Estimating cavity tree and snag abundance using negative binomial regression models and nearest neighbor imputation methods

Bianca N.I. Eskelson, Hailemariam Temesgen, and Tara M. Barrett

Abstract: Cavity tree and snag abundance data are highly variable and contain many zero observations. We predict cavity tree and snag abundance from variables that are readily available from forest cover maps or remotely sensed data using negative binomial (NB), zero-inflated NB, and zero-altered NB (ZANB) regression models as well as nearest neighbor (NN) imputation methods. The models were developed and fit to data collected by the Forest Inventory and Analysis program of the US Forest Service in Washington, Oregon, and California. For predicting cavity tree and snag abundance per stand, all three NB regression models performed better in terms of mean square prediction error than the NN imputation methods. The most similar neighbor imputation, however, outperformed the NB regression models in predicting overall cavity tree and snag abundance.

Résumé : Les données sur l'abondance des arbres creux et des chicots sont extrêmement variables et contiennent plusieurs observations nulles. Nous prédisons l'abondance des arbres creux et des chicots à partir de variables facilement disponibles dans les cartes du couvert forestier ou parmi les données obtenues par télédétection en utilisant des modèles de régression binomiale négative (BN), BN à excès de zéros (ZINB) et BN tronquée à zéro (ZANB), ainsi que des méthodes d'imputation par le plus proche voisin. Les modèles sont élaborés et ajustés aux données collectées par le programme d'analyse et d'inventaire forestiers du U.S. Forest Service dans les États de Washington, de l'Oregon et de la Californie. Les trois modèles de régression BN offraient une meilleure performance en terme d'erreur quadratique moyenne de prédiction que les méthodes d'imputation par le plus proche voisin pour prédire l'abondance des arbres creux et des chicots par peuplement. Cependant, l'imputation par le voisin le plus semblable donnait de meilleurs résultats que les modèles de régression BN pour prédire l'abondance globale des arbres creux et des chicots.

[Traduit par la Rédaction]

Introduction

In the past decades, traditional timber-oriented forest management has broadened to commodity production while managing forest resources in an ecologically sustainable manner (Fan et al. 2004). Derived benefits often include managing forests for wildlife, enhancing biodiversity, and protecting water quality (Sustainable Forestry Initiative 2002).

Snags (standing dead trees) are a significant structural component of many forest ecosystems (Harmon et al. 1986). They create nesting, foraging, and roosting habitat for a variety of wildlife species that depend on snags and large trees for survival and reproduction (Bate et al. 1999; Russell et al. 2006; Wisdom and Bate 2008). Snags are important for the maintenance of biodiversity (Shorohova and Tetioukhin 2004; Aakala et al. 2008), as many deadwood-dependent organisms are confined to snags during their life cycle (Nilsson et al. 2002). Snags also contribute to ecological processes and decay dynamics (Ganey and Vojta 2005). Episodic events (e.g., insect outbreaks, fire, snow- and wind-caused stem breakage) create large quantities of snags. Small-scale mortality caused, for example, by competition or suppression continuously creates smaller quantities of snags (Aakala et al. 2008).

Cavity trees contribute to diverse forest structure and wildlife habitat (Temesgen et al. 2008). Cavity trees are live trees or snags containing a hollow that provides wildlife species with shelter from the elements and protection from disturbance by predators and competitors (Carey 1983). Cavity trees provide many birds, mammals, reptiles, and amphibians with habitat for nesting, roosting, loafing, hibernating, and eating. They also provide escape cover and food storage locations (Carey 1983; Jensen et al. 2002).

Cavity-nesting birds and other wildlife species depend on an adequate and continuous supply of cavity trees and snags (Fan et al. 2003*a*). Timber harvest and human access (e.g., characterized by distance to nearest town or road) can have substantial effects on snag density (Wisdom and Bate 2008). Because cavity trees and snags are removed under an intensive timber management regime, the availability of cavity

Received 19 October 2008. Accepted 3 June 2009. Published on the NRC Research Press Web site at cjfr.nrc.ca on 11 September 2009.

B.N.I. Eskelson[1] and H. Temesgen. Oregon State University, Department of Forest Engineering, Resources and Management, 204 Peavy Hall, Corvallis, OR 97331-5706, USA.
T.M. Barrett. USDA Forest Service, Pacific Northwest Research Station, Forestry Sciences Lab, 3301 "C" Street, Suite 200, Anchorage, AK 99503-3954, USA.

[1]Corresponding author (e-mail: bianca.eskelson@oregonstate.edu).

trees is a concern in resource management and conservation (Fan et al. 2004). Maintaining snags in suitable abundance and stages of decay is critical to the preservation of biodiversity and the sustained functioning of forest ecosystems (DeLong et al. 2008).

