

Extending Ordinal Regression with a Latent Zero-Augmented Beta Distribution

Kathryn M. Irvine, T. J. Rodhouse & Ilai N. Keren

**Journal of Agricultural, Biological
and Environmental Statistics**

ISSN 1085-7117

JABES

DOI 10.1007/s13253-016-0265-2

Volume 21, Number 3

**ONLINE
FIRST**

Journal of Agricultural, Biological, and Environmental Statistics

Special Issue on Seismomatics: Space–Time Analysis of Natural or
Anthropogenic Catastrophes

Guest Editors: Jorge Mateu and Emilio Porcu



An Applied Journal Published by
The International Biometric Society
and The American Statistical Association

 **Springer**

13253 • ISSN 1085-7117
21(3) 403–600 (2016)

 **Springer**

Your article is protected by copyright and all rights are held exclusively by International Biometric Society. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".



Extending Ordinal Regression with a Latent Zero-Augmented Beta Distribution

Kathryn M. IRVINE, T. J. RODHOUSE, and Ilai N. KEREN

Ecological abundance data are often recorded on an ordinal scale in which the lowest category represents species absence. One common example is when plant species cover is visually assessed within bounded quadrats and then assigned to pre-defined cover class categories. We present an ordinal beta hurdle model that directly models ordinal category probabilities with a biologically realistic beta-distributed latent variable. A hurdle-at-zero model allows ecologists to explore distribution (absence) and abundance processes in an integrated framework. This provides an alternative to cumulative link models when data are inconsistent with the assumption that the odds of moving into a higher category are the same for all categories (proportional odds). Graphical tools and a deviance information criterion were developed to assess whether a hurdle-at-zero model should be used for inferences rather than standard ordinal methods. Hurdle-at-zero and non-hurdle ordinal models fit to vegetation cover class data produced substantially different conclusions. The ordinal beta hurdle model yielded more precise parameter estimates than cumulative logit models, although out-of-sample predictions were similar. The ordinal beta hurdle model provides inferences directly on the latent biological variable of interest, percent cover, and supports exploration of more realistic ecological patterns and processes through the hurdle-at-zero or two-part specification. We provide JAGS code as an on-line supplement. Supplementary materials accompanying this paper appear on-line.

Key Words: Beta regression; Cumulative link model; Grouped continuous; Hurdle model; Midpoint regression; Non-proportional odds; Plant abundance; Proportional odds model.

1. INTRODUCTION

We consider the analysis of ecological abundance data recorded as an ordered categorical response with each observation i assigned to one of $J + 1$ observed categories, specifically $C_i = j$ where $j = 0, 1, \dots, J$. In ecological and biological applications, outcomes mea-

K. M. Irvine (✉) US Geological Survey, Northern Rocky Mountain Science Center, Bozeman, MT, USA (E-mail: kirvine@usgs.gov). T. J. Rodhouse National Park Service, Upper Columbia Basin Network, Bend, OR, USA (E-mail: Tom_Rodhouse@nps.gov). I. N. Keren Washington Department of Fish and Wildlife, Olympia, WA, USA (E-mail: ilai.keren@dfw.wa.gov).

© 2016 International Biometric Society
Journal of Agricultural, Biological, and Environmental Statistics
 DOI: [10.1007/s13253-016-0265-2](https://doi.org/10.1007/s13253-016-0265-2)

sured on an ordinal scale differ in several ways from ordinal data commonly encountered within social sciences and medical research. In the latter, ordinal data arise from human surveys or questionnaires on, for example, health-related status outcomes in which a person's answers can be ordered (e.g., 0 = "poor," 1 = "moderate," and 2 = "good"; Congdon 2005). Although an underlying continuous scale can be envisioned for the health characteristic of interest (e.g., depression, pain, anxiety), it can never be measured directly. Consequently, in questionnaire-type data, the relative difference between poor and moderate versus moderate and good is unknown. In contrast, ecological researchers often record data as non-negative integers while in fact assessing an observable continuous variable, e.g., when visually estimating plant abundance (Braun-Blanquet 1932) or forest canopy cover (Schabenberger 1995). Procedurally, this approach reduces time spent gathering data and enables larger sample sizes. This type of grouped continuous ordinal data is naturally modeled with an underlying latent continuous variable (Y^*). In this paper, we introduce a flexible Bayesian model that enables ecologists to make inferences directly in terms of the latent continuous variable rather than cumulative categories, which can be cumbersome for non-statisticians to understand.

As a specific example, we focus on percent plant foliar or canopy cover, a frequently used proxy for plant abundance (Bonham 1989). Percentage coverage of a species or grouping of interest (e.g., all non-native species) encountered within a bounded quadrat is binned into a priori cover classes and recorded as an integer (Braun-Blanquet 1932; Daubenmire 1959; Milberg et al. 2008). The levels or classes do carry information regarding relative distance between categories. For example, using the Daubenmire (1959) scale (cover class 0: 0, cover class 1: >0–5% , cover class 2: >5–25 % , cover class 3: >25–50 % , cover class 4: >50–75 % , cover class 5: >75–95 % , and cover class 6: >95–100 %), a relative difference between cover class 1–2 is much smaller compared to cover class 2–3. In ecological applications where Daubenmire's (1959) or other pre-defined classification schemes are used, the cutpoints or thresholds that define categories are known. Consequently, the thresholds of the latent continuous variable do not need to be estimated, but rather can be considered fixed and known quantities.

When analyzing ecological abundance data, it is important to recognize that the lowest category is actually a meaningful zero. The zero class indicates the absence of a species, or when the observation process is imperfect, failure to detect. Species observations recorded as a zero class have an inherently different meaning than just being arbitrarily the lowest category (e.g., when "poor" is the lowest option on a questionnaire). For statistical modeling of plant distribution and abundance, having the ability to distinguish zeros (i.e., establishment) from higher cover classes (i.e., abundance) is essential (e.g., Duff et al. 2011). Zero-class observations indicate that a species either has not saturated its suitable habitat, such as when an invasive species is still early in its colonization phase, or that observations (e.g., quadrat locations) occurred in unsuitable habitat (Martin et al. 2005). Observations of cover classes >0 indicate abundance or population size, conditional on establishment. Modeling these two ecological processes explicitly and simultaneously within the same integrated modeling process is an efficient way for ecologists to explore basic ecological hypotheses as well as to pursue applied questions related to land management and conservation.

One popular approach for statistical modeling of ordered categorical responses is a cumulative logit or probit model (Agresti 2010). In these cumulative link models, the cumulative probabilities are modeled via a logit or probit link as a function of explanatory variables, e.g., $\text{logit}[Pr(C_i \leq j|\mathbf{X})] = \theta_j - \mathbf{X}\boldsymbol{\beta}$ where C_i is the assigned category for observation i with $J + 1$ observed categories, and $\boldsymbol{\theta}$ is the set of ordered cutpoints (threshold parameters) with $-\infty = \theta_0 < \theta_1 < \dots < \theta_J < \theta_{J+1} = \infty$ and $\{\boldsymbol{\beta}, \theta_1, \dots, \theta_J\}$ are estimated from data. The cumulative logit model or so-called proportional odds model (hereafter, POM) has been widely applied to solving problems in the biological sciences, especially in epidemiology (e.g., Ananth and Kleinbaum 1997) and the social sciences (Agresti 2010). Cumulative logit or probit models are encountered less frequently in the ecological sciences (Guisan and Harrell 2000). However, there are situations when it is needed, particularly in the study of vegetation abundance when observations of plant cover are binned into ordinal categories (e.g., Schabenberger 1995; Guisan and Harrell 2000; Coudun and Gegout 2007; Irvine and Rodhouse 2010).

Following Agresti (2010), the POM can be motivated by envisioning a discretization or grouping of an assumed logistic latent variable with cumulative distribution function $Pr(Y_i^* \leq y) = \frac{1}{[1 + \exp(-(y - \mu)/s)]}$ with mean μ and scale parameter set to one ($s = 1$). Since $C_i = j$ if $\theta_{j-1} < Y_i^* \leq \theta_j$, it follows that $\text{logit}(Pr(C_i \leq j|\mathbf{X})) = \text{logit}(Pr(Y_i^* \leq \theta_j|\mathbf{X})) = \theta_j - \mathbf{X}\boldsymbol{\beta}$. The latent variable conceptualization of POM aides in interpretation: $\boldsymbol{\beta}$ represents the location shift in the mean of the latent variable distribution (Agresti 2010; Irvine and Rodhouse 2010). The proportional odds assumption implies that the shift into a higher category is the same across the categories; in other words, all the cumulative logits have a common $\boldsymbol{\beta}$. However, when we explored empirical plant cover data (Sect. 2), there was a strong indication that this assumption was untenable, motivating our development of new tools for ordinal ecological data.

