RECOMMENDED READING

Gamerman, D. 1997. Sampling from the posterior distribution in generalized linear models. *Statistics and Computing* 7: 57-68.

Gray, B.R. 2005. Selecting a distributional assumption for modelling relative densities of benthic macroinvertebrates. *Ecological Modelling* 185: 1-12.

Kéry, M., 2010. Introduction to WinBUGS for ecologists. Academic Press. 302 p.

McCarthy, M.A., 2007. Bayesian methods for ecology. Cambridge University Press. 296 p.

Minnesota Department of Natural Resources. 2012. The emerald ash borer. Available online at http://www.dnr.state.mn.us/invasives/ terrestrialanimals/eab/index.html; last accessed 8 March 2013.

SAS Institute Inc. 2011. SAS/STAT(R) 9.3 user's guide. SAS Institute, Inc., Cary, NC.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F., and Lindenmayer, D.B. 1996. Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling* 88(1-3): 297-308.

USDA Forest Service. 2012. FIA DataMart: FIADB version 5.1. Available online at http://apps.fs.fed.us/fiadb-downloads/datamart.html; last accessed 11 October 2012.

Winkelmann R. 2008. Econometric analysis of count data. Springer, 333 pp.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Matthew Russell, PhD
Dept. of Forest Resources, University of Minnesota
1530 Cleveland Ave. N.
St. Paul, MN 55108
Email: russellm@umn.edu