Estimating Riparian Understory Vegetation Cover with Beta Regression and Copula Models

Bianca N. I. Eskelson, Lisa Madsen, Joan C. Hagar, and Hailemariam Temesgen

Abstract: Understory vegetation communities are critical components of forest ecosystems. As a result, the importance of modeling understory vegetation characteristics in forested landscapes has become more apparent. Abundance measures such as shrub cover are bounded between 0 and 1, exhibit heteroscedastic error variance, and are often subject to spatial dependence. These distributional features tend to be ignored when shrub cover data are analyzed. The beta distribution has been used successfully to describe the frequency distribution of vegetation cover. Beta regression models ignoring spatial dependence (BR) and accounting for spatial dependence (BRdep) were used to estimate percent shrub cover as a function of topographic conditions and overstory vegetation structure in riparian zones in western Oregon. The BR models showed poor explanatory power (pseudo-$R^2 \leq 0.34$) but outperformed ordinary least-squares (OLS) and generalized least-squares (GLS) regression models with logit-transformed response in terms of mean square prediction error and absolute bias. We introduce a copula (COP) model that is based on the beta distribution and accounts for spatial dependence. A simulation study was designed to illustrate the effects of incorrectly assuming normality, equal variance, and spatial independence. It showed that BR, BRdep, and COP models provide unbiased parameter estimates, whereas OLS and GLS models result in slightly biased estimates for two of the three parameters. On the basis of the simulation study, 93–97% of the GLS, BRdep, and COP confidence intervals covered the true parameters, whereas OLS and BR only resulted in 84–88% coverage, which demonstrated the superiority of GLS, BRdep, and COP over OLS and BR models in providing standard errors for the parameter estimates in the presence of spatial dependence. FOR. SCI. 57(3):212–221.

Keywords: shrub cover, beta regression, Gaussian copula, spatial copula

IN THE PAST DECADES THE IMPORTANCE of managing wildlife habitat, enhancing biodiversity, and protecting water quality in forest ecosystems while managing for sustainable timber production has become evident. Understory vegetation communities play a major role in all forest ecosystems (Suchar and Crookston 2010). In temperate forest ecosystems, most of the plant biodiversity is contained within the understory vegetation layers (Halpern and Spies 1995, Weisberg et al. 2003). The understory vegetation not only contributes to biodiversity and protects against erosion but also influences nutrient cycles and provides forage and cover for many wildlife species (Weisberg et al. 2003, Suchar and Crookston 2010). To use understory vegetation characteristics as biodiversity indicators or to assess habitat potential, predictive models for understory vegetation characteristics are needed (Suchar and Crookston 2010).

Predictions of understory vegetation characteristics such as abundance are inherently difficult. Vegetation abundance is often expressed as number of plant individuals, number of binary occurrences (presence/absence), plant cover, or biomass per unit area (Chen et al. 2008a), with plant cover being the most frequently used measure of abundance in vegetation surveys (Chen et al. 2008b). Typically, plant cover is visually assessed, resulting in a measure that is either continuous or ordinal if plant cover classes (e.g., Braun-Blanquet 1964 or Daubenmire 1959) are used (Damgaard 2009). Although plant cover is frequently collected in vegetation surveys, the theoretical and statistical bases underlying cover measures are not well understood (Chen et al. 2006). Vegetation abundance data are characterized by distributional features (e.g., bounded between 0 and 1, heteroscedastic error variance) that do not conform to the assumptions of standard statistical procedures (Damgaard 2009). The beta distribution can be appropriate for modeling cover data because it adequately describes the frequency distribution of cover for various individual species or plant communities (Pielou 1977, Bonham 1989, Chen et al. 2006, 2008a, 2008b, Damgaard 2009). Most of the work using the beta distribution has been done for examples of grasslands and crop fields (e.g., Chen et al. 2006, 2008a, 2008b). In recent years, beta regression has been applied in a variety of fields including forestry. For example, Korhonen et al. (2007) successfully estimated forest canopy cover with beta regression. However, to the authors' knowledge, no work exists that applies the beta distribution to describe the frequency distribution of understory vegetation cover in forested ecosystems.

Bianca N. I. Eskelson, Oregon State University, College of Forestry, Department of Forest Engineering, Resources and Management, 204 Peavy Hall, Corvallis, OR 97331—Phone: 541-737-9112; Fax: 541-737-3049; bianca.eskelson@oregonstate.edu. Lisa Madsen, Oregon State University—madsenl@onid.orst.edu. Joan C. Hagar, US Geological Survey—joan_hagar@usgs.gov. Hailemariam Temesgen, Oregon State University—hailemariam.temesgen@oregonstate.edu.

Acknowledgments: We thank Theresa Marquardt for her insights on the data used, Jacob Strunk and Nathan Chelgren for their insights on an earlier draft, and three anonymous reviewers and the associate editor for their helpful comments. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Manuscript received May 3, 2010, accepted September 17, 2010

Forest ecologists and wildlife biologists working in temperate conifer forests are particularly interested in modeling cover and distribution of shrubs. Shrubs comprise a major component of understory vegetation and provide critical food and cover resources for many wildlife species (Hagar 2007). Given the importance of shrubs to wildlife, the ability to accurately quantify and map shrub cover would greatly facilitate habitat management (Martinuzzi et al. 2009), but measurement of shrub cover is very laborious and costly. Because riparian zones include some of the most productive wildlife habitats in forest lands of western Oregon and Washington (Anthony et al. 1987), predictive models of shrub cover in riparian zones are of special interest.

The objectives of this article are to 1) use beta regression (with and without accounting for spatial dependence) for modeling shrub cover in riparian forests along headwater streams as a function of topographic conditions and overstory vegetation structure, 2) model shrub cover using a copula model that accounts for spatial dependence, 3) compare parameter estimates from five model types: beta regression (with and without dependence structure), copula models, and ordinary least-squares (OLS) models with logit-transformed response (with and without dependence structure), and 4) by means of a simulation study a) evaluate the performance of the five model types in terms of the parameter estimates they provide and b) demonstrate the importance of modeling existing spatial dependence.