Managers need to understand the nature of the cavity resource and the patterns of cavity tree abundance to effectively manage forest resources for ensuring viable populations of cavity-using wildlife (Carey 1983). Regional summaries of current deadwood amounts, including snags and down woody debris, are needed for broad-scale assessment of wildlife habitat (Ohmann and Waddell 2002). Snag abundance is frequently used to incorporate habitat requirements of cavity-nesting wildlife into management plans. Snag abundance, however, does not take into account live cavity trees (Allen and Corn 1990). Although the proportion of stems with cavities is often at least twice as high for snags as for live trees (e.g., Goodburn and Lorimer 1998; Fan et al. 2003*b*; Temesgen et al. 2008), cavities can be more common in live trees (Goodburn and Lorimer 1998), or the number of cavities in live trees can contribute substantially to the total number of cavity trees because live trees are more abundant than snags. Hence, cavity tree abundance, which considers both live trees and snags, should be used as an indicator in wildlife habitat models or to formulate wildlife tree retention recommendations in management plans.

Cavity tree abundance is highly variable even among forest stands that are similar in many other respects (Fan et al. 2004). This variability is due to the fact that cavity development is a relatively rare event governed by stochastic processes such as fire, insect attack, disease, and mechanical or chemical injury that can lead to tree death or injury (Carey 1983). Tree characteristics and stand attributes such as size, decay class, and species only play a partial role in cavity tree development (Fan et al. 2003*a*).

Recent studies have estimated cavity tree abundance at the stand level. For example, Temesgen et al. (2008) used nearest neighbor (NN) imputation and classification and regression tree (CART) methods to estimate cavity tree abundance. Fan et al. (2003*a*) estimated cavity tree abundance by stand age and basal area using CART and described the cavity tree density distribution within a cluster using the Weibull probability density function. They found that the proportion of stands with cavity trees increases with increasing stand age and increasing basal area. Fan et al. (2005) quantified the frequency and size distribution of cavity trees in seedling–sapling, pole, sawtimber, and old-growth stands based on plot data. Fan et al. (2004) simulated cavity tree dynamics under alternative harvest regimes. Most other studies concentrated on the distribution of cavity trees or snags at the individual-tree or species level. Fan et al. (2003*b*) explored factors associated with cavity tree abundance and developed models that can be used to predict relative frequency of cavity trees based on tree size, species, and decay class. Carey (1983) found that tree diameter measured at 1.37 m above the ground (DBH) and site index were good indicators for cavity tree abundance.

In the states of California, Oregon, and Washington, information on cavity tree and snag occurrence along with other information such as species, DBH, and height of individual trees is collected as part of the national inventory of public and private forests (the Forest Inventory and Analysis (FIA) Inventory). The FIA inventory uses an interpenetrating panel design with 10 panels in the western states, where all plots located in one of the 10 interpenetrating panels (10%) are measured each year (Brand et al. 2000). To estimate current cavity tree and snag abundance from paneled inventory data, the information on cavity tree and snag abundance in the current year needs to be updated for all unmeasured panels. It is of interest to be able to do this with variables that are readily available from forest cover maps or remotely sensed data (e.g., aerial photographs, satellite data, lidar). These variables will collectively be referred to as map label variables in this study. If cavity tree and snag abundance can be predicted from map label variables, accurate, spatially comprehensive, current, and very detailed information on cavity tree and snag abundance could be provided to managers and planners interested in assessing wildlife habitat in their forests (Temesgen et al. 2008).

Count distributions are useful to describe nonnegative integer values such as the number of snags or cavity trees per plot. A variety of count models are based on Poisson regression, which is the basic count model (Hilbe 2007, p. 8). The Poisson distribution assumes equidispersion and hence that the mean and variance are equal. The development of more general count models such as the negative binomial (NB) distribution, which do not assume equidispersion, has been driven by the fact that equality of mean and variance is rarely found in natural resource data. A variety of NB regression models have been developed to accommodate additional violations of distributional assumptions, such as no zeros or excess zeros in the data, which often occur in natural resource data (Hilbe 2007, pp. 8–10).

The objectives of this study are (1) to compare the suitability and predictive abilities of negative binomial regression models to estimate snag and cavity tree abundance using map label variables, and (2) to use distribution-free NN imputation methods to impute snag and cavity tree abundance and compare the imputation results with the results of NB regression models.

Negative binomial regression models

Count regression models are a subset of discrete-response regression models and aim to explain the number of occurrences or counts of an event. Poisson regression is the basic count model, and the Poisson distribution is characterized as

$$[1] \qquad P(Y = y) = \frac{\mu^y \mathrm{e}^{-\mu}}{y!}$$

where $y = 0, 1, 2, \ldots$; $\mu > 0$; the random variable y is the count response; and the parameter μ is the mean. A Poisson regression model is obtained by relating the mean μ to a vector of explanatory variables, $\boldsymbol{x}$, by $\mu = \mathrm{e}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}$, where $\boldsymbol{\beta}$ is a vector of regression coefficients to be estimated, and $\boldsymbol{x}^{\mathrm{T}}$ is the transpose of $\boldsymbol{x}$.

