Zero-Inflated Beta Distribution Regression Modeling

Becky TANG, Henry A. FRYE, Alan E. GELFAND, and John A. SILANDER

A frequent challenge encountered with ecological data is how to interpret, analyze, or model data having a high proportion of zeros. Much attention has been given to zero-inflated count data, whereas models for non-negative continuous data with an abundance of 0s are much fewer. We consider zero-inflated data on the unit interval and provide modeling to capture two types of 0s in the context of a Beta regression model. We model 0s due to missing by chance through left-censoring of a latent regression and 0s due to unsuitability using an independent Bernoulli specification. We extend the model by introducing spatial random effects. We specify models hierarchically, employing latent variables, and fit them within a Bayesian framework. Our motivating dataset consists of percent cover abundance of two plant families at a collection of sites in the Cape Floristic Region of South Africa. We find that environmental features enable learning about both types of 0s as well as positive percent cover. We also show that the spatial random effects model improves predictive performance. The proposed modeling enables ecologists to extract a better understanding of an organism’s absence due to unsuitability vs. missingness by chance, as well as abundance behavior when present.

Supplementary materials accompanying this paper appear online.

Key Words: Bayesian inference; Greater Cape Floristic Region; Hierarchical model; Hurdle model; Percent cover; Spatial random effects.

1. INTRODUCTION

A frequent and continuing challenge encountered with ecological data is how to interpret, analyze, or model data having a high proportion of zeros. This zero-inflation problem (Martin et al. 2005) is found in scenarios such as: the binary presence/absence of individual organisms, counts of individuals, ranked abundances as in many vegetation plot datasets, and measures of biomass or proportion of area occupied by individuals in plot-level survey
or remotely sampled data sets. A wealth of literature (e.g., Martin et al. 2005; Veech et al. 2016; Blasco-Moreno et al. 2019) asserts that zeros in a dataset can arise from multiple sources. Such prior studies have dichotomized zeros into “true” and “false” zeros; we avoid such value labels here, preferring stochastic interpretation.

While many statistical techniques have been proposed to address zero-inflation in discrete data, particularly for animal populations (Veech et al. 2016; Blasco-Moreno et al. 2019), less attention has been focused on continuous data. In particular, for count data, we remind the reader of the much-employed zero-inflated Poisson model (ZIP) (Lambert 1992) which introduces an additional point mass at zero. Models such as the zero-inflated negative binomial and other power series distributions, zero-inflated N-mixture models, and a general class of count transform models have also been developed (Hall 2000; Ghosh et al. 2006; Wenger and Freeman 2008; Veech et al. 2016; Siegfried and Hothorn 2020).

The zero-inflation challenge we focus on here is modeling the percent cover of plants within sampling plots yielding data on \([0, 1)\). Ecologists often assess plant community composition and abundance through visual assessments of percent cover, which has been shown to be a robust and repeatable measure of vegetation cover (Vanha-Majamaa et al. 2000). Percent cover is also often calculated by commonly used ordinal scales such as the Braun–Blanquet scale (van der Maarel 2007). In fact, our modeling can be applied to a variety of proportional cover data found in plant and animal surveys, e.g., marine invertebrates on hard substrate or coral reef communities (Bell and Galzin 1984), leaf cover estimates for plant phenological studies (Xie et al. 2018b,a), and biomass scaled as a proportion of area or organ mass (Jenkins et al. 2003). Absence, i.e., a zero, can be assumed to arise from one of three sources: unsuitability, random chance, or detection error. Unsuitability results from the biotic and abiotic factors impacting plant populations. If an area is suitable for a plant species but not found, we assume that this is due to random chance (i.e., it has not yet dispersed to that site). The explanation of failed detection of presence is not pursued here. In general, it requires detection/non-detection data from more than one independent visit to a sample unit. Further, it is extremely unlikely in our example system which is comprised almost exclusively of evergreen perennial species in an open shrubland. Lastly, we do not address the notion of “1-inflation” here.

Percent cover provides a challenge in that the data are continuous on \([0, 1)\). Though Beta regression (Ferrari and Cribari-Neto 2004) is a customary choice to model such data (Douma and Weedon 2019), it has only been recently elaborated in detail for the context of vegetation percent cover (Damgaard and Irvine 2019). However, as we clarify below, these models are not adequate for what is truly zero-inflated, continuous data.

The key issue here is that discrete distributions explicitly place point mass at 0, while continuous distributions do not. So, a different mechanism is required to create mass at 0. One method of introducing zeros for continuous data on \([0, 1)\) is through the left-censoring of a latent random variable at 0. For example, consider a continuous variable \(W\) with support on \((-1, 1)\). Assume that the actual observation is a left-censored at 0 realization of this variable. Then, the probability of observing a 0 becomes \(P(W \leq 0)\). This approach is similar to the Tobit model which has been long used in economics (Amemiya 1984) and applied in remote sensing (Peterson 2005). The Tobit model introduces a regression
for latent Gaussian variables and creates point mass at 0 through left-censoring at 0 (Chib 1992; Long and Long 1997).

A different approach to modeling zero-inflated data is a hurdle model (Mullahy 1986). Hurdle models explain a chosen measure of abundance given presence. In our setting, they model the positive percent coverages, ignoring the observed 0s. Separately, the observation at each site is given a binary response according to presence or absence, modeled through a binary regression. This zero-inflation for a Beta distribution, using site-specific independent Bernoulli variables, has been termed a zero-inflated Beta model in Ospina and Ferrari (2012) and Ospina and Ferrari (2010). They refer to this hurdle model as BEZI (BEta Zero Inflated). For observations on \([0, 1]\), neither of the foregoing models (BEZI or left-censored) is really a zero-inflated continuous data specification. In both cases, a single mass at 0 is introduced. We build upon the previous literature and propose a new model that accommodates two sources of 0 for percent cover, analogous to the zero-inflated Poisson model for counts.

For interpretation, the left-censored regression type of 0 will be referred to as absence by chance (Blasco-Moreno et al. 2019). The Bernoulli point mass at 0 will be referred to as absence due to unsuitability. Our contribution here is to formalize a zero-inflated Beta regression model for percent cover. This entails a regression within a left-censored Beta specification to capture absence by chance as well as a regression to capture absence due to unsuitability. Can we separate both sources of 0s? As we discuss below, the answer is yes given the use of informative priors for identifiability.

As the environmental covariates may not be sufficient to explain all the systematic variability in the data, introducing spatially correlated random effects is expected to provide improved model performance. Therefore, employing the geo-coded locations of the sampling sites, we offer a spatial random effects version of the foregoing specification. Spatial zero-inflated models have been proposed in the literature (e.g., Agarwal et al. 2002; Rathbun and Fei 2006), and we extend this work to our proposed zero-inflated Beta regression model. This raises the question of model comparison; is the spatial model preferred? We address this through novel model comparison for these zero-inflated distributions. As a case study, we consider two plant families (the Restionaceae and Crassulaceae) located across a suite of sample plots in the Cape Floristic Region (CFR) in South Africa.

The format of the paper is as follows: Sect. 2 describes the CFR dataset which motivates our modeling. Section 3 presents modeling details for the non-spatial case, offers model fitting details as well as model assessment and comparison approaches, and finishes with the spatial model. Section 4 offers simulation investigations to serve as a proof of concept and compare various models. Section 5 presents the results of analyses for the CFR data.