Beta Distribution and Beta Regression

The beta distribution is a two-parameter distribution that can accommodate various types of plant cover frequency distributions with "J," "L," "one-peak," "U," and "rectangular" shapes (Chen et al. 2006, Smithson and Verkuilen 2006). The beta distribution has been used in statistical ecology for many years (Chen et al. 2006) and is defined as follows:

$$f(y;\, p,\, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\, y^{p-1}(1-y)^{q-1}, \qquad (1)$$

where $0 < y < 1$, $p, q > 0$, and $\Gamma(\cdot)$ denotes the gamma function. p and q are shape parameters, with p pulling density toward 0 and q pulling density toward 1 (Smithson and Verkuilen 2006). Ferrari and Cribari-Neto (2004) proposed a different parameterization of the beta probability density function (pdf), by setting $\mu = p/(p+q)$ and $\phi = p + q$ (i.e., $p = \mu\phi$ and $q = (1-\mu)\phi$, where $0 < \mu < 1$ and $\phi > 0$:

$$f(y;\, \mu,\, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\, y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}. \qquad (2)$$

The shape parameters p and q as well as the parameters μ and ϕ can be used to express the mean and variance of y, respectively,

$$\mathrm{E}(y) = \frac{p}{p+q} = \mu, \qquad (3)$$

$$\mathrm{Var}(y) = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\mu(1-\mu)}{(1+\phi)}. \qquad (4)$$

Using the parameterization of the beta distribution described in Equation 2, Ferrari and Cribari-Neto (2004) introduced a beta regression model similar to the approach for generalized linear models (McCullagh and Nelder 1989), except that the distribution of the response is not a member of the exponential family. In the extended generalized linear model approach, $y_1, \ldots, y_n$ are independent random variables with each y_i following the density in Equation 2 with mean μ_i and precision ϕ. The beta regression model is obtained as follows:

$$g(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \eta_i, \qquad (5)$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^{\mathrm{T}}$ is a vector of k explanatory variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ is a $k \times 1$ vector of unknown regression parameters ($k < n$), η_i is a linear predictor, $g(\cdot)$ is a strictly increasing and twice differentiable link function that maps (0, 1) into the real line $\mathbb{R}$, and T is the transpose of a vector. A variety of link functions $g(\cdot)$ are available, but the logit link $g(\mu) = \log(\mu/(1-\mu))$ is particularly useful, in which case

$$\mu_i = \frac{1}{1 + \exp(-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}, \qquad (6)$$

with everything as defined above.

Measurements such as percent cover take on values on the open interval (0, 1), and the influence of explanatory variables on continuous responses bounded between 0 and 1 can be investigated with the beta regression proposed by Ferrari and Cribari-Neto (2004). OLS regression models with logit-transformed response variables have traditionally been used for this type of data. However, the logit-transformed OLS approach has been questioned, and Brehm and Gates (1993) suggested that the beta distribution should be favored over the normal distribution because it is theoretically and statistically more appropriate. Beta regression does not require the response to be transformed to take on values on the real line and therefore allows parameter interpretation in terms of the response in the original scale (Espinheira et al. 2008). The investigator can chose the link function, and if a logit link function is used to transform the mean response, the regression parameters can be interpreted in terms of the odds ratio, which is not possible for parameters from OLS regression, which uses a logit-transformed response (Ferrari and Cribari-Neto 2004). Nonconstant response variances are naturally accommodated into the beta regression model, because the variance of y_i is a function of μ_i (Equation 4) and, hence, a function of the values of the explanatory variables (Equation 6) (Ferrari and Cribari-Neto 2004). The underlying assumption of the beta regression model is that the response follows the beta law (Ferrari and Cribari-Neto 2004), hence allowing asymmetry of the response distribution (Espinheira et al. 2008). Measures on the (0, 1) interval typically display asymmetry; thus, inference based on the normality assumption of OLS with a logit-transformed response can be misleading (Ferrari and Cribari-Neto 2004). The logit transformation used in OLS

models will mitigate asymmetry but will not remove pronounced asymmetry. Kieschnick and McCullough (2003) compared several regression models for proportions observed on the open interval (0, 1) using economics and presidential election data and preferred the use of beta regression over the other regression models examined (linear and nonlinear OLS, additive logistic normal distribution, censored normal distribution, and simplex distribution).

Copula Models

A copula is a function that joins univariate marginal distributions into a multivariate distribution function. In other words, copulas are multivariate distribution functions whose one-dimensional margins are uniform on the interval (0, 1) (Nelsen 2006, p. 1). The multivariate Gaussian copula generalizes a multivariate normal dependence structure to non-normal marginals. If $y_1, \ldots, y_n$ are random variables with continuous marginal cumulative distribution functions (cdf's) F_i and pdf's f_i and Σ is a non-negative definite matrix with diagonal entries equal to 1, the multivariate Gaussian copula is the joint distribution function of $y_1, \ldots, y_n$ with specified marginals:

$$C(y; \Sigma) = \Phi_\Sigma[\Phi^{-1}\{F_1(y_1)\}, ..., \Phi^{-1}\{F_n(y_n)\}], \quad (7)$$

where Φ is the standard normal cdf and Φ_Σ is the multivariate normal cdf with covariance matrix Σ. The expression $\Phi^{-1}\{F_i(y_i)\}$ represents a normal transformation of y_i as a consequence of the probability integral transformation (Casella and Berger 2002, p. 54): $F_i(y_i)$ is uniform on (0, 1) and applying the inverse standard normal cdf Φ^{-1} to a uniform yields a standard normal random variable. Differentiating $C(y_1, \ldots, y_n)$ yields the joint pdf,

$$c(y; \Sigma) = |\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{z}^{\mathrm{T}}(\Sigma^{-1} - \mathbf{I}_n)\mathbf{z}\right)\prod_{i=1}^{n} f_i(y_i), \quad (8)$$

where $\mathbf{z} = [\Phi^{-1}\{F_i(y_i)\}, \ldots, \Phi^{-1}\{F_i(y_i)\}]^{\mathrm{T}}$ and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix.

The copula correlation matrix Σ models the correlation among the elements of z, which induces dependence among the y_i. Unless the y_i are themselves normally distributed, Σ is not their correlation matrix. If $y_1, \ldots, y_n$ are spatially referenced, the copula correlation matrix Σ can be given a spatial form. An exponential model with "decay" parameter θ is assumed so that the ijth element of the correlation matrix,

$$\Sigma_{ij}(\theta) = \begin{cases} \exp(-h_{ij}\theta), & i \neq j \\ 1, & i = j \end{cases} \quad (9)$$

where h_{ij} is the distance between the locations of y_i and y_j and $\theta > 0$. With this, the spatial Gaussian copula allows bringing non-normal distributions into the Gaussian geostatistical framework, where correlation completely describes dependence (Madsen 2009). When θ is large, Σ is approximately the identity matrix and the y_i are approximately independent. Decreasing θ corresponds to increasing the spatial dependence between the y_i. Therefore, small θ values indicate strong spatial dependence, whereas large θ values indicate weak spatial dependence. The scale of θ depends on the minimum distance between observations in a given study.