A consequence of the Poisson probability mass function (eq. 1) is that the mean and the variance are equal, that is $\mathrm{Var}[Y|x] = E[Y|x] = \mu$. When data do not fit the Poisson distribution, it is typically because of overdispersion, meaning the model's variance exceeds the mean value. The NB distribution, which can be derived as a gamma mixture of Poisson distributions, employs an extra parameter, α, that directly addresses the overdispersion in the Poisson models. The NB distribution is characterized as

$$[2] \qquad P(Y=y)=\frac{\Gamma(y+1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}$$

where $y = 0, 1, 2, \ldots$and $\mu > 0$. α represents the degree of overdispersion. The mean is μ, as in the Poisson model, but the variance is $\mu+\alpha\mu^2$, thus allowing the variance to exceed μ (Hilbe 2007, pp. 78, 80). NB regression models are obtained in the same way as Poisson regression models by relating the mean, μ, to a vector of explanatory variables, $\boldsymbol{x}$, by $\mu = e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}$.

Although the NB model is more flexible than the Poisson, there is no guarantee that it provides an adequate model for count data. Excess zeros or the absence of zeros in the data violates the distributional assumptions that apply equally to the Poisson and NB distributions. Other possible violations of the distributional assumptions occur when the data contain censored or truncated observations, or when the data are structured as panels (e.g., clustered and longitudinal data) (Hilbe 2007, pp. 11–13).

Zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB) regression models have been developed to account for data with a high percentage of zero counts (Lambert 1992; Welsh et al. 1996). Zero-inflated models are mixture models combining a count distribution with a point mass at zero. In zero-inflated models there are two sources of zeros: zeros come from either the count distribution or the point mass (Lambert 1992; Hall 2000). The ZINB model is defined as follows:

$$[3] \qquad P(Y=y) = \begin{cases} \pi + (1-\pi)f_{\text{count}}(0;\boldsymbol{x},\boldsymbol{\beta}) & \text{if } y=0 \\ (1-\pi)\, f_{\text{count}}(y;\boldsymbol{x},\boldsymbol{\beta}) & \text{if } y=1,2,\ldots. \end{cases}$$

where $\pi = f_{\text{zero}}(0;\, z,\, \boldsymbol{\gamma})$ is the probability of belonging to the point mass component, and $(1 - \pi)$ is the probability of belonging to the count distribution. z is a vector of explanatory variables used in the logistic model, and $\boldsymbol{\gamma}$ is a vector of regression coefficients to be estimated. $f_{\text{count}}(y;\, \boldsymbol{x},\, \boldsymbol{\beta})$ corresponds to the probability mass function of the NB distribution given in eq. 2, and $f_{\text{count}}(0;\, \boldsymbol{x},\, \boldsymbol{\beta}) = (1-(1+\alpha\mu)^{-(1/\alpha)})$, where α represents the degree of overdispersion, and μ is related to a vector of explanatory variables, $\boldsymbol{x}$, by $\mu = e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}}$, where $\boldsymbol{\beta}$ is a vector of regression coefficients to be estimated.

Another approach for dealing with excess zeros in the data is to model the response as having two states: a state in which no cavity trees or snags occur, and a state in which cavity trees or snags occur with varying levels of abundance (Welsh et al. 1996). The first state, a binary process that generates positive versus zero counts, is modeled by applying logistic regression. Given that cavity trees are observed, the number of cavity trees (the second state) can be modeled by a zero-truncated Poisson (ZTP) or zero-truncated NB (ZTNB) distribution. The process generating positive counts only commences after crossing a zero barrier or hurdle. The combined models are known as conditional models (Welsh et al. 1996) or are referred to as Poisson and NB hurdle models or as zero-altered Poisson (ZAP) and zero-altered NB (ZANB) models (Hilbe 2007). In the NB case, the combined regression model is defined as follows:

$$[4] \qquad P(Y=y) = \begin{cases} f_{\text{zero}}(0;z,\boldsymbol{\gamma}) & \text{if } y=0 \\ \left(1-f_{\text{zero}}(0;z,\boldsymbol{\gamma})\right) f_{\text{zt}}(y;\boldsymbol{x},\boldsymbol{\beta}) & \text{if } y=1,2,\ldots. \end{cases}$$

where $f_{\text{zero}}(0;\, z,\, \boldsymbol{\gamma})$ is the probability of a zero count, and $(1 - f_{\text{zero}}(0;\, z,\, \boldsymbol{\gamma}))$ is the probability of overcoming the hurdle. z is a vector of explanatory variables used in the logistic model, and $\boldsymbol{\gamma}$ is a vector of regression coefficients to be estimated. $f_{\text{zt}}(y;\boldsymbol{x},\boldsymbol{\beta}) = f_{\text{count}}(y;\boldsymbol{x},\boldsymbol{\beta})/(1-f_{\text{count}}(0;\boldsymbol{x},\boldsymbol{\beta}))$ is a ZTNB model, with $f_{\text{count}}(y;\, \boldsymbol{x},\, \boldsymbol{\beta})$ and $f_{\text{count}}(0;\, \boldsymbol{x},\, \boldsymbol{\beta})$ defined as above. All observations are used to fit $f_{\text{zero}}(0;\, z,\, \boldsymbol{\gamma})$, treating positive counts as 1's in the logistic regression framework. The data are separated into two subsets, using only data with positive counts to fit $f_{\text{zt}}(y;\, \boldsymbol{x},\, \boldsymbol{\beta})$. For more details on ZAP and ZANB models see Cameron and Trivedi (1998, pp. 123–128; 2005, pp. 544–546, 680, 681).