In this paper, we expand the field of latent variable ordinal modeling in two ways. First, we modify the underlying latent variable for ordered categories to better align with the continuous variable of interest, e.g., percent cover. Specifically, we develop our model with a latent beta distribution rather than logistic distribution. This allows for more intuitive ecological conclusions in terms of an explanatory variable's effect on species mean percent cover rather than on cover class cumulative probabilities. Second, a two-part model framework is introduced for ordinal responses with a meaningful zero category to allow for inferences on distinct ecological processes of interest: distribution (colonization or establishment) and abundance. Our work is similar to the two-part or hurdle-at-zero models developed for count data because we assume the mechanism for generating zeros is independent of the mechanism leading to non-zero cover classes (Heilbron 1994). Alternatively, in zero-inflated models, the assumption is that zeros are generated by a mixture of two processes: the so-called zero state that leads to only zero values and another state that leads to a mixture of zeros and non-zero values, e.g., Poisson for count data (Lambert 1992; Hall 2000) or a continuous distribution (Lachenbruch 2002; Moulton and Halsey 1995). In the absence of covariates, hurdle-at-zero and zero-inflated Poisson models are equivalent. For ordinal data, Kelley and Anderson (2008) developed a zero-inflated type model that assumes the zero-class category arises from a mixture of two processes: a point mass at zero and

a multinomial distribution. For a more in-depth discussion of latent variable modeling for categorical data analysis see [Agresti and Kateri \(2014\)](#).

Our model development is motivated by data from a US National Park Service vegetation monitoring program (Sect. 2). Visualizations of empirical observations from two ecologically distinct species, a deep-rooted perennial shrub and a non-native invasive annual grass, provided evidence of inconsistencies with the proportional odds assumption and indicated a potential need for alternative modeling approaches (Sect. 2). Consequently, we develop our ordinal beta hurdle model (hereafter referred to as OBHM, Sect. 3). Using a hurdle-at-zero or two-part construction, the zero cover class is modeled separately from the non-zero classes. By employing a zero-augmented beta distribution for the latent variable, we utilize the fact that there is a *known* biological variable of interest with a pre-defined discretization scheme. In Sect. 4, the ecological inferences based on our proposed OBHM are presented along with three other ordinal models. The four models are compared using a deviance information criterion (DIC) and their predictive performance is evaluated using a hold-out dataset. JAGS code for implementation is available at <https://irma.nps.gov/DataStore/Reference/Profile/2231866>. We conclude with a discussion and suggestions for future extensions to our proposed model (Sect. 5). Results from two simulation investigations are provided as electronic supplements (Supplemental Information S1 focuses on parameter estimation and Supplemental Information S2 explores the utility of using DIC for a formal model comparison technique).

2. MOTIVATING EXAMPLE

Our work is motivated by a sagebrush steppe vegetation monitoring program in the John Day Fossil Beds National Monument, a unit of the US National Park System. The monument is one of the few protected areas in the region containing representative Columbia Plateau sagebrush and bunchgrass steppe ([Rodhouse et al. 2014](#)), but is threatened by invasion from Eurasian annual grasses, especially cheatgrass (*Bromus tectorum*). Throughout the sagebrush steppe biome, widespread declines in the abundances of foundation shrubs, particularly big sagebrush (*Artemisia tridentata*) and native bunchgrasses, are occurring in the wake of over-exploitation, altered fire regimes, and weed invasion ([Davies et al. 2011](#)). We focus on cheatgrass and big sagebrush for the present study because they have distinctive physiognomies and life history strategies, and have very different establishment and population growth histories in the study area.

Plot-based categorical estimates of foliar canopy cover, by species, were obtained during June 2009 and June 2011 following methods developed by [Yeo et al. \(2009\)](#). Plant cover, by species, was estimated within 1169 quadrats using ocular estimation and recorded in the following [Daubenmire \(1959\)](#) cover classes: 0 %, >0–5 %, >5–25 %, >25–50 %, >50–75 %, >75–95 %, and >95 %. Plots were 1 m² quadrats placed across the Monument sampling frame at Universal Transverse Mercator (UTM) coordinates generated with a spatially balanced Generalized Random Tessellation Stratified (GRTS) random sampling algorithm ([Stevens and Olsen 2004](#)). New GRTS draws were generated each year and the same locations were never revisited. The spatial-balance properties of GRTS sup-

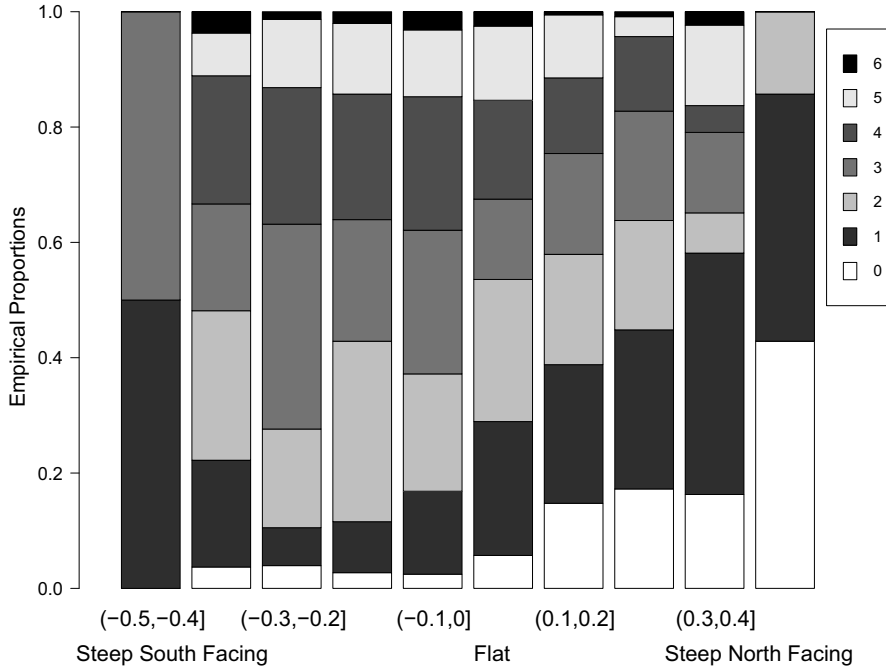
ported our assumption that observations were spatially and temporally independent, after accounting for key covariates, although we did not formally investigate this for the present study.

A suite of environmental predictor variables were chosen to reflect working hypotheses about the factors thought to be most influential for both establishment and abundance of sagebrush and cheatgrass. Specifically, human and livestock travel corridors are known to serve as vectors of annual grass invasion (Gelbard and Belnap 2003; Davies et al. 2013) and the nearest distance from plots to roads and to the Monument's boundaries was calculated as a means to represent this. Also, an indicator variable was created for plots that were located within the perimeters of prescribed burns and wildfires that occurred within the Monument, beginning in 1994, the year of the first well-documented wildfire. Finally, topographic heterogeneity (hereafter "topography") has been shown to be an important proxy for the influence of soil moisture and temperature on invasibility, with cooler, north-facing slopes being more resistant to cheatgrass invasion in some studies (Chambers et al. 2007; Reisner et al. 2013; Rodhouse et al. 2014). For topography, $\sin(\text{slope}) \cdot \cos(\text{aspect})$ was calculated using the 10-m digital elevation model (DEM), which provided an integrated measure of insolation or northness (Rodhouse et al. 2014). This variable ranges from -1 to 1 , with steep south-facing slopes being close to -1 , and steep north-facing slopes being close to 1 .

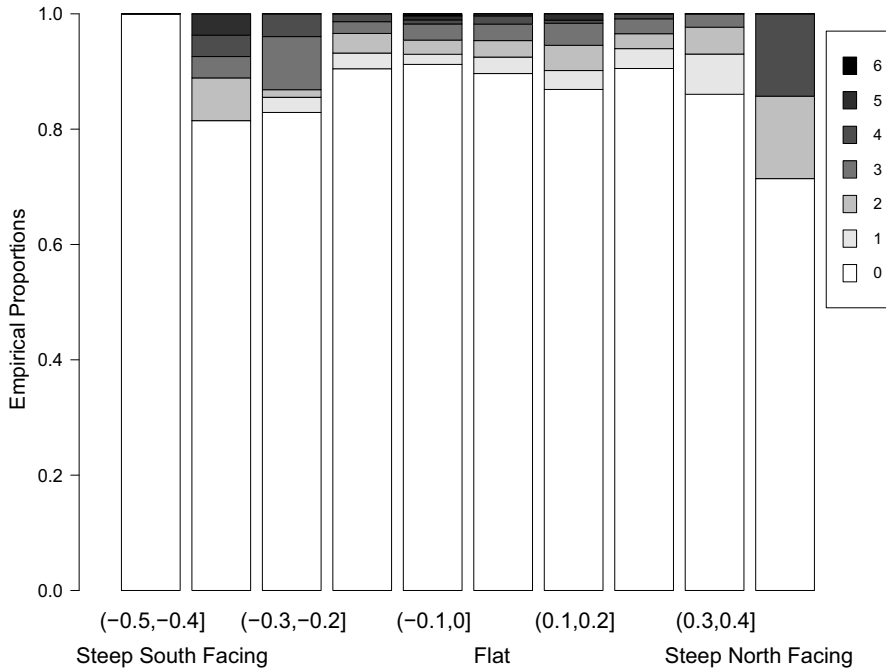
Our two focal species (big sagebrush and cheatgrass) present very different patterns of empirical cover class proportions along the topography gradient, particularly in terms of the quantity of plots with zero percent cover (Fig. 1). In the case of cheatgrass, there was evidence that the number of plots with zero cheatgrass cover increased as slopes became steeper and more northerly in aspect (Fig. 1a). Plots with higher abundance were most often observed on flat and south-facing slopes (Fig. 1a). In contrast, sagebrush was very rare across the study area and plots with higher cover class values were located in flat unburned areas (Fig. 1b).