Let $y_1, \ldots, y_n$ be beta random variables with pdfs as in Equation 2, mean μ_i (Equation 6), and precision parameter ϕ. Maximum likelihood estimates of $\boldsymbol{\beta}$, ϕ, and θ are obtained by numerically maximizing the log of expected likelihood with respect to $\boldsymbol{\beta}$, ϕ, and θ. From 8, the log expected likelihood is

$$\log L(\beta, \phi, \theta; y) = \log\left(|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{z}^{\mathrm{T}}(\Sigma^{-1} - \mathbf{I}_n)\mathbf{z}\right)\prod_{i=1}^{n} f_i(y_i)\right), \quad (10)$$

where $\mathbf{y}$ is the data vector, $\mathbf{z} = [\Phi^{-1}\{F_i(y_i)\}, \ldots, \Phi^{-1}\{F_i(y_i)\}]^{\mathrm{T}}$, and f_i is the beta pdf. Variance estimates are obtained by numerically approximating the Hessian matrix $\mathbf{H}$ at the maximum likelihood estimates.

The purpose of this analysis was to estimate the regression parameters $\boldsymbol{\beta}$. Therefore, the precision parameter (ϕ) and covariance parameter (θ) were considered as nuisance parameters and only used to account for dispersion and spatial dependence, respectively.

Methods

Case Study Data

Understory percent cover data of headwater streams were collected in 2006 on four sites managed by the Bureau of Land Management Density Management Study (DMS) in the Oregon Coast Range (Table 1) (for DMS details, see Cissel et al. 2006) as part of a larger study (see Marquardt

Table 1. General information on the four headwater sites

Site name	Reach no.	BLM district	Latitude	Longitude	Elevation (m)	Buffer	Stream slope (%)	Streamside slope (%)	Aspect of channel orientation (°)
Keel Mountain	18	Salem	44°31′41.0″N	122°37′55.0″W	745	Two-site potential tree heights	20	21	269
Bottom Line	13	Eugene	43°46′20.0″N	123°14′11.0″W	295	Two-site potential tree heights	18	66	323
Ten High	75	Eugene	44°16′50.0″N	123°31′06.0″W	520	Variable width	60	40	173
OM Hubbard	36	Roseburg	43°17′30.0″N	123°35′00.0″W	510	Variable width	19	40	71

BLM, Bureau of Land Management.

2010). Site locations ranged from west of Corvallis to north of Roseburg, Oregon, USA (range 43°17′30″N to 44°31′41″N and 122°37′5″W to 123°35′00″W). Stands were 40–70 years old when density and buffer treatments were applied. A buffer, which is a forested strip parallel to the stream, of width equal to the height of two site potential trees (146-m slope distance from the stream) was left on each side of the stream at two sites. At the remaining two sites a variable width buffer was applied to the streams, which had a minimum of 15-m slope distance from the stream and fluctuated based on sensitive areas (e.g., areas prone to landslides or areas with threatened species present) (Cissel et al. 2006). The four sites used in this study were moderate density retention sites for which 60–65% of the stand was thinned to 200 trees/ha (TPH), 10% was left in circular leave islands, and 15–20% was left unthinned in riparian buffers (Cissel et al. 2006, Chan et al. 2004). More detailed information can be found in Cissel et al. (2006, Appendix E).

A sampling block (72 m × 72 m horizontal distance) was randomly located along one headwater stream at each site. One 72-m axis of the block was oriented approximately parallel to the stream; the center of the block along this axis will be referred to as the center line. The second axis extended approximately 36 m (horizontal distance) upslope and perpendicular to the center line on each side of the stream. At the 32- and 68-m marks of the center line, understory vegetation plots were installed. Transects with 3-m spacing (horizontal distance) were laid out perpendicular to the center line at those points. In addition, transects with 10-m spacing (horizontal distance) were laid out perpendicular to the center line at two random points between 0 and 72 m at each site (Figure 1). The data that were collected along the four transects at either 3- or 10-m