Nearest neighbor imputation methods

NN imputation approaches are nonparametric donor-based methods where the imputed value is either a value that was actually observed for another item or unit or the average of values from more than one item or unit. These donors can be determined in a variety of ways. Forest attributes that are measured on all observations are referred to as *X* variables. *Y* variables are those forest attributes that are only measured on a subset of observations — in this case cavity tree and snag abundance. Observations with measured *X* and *Y* variables are called reference observations, and target observations are those that only have *X* variables measured. The similarity between target and reference data is determined with a distance metric defined in the multidimensional feature space of the *X* variables (LeMay and Temesgen 2005).

Methods

Data

Data for this study were obtained from the FIA inventory from Washington, Oregon, and California collected from 2001 to 2007. For details about the inventory, see Bechtold and Patterson (2005). Each field plot is composed of a cluster of four points, with each point being composed of two nested fixed-radius plots (subplot and macroplot) used to sample trees of different size (Bechtold and Scott 2005). Unique polygons (also called condition classes) on the FIA plot are distinguished by structure, management history, or forest type. Only data collected on the subplots were used for this study and summarized by condition class. The data set contained 5870 stands (or condition classes) for which all four subplots of the FIA plot belonged to the same condition class. The data covered a wide range of ground and map label variables (Tables 1 and 2).

Cavity presence was determined in the field by classifying each live tree or snag taller than 1.5 m and greater than 12.5 cm DBH into one of three categories: (1) no cavity

Table 1. Descriptive statistics for stands, n = 5870.

Variable	Min.	Mean	Median	Max.	SD
Cavity tree counts (no./stand)	0	0.58	0	13	1.20
Snag counts (no./stand)	0	3.40	2	77	5.59
% conifer	0	0.81	1	1	0.33
Mean stand age (years)	0	105	86	989	80
Elevation (m)	0	1127	1128	3366	679
Aspect (°)	0	162	161	360	114
Slope (%/100)	0	0.32	0.30	1.50	0.23

Note: SD, standard deviation.

Table 2. Number of stands in each site class and each height class, n = 5870.

	No. of stands
Site class (m^3·ha^{-1}·year^{-1})	
≥15.7	118
11.6–15.6	430
8.4–11.5	1051
5.9–8.3	915
3.5–5.8	1303
1.4–3.4	1124
0–1.3	929
Height class (m)	
0–9.99	1318
10–19.99	2130
20–29.99	1462
30–39.99	617
40–49.99	253
≥50	90

present, (2) cavity greater than 15.2 cm diameter present, and (3) cavity less than 15.2 cm diameter but no larger cavities present. Cavity presence was only recorded for trees with cavities that could, in the field crew's judgment, be used by wildlife such as birds or mammals. Cavity tree abundance is assumed to be additive from individual trees in a stand and is quantified as the number of cavity trees (both live trees and snags) per stand without apportioning it by species or species groups. Cavity tree abundance can be assumed to be under recorded, as field crews are more likely to miss cavities than record cavities that do not exist. Snag abundance is the number of standing dead trees per stand, and snags only included dead trees that leaned less than 45° from vertical.

While 71% of the cavities were observed in snags, 29% of the cavities were found in live trees. Live trees (87% of standing trees) are more abundant than snags (13% of standing trees), but only 0.73% of all live trees had cavities compared with 12.20% of the snags. Of the 5870 stands, 1802 and 3921 contained cavity trees and snags, respectively, resulting in large numbers of zero counts for both cavity tree and snag abundance (Fig. 1).

Mean stand age was used to represent stand development stage. The midpoint of five height classes was used (Table 2). Slope, aspect, elevation, and transformations of these three variables (Salas et al. 2008) as well as the midpoint of seven site classes (Table 2) represented site conditions. The percentage of conifer basal area and very broad forest type groups described general species composition. Pearson's correlation coefficient and significance between the continuous explanatory variables and cavity tree and snag abundance are displayed in Table 3. Four forest type groups were used: (1) Douglas-fir (1334), (2) fir – spruce – mountain hemlock (711), (3) other conifers (2637), and (4) hardwoods (1188). Four owner groups were distinguished: (1) Forest Service (2960), (2) other federal (653), (3) state and local government (354), and (4) private (1903). The group of private forest owners included corporations; nongovernmental conservation and natural resources organizations; unincorporated local partnerships, associations, and clubs; Native Americans; and individuals. The group of other federal forest owners included the National Park Service, the Bureau of Land Management, the Fish and Wildlife Service, the Department of Defense/Energy, and other federal owners. Boxplots of the owner and forest type groups for log-transformed cavity tree and snag abundance are displayed in Figs. 2 and 3.