### 2. THE DATASET

We consider percent cover for species within two plant families: Crassulaceae and Restionaceae in the Cape Floristic Region (CFR). The Crassulaceae found here are typically perennial succulents that contribute to the CFR’s diversity of succulent flora (Born et al. 2006). They are arid adapted plants; we would expect them to inhabit drier and warmer microhabitats within the CFR, likely in sites protected by fire. The Restionaceae are ever-
green reed-like plants that are an iconic component of the CFR. They are found in a wide range of habitats across the nutrient poor soils of the CFR (Linder 2001).

Our data were collected between 2010 and 2011 from plots concentrated in two regions of South Africa: 61 plots in the Cape Point section of Table Mountain National Park and 119 plots in the Baviaanskloof Mega-Reserve (Fig. 1). Following the relevé method (Ellenberg and Mueller-Dombois 1974), a well-established inventory protocol used to classify vegetation and inventory species (e.g., Schaminée et al. 2009; Dengler et al. 2011), each 10×5 m plot was sampled for its floristic composition using observations of percent cover. Histograms of the observed percents are presented in Fig. 2a. 49% of the observations for Crassulaceae and 31% of observations for Restionaceae are 0, arguing for our proposed introduction of zero-inflation.

The plots are associated with a variety of environmental variables georeferenced for each plot and largely derived from the WorldClim (Hijmans et al. 2005) and Schulze (Schulze 1997) climate datasets. We employ four of these variables in our regression modeling: mean annual evapotranspiration (1-min grid cell), mean annual precipitation (30 arc second), average minimum temperature for July (the austral winter; 1-min grid cell), and rainfall concentration index [1-min grid cell; scaled 0 (rainfall is evenly divided across all months) to 100 (rain only falls one month of the year)]. Figure S1 displays histograms of the environmental covariates, and Fig. 2b plots the positive percent covers of both species against the environmental covariates as well as the distribution of environmental covariates for observed 0s. As our observations are spatially referenced, we may also be interested in investigating if the data are spatially correlated for a given family. The percent covers in Baviaanskloof and Cape Point plotted by longitude-latitude are presented in Figs. 2c and S2, respectively. Throughout this work, “S” denotes Supplementary Information.
3. BUILDING THE MODEL

As we have discussed above, we propose a site-level model that will have a Beta distribution specification for the continuous observations. We first introduce a left-censored Beta regression model that admits zeros due to chance, and we investigate implications of the choice of a fixed parameter. We then add the ability to capture zeros that occur due to unsuitability at the site via a Bernoulli trial.

Figure 2. Exploratory analysis of observed percent cover for Crassulaceae (pink) and Restionaceae (navy). Where appropriate, orange corresponds to observed 0 percent cover (Color figure online).
We first introduce a 0 through left-censoring. In a general setting, for site \(i\), let the observation, \(Y_i = \max(0, W_i)\), where \(W_i\) is a continuous variable on \((-\infty, \infty)\) with distribution \(f_i\) so \(P(Y_i = 0) = P(W_i \leq 0) = \int_{-\infty}^{0} f_i(w) \, dw\). When \(f_i\) is a normal distribution this model is a Tobit model (Tobin 1958). In our context, we take \(f_i\) to be the Beta distribution. However, the Beta has bounded support on \((0, 1)\). Therefore, we propose extending the Beta distribution: if \(V \sim \text{Beta}(\alpha, \beta)\), then for \(0 < a < \infty\), \(W = (a + 1) V - a\) provides an extended or re-scaled Beta on \((-a, 1)\). The probability of a negative \(W\) provides our Beta mass at 0: \(P(W \leq 0) = P(V \leq \frac{a}{a+1} \equiv c)\) (Fig. 3a).

We follow a Bayesian approach for model fitting. We introduce a latent Beta variable \(V_i\) such that \(Y_i = \max\{0, W_i = (a + 1)V_i - a\}\). Then, for all \(Y_i > 0\), we immediately know \(V_i\). For \(Y_i = 0\), the associated \(V_i\) is less than or equal to \(c\) by construction. Thus, introducing the latent \(V_i\) allows us to have the “complete” data. In fitting the model, we work with a convenient reparameterization for Beta regression (Ferrari and Cribari-Neto 2004). We employ \(\mu_i = \frac{\alpha_i}{\alpha_i + \beta_i}\), the mean parameter and \(v_i = \alpha_i + \beta_i\), the so-called sample size parameter. The distribution of the extended Beta on \((-a, 1)\) is also parameterized in
Figure 4. Predicted and true probabilities of zero for simulated data for the LCEB model where \( \logit(\mu) = -0.5 - 0.5x_1 \) and \( \nu = 3 \). Each plot corresponds to data generated with different \( c_{\text{true}} \), with colors denoting posterior mean probabilities for models fit with different \( c_{\text{fixed}} \). For a given \( c_{\text{true}} \), predicted probabilities are generally quite similar across the \( c_{\text{fixed}} \). The notable exception is the under-prediction from the model fit with \( c_{\text{fixed}} = 0.1 \) to data generated with \( c_{\text{true}} = 0.9 \).

terms of \( \mu_i \) and \( v_i \). We model \( \logit(\mu_i) = X_i^T \delta \), and \( \log(v_i) = X_i^T \psi \), where \( X_i \) is a vector of environmental covariates associated with site \( i \).

With regard to addressing the effect of choice of \( a \), Appendix offers some analytical investigation. The choice of \( a = 1 \) mirrors the Tobit model by providing matching support above and below 0. Ideally, we would take \( a \) to be a model parameter and learn about its value during model fitting, along with \( \delta \) and \( \nu \). In implementation, this leads to identifiability issues: \( a \) (or equivalently, \( c \)) and the intercept \( \delta_0 \) for \( \mu \) compete to explain the probability of 0. One approach to addressing this issue would be to use an informative prior on \( c \) or \( \delta_0 \). However, we remark that the value of \( c \) does not have interpretation; it is a mechanism used to obtain 0s. Practically, this suggests holding \( c \), equivalently \( a \), fixed at a pre-specified value during model fitting and then doing out-of-sample model comparison.

In this regard, and following Appendix, through simulations we briefly explore sensitivity of inference to the choice of \( a \). The simulations are performed by generating data under a \( c_{\text{true}} \), fitting the censored model with several different values of \( c_{\text{fixed}} \), and examining how recovery of the covariate effects and probabilities of 0 are affected. The regression for \( \mu \) is \( \logit(\mu) = \delta_0 - 0.5x_1 \), with different \( \delta_0 \) and \( \nu \) in each simulation. We evaluate credible interval coverage of \( \delta_1 \) and root mean square predictive error (RMSPE) of the probabilities of 0 across 100 simulations for each pair \( (c_{\text{true}}, c_{\text{fixed}}) \). We find that generally, recovery of the probability of 0 is robust to the choice of fixed \( a \); predicted probabilities on the held-out test set tend to yield similar RMSPEs (Fig. S3a). The largest differences in predictions and RMSPEs occur when \( c_{\text{true}} \) is small and \( c_{\text{fixed}} \) is large, or vice versa (Fig. S4).
Empirical coverage of the credible intervals for $\delta_1$ achieves or is slightly below nominal when $c$ is fixed at the true value, but tends to decrease as the difference between $c_{\text{true}}$ and $c_{\text{fixed}}$ increases (Fig. S3a). This decreasing coverage is exacerbated when data are generated with smaller intercept $\delta_0$, holding everything else fixed (e.g., Simulations 1–3, Simulations 4–6). This is due to the larger number of 0s in the data induced by the smaller intercept. However, in almost all pairs of $(c_{\text{true}}, c_{\text{fixed}})$, the sign of $\delta_1$ is correctly estimated (Fig. S4). Therefore, even if $c$ is fixed incorrectly, we expect to recover the correct covariate relationships.