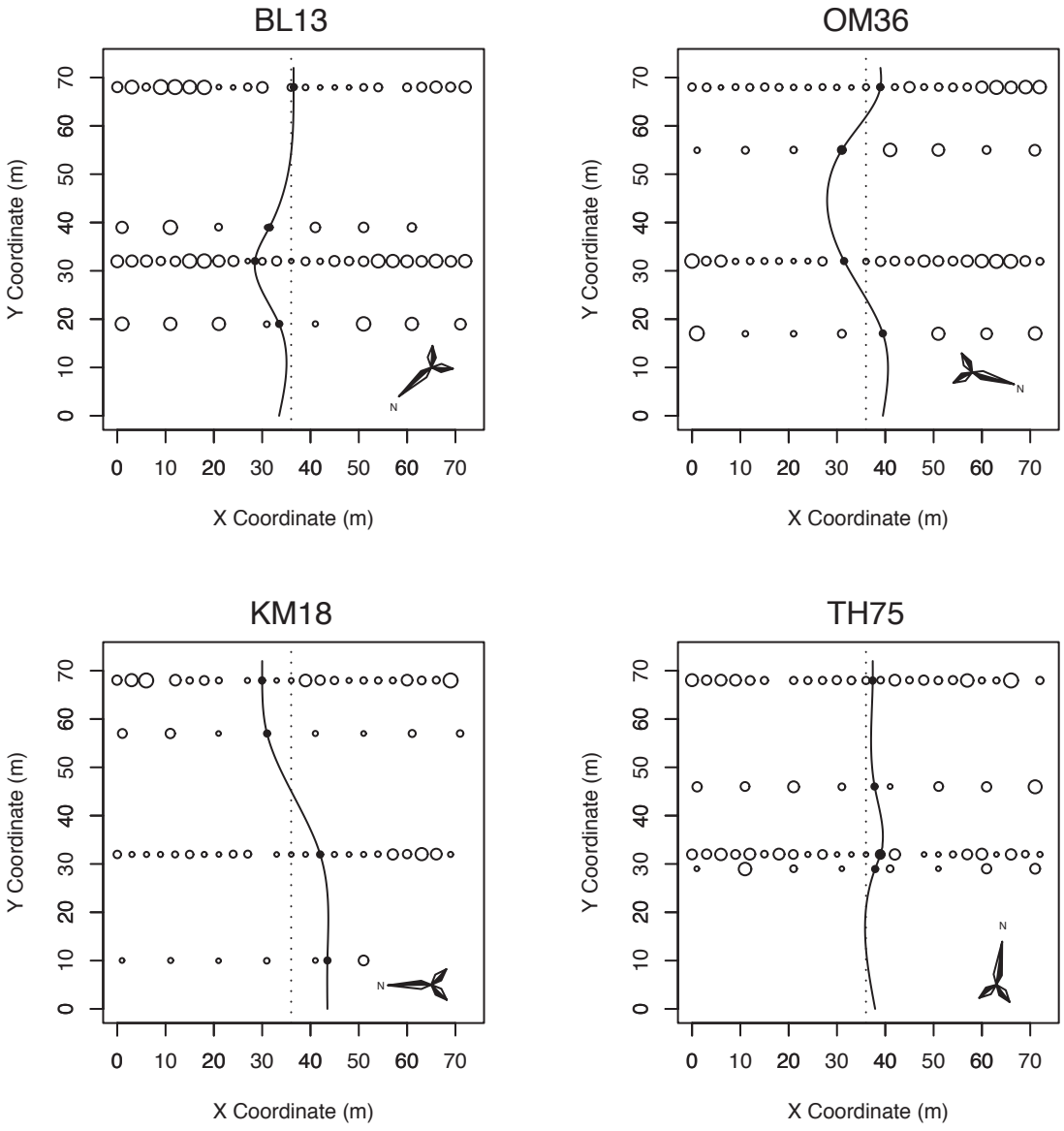

Figure 1. Locations of the vegetation plots on the four sites. Open circles, vegetation plot location, size proportional to observed percent shrub cover within site; dotted line, center line; solid circles: observed stream location; solid line: spline interpolated stream course.

spacing are described in detail in the following. More detailed information on the sampling design can be found in Marquardt (2010).

At each sample point, percent cover of shrubs (perennial woody plants <1.4 m) was visually determined to the nearest 5% on 1 m × 1 m plots. At the same points, the overstory density measures TPH and basal area per hectare (BA/ha in m^2/ha) were calculated on the basis of a variable radius plot with basal area factor 8. For each plot center, the horizontal distance to the stream (DTS in m) and the height above the stream (HAS in m) was recorded. Leaf area index (LAI in m^2 foliage/m^2 ground) was measured at each plot center using hemispherical detection of canopy light transmittance (plant canopy analyzer, model LAI-2000, Li-Cor Biosciences, Lincoln, NE) (Table 2).

Simulation Study Data

To illustrate the consequences of ignoring violations of the assumptions of normality, equal variance, and spatial independence, we simulated 500 data sets of size $n = 248$ to mimic the observed shrub data. Spatial locations, DTS, and LAI agree with the actual data from the case study described above. We replaced each observed response with a simulated response from a beta distribution with mean and variance as provided in Equations 3 and 4 where

$$\mu_i = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \mathrm{DTS}_i + \beta_2 \mathrm{LAI}_i))}. \quad (11)$$

True parameters are $(\beta_0, \beta_1, \beta_2, \phi) = (-0.34, 0.037, -0.24, 2.6)$, approximately the estimates from fitting the above model to the case study data.

We achieve spatial dependence in the simulated responses by first simulating spatially dependent standard normal random variables Z with a covariance matrix determined by the exponential model given in Equation 9, where h_{ij} is the distance between locations i and j and $\theta = 0.31$, approximately the estimate from the copula model fit to the case study data. Each normal Z is transformed to a beta response Y as $Y_i = F_i^{-1}[\Phi(Z_i)]$, where Φ is the standard normal cdf and F_i^{-1} is the beta cdf with mean and variance as given in Equations 3 and 4, respectively. This transformation reverses the probability integral transformation of Equation 7 (Casella and Berger 2002, p. 54). In addition to $\theta = 0.31$, which was obtained as an estimate from the case study data, we simulated 500 data sets of size $n = 248$ as described above for a variety of θ values (0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, infinity), ranging from $\theta = 0.01$ = strong spatial dependence to θ = infinity = no spatial dependence.

Models for Case Study Data

Beta regression (BR) models were fit to a variety of models with different sets and combinations of explanatory variables. Explanatory variables described the topographic conditions, DTS, HAS, slope, aspect, and slope-aspect transformations [slope × cosine(aspect), slope × sine(aspect)], as well as the overstory vegetation structure, LAI, BA/ha, and TPH. Interactions of DTS and HAS with LAI were also included in the models. From these models, the three models with the smallest Bayesian information criterion (BIC) values were selected for further analysis. The pseudo-R^2 value, which is the squared correlation of the linear predictor and link-transformed response, was calculated and compared. Because percent shrub cover included the extremes 0 and 1, the following transformation that is commonly used in practice (Smithson and Verkuilen 2006) was used:

$$y^* = (y(n-1) + 0.5)/n, \quad (12)$$

where n is the sample size.

OLS regression with a logit-transformed response and copula (COP) models were fit using the sets of explanatory variables of the three models that were picked based on the BR. To account for the dependence structure due to the sampling design, an exponential spatial covariance structure was incorporated into the OLS and BR models, which resulted in a generalized least squares (GLS) and a BR model with exponential dependence structure (BRdep), respectively.

For each model type (OLS, GLS, BR, BRdep, and COP) and set of explanatory variables (1–3), the mean squared prediction error (MSPE) and absolute bias (AB) were reported:

$$\mathrm{MSPE} = \sum_{i=1}^{n} \frac{(\text{predicted} - \text{observed})^2}{n}, \quad (13)$$

$$\mathrm{AB} = \sum_{i=1}^{n} \frac{(\text{predicted} - \text{observed})}{n}, \quad (14)$$

Analysis of Simulation Study Data

OLS, GLS, BR, BRdep, and COP models were fit to the 500 simulated data sets for each θ value using DTS and LAI

Table 2. Summary of topographic and overstory vegetation attributes used in the case study

Variable	Minimum	Mean	Maximum	SD
% shrub cover	0.008	0.37	0.992	0.27
DTS (m)	0.0	18.9	43.5	11.2
HAS (m)	−0.4	6.9	23.4	5.4
LAI (m^2/m^2)	1.21	3.86	6.53	1.21
BA/ha (m^2/ha)	0	63.29	208	39.07
TPH	0	597	4118	534
Slope (%)	1	38	88	19

DTS, horizontal distance to stream; HAS, height above stream; LAI, leaf area index; BA/ha, basal area per hectare; TPH, trees per hectare.

as explanatory variables. On the basis of the 500 estimates for each parameter, the bias of the parameter estimates was calculated for each method:

$$\text{Bias} = \frac{\sum_{l=1}^{500} \hat{\beta}_{kl}}{500} - \beta_k, \quad (15)$$

where β_k is the true parameter with $k = 0, 1, 2$ and $\hat{\beta}_{kl}$ is the estimate for β_k based on the lth simulated data set with $l = 1, \ldots, 500$.