Negative binomial (NB) regression models

The complete data set was used to fit NB, ZINB, and ZANB models in R using a maximum likelihood estimator (R Development Core Team 2008; Zeileis et al. 2008). Snag abundance and cavity tree abundance were used as response variables. Map label variables related to site, ownership, forest development stage, and general species composition were used as explanatory variables. Simple interactions and second-order terms of the continuous explanatory variables were included in the model selection.

Explanatory variables that did not contribute significantly in explaining variation were dropped from the NB, ZINB, and ZANB models. Nested and non-nested models were compared using Akaike's information criterion (AIC; Akaike 1973, 1974) and Schwarz's Bayesian information criterion (BIC, Schwarz 1978):

$$[5] \quad \mathrm{AIC} = -2\ \ln(L) + 2p$$

$$[6] \quad \mathrm{BIC} = -2\ \ln(L) + p\ \ln(n)$$

where p is the number of parameters estimated in the model, n is the number of observations in the modeling data set, and $\ln(L)$ is the natural logarithm of the maximum likelihood of the model. The parameter estimates as well as the AIC and BIC values for the cavity tree and snag models are shown in Tables A1 and A2, respectively.

The performance of the models chosen for NB, ZINB, and ZANB for predicting cavity tree and snag abundance was examined using leave-one-out cross-validation. To as-

Fig. 1. Frequency distribution of stands with up to 25 counts of cavity trees (left) and snags (right), n = 5870.

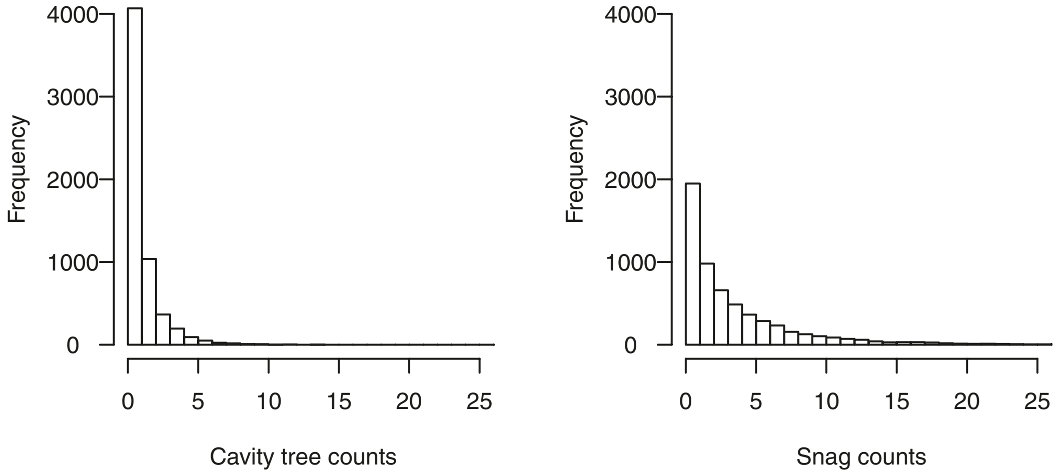
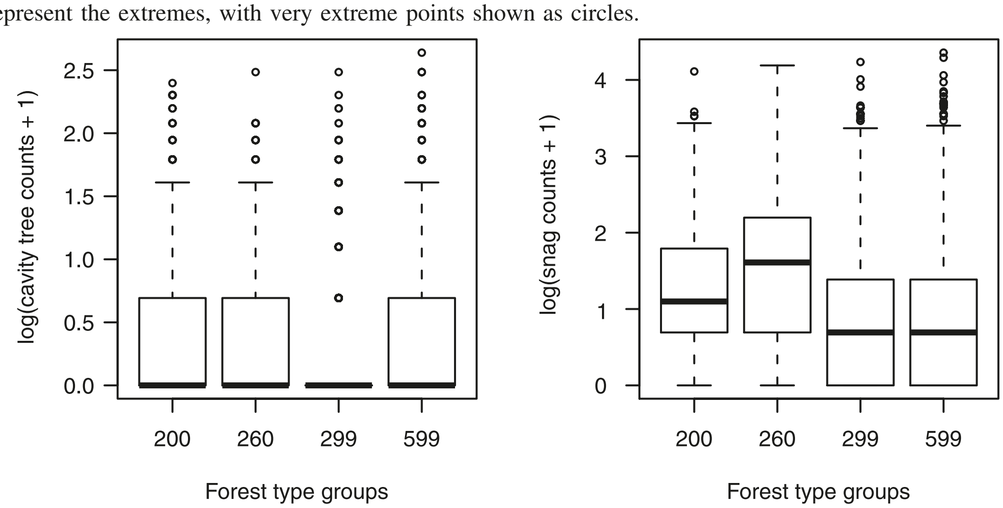


Table 3. Pearson's correlation coefficient and significance (in parentheses) between the continuous explanatory variables and cavity tree and snag abundance, n = 5870.