From the simulations, unsurprisingly, we find that fixing $c$ to be large when the data contain a high proportion of 0s leads to better performance, and similarly for small $c$ with a smaller proportion of 0s (Fig. 4). Inference is the least robust when the magnitude of the difference between $c_{\text{true}}$ and $c_{\text{fixed}}$ is large. Thus, if the modeler seeks to choose a single value \textit{a priori}, perhaps fixing $c = 0.5$ ($a = 1$) is the most natural option. Again, we suggest analyzing the data with differences choices of $c$ and performing sensitivity analysis to determine the best fitting model. For the remainder of the text, we refer to this left-censored extended Beta with single-zero model as “LCEB.” In the following subsection, we build off the LCEB to develop a model that incorporates two sources of zero.

### 3.2. Zero-Inflated Beta Modeling: The Non-spatial Case

Let $Z_i$ be a Bernoulli variable with $P(Z_i = 1) = \pi_i$ and let $f_i$ be a density on $(0, 1)$. Then, suppose $Y_i = 0$ if $Z_i = 1$, and $Y_i \sim f_i(y)$ if $Z_i = 0$. Here, $P(Y_i = 0) = \pi_i$. This is the BEZI model of Ospina and Ferrari (2010). It is a hurdle model analogous to the setting where $Y_i$ is a count variable and $f_i$ is the Poisson distribution truncated at 1 (Mullahy 1986).

If $Z_i = 0$, then let $Y_i$ be a realization of the left-censored extended Beta regression model described in Sect. 3.1: $Y_i = \max(0, W_i)$ where $W_i = (a + 1)V_i - a$ and, using the Ferrari and Cribari-Neto (2004) parameterization, $V_i \sim \text{Beta}(\mu_i, v_i)$. Now, $P(Y_i = 0) = P(Z_i = 1) + P(Z_i = 0 \cap W_i \leq 0) = \pi_i + (1 - \pi_i) \int_0^a f_i(w) \, dw$ (so $P(Y_i > 0) = (1 - \pi) \int_0^1 f_i(w) \, dw$).

We immediately see the two sources of 0s as well as the parallel construction with the familiar zero-inflated count model (see Fig. 3b for a comparison between LCEB and the proposed zero-inflated Beta model). In this regard, we will refer to $(1 - \pi_i) \int_0^1 f_i(w) \, dw$ as the probability of absence by chance with $\pi_i$ the inflated probability of absence which we ascribe to unsuitability.

We can identify the two sources of 0s if $a$ is held fixed because the $\mu_i$ and $v_i$ are informed by all of the data, while only the 0’s inform about $\pi_i$. However, informative prior information is needed to capture the relative magnitudes of the chances of each source. In the absence of further knowledge, this is addressed via out-of-sample model comparison. We refer to our proposed zero-inflated Beta model as $M_1$.

Again, following the above, for all $Y_i > 0$, we immediately know $V_i$. For $Y_i = 0$, we know that if the observation arose from the Beta, by construction, the associated $V_i$ is less than or equal to $\frac{a}{1 + a} = c$. Our zero-inflated model takes the form

$$\begin{align*}
P(Y_i = 0|\pi_i, \mu_i, v_i) &= \pi_i + (1 - \pi_i)P(V_i \leq c|\mu_i, v_i) \\
f(y_i|\pi_i, \mu_i, v_i) &= (1 - \pi_i)f_{v_i}(v_i|\mu_i, v_i), \quad y_i > 0. \tag{1}
\end{align*}$$
Here, a latent indicator variable $Z_i$ is introduced to determine the source of an observation: If $Z_i = 1$, then the associated $Y_i$ arises as a 0 from the degenerate process, and if $Z_i = 0$, then $Y_i$ is a realization of the extended Beta. Then, we have that $P(Y_i = 0, Z_i = 1 | \pi_i, \mu_i, \nu_i) = \pi_i$ and $P(Y_i = 0, Z_i = 0 | \pi_i, \mu_i, \nu_i) = (1 - \pi_i) P(V_i \leq c | \mu_i, \nu_i)$.

Given the parameters $\pi_i, \mu_i, \nu_i$, and the latent $Z_i$, the likelihood is:

$$L(\pi, \mu, \nu, Z; Y) = \prod_{i=1}^{n} \pi_i^{z_i} \cdot \left( (1 - \pi_i) f_{y_i}(y_i | \mu_i, \nu_i) \right)^{1-z_i}. \quad (2)$$

We model $\text{logit}(\pi_i) = G_i^T \gamma$, and as in the case of the single-zero left-censored Beta, $\text{logit}(\mu_i) = X_i^T \delta$, and $\text{logit}(\nu_i) = \Phi_i^T \psi$ where $X_i, G_i, \Phi_i$. The context will determine the commonalities between $X_i$ and $G_i$ with regard to explaining $\pi$ and the parameters of the Beta distribution. Then, we wish to infer about the regression coefficients, $\gamma, \delta, \nu$, as well as the latent $Z_i$ associated with $Y_i = 0$. With priors, the posterior distribution is proportional to

$$\prod_{i=1}^{n} f_{Y_i | Z_i, \pi, \mu, \nu}(Y_i | Z_i, \pi, \mu, \nu) \cdot \prod_{i=1}^{n} f_{Z_i | \gamma}(Z_i | \gamma) \cdot f_Y(\gamma) \cdot f_\delta(\delta) \cdot f_\psi(\psi). \quad (3)$$

Prior choices follow in the next subsection.

### 3.3. Model Fitting

We use a Gibbs sampler with Metropolis updates for the $\delta, \gamma, \psi$. If $Y_i > 0$, we immediately have $Z_i$ and $V_i$. Then, for all $i$ where $Y_i = 0$, at every iteration within the sampler we sample $Z_i$ from its Bernoulli full conditional with success probability $P(Z_i = 1 | Y_i = 0) = \pi_i / (\pi_i + (1 - \pi_i) P(V_i \leq c))$. If $Z_i = 0$, we sample an associated latent $V_i$ from its conditional Beta distribution restricted to the interval $(0, c)$, analogous to sampling from truncated Normal variables in Bayesian Tobit modeling (Chib 1992). Given $Z$ and $V$, the parameters $\gamma$ are independent of the parameters $\delta, \psi$. In the sequel, we assume $\nu_i = \nu$, i.e., a common sample size parameter across sites. This is in accord with Damgaard and Irvine (2019) who use the parameterization, $\mu_i$ with common shape parameter $\delta = 1/(\nu + 1)$.

Distinguishing the two types of zeros in the data is challenging. The magnitudes of $\pi_i$ and $\mu_i$ are strongly influenced by their intercepts $\gamma_0$ and $\delta_0$, respectively. A large positive $\gamma_0$ will result in $\pi_i$ close to one, encouraging unsuitability absence; a large negative $\delta_0$ will result in $\mu_i$ small, encouraging absence by chance. Recall that the probability of a zero at location $i$ is $\pi_i + (1 - \pi_i) P(V_i \leq c | \mu_i, \nu_i)$. The second term leads to difficulty in separating the source of a zero due to weak identifiability of the intercepts. Rather than removing an intercept from $\pi_i$ or $\mu_i$, we choose to adopt a strong prior on one of the intercepts.