The confidence interval for each parameter estimate was calculated as

$$\hat{\beta}_{kl} \pm 1.96 * \text{SE}(\hat{\beta}_{kl}), \quad (16)$$

where $\text{SE}(\hat{\beta}_{kl})$ is the standard error of $\hat{\beta}_{kl}$. For each method, the percent coverage of the true parameter was calculated based on the confidence intervals. The regression parameters $\boldsymbol{\beta}$ have different interpretations in the OLS and GLS models than in the BR, BRdep, and COP models (for details, see Discussion). Equation 15 of Espinheira et al. (2008) can be used to calculate the OLS/GLS parameters in terms of $(\beta_0, \beta_1, \beta_2) = (-0.34, 0.037, -0.24)$ from the BR, BRdep, and COP models. The vector of true parameters used to calculate bias and confidence coverage for the OLS and GLS models will be referred to as $\boldsymbol{\gamma}$ and equals $(\gamma_0, \gamma_1, \gamma_2) = (-0.505, 0.056, -0.369)$.

BRdep and GLS models were fit with PROC GLIMMIX in SAS (SAS Institute, Inc. 2008). All other analyses were performed in R 2.10.1 (R Development Core Team 2009), and the beta regression was implemented with the betareg R package version 2.2-2 (Cribari-Neto and Zeileis 2010).

Results

Case Study

The preferred BR model (model 1) with regards to BIC (−118.77) had only two explanatory variables (DTS and LAI). The model with the second smallest BIC value (−116.28) also included slope, aspect, slope-aspect transformations, HAS, and the HAS × LAI interaction as explanatory variables (model 2). The model with the third smallest BIC value (−114.92) was like model 2 without HAS and the HAS × LAI interaction as explanatory variables (model 3). The BIC values for the corresponding COP models were −141.83, −131.32, and −128.13, respectively. The OLS and GLS models had large positive BIC values corresponding to 999.03, 1042.93, and 1051.23 for OLS models 1 through 3, respectively, and 949.22, 977.29, and 975.52 for GLS models 1 through 3.

Pseudo-R^2 values for the three BR models increased with complexity of the model with 0.26 for BR1, 0.30 for BR3, and 0.34 for BR2. Addition of BA/ha and TPH as explanatory variables to the models increased the pseudo-R^2 values but decreased the BIC values. DTS explained more variation than HAS by itself. DTS and HAS were linearly correlated (correlation coefficient = 0.73). For the simulated data with $\theta = 0.31$, the pseudo-R^2 values ranged from 0.09 to 0.39 with mean 0.22.

MSPE was largest for OLS and GLS models for all three sets of explanatory variables, whereas the MSPE values were slightly smaller and essentially the same for BR, BRdep, and COP models (Table 3). For all five model types, MSPE was largest for the simplest model (model 1) and smallest for the most complex model (model 2) (Table 3). BR, BRdep, and COP models provided unbiased predictions for all three sets of explanatory variables, whereas the predictions based on OLS and GLS exhibit a negative bias (Table 3).

Simulation Study

Based on the simulated data with $\theta = 0.31$, the OLS and GLS models provided unbiased parameter estimates for β_0 and biased parameter estimates for β_1 and β_2, whereas the BR, BRdep, and COP models provided unbiased estimates for all three parameters (Table 4). The OLS and BR models provided the worst confidence coverage of the true parameters, ranging between 84 and 89%. For the GLS model, the confidence coverage is 97% for β_0 and 93 and 94% for β_1 and β_2, respectively, whereas 94–96% of the confidence intervals covered the true parameter for the BRdep and COP models (Table 4).

For very strong and very weak dependence, the BRdep model occasionally had convergence issues. For example, for $\theta = 0.01$ (very strong dependence), 5.4% of the simulations did not converge, and for $\theta = 1$ (weak dependence), 9.4% of simulations did not converge. The percent coverage of the confidence intervals is reported for the simulated data sets for which the BRdep model converged.

The GLS, BRdep, and COP models provided approximately 95% confidence coverage for all three parameter estimates for $\theta \geq 0.1$. For $\theta = 0.01$ or $\theta = 0.05$, which correspond to very strong spatial dependence, GLS, BRdep,

Table 3. Mean squared prediction errors (MSPE) and absolute bias (AB) for each model type and three sets of explanatory variables

	MSPE			AB		
Model	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
OLS	0.064	0.051	0.056	−0.0579 ($P < 0.001$)*	−0.0463 ($P = 0.001$)	−0.0532 ($P < 0.001$)
GLS	0.063	0.051	0.056	−0.0556 ($P < 0.001$)	−0.0409 ($P = 0.004$)	−0.0516 ($P < 0.001$)
BR	0.056	0.048	0.050	−0.0014 ($P = 0.927$)	−0.0006 ($P = 0.965$)	−0.0017 ($P = 0.908$)
BRdep	0.056	0.047	0.050	0.0014 ($P = 0.924$)	0.0027 ($P = 0.847$)	0 ($P = 0.998$)
COP	0.056	0.048	0.051	0.0029 ($P = 0.846$)	0.0055 ($P = 0.693$)	0.0021 ($P = 0.886$)

Model 1 explanatory variables: DTS, LAI; model 2 explanatory variables: DTS, LAI, HAS, HAS × LAI, slope, aspect, slope-aspect transformations; model 3 explanatory variables: DTS, LAI, slope, aspect, slope-aspect transformations.
* P value for a t test testing whether the bias is significantly different from 0.

Table 4. Bias of parameter estimates and % coverage of confidence intervals for simulated data with $\theta = 0.31$

Model	Bias			% Coverage of confidence intervals		
	β_0	β_1	β_2	β_0	β_1	β_2
OLS	−0.0449 ($P = 0.089$)*	0.0055 $P < 0.001$)	−0.0257 ($P < 0.001$)	89	84	86
GLS	−0.0351 ($P = 0.1615$)	0.0054 ($P < 0.001$)	−0.0292 ($P < 0.001$)	97	93	94
BR	−0.0098 ($P = 0.543$)	0.0002 ($P = 0.508$)	0.0008 ($P = 0.807$)	87	85	88
BRdep	−0.0001 ($P = 0.995$)	0.0003 ($P = 0.366$)	−0.0024 ($P = 0.493$)	95	96	94
COP	−0.0065 ($P = 0.668$)	0.0004 ($P = 0.252$)	−0.0007 ($P = 0.819$)	95	95	94

True regression parameters are $\beta_0 = -0.34$, $\beta_1 = -0.037$, and $\beta_2 = -0.24$, for the BR, BRdep, and COP models and (γ_0, γ_1, γ_2) = −0.505, 0.056, −0.369) for the OLS and GLS models

* P value for a t test testing whether the bias is significantly different from 0.

and COP models tended to result in lower confidence coverage of 77–94%, with GLS having the smallest confidence coverage for β_1 and β_2 among the three models. With fairly high confidence coverage of 97% for β_1, the COP model provided an exception when $\theta = 0.01$ (Figure 2).