A graphical approach adopted from Kim (2003) was used to assess the appropriateness of assuming a common β for all cumulative probabilities when fitting a POM. Separate logistic regression models were fit to each of the cumulative probabilities (Venables and Ripley 2002). The estimates using the POM model were included for comparison (the negative of the estimates were plotted due to the parameterization used in the ordinal package in R; Christensen 2014). POM estimates provide a smoothed version of the estimates based on fitting separate models for each cumulative probability (Figs. 2, 3). Presumably, if the point estimates differ dramatically, one should investigate non-proportional odds models for analysis.

In our case, there was evidence that the zero class for cheatgrass was linearly associated with topography (denoted as $Pr[C \leq 0]$; quadratic coefficient interval spans 0, labeled Topography*Topography right-side Fig. 2a). In contrast, the non-zero cumulative probabilities for cheatgrass suggested a quadratic association with topography (some intervals did not span 0, Topography*Topography right-side Fig. 2a) except $Pr(C \leq 5)$ because the larger uncertainty was due to lower sample size ($n = 24$ for cover class 6). Also, the probability of cheatgrass absence ($Pr[C \leq 0]$) was greater in unburned areas compared to burned areas, whereas non-zero cumulative probabilities were similar in burned and unburned areas (intervals spanned zero and point estimates close to zero, Fig. 2b).



(a) Cheatgrass (*Bromus tectorum*)



(b) Big sagebrush (*Artemisia tridentata*)

Figure 1. Cover class empirical cumulative proportions for two species along topography gradient. The bars are stacked such that both the cover class proportions and cumulative proportions are visible (e.g., cover class ≤ 1 and cover class = 1)

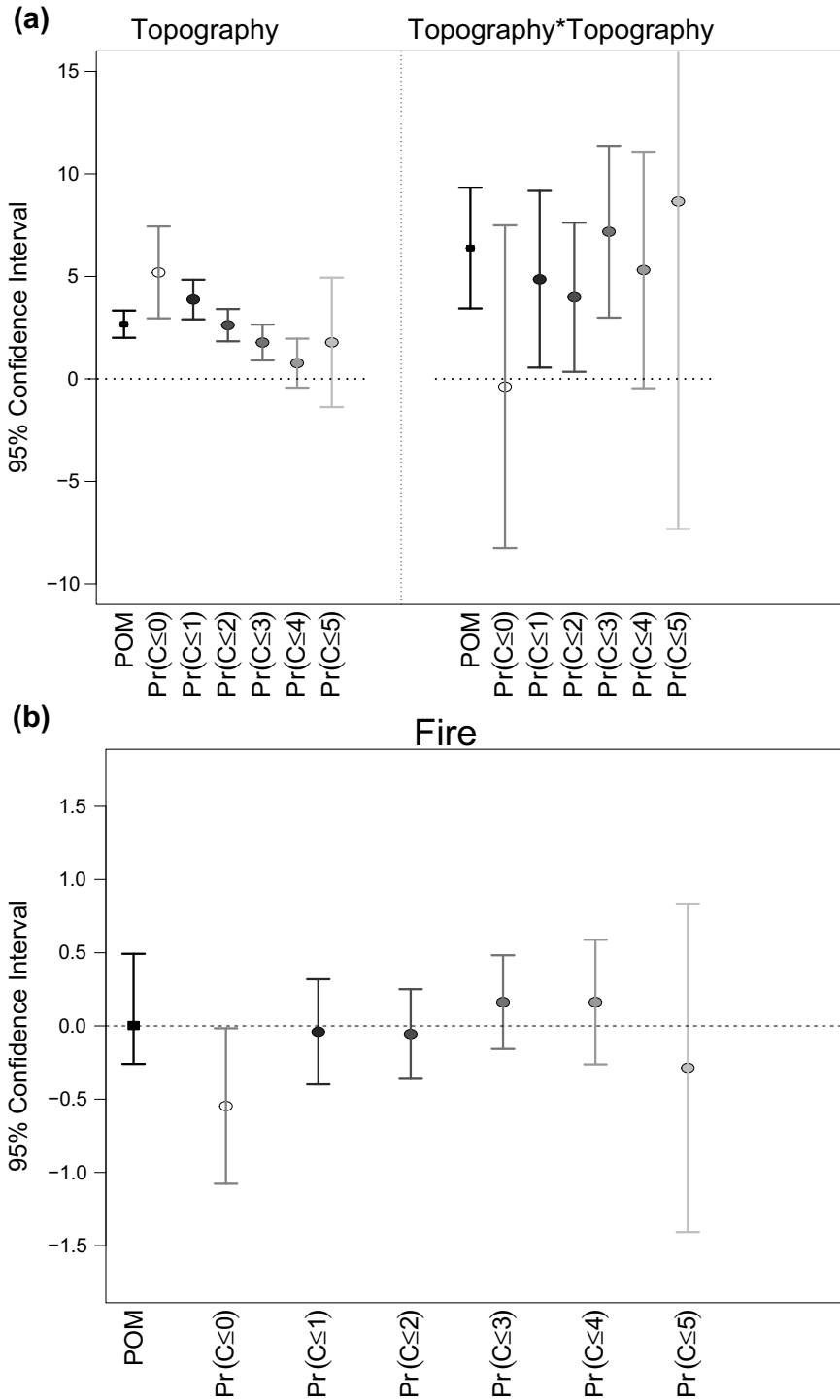


Figure 2. Exploratory plot investigating proportional odds assumption for association between cheatgrass and explanatory variables: Topography (**a-left side**), Topography*Topography (**a-right side**), and Fire (**b**). 95 % confidence intervals for regression estimates based on POM and separate LOGIT models for each cumulative probability, denoted as $Pr(C \leq j)$ where C equals cover class and $j = 0, \dots, 5$

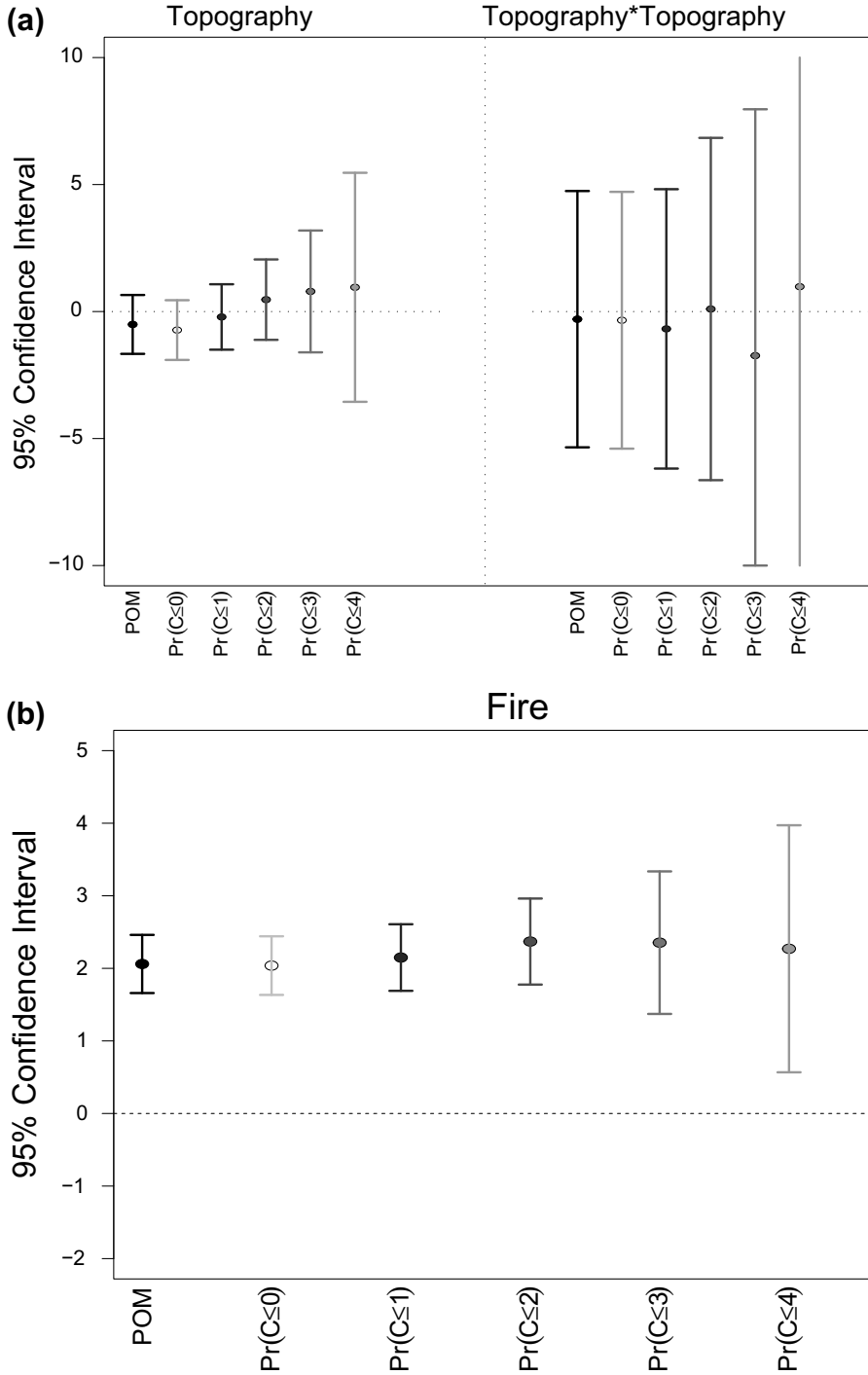


Figure 3. Exploratory plot investigating proportional odds assumption for association between big sagebrush and explanatory variables: Topography (**a-left side**), Topography*Topography (**a-right side**), and Fire (**b**). 95 % confidence intervals for regression estimates based on POM and separate LOGIT models for each cumulative probability, denoted as $Pr(C \leq j)$ where C equals cover class and $j = 0, \dots, 4$

But unlike our investigation for cheatgrass (Fig. 2), the separate cumulative logit point estimates for sagebrush were all fairly consistent and similar to the pooled POM point estimate (Fig. 3). For sagebrush, some of the cumulative logits were not plotted because of low sample sizes ($n = 7$ for cover classes 5 and 6 pooled together). As was evident for cheatgrass, the sparsity in the higher cover classes led to wider confidence intervals. For sagebrush, there was possibly a weak linear positive association between higher cover classes and the topographic gradient and negative linear association with lower cover classes (all the confidence intervals for the quadratic term included zero, denoted as Topography*Topography right-side Fig. 3a). Interestingly, there was an indication that sagebrush cumulative probabilities differed between burned and unburned areas (Fig. 3b). In summary, our graphical investigation suggested that an alternative strategy was needed to explicitly distinguish the ecological processes that generate zeros (i.e., distribution) from those governing non-zero classes (i.e., abundance), and that a non-proportional odds model may also improve model-data agreement.