This leads to two questions: which intercept, and what prior? The natural choice is the intercept $\gamma_0$ for the incidence of inflated zeros; an ecologist may have a general belief of how (un)favorable a location is for a given species, but may have less knowledge about the expected percent cover. If so, a tight prior centered at the modeler’s belief for $\gamma_0$ can be used. We take this approach in our simulations. However, we suggest performing modeling
comparison or sensitivity analysis with different priors to select a model. Our proposed metrics for modeling comparison follow in the next subsection. We use non-informative or weakly informative priors for all the remaining parameters (see Table S1 for full list).

3.4. MODEL ASSESSMENT AND COMPARISON APPROACHES

If the source of a zero is known—as with the simulated data below—one measure of model assessment is through classification of the type of zero. We use receiver operating characteristic (ROC) curves and the corresponding area under the curve (AUC) for the true source of a zero versus the predicted probability of arising from that source. In practice, we will not know the true source of a zero. Instead, we assess and compare models in terms of their ability to distinguish zero and positive observations. Another metric is Tjur’s $R^2$ statistic, a coefficient of discrimination often used to evaluate logistic regression prediction (Tjur 2009). This statistic is calculated as $R^2 = \bar{\hat{p}}_1 - \bar{\hat{p}}_0$, where $\bar{\hat{p}}_1$ and $\bar{\hat{p}}_0$ are the average of fitted values for successes and failures. Larger AUC and $R^2$ indicate superior classification performance.

The continuous ranked probability score (CRPS) which compares the entire predictive cumulative distribution function, $F(x)$, to an observed held-out data point (Gneiting and Raftery 2007) offers further model comparison. For a continuous CDF, CRPS = \int (F(x) - 1(y < x))^2 \, dx$, where $y$ is the observed value. For a discrete distribution, the ranked probability score replaces the integral with a sum. A smaller (C)RPS indicates that the predictive distribution is more concentrated around the observation and is thus preferred. In the case of our proposed distribution, we have a continuous predictive CDF with a point mass at zero. Let $p_0$ denote the mass at 0. Then, our score takes the form:

$$(C)RPS = \begin{cases} p_0^2 + \int_0^{\infty} (F(x) - 1(y < x))^2 \, dx, & y > 0 \\ (1 - p_0)^2 + \int_{-\infty}^{0} (1 - F(x))^2 \, dx, & y = 0. \end{cases}$$

Thus, we can obtain the (C)RPS for the full distribution, denoted as CRPS$_f$. How well the predictive distribution approximates the positive observations may also be of interest. In this case, we omit the zero observations from the calculation and only use the positive data. We refer to this resulting score as CRPS$_h$. We use Monte Carlo integration to evaluate the integrals (Supplement S2).

Because evaluation using our ranked probability score requires separating observations based on whether they are zero or positive, the concentration of the predictive distribution depends on the number of sources of zeros that are being modeled. For example, it would be inappropriate to use our CRPS$_f$ to compare our proposed model with the BEZI model. However, if we wish to perform sensitivity analysis to assess the effect of prior choice in our proposed model, then the use of CRPS$_f$ would be appropriate to compare models.

3.5. ZERO-INFLATED BETA MODELING: THE SPATIAL CASE

With observations recorded at geo-coded (point-referenced) locations, spatial dependence can be introduced using random effects. Within our zero-inflated Beta regression
model, we can insert these spatial random effects in two places. The first is in the mean specification, e.g., \( \text{logit}\mu(s_i) = X_i^T\delta + \eta(s_i) \), where \( \eta(s_i) \) is the random effect at the geocoded location \( s_i \). In this case, the random effect impacts both the positive percent covers and the classification of the zero source. The second introduces a spatial random effect \( \tau(s_i) \) into the probability that a site is unfavorable, e.g., \( \text{logit}\pi(s_i) = G_i^T\gamma + \tau(s_i) \). Then, space impacts the classification of the zero source. Given an unobserved set of locations and associated predictors, we can perform spatial prediction for percent covers at these locations.

We model the set of spatial random effects \( \eta(s) \) using a Gaussian process centered at 0 with exponential covariance function, \( C(s_i, s_i + h) = \sigma^2 e^{-\phi||h||} \). As shown by Zhang (2004), the product of the spatial variance \( \sigma^2 \) and decay term \( \phi \) is identifiable, but the individual terms are not. With greater interest in learning about spatial variability, we set \( \phi \) to be fixed during model fitting and estimate \( \sigma^2 \) (see Banerjee et al. 2014 in this regard). The posteriors for both spatial models are provided in Supplement S4.

We fit the \( \eta(s) \) (\( \tau(s) \)) with an elliptical slice sampler (Murray et al. 2010). As mentioned previously, we hold the spatial decay parameter fixed and estimate the spatial variance \( \sigma^2 \) for identifiability (Zhang 2004). We take a weakly informative prior Inverse Gamma(0.5, 0.5) prior on \( \sigma^2_\eta \) (\( \sigma^2_\tau \)). Introducing a random effect in the regression to create \( \pi(s_i) \) or \( \mu(s_i) \) may lead to instability when estimating the corresponding intercept. Therefore, we suggest using an informative prior for the intercept in the regression where the random effect is introduced, and performing model assessment and comparison to determine the best model.

For the remainder of the paper, let \( \mathcal{M}_2 \) denote the model where the spatial random effect \( \eta(s_i) \) is inserted into the mean for the extended Beta, and \( \mathcal{M}_3 \) the model where the random effect \( \tau(s_i) \) is included in the probability of unsuitability. Including the spatial process in models \( \mathcal{M}_2 \) and \( \mathcal{M}_3 \) leads to increased computational time compared with the non-spatial \( \mathcal{M}_1 \), especially as the size of the dataset increases (ex. about 0.5 vs. 3 minutes for \( n = 300 \) using a server node with one core Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50 GHz; Fig. S6).

4. SIMULATION

4.1. \( \mathcal{M}_1 \): A PROOF OF CONCEPT

We generate data under our proposed zero-inflated Beta model \( \mathcal{M}_1 \) and examine how well we recover the truth. Data are generated and models fit with \( a = 1 \). We include a \( N(0, 1) \) covariate for \( \pi_i \), and a separate \( N(0, 1) \) variable for \( \mu_i \) and have a constant \( \nu_i = \nu \). For the intercept in unsuitability, we use an informative Normal prior centered at the true \( \gamma_0 \) with 0.25 standard deviation. We run the samplers for 7500 iterations after 2500 burn-in and thin to retain every fifth sample. We examine traceplots and autocorrelation to assess convergence. All analyses were conducted using R (Version 3.6.1) (R Core Team 2013).

As noted above, the intercepts in \( \pi_i \) and \( \mu_i \) largely control the amount of zeros in the data, as well as the proportions of the sources of zeros: Simulation 1.1 has equal proportions, Simulation 1.2 has more zeros arising from the Beta, and Simulation 1.3 has more zeros arising due to unsuitability. Plots of ROC curves demonstrate that our model is able to perform much better than random guessing in separating the two types of zeros (Fig. S5).
4.2. Model Comparison: Non-spatial Models

We compare BEZI, LCEB, and $M_1$ by generating train and test data under each of these models, fitting all three models on the sets of train data, and predicting on the sets of test data. In the case of LCEB, and $M_1$, $a$ is fixed at 1. We use Tjur’s $R^2$ and AUC to evaluate the ability of each model to distinguish zero and positive observations, and CRPS$_h$ to compare each model’s predictive distributions for the positive observations. Comparisons are performed for varying amounts of 0s in the data and, for data generated under $M_1$, with different proportions for the sources of 0s by changing $\delta_0$ and $\gamma_0$. We perform 50 replications for each data-generating model and set of $\delta_0$ and $\gamma_0$, and present results averaged across the 50 runs. More details on the simulations are provided in Supplement S3.