Attribute	% conifer	Mean stand age (years)	Elevation (m)	Slope (%/100)	Aspect (°)	Site index ($m^3 \cdot ha^{-1} \cdot year^{-1}$)	Height (m)
Cavity tree counts	–0.063	0.103	–0.095	0.101	0.033	–0.082	0.193
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.012)	(<0.001)	(<0.001)
Snag counts	0.025	0.060	0.017	0.046	0.050	–0.098	0.191
	(0.058)	(<0.001)	(0.186)	(0.001)	(0.001)	(<0.001)	(<0.001)

Fig. 2. Number of log transformed cavity tree counts (left) and snag counts (right) against forest type groups: 200, Douglas-fir; 260, fir – spruce – mountain hemlock; 299, other conifers; 599, hardwoods. Boxes show the data between the quartiles, thick lines represent the median, "whiskers" represent the extremes, with very extreme points shown as circles.

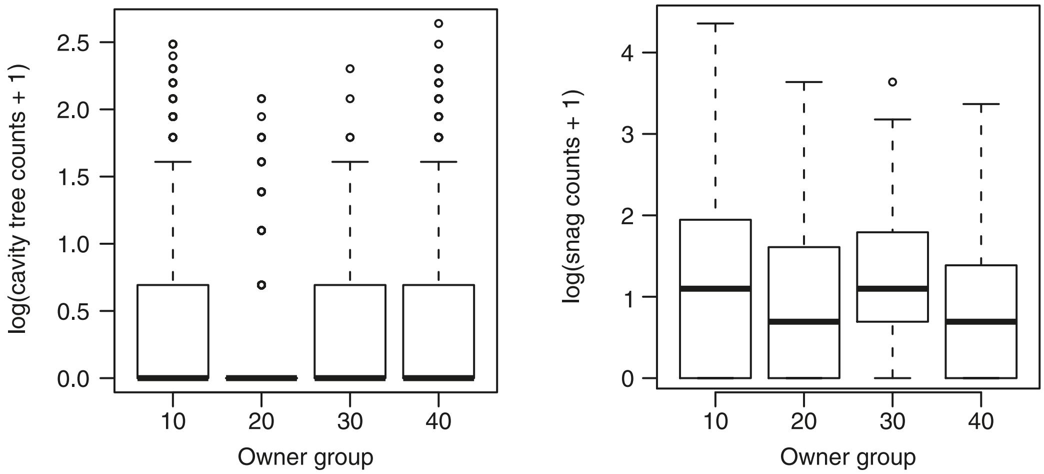


sess the adequacy of the models for predicting the overall counts of cavity trees and snags, an approximate χ^2 goodness-of-fit test was calculated using observed and predicted counts from the leave-one-out cross-validation of cavity trees and snags, respectively:

$$[7] \qquad \chi^2 = \sum_{k=1}^{m} \frac{\left[\#(y_i = k) - \sum_i \Pr(y_i = k)\right]^2}{\sum_i \Pr(y_i = k)}$$

where # denotes the frequency of observations y_i in count class k across the data set, and $\Pr(y_i = k)$ is the predicted probability that an observation belongs to count class k. This statistic is χ^2 distributed with (m – 1) degrees of freedom. The number of count classes, m, was 14 and 78 for cavity tree and snag counts, respectively. For cavity trees the largest count class, m = 14, included counts of 13 and up (≥13). For snags the largest count class, m = 78, included counts of 77 and up (≥77). The reliability of the χ^2 statistic is questionable, since many observed frequencies were either zero or smaller than 5. Hence, diagnostic plots, which plot the differences between predicted and observed proportions against the count classes, k, were used to detect any predictive bias and assess goodness-of-fit (Lambert 1992;

Fig. 3. Number of log-transformed cavity tree counts (left) and snag counts (right) against owner groups: 10, Forest Service; 20, other federal; 30, state and local government; 40, private. Boxes show the data between the quartiles, thick lines represent the median, "whiskers" represent the extremes, with very extreme points shown as circles.



Fortin and DeBlois 2007). The difference, d_k, between predicted proportion and observed proportion is computed as follows:

$$[8] \qquad d_k = \sum_{i=1}^{n} \left(\frac{\Pr(y_i = k)}{n} \right) - \left(\frac{\#(y_i = k)}{n} \right)$$

where n is the number of observations in the data set, and the other parameters are as defined above. As suggested by Fortin and DeBlois (2007), the sum of absolute d_k's was defined as w and used as an index of the goodness-of-fit of the different NB regression models and the imputation methods.