3. ORDINAL BETA HURDLE MODEL

In situations involving plant cover measurements made on the continuous percent scale, the use of a beta distribution has become recommended practice (Ferrari and Cribari-Neto 2004; Chen et al. 2008a,b; Damgaard 2009, 2012) and it is likely that specifying a latent beta distribution would improve ordinal vegetation modeling, as we explore here. We extend the work of Tamhane et al. (2002), who originally proposed a latent beta distribution for ordinal data, by developing a regression setting within a Bayesian computational and inferential framework. Although we focus on applications to vegetation studies, our approach can be used for any ordinal data arising from a discretized or grouped continuous variable on any interval that is directly measurable. Our method becomes particularly interesting when there is a meaningful zero class in the measurement scheme and it has broad applicability in ecological and environmental settings.

We use the bracket notation to denote a density or distribution following Gelfand and Smith (1990). In our setting, the observed data are of the grouped continuous type of ordinal response and can be considered a realization from a multinomial distribution with $J + 1$ categories, $[C_i | \mathbf{p}] = MULT(1, p_{0i}, \dots, p_{Ji})$, for $i = 1 \dots n$ observations (here, descriptions of areal 1m^2 plots) with category levels of $0, 1, \dots, J$ and $Pr(C_i = j) = p_{ji}$. In our case, following Daubenmire (1959), there were 7 pre-defined categories ($J = 6$), but other classification schemes are possible (e.g., Milberg et al. 2008). The novelty in our approach is specifying the latent variable (Y_i^*) distribution as a mixture of a continuous beta distribution and a point mass at 0, a zero-augmented beta distribution (Ospina and Ferrari 2010), as follows:

$$[Y_i^* | \alpha, \mu_i, \phi] = \begin{cases} \alpha, & \text{if } Y_i^* = 0, \\ (1 - \alpha)[Y_i^* | \mu_i, \phi], & \text{if } Y_i^* \in (0, 1), \end{cases}$$

where $[Y_i^* | \mu_i, \phi]$ is the beta distribution as parameterized for regression analysis by Ferrari and Cribari-Neto (2004) with $E(Y_i^*) = \mu$ and $Var(Y_i^*) = \frac{\mu(1-\mu)}{\phi+1}$.

The mean, μ , and mixture parameter, α , can be extended to include explanatory variables of interest using $\text{logit}(\mu_i) = \mathbf{X}_i \boldsymbol{\beta}$ and $\text{logit}(\alpha_i) = \mathbf{W}_i \boldsymbol{\gamma}$. For a given predictor k , the

interpretation of $\exp(\beta_k)$ is in terms of a multiplicative change in the odds of the proportion of a plot covered by a species with a unit change in X_k , given all other predictors are held at a fixed value. For a given predictor l that is related to the presence or absence of a species, the interpretation of $\exp(\gamma_l)$ is the effect of a one-unit change in W_l on the odds of a zero class, given all other predictors are held at a fixed value. Importantly, for a more intuitive interpretation, the inverse logit transformation of $\mathbf{X}_i\boldsymbol{\beta}$ provides inferences regarding the mean of the latent abundance process or the proportion of a plot covered (e.g., percent areal cover/100). Similarly, the probability of absence (zero class) can be explored using the inverse logit transformation of $\mathbf{W}_i\boldsymbol{\gamma}$. Larger values for the precision parameter, ϕ , produce less variability in percent cover values.

Then, following the usual latent variable motivation for an ordinal observation, the observed cover class category is a discretized version of Y_i^* for observation i as follows:

$$C_i = \begin{cases} 0, & \text{iff } Y_i^* = 0, \\ j, & \text{iff } \theta_{j-1} < Y_i^* \leq \theta_j \text{ for } j = 1, \dots, J. \end{cases} \quad \text{In our case, } \boldsymbol{\theta} \text{ are a set of known parameter values based on the Daubenmire scale (1959).}$$

The category probabilities still meet the assumption of $\sum_{j=0}^J p_j = 1$ because

$$Pr(C_i = 0) = p_{0,i} = \pi_{0,i}(\mathbf{W}) \quad (1)$$

$$= \exp(\mathbf{W}_i\boldsymbol{\gamma}) / (1 + \exp(\mathbf{W}_i\boldsymbol{\gamma})) \quad (2)$$

$$\text{and for } j = 1, \dots, J \quad (3)$$

$$Pr(C_i = j) = p_{j,i} = [1 - \pi_{0,i}(\mathbf{W}_i)]\pi_{j,i}(\mathbf{X}_i) \quad (4)$$

$$= [1 - \pi_{0,i}(\mathbf{W}_i)] \int_{\theta_{C_i-1}}^{\theta_{C_i}} [Y_i | \mu_i, \phi] dY_i, \quad (5)$$

where $[Y_i | \mu_i, \phi]$ is the beta density (Ferrari and Cribari-Neto 2004). The parenthetical $\mathbf{W}_i$ and $\mathbf{X}_i$ denote that the probabilities are modeled as a function of explanatory variables, and may differ for the zero class and non-zero classes.

Under this formulation, the likelihood can be written in terms of the category probabilities derived above:

$$[\mathbf{C} | \boldsymbol{\gamma}, \boldsymbol{\mu}, \phi, \boldsymbol{\theta}] = \prod_{i=1}^n \pi_{0,i}^{\mathbb{1}(C_i=0)} \left((1 - \pi_{0,i}) \int_{\theta_{C_i-1}}^{\theta_{C_i}} [Y_i | \mu_i, \phi] dY_i \right)^{1-\mathbb{1}(C_i=0)},$$

where $\mathbb{1}(C_i = 0)$ is an indicator variable for the zero cover class observations. Importantly, this likelihood can be factored into two parts as follows,

$$\left(\prod_{i=1}^n \pi_{0,i}^{\mathbb{1}(C_i=0)} (1 - \pi_{0,i})^{1-\mathbb{1}(C_i=0)} \right) \left(\prod_{k=1}^{n-n_0} \int_{\theta_{C_k-1}}^{\theta_{C_k}} [Y_k | \mu_k, \phi] dY_k \right),$$

where k indexes the non-zero observations, n is the total number of observations, and n_0 is the number of zero-class observations. This simplification allows for parameter estimation to proceed independently for the zero class and non-zero classes. Assuming independent priors on the parameters, the posterior distribution can be written as

$[\gamma, \mu, \phi | \mathbf{C}, \theta] \propto [\mathbf{C} | \gamma, \mu, \phi, \theta][\gamma][\mu][\phi]$. Incorporating the latent variable, similar to the work by [Albert and Chib \(1993\)](#), the joint posterior of the parameters and the latent variable can be written as $[Y^*, \beta, \phi, \gamma | \mathbf{C}, \theta] \propto [\mathbf{C} | Y^*, \theta][Y^* | \beta, \phi, \gamma][\beta][\phi][\gamma]$. Then, exploiting the two-part factorization allows for re-expression of the joint posterior as $\propto [\mathbb{1}(C = 0) | \gamma][\gamma][C_1, \dots, C_6 | Y, \theta][Y | \phi, \beta][\phi][\beta]$, with C_j denoting total cell frequencies (e.g., $\sum_i \mathbb{1}(C_i = j)$ for $j > 0$) and Y representing the beta latent variable with $Y \in (0, 1)$. Again, this is a two-part model in which we assume the zero-class observations stem from one ecological process that is independent of the process leading to non-zero cover classes (abundance).

To make our methods more accessible to practitioners, the available MCMC algorithms within JAGS ([Plummer 2003](#)) were utilized for inference. The model for the zero process was coded as $Z_i \sim \text{Bernoulli}(\alpha_i)$, where $Z_i = \mathbb{1}(C_i = 0)$ for $i = 1, \dots, n$, $\text{logit}(\alpha_i) = \mathbf{W}_i \boldsymbol{\gamma}$, and diffuse normal priors on γ . Creating an indicator for the response in the zero process model and then modeling the non-zeros separately was introduced for count data by [Heilbron \(1994\)](#).