When the data-generating model is $M_1$, the true model outperforms the other models using all three prediction metrics (Table 1). The higher $R^2$ and AUC indicate its superior ability in separating the positive observations from the zero observations. As evidenced by the lower CRPS$_h$, $M_1$’s predictive distribution for positive observations is more concentrated around the truth in all scenarios. An additional simulation reveals that when data are generated under $M_1$, the true model does a much better job of capturing the mean of the extended Beta as well as the overall probability of 0 (Fig. 5). These findings reveal that the BEZI and LCEB models cannot capture all the 0s well enough when data are generated with two sources. That is, the regression for $\pi$ in BEZI cannot explain all the 0s, and the regressions for $\mu$ and $\nu$ in LCEB cannot accommodate both the positive percent covers and the large incidence of 0s.

When BEZI is the data-generating model, we see that model $M_1$ performs very similar to and sometimes better than the true model (Table S3). In particular, when $\delta_0$ is large and positive (i.e., the mean of the positive percent covers is large), the model $M_1$ sometimes outperforms the true model. However, when $\delta_0$ is negative, the BEZI model offers prediction performance. This can be attributed to the dual role that $\delta_0$ plays in $M_1$, as $\delta_0$ helps determine the mean of the positive percent covers and also contributes to the probability of 0. Under the BEZI model, $\delta_0$ only has a single role. Thus, when $\delta_0$ is negative, the probability of 0 must increase under $M_1$ if the model wishes to capture the lower positive percent covers.

If the left-censoring model LCEB generates the data, we find that $M_1$ outperforms or performs just as well as the true model in almost all scenarios (Table S4). Additionally, the BEZI model often performs just as well. This is explained by the extra $\pi$ parameter in both the $M_1$ and BEZI models, as it provides flexibility to accommodate the behavior of 0s generated under the censor-only model. That is, the likelihood under LCEB is $L(a, \delta, \psi; \{Y_i\})$, while under BEZI it is $L(\gamma, \delta, \psi; \{Y_i\})$ and under $M_1$ it is $L(a, \gamma, \delta, \psi; \{Y_i\})$. For this reason, LCEB has poor performance for data generated under BEZI and $M_1$, precisely because it has fewer parameters and thus less flexibility (Tables 1, S3).

4.3. Model Comparison: Spatial vs Non-spatial Zero-Inflated Models

Again, $M_1$ is the non-spatial zero-inflated model we developed in Sect. 3.2, and $M_2$ and $M_3$ are the spatial versions with space in the either the mean or probability of site unsuitability, respectively. To compare the spatial and non-spatial versions, we conduct
Table 1. Simulation for data generated under the proposed zero-inflated Beta model \( \mathcal{M}_1 \)

| (% deg. 0, % Beta 0) | \( R^2 \) | AUC | CRPS \( h \) |
|----------------------|-------|-----|-----------|
|                      | BEZI  | LCEB | \( \mathcal{M}_1 \) | BEZI  | LCEB | \( \mathcal{M}_1 \) | BEZI  | LCEB | \( \mathcal{M}_1 \) |
| 18%, 5%              | 0.046 | 0.033 | 0.064 | 0.050 | 0.062 | 0.590 | 0.656 | 0.085 | 0.115 | 0.058 |
| 12%, 12%             | 0.016 | 0.092 | 0.096 | 0.074 | 0.574 | 0.647 | 0.667 | 0.138 | 0.083 | 0.073 |
| 5%, 19%              | 0.002 | 0.160 | 0.168 | 0.561 | 0.746 | 0.750 | 0.144 | 0.080 | 0.077 |
| 38%, 15%             | 0.074 | 0.034 | 0.097 | 0.634 | 0.604 | 0.672 | 0.125 | 0.150 | 0.083 |
| 22%, 26%             | 0.007 | 0.152 | 0.158 | 0.563 | 0.716 | 0.719 | 0.148 | 0.097 | 0.084 |
| 13%, 37%             | 0.003 | 0.193 | 0.198 | 0.560 | 0.752 | 0.753 | 0.145 | 0.096 | 0.087 |
| 71%, 9%              | 0.046 | 0.025 | 0.051 | 0.600 | 0.594 | 0.647 | 0.151 | 0.191 | 0.086 |
| 39%, 40%             | 0.014 | 0.125 | 0.135 | 0.579 | 0.729 | 0.748 | 0.140 | 0.124 | 0.099 |
| 15%, 56%             | 0.002 | 0.183 | 0.187 | 0.550 | 0.772 | 0.771 | 0.135 | 0.104 | 0.094 |

\( R^2 \) and AUC for classifying between zero and positive, and CRPS \( h \) for predictive distribution of positive observations for various simulations. Simulations vary by amount of zeros in the data, as well as the proportions of zeros arising from each source (% unsuitable, % random chance). Bold values indicate the best performer.

(a) Parallel histograms for predictions of positive and zero observations. Top: distribution of fitted values for the zero observations. Bottom: distribution for the positive observations. Histograms with considerable overlap demonstrate that the model does not have much discriminatory power.

(b) Plots of true vs predicted means \( \mu \) and probability of zero under the three models. Blue points lying close to black \( y = x \) lines indicate that the true model \( \mathcal{M}_1 \) is able to recover the values well.

Figure 5. Comparisons between BEZI, LCEB, and \( \mathcal{M}_1 \) for data generated under \( \mathcal{M}_1 \) with \( a = 1, \logit(\pi_i) = -1.25 + 0.75\gamma_1, \logit(\mu_i) = 0.50 - 0.50\delta_1, \) and \( \psi = 1.50. \) Resulting data have 54% zero observations, split evenly between unsuitable and chance. \( \mathcal{M}_1 \) fit with prior \( \gamma_0 \sim N(-1.25, 0.25) \). Blue, red, and green represent \( \mathcal{M}_1, \) LCEB, and BEZI models, respectively.

Simulations generating \( n_{\text{train}} = 400 \) observations from the spatial model, fitting the data to the true model and \( \mathcal{M}_1 \), and predicting on \( n_{\text{test}} = 200 \) points. The spatial region is a square on \( (0, 50) \times (0, 50) \), and locations are generated randomly uniformly. Data are generated and the models fit using \( a = 1, \sigma^2 = 1, \) and \( \phi = 20, \) with single independent \( N(0, 1) \) covariates in \( \pi(s_i) \) and \( \mu(s_i) \). We use a Normal prior centered at the truth with standard deviation 0.25 for the appropriate intercept when fitting either the true model (\( \mathcal{M}_2 \) or \( \mathcal{M}_3 \)) and \( \mathcal{M}_1 \). When fitting \( \mathcal{M}_2 \) or \( \mathcal{M}_3 \), \( \phi \) is held fixed at one-third the maximum observed distance.
We first compare $\mathcal{M}_2$ with $\mathcal{M}_1$. Simulations 2.1–2.3 in Supplemental Information compare the parameter recovery of the two models for $\sigma^2_\eta = \{0.5, 1, 2\}$. Table S5 of posterior means and 95% credible intervals reveals the ability of $\mathcal{M}_2$ to recover the true parameters, including the spatial variance. Figure S7 shows that we are able to estimate the spatial surface well. In contrast, we find $\mathcal{M}_1$ often fails to capture the truth for the parameters involved in the Beta regression as $\sigma^2_\eta$ increases. We then evaluate the predictive performance of the two models by generating data under $\mathcal{M}_2$ with varying proportions of 0s, and evaluating $R^2$, AUC, and CRPS when fitting the data with $\mathcal{M}_2$ and $\mathcal{M}_1$, averaged across 50 runs (Table S6). $\mathcal{M}_2$ always outperforms in terms of AUC, $R^2$, and CRPS$_f$. Interestingly, when the data consist of a large proportion of zeros (> 70%, bottom right cells in Table S6), the incorrect model actually performs better in terms of CRPS$_f$. In these scenarios, the prevalence of $Y_i > 0$ is sparse which inhibits learning the spatial surface. For a $Y_i = 0$, whether $\mu(s_i)$ receives a random effect depends on the iteration’s Bernoulli trial for the corresponding $Z_i$. We suspect this uncertainty leads to some model instability and reflects the weakness of $\mathcal{M}_2$ in separating the two sources with a large number of zeros in the data.