Nearest neighbor (NN) imputation

Two imputation methods were employed in this study using the yaImpute package in R (Crookston and Finley 2008): (1) MSN — most similar neighbor (Moeur and Stage 1995); and (2) RF — randomForest (Breiman 2001; Crookston and Finley 2008).

In the MSN procedure (Moeur and Stage 1995), the distance metric is of quadratic form:

$$[9] \qquad d_{ij}^2 = (\boldsymbol{X}_i - \boldsymbol{X}_j)\mathbf{W}(\boldsymbol{X}_i - \boldsymbol{X}_j)'$$

where $\boldsymbol{X}_i$ is the $(1 \times p)$ vector of X variables for the ith observation unit, $\boldsymbol{X}_j$ is the $(1 \times p)$ vector of X variables for the jth reference observation unit, and $\mathbf{W}$ is a $(p \times p)$ symmetric matrix of weights. $\mathbf{W}$ is derived from canonical correlation analysis, which makes use of the relationships between X and Y variables so that stronger correlations result in higher weights for a particular X (LeMay and Temesgen 2005).

The RF method is a CART method (Breiman 2001). The data and variables are randomly and iteratively sampled to generate a large group, or forest, of classification and regression trees. For RF two observations are considered similar if they tend to end up in the same terminal nodes in a forest of classification and regression trees. The distance measure is one minus the proportion of trees where a target observation is in the same terminal node as a reference observation (Crookston and Finley 2008; Hudak et al. 2008).

Cavity tree and snag abundance were used as Y variables, as well as square root ($Y^{0.5}$), inverse ($1/(Y + 1)$), and logarithmic ($\ln(Y + 1)$) transformations of these variables. The X variables used for the NN imputation are mean stand age, percentage of conifers, height class midpoint, site class midpoint, elevation (EL), EL^2, slope, slope $\times$ cosine(aspect), and slope $\times$ sine(aspect).

The predictive performance of the NN imputation methods was evaluated by using leave-one-out cross-validation. One stand was used as the target stand, and the remaining 5869 stands served as reference stands. This procedure was repeated for all stands of the four Y variable sets (untransformed data and three transformations of cavity tree and snag abundance) to calculate the mean difference between imputed and observed values (often called bias) and the root mean squared error (square root of the mean squared difference; RMSE; Rawlings et al. 1998, p. 444):

$$[10] \qquad \text{bias} = \frac{\sum_{i=1}^{n} (\text{imputed}_i - \text{observed}_i)}{n}$$

$$[11] \qquad \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\text{imputed}_i - \text{observed}_i)^2}{n}}$$

where n is the number of stands.

Bias and percent RMSE (expressed as percentage of the mean) were used to select three imputation approaches that were then used to compare NB regression models with imputation methods.

Comparison of NB regression models and NN imputation methods

To visually compare NB regression models with NN imputation methods on a per-stand basis, the results of the

Table 4. Bias and relative root mean square error (% RMSE) for the *Y* variables (cavity tree abundance and snag abundance) and the square root, inverse, and logarithmic (ln) transformations of the *Y* variables.

		Cavity trees		Snags	
Method	Response	Bias	% RMSE	Bias	% RMSE
MSN	Y	–0.01	1.62	–0.05	7.55
	$Y^{0.5}$	0.02	1.68	0.05	7.67
	$1/(Y + 1)$	0.00	1.67	0.02	7.73
	$\ln(Y + 1)$	0.01	1.64	0.01	7.58
RF	Y	–0.14	1.52	–0.34	6.85
	$Y^{0.5}$	–0.13	1.54	–0.32	6.93
	$1/(Y + 1)$	–0.13	1.53	–0.25	7.03
	$\ln(Y + 1)$	–0.13	1.53	–0.32	6.87

Note: MSN, most similar neighbor method; RF, randomForest imputation method.

leave-one-out cross-validation were used to plot histograms of the prediction errors (PE). PE is computed as follows:

[12] $$\text{PE} = \text{predicted}_i - \text{observed}_i$$

where observed$_i$ is the observed cavity tree or snag count for the *i*th stand. For the NB, ZINB, and ZANB models, predicted$_i$ is the fitted count response for the *i*th stand, which is based on both the count and zero components for the ZINB and ZANB models. For the NN imputation methods, predicted$_i$ is the imputed count for the *i*th stand. Positive and negative PE values indicate overestimation and underestimation, respectively.

Results

The bias (mean difference) for cavity tree and snag abundance was close to zero for the untransformed data and the three transformations when the NN was imputed using MSN. RF imputation resulted in small negative bias for estimating cavity tree and snag abundance using the untransformed data or any of the three transformations. Bias was smaller for MSN than for RF when the untransformed data or either of the transformations of the *Y* variables was used (Table 4).

For both MSN and RF, the relative RMSE for cavity tree and snag abundance was smallest for the untransformed data followed by the log-transformed data. Relative RMSE for cavity tree and snag abundance was smaller for RF than for MSN when the untransformed data or any of the transformations was used as *Y* variables (Table 4).