The more complicated parameters for inference are those related to the latent beta distribution parameters, β and ϕ . However, the data augmentation methods originally proposed by [Albert and Chib \(1993\)](#) for Gibbs sampling of the latent Gaussian variable for ordinal data can be harnessed (for examples and WinBUGS code see [Congdon 2005](#)). Our modification was to specify a truncated beta distribution for $[Y_i | \beta, \phi, \theta, C_i = j]$ as opposed to a Gaussian one as in [Albert and Chib \(1993, Eq. 17\)](#). The constrained sampling of the latent variable (Y) depends on the observed category. For example, if $C_i = 2$, then Y_i falls within the interval (θ_1, θ_2) , $Y_i \sim \text{Beta}(\mathbf{X}_i \boldsymbol{\beta}, \phi)$ truncated on the left by θ_{C_i-1} and the right by θ_{C_i} . To sample from the truncated beta distribution for a given observation in JAGS, the dinterval function was used (Code available at <https://irma.nps.gov/DataStore/Reference/Profile/2231866>). The iterative sampling from $[\beta | \phi, Y]$ and $[\phi | \beta, Y]$ then proceeds as in a Bayesian beta regression model (WinBUGS implementation developed by [Branscum et al. 2007](#)).

In our model formulation, the cutpoint or threshold parameters (θ) are set to known values based on the cover class scheme used in data collection. A diffuse normal prior for $[\beta]$ and a diffuse prior for the precision parameter, $[\phi] = \text{Gamma}(0.1, 0.1)$, were specified. Random and dispersed starting values were generated from the prior distributions for β , γ , and ϕ to initialize the chains. The corresponding midpoint values on the proportion cover scale for each observation were used as the starting values for the unobserved beta latent variable.

By modeling the mean of the beta latent variable as a function of predictors, substantial heterogeneity in the category probabilities can be accommodated (Fig. 4). Our model can be considered another type of location–dispersion ordinal model ([Agresti 2010](#)) because the variance of the latent variable distribution is a function of μ and ϕ . In other words, even with a fixed value of ϕ , the variance changes with μ ; it is not fixed at one value as is the assumption in cumulative logit and probit models. The key benefit to our model is the appealing interpretation in terms of the biological (latent) variable of interest, percent cover, and the ability to explore more realistic ecological processes and questions through the hurdle-at-zero or two-part specification. We anticipate that hierarchical extensions to our OBHM can

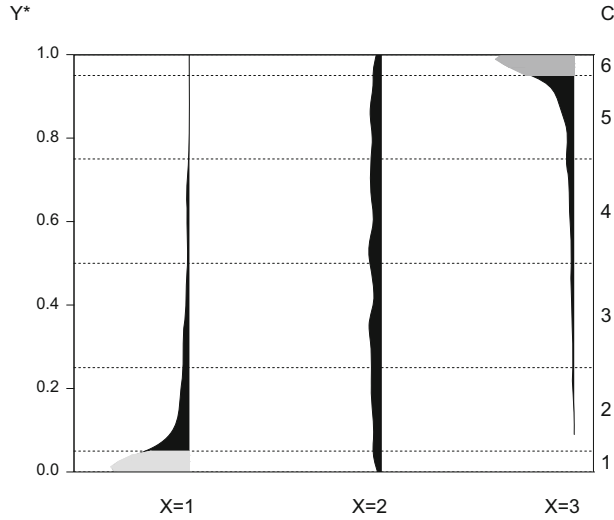


Figure 4. Conceptual example displaying discretizing or grouping a beta-distributed random variable (Y^*) into 6 categories (C) based on known cutpoints (represented by dashed horizontal lines). One realization generated from a beta distribution with mean (μ) varying at different levels of a discrete predictor X . The lighter gray shaded area represents the empirical proportion for category 1 when $X = 1$ (beta parameters; $\text{logit}(\mu) = -2, \phi = 2$); when $X = 2$ the latent beta (beta parameters $\text{logit}(\mu) = 0, \phi = 2$) is equivalent to a Uniform (0,1) and the logit-transformed proportions are distributed as logistic; the darker gray shaded area represents the empirical proportion of category 6 when $X = 3$ (beta parameters; $\text{logit}(\mu) = 2, \phi = 2$)

be developed that model spatial correlation among observations and measurement errors in the observation process (e.g., imperfect detection), if needed.

4. COMPARISON OF BAYESIAN ORDINAL MODELS FIT TO STEPPE VEGETATION DATA

Four different approaches for Bayesian modeling of both the presence and abundance of big sagebrush and cheatgrass as a function of explanatory variables were explored (JAGS code provided at <https://irma.nps.gov/DataStore/Reference/Profile/2231866>, Plummer 2003). A randomly selected 25 % of the entire dataset was used for a hold-out or validation dataset ($n = 292$), the remaining 877 observations were used for parameter estimation (or training data). For Bayesian estimation, four chains each initialized with dispersed starting values ran for 10,000 iterations with a burn in period of 5000. Each chain was thinned by 4 (although Link and Eaton 2012 point out thinning may be unnecessary in some cases) to produce a total of 5000 posterior draws for each parameter. For numerical stability, the explanatory variable measuring distance to boundary was divided by 1000 and distance to roads divided by 100. A graphical investigation was used for evidence of proper MCMC chain mixing and consistency of means and standard deviation estimates across MCMC chains, along with the Gelman–Rubin statistic ($\hat{R}$) as a convergence diagnostic.

First, a multiple linear regression model was fitted using the midpoints of the categories as the response variable (denoted as MLR). This model was included for illustration purposes

because it is still a common approach, even though many have pointed out the inadequacies of doing so (Agresti 2010; Irvine and Rodhouse 2010; Higgs and Ver Hoef 2012; Larrabee et al. 2014). The continued use of this incorrect method by practitioners reflects the desire for interpretation on the percent cover scale. Diffuse normal priors ($[\beta] = \text{Normal}(0, 0.0001)$) on regression parameters and a diffuse gamma prior for the residual variance were specified.

Second, the cumulative logit model with proportional odds assumption was estimated for all the observed categories (POM). Bayesian estimation of the cumulative logit model followed Congdon (2005) and Agresti (2010). The no-intercept constraint in the latent variable mean structure (μ) was used; however, other constraints could be implemented (Congdon 2005, p. 237). A similar data augmentation approach as that in our OBHM was employed, but constrained sampling of a logistic was specified as opposed to a beta latent variable (as described in Sect. 3). Diffuse normal priors on regression parameters for the mean were used. Standard uniform priors on the probability-scale for the threshold parameters were as follows: $[\xi_j] = \text{Uniform}(0, 1)$ and $\theta_j = \text{logit}(\xi_j)$ for $j = 1, \dots, 6$. The *sort* function within JAGS achieves the stochastic ordering of the threshold parameters (Plummer 2015). This prior distribution specification in JAGS is equivalent to specifying a truncated multivariate logistic prior that respected the ordering of the θ 's (on the logit-scale). This is a vaguely informative prior, but prior knowledge about the cumulative probabilities can be incorporated by adjusting the bounds on the uniform priors. Our approach is similar to the truncated multivariate normal prior distribution suggested for the cumulative probit model (Agresti 2010, p. 328). There are other choices for the prior distribution and constraints on the threshold parameters to improve convergence issues related to observing sparse categories and to incorporate prior information (Ishwaran 2000; Fahrmeier and Tutz 2001; Congdon 2005, p. 238; Higgs and Ver Hoef 2012; Schliep and Hoeting 2013). Code is provided as an electronic supplement (<https://irma.nps.gov/DataStore/Reference/Profile/2231866>).

Third, a simpler ordinal hurdle-at-zero model was constructed that models the zero class separately with a logistic regression and then models the non-zero cover classes $\{1, \dots, 6\}$ with a cumulative logit model as described above (denoted OLHM for ordinal logistic hurdle model). Specifying this model or the OBHM, assumes that the zero-class observations are generated by a process independent of that which leads to observations of non-zero cover classes. The main distinction between OBHM and OLHM is that for OLHM we assume the latent abundance variable after a logit transformation ($\text{logit}[\text{percent cover}]$) is consistent with a logistic distribution. Finally, a simulation study investigated whether there was a benefit to specifying a beta compared to a logistic latent variable under biologically realistic scenarios (Supplement S1). The simulation study focused on the latent variable, as this was the main difference between our two hurdle-at-zero models: OLHM and OBHM.

A Deviance Information Criterion (DIC) was used to formally compare the relative fit of MLR, POM, OLHM, and OBHM to the two focal species datasets (Spiegelhalter et al. 2002; Plummer 2008). DIC was calculated as follows, $DIC = \bar{D} + p_D$ where $\bar{D} = E(D|Y)$ is a measure of model fit, $p_D = \bar{D} - D\{\bar{\theta}(Y)\}$ is a measure of model complexity, and $D\{\bar{\theta}(Y)\}$ is the deviance evaluated at the posterior means of the parameter set (θ). We assessed the ability of DIC to select the true data-generating model in a simulation study (Supplement S2), because there was debate on the reliability of DIC with mixture models generally

(Plummer 2008; Millar 2009; Neelon et al. 2010), and there was no published guidance for ordinal hurdle-at-zero models.

A comparison of POM, OLHM, or OBHM cover class predictions to the observed cover classes in the hold-out dataset provided an empirical predictive performance assessment. For each of 100 posterior draws (indexed i) from the entire parameter set (e.g., for OBHM β , α , and ϕ), the predicted category probabilities were calculated for each hold-out observation i' , $\{p'_{0,i'}, \dots, p'_{6,i'}\}$ based on their observed predictors (e.g., Fire, Topography, Distance to Roads, ...). The predicted category, defined as the category with the maximum posterior mean probability ($\max\{\bar{p}_{0,i'}, \dots, \bar{p}_{6,i'}\}$), compared to the observed category provided an overall assessment of predictive ability. A posterior distribution for Kendall's τ indicated a measure of concordance between the observed and predicted cover class categories over the 100 posterior draws (Agresti 2010, p. 189).