We also generate data where the spatial effect is added to the regression for $\pi(s_i)$ and compare $\mathcal{M}_3$ with $\mathcal{M}_1$ (Simulations 3.1–3.3, Supplement). We find that $\mathcal{M}_3$ and $\mathcal{M}_1$ perform very similarly in terms of parameter recovery (Table S7, S10, third column). Figure S9 displays the estimated spatial surface, and Fig. S10 presents plots of comparisons from these simulations. The results of simulations comparing predictive performance under differing amounts of zeros are presented in Table S8. We find that $\mathcal{M}_3$ is marginally superior to the non-spatial model in all scenarios across the four metrics, except when there are a few number of zeros in the data (bottom right cell of Table S8). In these cases, only a fraction of the already low-prevalence 0s will arise due to unsuitability. Therefore, the estimate of $\pi_i$ parameter in the non-spatial model is only marginally impacted by the lack of spatial process, which serves to inflate/deflate the probability of zero.

5. APPLICATION TO CAPE FLORISTIC REGION DATA

5.1. $\mathcal{M}_1$: RESULTS FOR THE CFR DATA

We fit our proposed model to the percent cover data for two plant families, namely Crassulaceae and Restionaceae. We use average annual potential evaporation and the rainfall concentration index as predictors for unsuitability. These predictors capture the overall aridity of a site which should provide a hard threshold (particularly for the succulent Crassulaceae) as to whether certain species within the two families could establish. We use average annual precipitation and minimum temperature in July (austral winter) to estimate the mean of the Beta distribution. If a site is suitable, these variables should capture factors that would attenuate plant abundance through water availability (i.e., precipitation) or amount plants could grow in winter (proxied by the minimum winter temperature). All covariates were centered and scaled. To center the prior for the intercept in the degenerate probability, for various intercept values we fit the model to 130 points and computed the average CRPS$_f$ on the 50 held-out points. Additionally, at this stage, we fix $a = 1$, thus a priori giving equal
support above and below 0. We present the models with the lowest CRPS$_f$; centering $\gamma_0$ at different values does not impact significance nor sign of the remaining coefficients.

Table 2a displays posterior means and 95% credible intervals for both plant families. For Crassulaceae, the large negative $\gamma_0$ intercept ($-1.345$) suggests that fewer zeros arose from unsuitability; most sites were climatically suitable. The negative coefficient for mean annual potential evaporation indicates that for fixed rainfall concentration, higher evaporation increases the suitability of a location, reasonable given that Crassulaceae are succulent plants adapted for drier conditions. The effect of rainfall concentration on unsuitability was insignificant. Higher annual mean precipitation and colder winter temperatures both tended to favor higher Crassulaceae percent cover, holding the other covariate constant. The Restionaceae typically had results opposite to those of Crassulaceae, and there was evidence that more concentrated rainfall throughout the year at a site led to higher suitability compared with sites with similar rainfall.

We hypothesize that opposing results between the two families are driven by both fire dynamics and subregion-specific climate influences. While members of the Restionaceae can inhabit a wide range of moisture availability levels, restio-dominated fynbos primarily occur on warmer north-facing slopes on drought-prone soils (Rebelo et al. 2006). The Restionaceae are also a fire-adapted family, e.g., members germinate in response to smoke (Brown et al. 1994) and have reseeding or resprouting life cycles (Wüest et al. 2016), that often ‘carry’ fires in fynbos (Cowling and Holmes 1992). Our results agreed with these generalizations in that higher Restionaceae abundances were associated with lower rainfall and higher winter temperatures. Within the Baviaanskloof subregion, an additional dynamic of vegetation turnover may play a role in the Restionaceae suitability patterns we observed. Grassy fynbos, i.e., fynbos dominated by C4 grasses, begins to outcompete restio-dominated fynbos in the eastern cape, particularly on drier north-facing slopes. In this context, fynbos tends to occur in wetter areas (Rebelo et al. 2006). This may explain why we observed lower suitability for areas with higher potential evaporation, despite the favor of drier sites for abundance. On the other hand, Crassulaceae are likely “fire avoiders” and may have higher abundances in rockier areas that act as outcropping island microsites (Cousins et al. 2016). Higher Crassulaceae abundance associated with wetter, cooler sites may also reflect a broader signal of lower fire frequency in these areas.

5.2. $M_2$ and $M_3$: Results for a Subset of the CFR Data

Baviaanskloof and Cape Point are disjoint and too far apart geographically to envision a common spatial process for the random effects. So, we perform spatial analysis only on Baviaanskloof as two-thirds ($n = 119$) of the total observations are located in this subregion. The analyses presented here are for Crassulaceae; similar analyses for Restionaceae are provided in Supplement (Tables S9 and S11 and Fig. S11).

In fitting the two spatial models to the Crassulaceae data in Baviaanskloof, we fixed $\alpha = 1$ and experimented with different informative priors for the respective intercepts. Similar to Sect. 5.1, we ultimately select a prior which leads to the best CRPS$_f$. We began with the same covariates used to fit the non-spatial $M_1$ in Sect. 5.1. When fitting $M_2$, minimum July temperature was no longer significant for $\mu(s_i)$. For $M_3$, average annual evaporation was
Table 2. Posterior means and 95% credible intervals for coefficients from fitting various zero-inflated Beta models to CFR data

|                | Crassulaceae Mean | 2.5% | 97.5% | Restionaceae Mean | 2.5% | 97.5% |
|----------------|-------------------|------|-------|-------------------|------|-------|
| (a) Posterior summaries from non-spatial model $M_1$ fit independently for each species across both Baviaanskloof and Cape Point. Priors for $\gamma_0$ for both species are centered at $-0.75$. | | | | | | |
| $\gamma_0$: Intercept | -1.345 | -1.765 | -0.876 | -1.0058 | -1.2453 | -0.7412 |
| $\gamma_1$: mean_ann_pan | -1.269 | -1.960 | -0.377 | 0.884 | 0.294 | 1.630 |
| $\gamma_2$: rainfall_conc | - | - | - | -0.676 | -1.210 | -0.015 |
| $\delta_0$: Intercept | 0.238 | 0.041 | 0.473 | 1.685 | 1.420 | 2.115 |
| $\delta_1$: mean_ann_precip | 0.800 | 0.625 | 0.998 | -0.896 | -1.233 | -0.505 |
| $\delta_2$: min_temp_july | -0.396 | -0.694 | -0.061 | 0.419 | 0.269 | 0.583 |
| $\psi$ | 1.023 | 0.790 | 1.306 | 1.360 | 1.036 | 1.776 |