When spatial dependence was strong, the OLS and BR models provided poor confidence coverage for all three parameter estimates. The confidence coverage increased with increasing θ, corresponding to decreasing dependence, and leveled off between 92 and 94% confidence coverage at approximately $\theta = 0.6$. The confidence coverage of the OLS model compared with that of the GLS model tended to be slightly larger for β_0, slightly smaller for β_1, and identical for β_2 (Figure 2).

Discussion

Pseudo-R^2 is a measure of explained variation (Ferrari and Cribari-Neto 2004), and, hence, the low pseudo-R^2 values of the BR models, ranging from 0.26 to 0.34, suggested a lot of unexplained variation, which is consistent with previous studies. It has been shown repeatedly that predicting percent shrub cover in forested ecosystems is inherently difficult. Suchar and Crookston (2010) reported adjusted R^2 values of 0.22 and 0.24 for their percent shrub cover models. Kerns and Ohmann (2004) achieved R^2 values between 0.14 and 0.49 for percent shrub cover models for different ownership groups. Even though general trends such as increases in percent shrub cover with increases in DTS and HAS and decreases in percent shrub cover with increases in LAI, BA/ha, and TPH are apparent in the data, the linear correlation coefficients are fairly small ($\leq |0.4|$) due to high variability in the data. Suchar and Crookston (2010) argue that shrub and herb cover are more heterogeneous than overstory cover attributes and that finer scale environmental conditions such as soil nutrient content, moisture, and allelophatic effects might be needed to effectively predict understory characteristics. Therefore, explanatory variables at a finer scale than those available in this study might be needed to improve the amount of explained variation for percent shrub cover models. Light detection and ranging (LiDAR)-derived metrics have successfully been used for mapping the presence/absence of understory shrub species in forested landscapes (Martinuzzi et al. 2009). The use of LiDAR metrics for modeling percent shrub cover in forested landscapes should be explored. Because LiDAR technology enables precise three-dimensional maps of vegetation structure, it may be possible to simply obtain census measurements of understory cover by using LiDAR measurements.

The three sets of explanatory variables for which the five model types were compared were chosen on the basis of BIC values from the BR models. Because of different model fitting techniques, BIC values from all five model types are not directly comparable. For example, the BIC values from the BRdep models, fit with PROC GLIMMIX in SAS, are based on the quasi-likelihood and therefore cannot be compared with the BIC values from the BR and COP models that are based on the maximum likelihood. The COP models provided smaller BIC values than the BR models, indicating a better model fit, apparently because the BR model ignores spatial dependence, whereas the COP model accounts for it. Likewise, the GLS models had smaller BIC values than the OLS models, which ignore the spatial dependence present in the case study data. According to Raftery (1995) and Kass and Raftery (1995), a difference in BIC values (ΔBIC) of ≤ 2 between models is "not worth more than a bare mention" and a $\Delta\text{BIC} > 10$ implies very strong evidence that the models are different. Hence, based on ΔBIC, there is strong evidence that the COP models are superior to the BR models, followed by the GLS and OLS models. Within each model type, ΔBIC indicates that model 2 is not significantly different from model 3. However, for OLS, GLS, and COP models, model 1 is significantly different from model 2.

Based on the case study results, the OLS and GLS models were inferior to the BR, BRdep, and COP models, because they resulted in the largest MSPE values and negatively biased predictions. This finding suggests that BR, BRdep, and COP models based on the beta distribution may be more appropriate for modeling percent shrub cover than OLS regression with a logit-transformed response, which has already been argued for other responses that take on values in the open interval (0, 1) (e.g., Kieschnick and McCullough 2003, Smithson and Verkuilen 2006). Asymmetry was low in the case study data, and it can be expected that the difference in performance of the OLS and GLS models compared with that of the BR, BRdep, and COP models will be more pronounced if the assumption of normality of residuals was violated. The COP model provided no improvement compared with the BR model in terms of MSPE, because we simply inserted explanatory variables and estimated betas into Equation 11, thus ignoring spatial information from nearby points in the prediction. Prediction

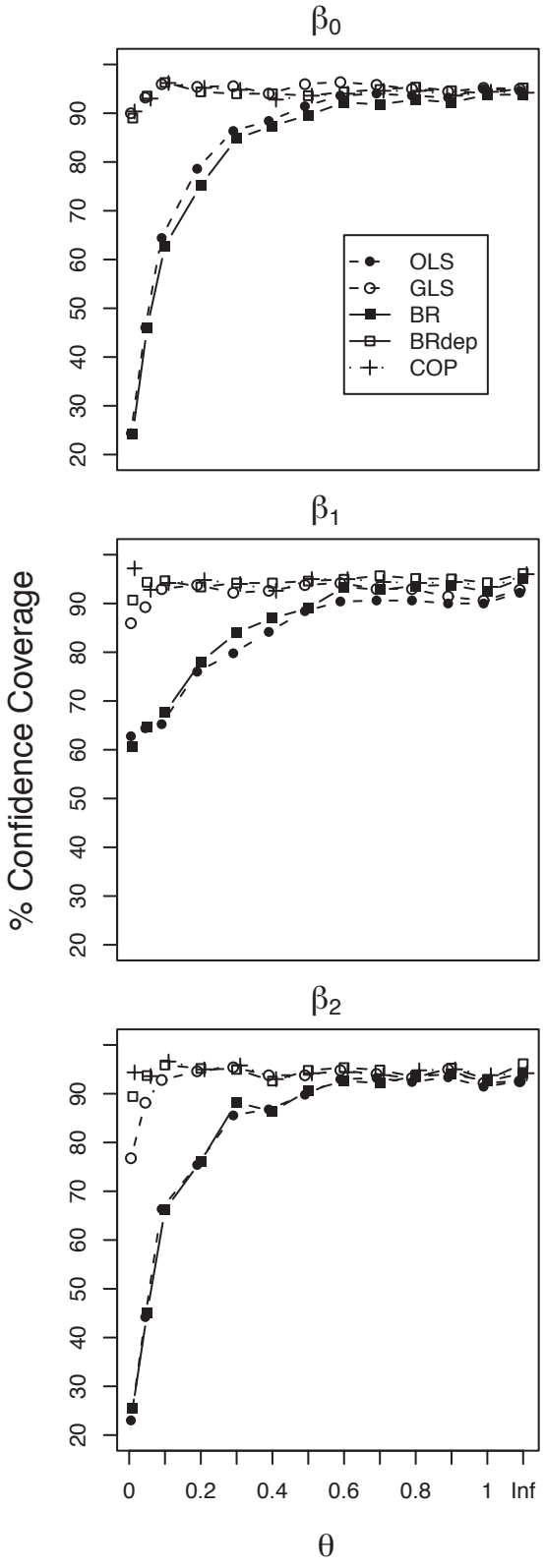

Figure 2. Percent confidence coverage of the parameter estimates β_0 (top), β_1 (middle), and β_2 (bottom) for OLS, GLS, BR, BRdep, and COP models for the simulated data over all θ values. θ = 0.01 = strong spatial dependence; θ = infinity = no spatial dependence.

at unobserved locations using the COP model is a topic of current research that needs further investigation.