4.1. CHEATGRASS (*Bromus tectorum*)

The statistical inferences regarding the environmental predictors associated with cheatgrass percent cover differed dramatically between the hurdle-at-zero and non-hurdle cheatgrass models (Table 1). The magnitude of OBHM-abundance estimates were larger for some explanatory variables, and also more precise than those from the POM and OLHM (comparison of POM, OBHM, and OLHM with response $\{1, \dots, 6\}$ in Table 1). There was no evidence for an association between fire and cheatgrass using MLR or POM, contrary to well-established ecological knowledge (e.g., D'Antonio and Vitousek 1992; Table 1). However, consistent with our diagnostic plots in Fig. 2 and published literature, we concluded

Table 1. Comparison of Bayesian posterior distribution means and standard deviations (SD) from four different approaches for modeling cheatgrass: using midpoint values as the response in a Bayesian Multiple Linear Regression model (MLR); Bayesian proportional odds model with response all observed cover classes (POM); Bayesian ordinal logistic hurdle model (OLHM) which is a two-part model with logistic latent variable for cover classes $\{1, \dots, 6\}$ and a separate logistic regression for zero cover class; and ordinal beta hurdle model (OBHM) which assumes a beta distribution for the non-zero cover class probabilities and a separate logistic regression for the zero cover class

Model	p_D	Response	Topography		Topography ²		Fire		Dist. to Rds.	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
MLR	4.0	$\{0, \dots, 0.975\}$	-0.321	0.061	-0.821	0.281	NS		NS	
POM	8.9	$\{0, \dots, 6\}$	-2.582	0.375	-6.679	1.683	NS		-0.019	0.014
OLHM	3.0	$I(Y = 0)$	5.699	0.915	NS		-0.666	0.309	NS	
	7.7	$\{1, \dots, 6\}$	-1.87	0.397	-6.85	1.942	NS		-0.028	0.014
OBHM	3.0	$I(Y = 0)$	5.704	0.922	NS		-0.686	0.296	NS	
	3.6	$\{1, \dots, 6\}$	-1.471	0.295	-11.244	1.195	NS		-0.016	0.010

We included in all models a linear and quadratic term for topography variable (denoted as Topography and Topography²), an indicator for burned areas (denoted Fire), and the distance to road divided by 100 (denoted as Dist. to Rds.). NS represents the situation when the given variable's 90 % posterior interval included 0 and a reduced model excluding that variable was fitted for inferences. Note for both OLHM and OBHM the abundance portion of the model has response listed as $\{1, \dots, 6\}$. The estimated posterior mean for ϕ was 1.44 with SD 0.072 in OBHM. p_D is the effective number of parameters for each model

based on estimates from a hurdle-at-zero model (OLHM or OBHM) that burned areas had a higher probability of cheatgrass presence (corresponds to OLHM and OBHM with response $I(Y = 0)$ in Table 1). Interestingly, there was no evidence that cheatgrass abundance differed in burned and unburned plots suggesting that fire may increase the invasibility of a landscape by cheatgrass, but that other factors may contribute to persistence and growth (Table 1 OBHM and OLHM with response denoted $\{1, \dots, 6\}$).

Proximity to corridors of disturbance has long been identified as a factor in weed invasions, but there was only weak support for this hypothesis (distance to road estimates Table 1). Notably, our OBHM estimates had smaller uncertainties than those estimated from the other models. There was no evidence that distance to roads was associated with the probability of cheatgrass absence (OLHM and OBHM with response $I(Y = 0)$ in Table 1). This pattern and the ubiquity of cheatgrass in the Monument study area (only 88 of 1169 plots in the zero class) underscores the late stage of the invasion process there.

The estimated probability of cheatgrass absence was linearly associated with topography (OLHM and OBHM with response $I(Y = 0)$ Table 1), consistent with diagnostic plots (Sect. 2). Interestingly, there was evidence for a quadratic association between cheatgrass abundance and topography, with the highest percent cover observed on flat sites (POM and abundance portions of OLHM and OBHM, Table 1).

Although we calculated DIC for MLR (DIC = 306), we felt it did not provide a meaningful measure for relative comparison to the other models and should not be used in practice based on our simulation study (Supplement S2). Among the other models, OLHM had the lowest DIC with a total value of 3088 and was 10 units less than POM. The DIC values for the zero-component of OBHM and OLHM were identical (DIC = 415), as expected, and there was no apparent benefit for using the beta latent variable over the logistic in this case (OBHM-abundance DIC = 2757 compared to OLHM-abundance DIC = 2672). For cheatgrass data, our diagnostic visualizations (Sect. 2) and formal model comparison results indicated the need for a hurdle-specification. Our simulation study also supported the use of DIC to select among hurdle-at-zero and non-hurdle models (OLHM or OBHM vs. POM Supplement S2).

Predictive performance of POM, OLHM, and OBHM was similar for cheatgrass, with 55–60 % of predictions within 1 cover class. Each of the three models also provided 23 % accuracy based on the hold-out dataset. The first and third quantiles of the posterior distribution of Kendall's τ based on predictions from POM, OLHM, and OBHM were 18–24, 15.7–22.5, and 16.6–19.2 %, respectively.

4.2. BIG SAGEBRUSH (*Artemisia tridentata*)

The ecological conclusions concerning factors that influence distribution and abundance of big sagebrush in the Monument varied substantially among the different models (Table 2). There was no meaningful biological association of percent sagebrush cover with fire based on MLR Bayesian estimates. Importantly, using a hurdle-at-zero model, we found that abundance increased in the absence of fire and the probability of sagebrush being absent increased in burned plots (Table 2 OBHM and OLHM results). This finding was consistent with the documented intolerance of big sagebrush to fire and its slow recovery, particularly

Table 2. Comparison of Bayesian posterior distribution means and standard deviations (SD) for modeling big sagebrush using four different approaches: midpoint values as the response in a Bayesian multiple linear regression model (MLR); Bayesian proportional odds model with response all observed cover classes (POM); Bayesian ordinal logistic hurdle model (OLHM) which is a two-part model with logistic latent variable for cover classes $\{1, \dots, 6\}$ and a separate logistic regression for zero cover class; and ordinal beta hurdle model (OBHM) which assumes a beta distribution for the non-zero cover class probabilities and a separate logistic regression for the zero cover class

Model	p_D	Response	Topography		Topography ²		Fire		Dist. to Bd.	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
MLR	3.0	$\{0, \dots, 0.975\}$	NS		NS		-0.086	0.008	NS	
POM	7.2	$\{0, \dots, 6\}$	0.515	0.563	NS		-2.147	0.184	NS	
OLHM	4.0	$I(Y = 0)$	-0.949	0.698	NS		2.086	0.231	-0.533	0.262
	7.6	$\{1, \dots, 6\}$	-0.687	1.037	-5.79	4.853	-1.12	0.381	NS	
OBHM	4.0	$I(Y = 0)$	-0.925	0.607	NS		1.996	0.202	-0.353	0.231
	4.2	$\{1, \dots, 6\}$	-0.920	0.659	-8.787	2.571	-1.099	0.186	NS	

All models included a linear and quadratic term for topography variable (denoted as Topography and Topography²), an indicator for burned areas (denoted Fire), and the distance to boundary variable divided by 1000 (denoted as Dist. to Bd.). NS represents the situation when the given variable's 90 % posterior interval included 0 and a reduced model excluding that variable was fitted for inferences. Note for both OLHM and OBHM the abundance portion of the model has response listed as $\{1, \dots, 6\}$. The estimated posterior mean for ϕ was 2.13 with SD = 0.33 in OBHM. p_D is the effective number of parameters for each model

in large burned expanses where seed dispersal was limited (Miller et al. 2013). The estimated effect size for fire using OBHM was larger in magnitude and had less uncertainty compared to OLHM (OBHM and OLHM with response $\{1, \dots, 6\}$ Table 2). We found no meaningful associations between distance to boundary and sagebrush presence or abundance using any of the models (Table 2).

The estimates based on OBHM and OLHM suggested a quadratic association with topography, indicating that sagebrush was most abundant in flat areas with deep soil accumulations (Table 2). This was consistent with well-established ecological knowledge of sagebrush (Miller et al. 2013) as a deep-rooted perennial shrub. However, OLHM parameter estimates had larger uncertainty relative to OBHM. Conversely, the estimated relationship between topography and sagebrush abundance was non-existent using MLR and weakly linearly decreasing using POM (Table 2). The POM point estimates for fire and topography were consistent with the non-Bayesian estimates (Sect. 2, Fig. 3).

For sagebrush, OLHM had the lowest DIC criteria (DIC = 840) of the three viable models and was only 2 units less than POM (DIC = 842). However, based on plotting the indicator for fire versus POM latent residuals (calculated as $\overline{M_i} - \exp(X\beta^t)/(1 + \exp(X\beta^t))$ where β^t is the t posterior draw and M_i is the midpoint value for observation i), there was strong evidence for the need to separate the sagebrush zero-class observations from the non-zero cover classes. Similar to our cheatgrass findings, there was no (or very weak) support for OBHM (DIC = 852) compared to OLHM (DIC = 840). In terms of our predictive assessment, POM, OLHM, and OBHM all generated predicted values of zero cover class for the hold-out data resulting in 90.4 % accuracy and only 9.6 % were missclassified as zero cover class. Similar to the model-building dataset, the hold-out dataset was dominated by zeros with 264 of 292 observations with zero sagebrush percent cover.