|                | $M_2$ Mean | 2.5% | 97.5% | $M_3$ Mean | 2.5% | 97.5% |
|----------------|------------|------|-------|------------|------|-------|
| (b) Posterior summaries from the spatial zero-inflated Beta models $M_2$ and $M_3$ fit to the Crassulaceae percent cover data in Baviaanskloof. $M_2$ includes the spatial effect in $\mu_i$, with the prior for $\delta_0$ centered at 1. $M_3$ includes the effect in $\pi_i$, with the prior for $\gamma_0$ centered at $-0.75$. For both models, $\phi$ fixed at one-third the maximum observed distance (14.8 km). | | | | | | |
| $\gamma_0$: Intercept | -0.951 | -1.326 | -0.572 | -1.080 | -1.642 | -0.377 |
| $\gamma_1$: mean_ann_pan | -0.787 | -1.395 | -0.147 | - | - | - |
| $\delta_0$: Intercept | 0.956 | 0.658 | 1.271 | 0.574 | 0.312 | 0.855 |
| $\delta_1$: mean_ann_precip | 1.222 | 0.903 | 1.677 | 1.140 | 0.819 | 1.504 |
| $\delta_2$: min_temp_july | - | - | - | -0.405 | -0.683 | -0.056 |
| $\psi$ | 1.628 | 1.230 | 2.041 | 0.920 | 0.647 | 1.292 |
| $\sigma^2$ | 1.510 | 1.102 | 2.047 | 0.920 | 0.647 | 1.292 |

For all the following, $c$ was fixed at 0.5. The $\gamma$ coefficients are for probability of unsuitability; the $\delta$ and $\psi$ coefficients are for the Beta regression. Dashes correspond to intervals that included 0.

no longer significant for $\pi(s_i)$. Otherwise, the ecological results and interpretations found in Sect. 5 still hold (Table 2b). The estimated spatial variance under $M_2$ is larger than that of $M_3$ (see Fig. 6 for estimates of the random effects).

5.3. Model Comparison for the CFR Data

Lastly, to compare the spatial and non-spatial models, we randomly split the data into a training set of 99 locations and a test set of 20 locations and fit five models—BEZI, $M_0$, $M_1$, $M_2$, and $M_3$. The estimated spatial variance under $M_2$ is larger than that of $M_3$ (see Fig. 6 for estimates of the random effects).
Figure 6. Posterior means of spatial random effects under the two spatial models $M_2$ (left) and $M_3$ (right) fit on the Crassulaceae data in Baviaanskloof. Please see Table S9 for parameter estimates.

$M_1$, $M_2$, and $M_3$—using the covariates that were found significant for each respective model. Additionally, we fix $c$ at three different values: $\{0.1, 0.3, 0.5\}$ for all of the $M$ models in order to assess whether the data are sensitive to the choice of $c$. We compare the models using the test data by evaluating Tjur’s $R^2$ and AUC for all five models and the two CRPS metrics for $M_1$-$M_3$. We repeat this process thirty times and average the results. Table S10 shows that the model with the spatial effect in the mean, $M_2$, with $a = 1$ outperforms the other four models in all the comparison metrics. This could be related to the estimated larger spatial variance under $M_2$. Changing $c$ leads to minor differences in predictive performance. The BEZI model exhibits the lowest performance in classifying between zero and positive percent covers for the data (lowest $R^2$ and AUC, Table S10). This suggests that modeling two sources of zeros is more appropriate for the data, with the inclusion of spatial dependence for $\mu_i$ yielding the best fit. Comparisons for Restionaceae are presented in Table S11.

6. CONCLUSIONS

Modeling zero-inflated percent cover data has been a statistical challenge for ecologists. While many approaches have been proposed, we have argued that these do not satisfactorily address the zero-inflation issue. We have developed a left-censored approach to obtain zeros for data on the unit interval and offered a zero-inflated model for percent or proportion data which parallels such modeling for count data. The specifications enable better understanding of the incidence of absence as well as the explanatory contribution of environmental unsuitability versus absence by chance. Given that ecological data are often spatial, we have supplied spatial models to enhance explanation and enable prediction to unobserved sites with known environmental attributes. Better ecological insights emerge, particularly illustrated by our example with plant percent cover data.

Future work, in progress, considers joint zero-inflation. For example, we can examine percent cover for a pair of species at a site to allow deeper focus on biotic factors (i.e., interspecific competitive interactions) that influence site suitability and relative species abun-
dances. Given the sum constraint at a site to at most 1, this framework anticipates negative association between pairs of species.

**Declarations**

**Data availability** The percent cover data that support the finding of this study are available from the author’s GitHub. The hyperlink is currently omitted due to the double-blind review process.

\[ \text{Received January 2022. Revised May 2022. Accepted August 2022. Published Online September 2022.} \]

**APPENDIX**

Here, we offer analytical insight into the effect of the choice of support for the latent variable, \( W \) (suppressing the site subscript). By letting \( W_a = -a + (1 + a)V \), where \( V \sim \text{Beta}(\mu, \nu) \), \( W \) has support \((-a, 1)\). Again, \( a = 1 \) provides matching support above and below 0. Then, \( P(W_0 \leq 0) = P(V \leq \frac{a}{a+1} \equiv c(a)) \). Though \( W_a \) is a linear transformation of \( W_1 \), there is no linear transformation of the Beta distribution that moves \( c \) to .5, corresponding to \( a = 1 \). We cannot “relocate” \( \mu \) for general \( a \) to \( \mu \) for \( a = 1 \). The shape of the Beta changes with \( \mu \).

It is more convenient to work with \( c \) since we seek the effect of changing \( \mu \) for a fixed \( v \) and \( c \) is on the support of \( V \). We seek to understand the behavior of \( P(V \leq c|\mu, v) \). First, for a fixed value \( p \) of this probability and also a fixed \( v \), qualitatively, \( c \uparrow \mu \). Pursuing this analytically, consider the implicit function, \( F(c_{p,v}(\mu); \mu, v) = p \), where \( F \) denotes a Beta cdf. Plots of the function, \( c_{p,v}(\mu) \) vs. \( \mu \), from \((0, 1)\) to \((0, 1)\), appear sigmoidal, i.e., with asymptotes at 0 and at 1 (Figure S12).

\( S \)-shaped curves are often characterized by an ordinary differential equation. For example, the generalized logistic functions (Richards’ curves) are a rich class including the Gompertz and Weibull with such an attractive characterization as well as an explicit solution (a convenient current site is https://en.wikipedia.org/wiki/Generalisedlogisticfunction). Here, we produce the differential equation which \( c_{p,v}(\mu) \) satisfies. It is not an ordinary differential equation, and in general, it has no closed-form solution since \( F(c; \mu, v) \) has no closed form.