The focus of the statistical analysis was to estimate regression parameters β_0, β_1, and β_2, whereas the simulation study demonstrated the effect of ignoring within-site spatial dependence. Therefore, it is important to consider the distinctions between the models with respect to the interpretation of the regression parameters and with respect to the way they model dependence.

BR, BRdep, and COP all model percent cover with a beta distribution and allow a heteroscedastic response in accordance with Equation 4. Each response y_i has a mean given by Equation 5 with a logit link function. With this choice of link, the coefficient β_k of one of the predictors, say DTS, can be interpreted in terms of the odds ratio. In particular, if y_i and y_j have identical predictor values except for $\text{DTS}_i = \text{DTS}_j + 1$, then e^{β_k} is the ratio of $\mu_i/(1 - \mu_i)$ to $\mu_j/(1 - \mu_j)$.

BRdep models dependence as correlation between the y_i. Because correlation is specifically linear dependence, this characterization can be problematic for non-normal distributions because the maximum possible correlation may be much smaller than 1, making it difficult to interpret the strength of dependence (see, e.g., Madsen and Dalthorp 2007).

COP models monotone dependence rather than correlation among the y_i by means of the copula correlation matrix Σ. As an element of Σ ranges from -1 to 1, the dependence between the corresponding pair of y_i ranges from perfect negative dependence to perfect positive dependence (Joe 2001). If Σ is the identity matrix, the COP model coincides with BR.

OLS assumes independence and homoscedasticity, whereas GLS models dependence as residual covariance between logit-transformed percentages and can accommodate heteroscedasticity on the logit scale. Both OLS and GLS assume normality of residuals, and if this assumption is reasonable, then modeling dependence as residual covariance is sensible. However, because covariance is correlation scaled by the standard deviations, GLS may have the same drawbacks as BRdep in terms of modeling dependence when the normality assumption is not met.

In contrast to the marginal beta models, OLS and GLS model the logit-transformed percentages as a linear function of the regression parameters. The coefficient β_k of DTS is the change in mean logit-transformed percentages when DTS is increased by 1 unit and all other predictor variables are held constant. Alternatively, e^{β_k} is the ratio of the median $y_i/(1 - y_i)$ to the median $y_j/(1 - y_j)$, where $\text{DTS}_i = \text{DTS}_j + 1$ and all other predictors are held constant. The awkwardness of this interpretation compared with the more natural odds ratio makes the marginal beta models more attractive.

GLS, BRdep, and COP models account for spatial dependence, whereas the OLS and BR models ignore it. Ignoring spatial dependence in a model when it is present in the data can affect the precision of the regression estimates and result in invalid tests of significance. The simulation study data were created to incorporate spatial dependence with an exponential decay model within each site. When the simulated data are dependent, the effective sample size is smaller than $n = 248$. The GLS, BRdep, and COP models

adjust the standard errors to account for the reduced effective sample size. The OLS and BR models ignore dependence and thus their standard errors are too small, leading to reduced confidence coverage. Accordingly, the confidence coverage of the GLS, BRdep, and COP models exceeded that of the OLS and BR models in the simulation study, which demonstrated the superiority of the GLS, BRdep, and COP models over the OLS and BR models when spatial dependence is present. When the data are spatially independent, the BR, BRdep, and COP models as well as the OLS and GLS models, respectively, yield almost identical results, providing unbiased parameter estimates and approximately 95% confidence coverage for the parameters. The simulation study showed that for the given data, θ of approximately ≥ 0.6 was equivalent to very weak spatial dependence, making it unnecessary to include spatial dependence structure in the model for $\theta \geq 0.6$. Because the scale of θ depends on the minimum distance between observations (3 m in the case study presented and simulation data), the θ value for which spatial dependence can be ignored will differ for studies with a different sampling design.

The simulation study also showed that even the GLS, BRdep, and COP models provide fairly low confidence coverage when very strong spatial dependence exists. Of the three model types, the GLS model resulted in the lowest confidence coverage among the three model types, suggesting that the use of BRdep and COP models should be preferred over the GLS model in the presence of very strong spatial dependence. A possible explanation for the reduced coverage with BRdep and COP models under very strong spatial dependence is that the standard errors assume a large sample, and effective sample size decreases as dependence increases.

The BR model accounts for the bounded nature of vegetation abundance data, and the BRdep as well as the COP model, which was introduced in this study, also account for spatial dependence in the data structure. Another common problem faced by researchers who deal with vegetation abundance data is the presence of excess zeros or a point mass at zero. If the data are zero-inflated, it might be necessary to fit zero-inflated BR and COP models. Cook et al. (2008) presented zero-inflated beta regression in the context of the analysis of corporate capital structure decisions and found that the zero-inflated model outperformed other standard methods that ignored the point mass at zero. Adapting the COP model to accommodate a zero-inflated response, while spatial dependence is modeled, is a topic for future research.

Conclusions

BR and COP models based on the beta distribution can be used to estimate percent shrub cover in forested landscapes. The amount of unexplained variation in the model is generally large when percent shrub cover responds to processes and conditions that occur at a finer scale than the available explanatory variables. Future researchers should focus on improving the model fit of understory vegetation cover models by eliminating the scale issue between response and explanatory variables.

OLS and GLS models with log-transformed response provided biased model predictions and larger MSPE than the BR, BRdep, and COP models, which are based on the beta distribution. Hence, the use of OLS models for modeling shrub cover data bounded between 0 and 1 is not recommended. An additional drawback of the OLS and GLS models with log-transformed response is the interpretation of the parameter estimates that is not straightforward on the original scale of the response.

The COP model introduced here, which accounts for spatial dependence, resulted in 95% confidence coverage and, hence, provided better confidence coverage than the BR model when spatial dependence was present in the data. The COP and BRdep models both account for spatial dependence and result in the same confidence coverage when spatial dependence is present. Because BRdep models dependence as correlation, which is strictly linear dependence, interpreting the strength of dependence can be difficult using this model. The COP model does not have this problem because it models monotone dependence rather than correlation.

When the spatial dependence is very strong, even the GLS, BRdep, and COP models result in smaller confidence coverage than 95%, with the GLS model providing the worst confidence coverage among the three model types.