5. DISCUSSION

Ecological abundance data are often recorded in ordinal classes even though the latent continuously distributed abundance variable is of primary interest. In the case of vegetation surveys, visually estimating cover classes compared to measuring percentage areal cover as a continuous variable leads to reduced time in the field, can yield lower observer variability compared with other methods, and can enable accumulations of larger sample sizes (Irvine and Rodhouse 2010 and citations therein). However, the appropriate use of ordinal regression in these applications has lagged behind other domains (Schabenberger 1995; Guisan and Harrell 2000; Stroup 2014), although see Rodhouse et al. 2014 for a recent example. We suspect part of this delay is due to unfamiliarity with interpreting odds ratios of cumulative category probabilities, as is done with POM and OLHM.

In this paper, we have expanded the ordinal toolkit by developing a Bayesian model that specifies a beta-distributed latent variable which allows for a more intuitive interpretation in terms of percent cover (the abundance variable of interest) rather than on the ordinal cover classes. In addition to the interpretational benefits, the OBHM uses all the available information regarding the known cutpoint values. Although our ordinal beta hurdle model was applied to plant cover data, our approach is actually applicable to any situation where the biological variable of interest is measured on an interval. In order to facilitate rapid assimilation of our model into ecological practice, we fitted our Bayesian models using freely available JAGS MCMC software, and provide code as a supplement for practitioners (available at <https://irma.nps.gov/DataStore/Reference/Profile/2231866>).

Ecology is often summarily described as the study of distribution and abundance of organisms (Andrewartha and Birch 1954), underscoring the need for making integrated statistical modeling methods available and tractable for use by ecologists. The over-abundance of zeros (i.e., absence or failure to detect a species) presents an interesting challenge to address and zero-inflated and zero-augmented models have been developed for some data types (binary, counts, Gaussian continuous; Martin et al. 2005; Wenger and Freeman 2008). In this paper, we introduced zero-augmented model options for ordinal responses. We found the independent processes assumption of the hurdle-at-zero model more compelling for our plant species datasets, but note that the zero-inflated approach may also be worth pursuing for other situations (e.g., Kelley and Anderson 2008).

In our empirical examples, we found striking differences in conclusions drawn from our hurdle-at-zero models that would not have been apparent otherwise. Specifically, cheat-grass absence (conversely presence or establishment) was more likely in unburned plots and on cooler north-facing slopes. The hurdle-at-zero models allowed for distinguishing these distinct patterns from those associated with abundance. Surprisingly, the higher cover classes were not more likely in burned plots as might have been expected, and were in fact increasingly likely on flat areas, reflecting a quadratic form of relationship. Similar kinds of subtle but important insights were also evident for big sagebrush, most notably regarding the species' association with flat, deep-soiled playas and benches in the study area. These nuances would have been missed by using a POM and even more misleading conclusions would have been drawn based on MLR estimates. In our case, using exploratory data analysis techniques and DIC for model comparison provided multiple supporting lines of evidence

that a hurdle-specification was needed for modeling cheatgrass and sagebrush distribution and abundance.

One consideration is that by specifying a two-part model, the abundance portion of the model can suffer from diminishing sample size (Lachenbruch 2002). For example, with sagebrush there were 1040 zeros and only 129 non-zeros recorded. The OBHM outperformed the other ordinal models in terms of less uncertainty in posterior distributions, but sample size and a very sparse distribution (e.g., rarity) could lead to estimation instability. Our simulations suggested that even for smaller sample sizes ($n = 64$) specifying a beta latent variable for the non-zero cover classes was substantially less biased than assuming a logistic latent variable under non-constant variance (OLHM vs OBHM). The slight bias observed for scenarios in which many observations are close to 0 or 1 is an artifact of beta regression generally and not specific to our ordinal approach (Kosmidis and Firth 2010; Gruen et al. 2012). Furthermore, in the case of constant variance (i.e., generating ordinal data by aggregating a logistic latent variable), specifying a beta latent variable would produce, on average, slightly negatively biased estimates. But the bias was much greater for the POM slope coefficient estimates when the latent variable was not specified as the logistic. Across all the scenarios we investigated, the OBHM displayed less uncertainty in the estimates consistent with our data analysis results. Some of the observed reduction in uncertainty with OBHM was likely because cutpoint parameters were fixed to known values and did not need to be estimated (as reflected in reduction of p_D for OBHM compared to OLHM Tables 1 and 2).

Although in our empirical example, OBHM was not the lowest ranked model using DIC for either species, we feel the use of OBHM holds appeal as it provides a more straightforward interpretation and link to the biological variable of interest, continuous percent cover. For sagebrush, the estimated posterior mean for ϕ was 2.13 (SD = 0.33) and for cheatgrass it was only 1.44 (SD = 0.072). Based on our simulation study, specifying a larger ϕ value led to a clearer visual separation between datasets generated by discretizing a beta versus logistic latent variable (Supplemental Information S1 Figure A1). Based on additional simulations (not shown), our empirically estimated ϕ values were consistent with posterior estimates based on data generated by discretizing a logistic latent variable (average posterior mean was 2 with minimum posterior mean = 1.75 and maximum posterior mean = 2.7 based on 95 simulated datasets). We suspect for datasets with larger estimated ϕ values there will be more support for using OBHM for both prediction and explanation of ecological patterns. Even in our empirical examples, there were substantial gains in terms of less uncertainty and a clearer interpretation that may support using OBHM for ecological conclusions in some applications.

Occupancy models and other approaches that account for detectability of a species are examples of zero-inflated models (Mackenzie et al. 2006). In these applications, the zeros are more appropriately considered as a mixture of false zeros because the species was imperfectly observed due to detection errors and true absences (Martin et al. 2005). In the monitoring program that supplied our motivating example, training of observers was considered paramount and the probability of false zeros assumed to be 0. However, we anticipate expanding the observation process component of our model to account for detection errors within the zero-class observations and possibly category classification errors in future (e.g.,

Royle and Link 2005; Holland and Gray 2011). A double-observer field study was conducted with our monitoring project (Esposito et al. 2016) which provided some information about the observation process that could inform such an extension.

In our study, the GRTS sampling design algorithm increases spatial dispersion among points and, consequently, we proceeded with the assumption of spatial independence and accounted for potential spatial structure by including spatially explicit predictors (e.g., topography and distance to roads and boundaries). But another extension to our OBHM approach could be to modify the beta latent process component of the model by way of including a random effect in the mean structure to account for possible spatial or temporal dependence among observations. Another option might be to assume a latent spatial process appropriate for modeling spatially correlated continuous proportions (Eskelson et al. 2011; Feng et al. 2015) in lieu of our zero-augmented beta distribution (Sect. 3). This is essentially the same approach developed for a Bayesian spatial cumulative probit model; a Gaussian-distributed variable with assumed spatial variance–covariance structure was used as the latent variable (e.g., Higgs and Hoeting 2010; Schliep and Hoeting 2013). Although beyond the current study scope, these and other extensions (e.g., multivariate approaches to account for species interactions) provide exciting opportunities for further statistical method development for ordinal ecological data applications. Our beta latent variable construction could be easily incorporated into such future work.

ACKNOWLEDGEMENTS

We thank Dr. Megan D. Higgs for early discussions on this work and her assistance with WinBUGS code for clipping latent distributions. Dr. Brian Gray provided encouragement and interest in this work and we are appreciative. We also thank Dr. Andrew Hoegh, two anonymous reviewers', and our associate editor's comments and suggestion for revising our paper. The work by K. M. Irvine was funded through an Interagency Agreement P12PG70586 with the National Park Service. T. J. Rodhouse was funded by Upper Columbia Basin Network Inventory and Monitoring Program of the National Park Service. I. N. Keren's participation was secured by an interagency agreement with Montana State's Institute on Ecosystems with funding by North Central Climate Science Center. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

[Received June 2016. Accepted August 2016.]

REFERENCES

- Agresti, A. (2010), *Analysis of ordinal categorical data* John Wiley and Sons, Hoboken, NJ, USA.
- Agresti, A., and Kateri, M. (2014), Some Remarks on Latent Variable Models in Categorical Data Analysis. *Communications in Statistics - Theory and Methods*, **43**, 801–814.
- Albert, J. H., and Chib, S. (1993), Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.
- Ananth, C. V., and Kleinbaum, D. G. (1997), Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology*, **26**, 1323–1333.
- Andrewartha, H. G., and Birch, L. C. (1954), *The distribution and abundance of animals*. University of Chicago Press, Chicago, Illinois, USA.