We have \( p = F(c_{p,v}(\mu); \mu, v) = \int_{0}^{c_{p,v}(\mu)} f(x; \mu, v) \, dx \), where \( f \) is the Beta\((\mu, v)\) density. Using Leibniz’s rule, we take the derivative of both sides yielding

\[
\frac{d}{d\mu} f(c_{p,v}(\mu); \mu, v) \, dc_{p,v}(\mu) + \int_{0}^{c_{p,v}(\mu)} \frac{\partial f(x; \mu, v)}{\partial x} \, dx = 0.
\]

Solving, we have the differential equation

\[
dc_{p,v}(\mu)/d\mu = -\int_{0}^{c_{p,v}(\mu)} \frac{\partial f(x; \mu, v)}{\partial x} \, dx / f(c_{p,v}(\mu); \mu, v).
\]

Restoring the \((\alpha, \beta)\) notation, the integral in the numerator exists if \( \alpha > 1 \) and \( \beta > 1 \) (implying \( v > 2 \)); otherwise \( c_{p,v}(\mu) \) is not differentiable. If the integral exists, we have
\[- \int_0^{c_{p,v}(\mu)} \frac{\partial f(x; \mu, \nu)}{\partial x} \, dx = (\alpha + \beta + 1)[F(c_{p,v}(\mu); \alpha, \beta - 1) - F(c_{p,v}(\mu); \alpha - 1, \beta)]. \quad (5)\]

Clearly, this expression is \(> 0\), and since the denominator is positive, the monotone increasing behavior of \(c_{p,v}(\mu)\) is demonstrated. The support for the existence over \((\nu, \mu) \in (0, \infty) \times (0, 1)\) satisfies \(\mu > 1/\nu\) and \(\mu < 1 - 1/\nu\).

For \(c_{p,v}(\mu)\) to be \(S\)-shaped, we need to show that it has exactly one point of inflection, i.e., that \(d^2 c_{p,v}(\mu)/d\mu\) increases up to a point and then decreases. Again using Leibniz’s rule, we can take the derivative of \(F(c_{p,v}(\mu); \alpha, \beta - 1) - F(c_{p,v}(\mu); \alpha - 1, \beta)\). This derivative exists if \(\alpha > 2\) and \(\beta > 2\). So, now \(\nu > 4\) and the support for the existence of the second derivative is \(\mu > 2/\nu\) and \(\mu < 1 - 2/\nu\). Omitting details, in special cases we can demonstrate that this derivative starts positive and then becomes negative, implying a unique point of inflection.

Finally, to return to sensitivity to choice of \(c\), we fit \(S\)-shaped curves to \(c_{p,v}(\mu)\) for various \(p\) and \(\nu > 4\). In the literature, the support of the \(S\)-shaped curves is always on \(R^1\), while we need support \((0, 1)\). We propose to use the logit function \(g_1(x) = \ln \frac{x}{1-x}\) from \((0, 1)\) to \(R^1\). Then, we apply the generalized logistic function, \(g_2(t) = (1 + \beta_0 e^{-\beta_1 t})^{-1/\gamma}\).

Plugging in, we obtain the class \(g_2(g_1(x)) = \left(1 + \beta_0 \left(\frac{1-x}{x}\right)^{\beta_1}\right)^{-1/\gamma}\). This three-parameter function is straightforward to fit to \(c_{p,v}(\mu)\). Implementing the curve fitting over a range of \(p\) and \(\nu\) (Fig. S13), we see that the curves are essentially indistinguishable implying that there is little sensitivity to choice of \(a\).

REFERENCES

Agarwal DK, Gelfand AE, Citron-Pousty S (2002) Zero-inflated models with application to spatial count data. Environ Ecol Stat 9(4):341–355
Amemiya T (1984) Tobit models: a survey. J Econom 24(1–2):3–61
Banerjee S, Carlin BP , Gelfand AE (2014) Hierarchical modeling and analysis for spatial data. CRC Press, Boca Raton
Bell JD, Galzin R (1984) Influence of live coral cover on coral-reef fish communities. Mar Ecol Prog Ser 15:265–274
Blasco-Moreno A, Prez-Casany M, Puig P, Morante M, Castells E (2019) What does a zero mean? Understanding false, random and structural zeros in ecology. Methods Ecol Evol 10(7):949–959
Born J, Linder HP , Desmet P (2006) The greater Cape Floristic Region: greater Cape Floristic Region. J Biogeogr 34(1):147–162
Brown NAC, Jamieson H, Botha PA (1994) Stimulation of seed germination in South African species of Restionaceae by plant-derived smoke. Plant Growth Regul 15(1):93–100
Chib S (1992) Bayes inference in the tobit censored regression model. J Econom 51(1–2):79–99
Cousins SR, Witkowski ETF, Pfab MF (2016) Beating the blaze: Fire survival in the fan aloe (Kumara plicatilis), a succulent monocotyledonous tree endemic to the Cape fynbos, South Africa. Austral Ecol 41(5):466–479
Cowling RM, Holmes PM (1992) Flora and vegetation. The ecology of fynbos: nutrients, fire and diversity. Oxford University Press, Cape Town, pp 23–61
Damgaard C, Irvine K (2019) Using the Beta distribution to analyse plant cover data. J Ecol 107:2747–2759
Dengler J, Jansen F, Glöckler F, Peet RK, De Cáceres M, Chytrý M, Ewald J, Oldeland J, Lopez-Gonzalez G, Finch M et al (2011) The global index of vegetation-plot databases (GIVD): a new resource for vegetation science. J Veg Sci 22(4):582–597
Douma JC, Weedon JT (2019) Analysing continuous proportions in ecology and evolution: a practical introduction to Beta and Dirichlet regression. Methods Ecol Evol 10(9):1412–1430

Ellenberg H, Mueller-Dombois D (1974) Community sampling: the Relevé method. Aims and methods of vegetation ecology. John Wiley & Sons, New York

Ferrari S, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. J Appl Stat 31(7):799–815

Ghosh SK, Mukhopadhyay P, Lu J-CJ (2006) Bayesian analysis of zero-inflated regression models. J Stat Plan Inference 136(4):1360–1375

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol J R Meteorol Soc 25(15):1965–1978

Jenkins JC, Chojnacky DC, Heath LS, Birdsey RA (2003) National-scale biomass estimators for United States tree species. For Sci 49(1):12–35

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(4):1030–1039

Linder H (2001) The African Restionaceae: an interactive key to the species (on CD ROM), vol 20

Mullahy J (1986) Specification and testing of some modified count data models. J Econom 33(3):341–365

Peterson EB (2005) Estimating cover of an invasive grass (Bromus tectorum using tobit regression and phenology derived from two dates of Landsat ETM+ data. Int J Remote Sens 26(12):2491–2507

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Siegfried S, Hothorn T (2020) Count transformation models. Methods Ecol Evol 11(7):818–827

Tjur T (2009) Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. Am Stat 63(4):366–372

van der Maarel E (2007) Transformation of cover-abundance values for appropriate numerical treatment—alternatives to the proposals by Podani. J Veg Sci 18:767–770
Vanha-Majamaa I, Salemaa M, Tuominen S, Mikkola K (2000) Digitized photographs in vegetation analysis—a comparison of cover estimates. Appl Veg Sci 3:89–94

Veech JA, Ott JR, Troy JR (2016) Intrinsic heterogeneity in detection probability and its effect on N-mixture models. Methods Ecol Evol 7(9):1019–1028

Wenger SJ, Freeman MC (2008) Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. Ecology 89(10):2953–2959

Wüest RO, Litsios G, Forest F, Lexer C, Linder HP, Salamin N, Zimmermann NE, Pearman PB (2016) Resprouter fraction in Cape Restionaceae assemblages varies with climate and soil type. Funct Ecol 30(9):1583–1592

Xie Y, Civco DL, Silander JA (2018a) Species-specific spring and autumn leaf phenology captured by time-lapse digital cameras. Ecosphere 9(1). https://onlinelibrary.wiley.com/doi/10.1002/ecs2.2089

Xie Y, Wang X, Wilson AM, Silander JA (2018b) Predicting autumn phenology: how deciduous tree species respond to weather stressors. Agric For Meteorol 250–251:127–137. https://linkinghub.elsevier.com/retrieve/pii/S0168192317306792

Zhang H (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J Am Stat Assoc 99(465):250–261

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.