Although the motivation for this study was to model percent shrub cover in riparian forests, the BR, BRdep, and COP models are general and can be applied to other vegetation communities or other types of data that are bounded to the open interval $(0, 1)$. The BR, BRdep, and COP models should be extended so that they allow accounting for zero inflation, which is frequently observed in vegetation studies.

Literature Cited

ANTHONY, R.G., E.C. MESLOW, AND D.S. DECALESTA. 1987. The role of riparian zones for wildlife in Westside Oregon forests—What we know and don't know. National Council of the Paper Industry for Air and Stream Improvement. *Tech. Bull.* 514:5–12.

BONHAM, C.D. 1989. *Measurements for terrestrial vegetation.* John Wiley and Sons, New York, NY. 338 p.

BRAUN-BLANQUET, J. 1964. *Pflanzensoziologie; Grundzüge der Vegetationskunde*. 3rd ed. Springer, Vienna, Austria.

BREHM, J., AND S. GATES. 1993. Donut shops and speed traps: Evaluating models of supervision on police behavior. *Am. J. Polit. Sci.* 37(2):555–581.

CASELLA, G., AND R.L. BERGER. 2002. *Statistical inference.* Duxbury, Pacific Grove, CA. 660 p.

CHAN, S.S., P.D. ANDERSON, J.H. CISSEL, L. LARSEN, AND C. THOMPSON. 2004. Variable density management in riparian reserves: Lessons learned from an operational study in managed forests of western Oregon, USA. *For. Snow Landsc. Res.* 78(1/2):151–172.

CHEN, J., M. SHIYOMI, Y. YAMAMURA, AND Y. HORI. 2006. Distribution model and spatial variation of cover in grassland vegetation. *Grassland Sci.* 52:167–173.

CHEN, J., M. SHIYOMI, Y. HORI, AND Y. YAMAMURA. 2008a. Frequency distribution models for spatial patterns of vegetation abundance. *Ecol. Model.* 211:403–410.

CHEN, J., M. SHIYOMI, C.D. BONHAM, T. YASUDA, Y. HORI, AND Y. YAMAMURA. 2008b. Plant cover estimation based on the

beta distribution in grassland vegetation. *Ecol. Res.* 23:813–819.

CISSEL, J.H., P.D. ANDERSON, D. OLSON, K.P. PUETTMANN, S. BERRYMAN, S.S. CHAN, AND C. THOMPSON. 2006. *BLM density management and riparian buffer study: Establishment report and study plan.* US Geological Survey, Scientific Investigations Report 2006-5087. 151 p.

COOK, D.O., R. KIESCHNICK, AND B.D. MCCULLOUGH. 2008. Regression analysis of proportions in finance with self selection. *J. Empir. Finan.* 15:860–867.

CRIBARI-NETO, F., AND A. ZEILEIS. 2010. Beta regression in R. *J. Statist. Softw.* 34(2):1–24.

DAMGAARD, C. 2009. On the distribution of plant abundance data. *Ecol. Inform.* 4:76–82.

DAUBENMIRE, R.F. 1959. Canopy coverage method of vegetation analysis. *Northwest Sci.* 33:39–64.

ESPINHEIRA, P.L., S.L.P. FERRARI, AND F. CRIBARI-NETO. 2008. On beta regression residuals. *J. Appl. Statist.* 35(4):407–419.

FERRARI, S.L.P., AND F. CRIBARI-NETO. 2004. Beta regression for modelling rates and proportions. *J. Appl. Statist.* 31(7):799–815.

HAGAR, J.C. 2007. Wildlife species associated with non-coniferous vegetation in Pacific Northwest conifer forests: A review. *For. Ecol. Manag.* 246:108–122.

HALPERN, C.B., AND T.A. SPIES. 1995. Plant species diversity in natural and managed forests of the Pacific Northwest. *Ecol. Appl.* 5(4):913–934.

JOE, H. 2001. *Multivariate models and dependence concepts.* Chapman & Hall/CRC, London, UK. 424 p.

KASS, R.E., AND A.E. RAFTERY. 1995. Bayes factors. *J. Am. Statist. Assoc.* 90(430):773–795.

KERNS, B.K., AND J.L. OHMANN. 2004. Evaluation and prediction of shrub cover in coastal Oregon forests (USA). *Ecol. Indic.* 4:83–98.

KIESCHNICK, R., AND B.D. MCCULLOUGH. 2003. Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Statist. Model.* 3:193–213.

KORHONEN, L., K.T. KORHONEN, P. STENBERG, M. MALTAMO, AND M. RAUTIAINEN. 2007. Local models for forest canopy cover with beta regression. *Silva Fenn.* 41(4):671–685.

MADSEN, L. 2009. Maximum likelihood estimation of regression parameters with spatially discrete data. *J. Agr. Biol. Environ. Stat.* 14:375–391.

MADSEN, L., AND D. DALTHORP. 2007. Simulating correlated count data. *Environ. Ecol. Statist.* 14:129–148.

MARQUARDT, T. 2010. *Accuracy and suitability of several stand sampling methods in riparian zones.* MS thesis. Oregon State Univ., Corvallis, OR. 77 p.

MARTINUZZI, S., L.A. VIERLING, W.A. GOULD, M.J. FALKOWSKI, J.S. EVANS, A.T. HUDAK, AND K.T. VIERLING. 2009. Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. *Remote Sens. Environ.* 113: 2522–2546.

MCCULLAGH, P., AND J.A. NELDER. 1989. *Generalized linear models.* 2nd ed. *Monographs on statistics and probability 37*. Chapman & Hall, London, UK. 511 p.

NELSEN, R.B. 2006. *An introduction to copulas*. Springer, New York, NY. 239 p.

PIELOU, E.C. 1977. *Mathematical ecology.* John Wiley & Sons, New York, NY. 384 p.

R DEVELOPMENT CORE TEAM. 2009. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

RAFTERY, A.E. 1995. Bayesian model selection in social research. *Sociol. Methodol.* 25:111–163.

SAS INSTITUTE, INC. 2008. *SAS/STAT 9.2 user's guide*. Cary, NC: SAS Institute, Inc.

SMITHSON, M., AND J. VERKUILEN. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods* 11(1):54–71.

SUCHAR, V.A., AND N.L. CROOKSTON. 2010. Understory cover and biomass indices predictions for forest ecosystems of the Northwestern United States. *Ecol. Indic.* 10:602–609.

WEISBERG, P.J., C. HADORN, AND H. BUGMANN. 2003. Predicting understory vegetation cover from overstory attributes in two temperate mountain forests. *Forstwissenschaft. Centralblatt.* 122:273–286.