- Bonham, C. D. (1989), *Measurements for terrestrial vegetation* John Wiley and Sons, New York, NY.
- Braun-Blanquet, J. (1932), *Plant sociology. The study of plant communities*. McGraw-Hill, New York, NY, US.
- Branscum, A. J., Johnson, W. O., and Thurmond, M. C. (2007), Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian New Zealand Journal of Statistics*, **49**, 287–301.
- Chambers, J. C., Roundy, B. A., Blank, R. R., Meyer, S. E., and Whittaker, A. (2007), What makes Great Basin sagebrush ecosystems invisable by *Bromus tectorum*? *Ecological Monographs*, **77**, 117–145.
- Chen, J., Shiyomi, M, Hori, Y., and Yamamura, Y. (2008a), Frequency distribution models for spatial patterns of vegetation abundance. *Ecological Modelling*, **211**, 403–410.
- Chen, J., Shiyomi, M., Bonham, C. D., Yasuda, T., Hori, Y., and Yamamura, Y. (2008b), Plant cover estimation based on the beta distribution in grassland vegetation. *Ecological Research*, **23**, 813–819.
- Christensen, R. H. B. (2014), Ordinal:Regression models for ordinal data. R package version 2014.12-22. Available at <http://cran.r-project.org/web/packages/ordinal/index.html>.
- Congdon, P. (2005), *Bayesian models for categorical data*. John Wiley and Sons, Hoboken, NJ, USA.
- Coudun, C. and Gegout, J.-C. (2007). Quantitative prediction of the distribution and abundance of *Vaccinium myrtillus* with climatic and edaphic factors. *Journal of Vegetation Science*, **18**, 517–524.
- Damgaard, C. (2009), On the distribution of plant abundance data. *Ecological Informatics*, **4**, 76–82.
- Damgaard, C. (2012), Trend analyses of hierarchical pin-point cover data. *Ecology*, **93**, 1269–1274.
- Daubenmire, R. F. (1959), A canopy-coverage method. *Northwest Science*, **33**, 43–64.
- Davies, K. W., Boyd, C. S., Beck, J. L., Bates, J. D., Svejcar, T. J., and Gregg, M. A. (2011), Saving the sagebrush sea: an ecosystem conservation plan for big sagebrush plant communities. *Biological Conservation*, **144**, 2573–2584.
- Davies, K. W., Nafus, A. M., and Madsen, M. D. (2013), Medusahead invasion along unimproved roads, animal trails, and random transects. *Western North American Naturalist*, **73**, 54–59.
- Duff, T. J., Bell, T. L., and York, A. (2011), Patterns of plant abundances in natural systems: is there value in modelling both species abundance and distribution?. *Australian Journal of Botany*, **59**, 719–733.
- Eskelson, N. I., Madsen, L., Hagar, J. C., and Temesgen, H. (2011), Estimating riparian understory vegetation cover with beta regression and copula models. *Forest Science*, **57**, 212–221.
- Esposito, D. M., Shanahan, E., and Rodhouse, T. J. (2016), UCBN and GRYN Sagebrush Steppe Vegetation Monitoring: Double Observer Study 2015. John Day Fossil Beds National Monument-Clarno Unit and City of Rocks National Reserve. Natural Resource Reporting Series NPS/UCBN/NRR-2016/1052. National Park Service, Fort Collins, Colorado.
- Fahrmeier, L. and Tutz, G. (2001), *Multivariate statistical modelling based on generalized linear models*. Springer, Berlin.
- Feng, X., Zhu, J., and Steen-Adams, M. M. (2015), On regression analysis of spatial proportional data with zero/one values. *Spatial Statistics*, **14**, 452–471.
- Ferrari, S. L. P. and Cribari-Neto, F. (2004), Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799–815.
- Gelbard, J. L. and Belnap, J. (2003), Roads as conduits for exotic plant invasions in a semiarid landscape. *Conservation Biology*, **17**, 420–432.
- Gelfand, A. E. and Smith, A. F. M. (1990), Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.
- Gruen, B., Kosmidis, I., and Zeileis, A. (2012), Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software*, **48**, 1–25.
- Guisan, A. and Harrell, F. E. (2000), Ordinal response regression models in Ecology. *Journal of Vegetation Science*, **11**, 617–626.
- Hall, D. B. (2000), Zero-inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*, **56**, 1030–1039.

- Heilbron, D. C. (1994), Zero-altered and other Regression Models for Count Data with Added Zeros. *Biometrical Journal*, **36**, 531–547.
- Higgs, M. D. and Hoeting, J. A. (2010), A clipped latent variable model for spatially correlated ordered categorical data. *Computational Statistics and Data Analysis*, **54**, 1999–2011.
- Higgs, M. D. and Ver Hoef, J. M. (2012), Discretized and Aggregated: modeling dive depth of Harbor Seals from Ordered Categorical data with temporal autocorrelation. *Biometrics*, **68**, 965–974.
- Holland, M. D. and Gray, B. R. (2011), Multinomial mixture model with heterogeneous classification probabilities. *Ecological and Environmental Statistics*, **18**, 257–270.
- Irvine, K. M. and Rodhouse, T. J. (2010), Power analysis for trend in ordinal cover classes: implications for long-term vegetation monitoring. *Journal of Vegetation Science*, **21**, 1152–1161.
- Ishwaran, H. (2000), Univariate and multivariate ordinal cumulative link regression with covariate specific cutpoints. *The Canadian Journal of Statistics*, **28**, 715–730.
- Kelley, M. E. and Anderson, S. J. (2008), Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. *Statistics in Medicine*, **27**, 3674–3688.
- Kim, J.-H. (2003), Assessing practical significance of the proportional odds assumption. *Statistics and Probability Letters*, **65**, 233–239.
- Kosmidis, I. and Firth, D. (2010), A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, **4**, 1097–1112.
- Lachenbruch, P. A. (2002), Analysis of Data with Excess Zeros. *Statistical Methods in Medical Research*, **11**, 297–302.
- Lambert, D. (1992), Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1–14.
- Larrabee, B., Scott, H. M., and Bello, N. M. (2014), Ordinary least squares regression of ordered categorical data: inferential implications in practice. *Journal of Agricultural, Biological, and Environmental Statistics*, **19**, 373–386.
- Link, W. A. and Eaton, M. J. (2012) On thinning of chains in MCMC. *Methods in Ecology and Evolution*, **3**, 112–115.
- Mackenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., and Hines, J. E. (2006), *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier Academic Press, Burlington, MA, USA.
- Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005), Zero tolerance ecology: improving ecological inference by modeling the source of zero observations. *Ecology Letters*, **8**, 1235–1246.
- Milberg, P., Bergstedt, J., Fridman, J., Gunnar, O., Westerberg, L. (2008), Observer bias and random variation in vegetation monitoring data. *Journal of Vegetation Science*, **19**, 633–644.
- Millar, R. B. (2009), Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics*, **65**, 962–969.
- Miller, R. F., Chambers, J. C., Pyke, D. A., Pierson, F. B., and Williams, C. J. (2013), A review of fire effects on vegetation and soils in the Great Basin region: response and ecological site characteristics. RMRS GTR-308. USDA Forest Service, Rocky Mountain Research Station, Fort Collins, Colorado, USA.
- Moulton, L. H. and Halsey, N. A. (1995), A Mixture Model with Detection Limits for Regression Analyses of Antibody Response to Vaccine. *Biometrics*, **51**, 1570–1578.
- Neelon, B. H., O'Malley, A. J., and Normand, S.-L. T. (2010), A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, **10**, 421–439.
- Ospina, R. and Ferrari, S. L. (2010), Inflated beta distributions. *Statistical Papers*, **51**, 111–126.
- Plummer, M. (2003), JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, K. Hornik, F. Leisch, and A. Zeileis (eds.) Vienna, Austria. Available at: <http://www.ci.tuwien.ac.at/Conferences/DSC-2003/>

- Plummer, M. (2008), Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–539.
- Plummer, M. (2015), JAGS Version 4.0.0 user manual. Available at <https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/>.
- Reisner, M. D., Grace, J. B., Pyke, D. A., and Doescher, P. S. (2013), Conditions favouring *Bromus tectorum* dominance of endangered sagebrush steppe ecosystems. *Journal of Applied Ecology*, **50**, 1039–1049.
- Rodhouse, T. J., Irvine, K. M., Sheley, R. L., Smith, B. S., Hoh, S., Esposito, D. M., and Mata-Gonzalez, R. (2014), Predicting foundation bunchgrass species abundances: model-assisted decision-making in protected-area sagebrush-steppe. *Ecosphere*, **5**, art208.
- Royle, J. A. and Link, W. A. (2005), A general class of multinomial mixture models for anuran calling survey data. *Ecology*, **86**, 2505–2512.
- Schabenberger, O. (1995), The use of ordinal response methodology in Forestry. *Forest Science*, **41**, 321–336.
- Schliep, E. M. and Hoeting, J. A. (2013), Multilevel latent Gaussian process model for mixed discrete and continuous multivariate response data. *Journal of Agricultural, Biological, and Environmental Statistics*, **18**, 492–513.
- Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002), Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, **64**, 583–639.
- Stevens, D. L., and Olsen, A. R. (2004), Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, **99**, 262–278.
- Stroup, W. W. (2014), Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal*, **106**, 1–17.
- Tamhane, A. C., Ankenman, B. E., and Yang, Y. (2002), The beta distribution as a latent response model for ordinal data (I): estimation of location and dispersion parameters. *Journal of Statistical Computation and Simulation*, **72**, 473–494.
- Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
- Wenger, S. J. and Freeman, M. C. (2008), Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, **89**, 2953–2959.
- Yeo, J. J., Rodhouse, T. J., Dicus, G. H., Irvine, K. M., and Garrett, L. K. (2009), Sagebrush steppe vegetation monitoring protocol. Upper Columbia Basin Network. Version 1.0. Natural Resource Report NPS/UCBN/NRR–2009/142. National Park Service, Fort Collins, CO, USA.