Based on the relative RMSE, MSN and RF imputation with the untransformed data as *Y* variables were chosen to be compared with the NB regression models. MSN imputation with the log-transformed data as *Y* variables was chosen as the third method to be compared with the NB regression models because it resulted in the second smallest relative RMSE among the MSN approaches and in smaller bias than any of the RF approaches.

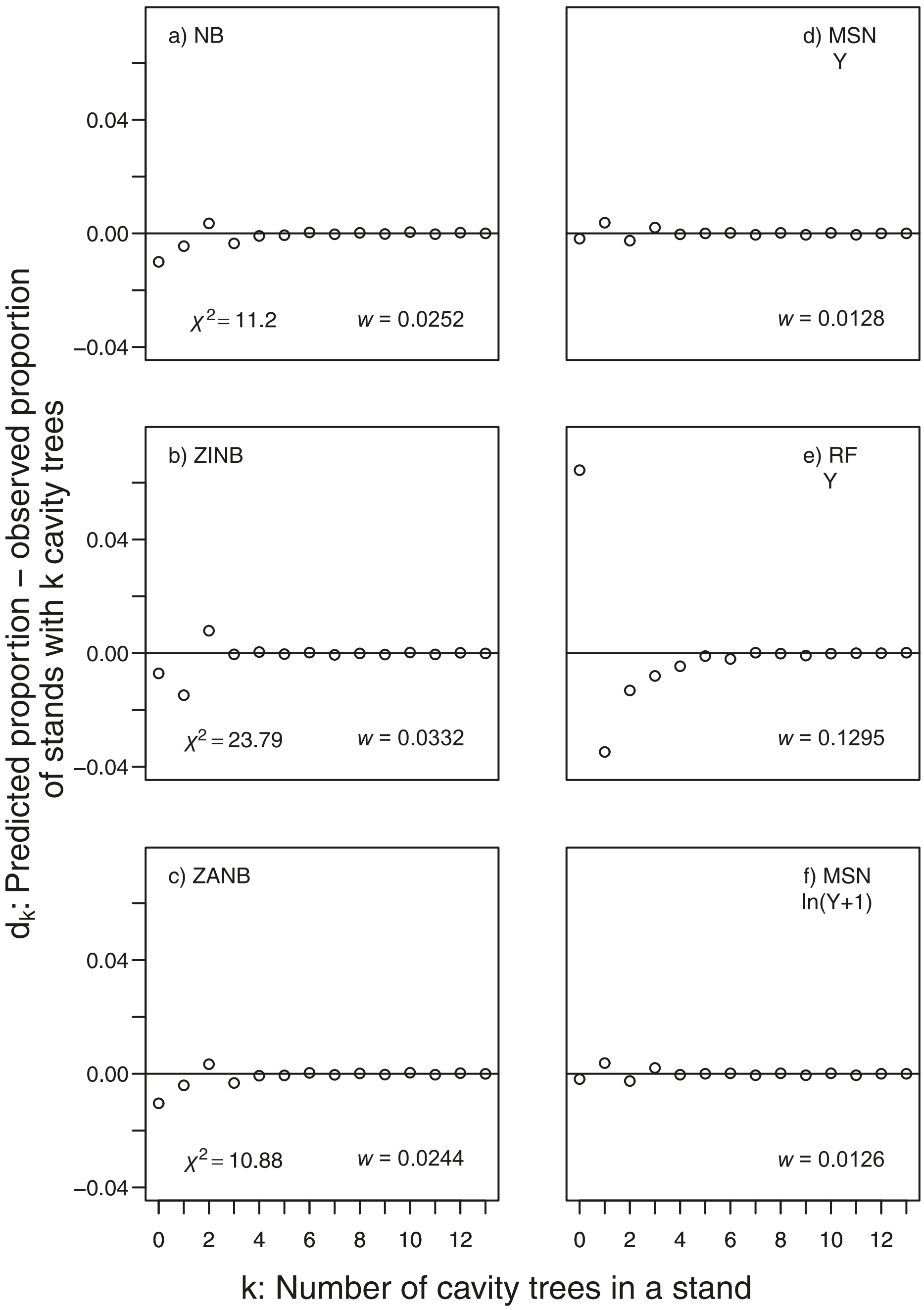
For estimating cavity tree abundance, the ZINB model had the largest χ^2 goodness-of-fit statistic, with a value of 23.79 compared with values of 11.20 and 10.88 for the NB and ZANB models, respectively. The lower the χ^2 statistic, the better the model fit. The critical χ^2 statistic was 22.36 for the probability of a greater value = 0.05 and 13 degrees of freedom. The χ^2 goodness-of-fit statistic of the ZINB model slightly exceeded the critical χ^2 statistic, suggesting that the ZINB model might be inadequate for estimating cavity tree abundance. There is no evidence that the NB and ZANB models are inadequate for estimating cavity tree abundance.

In addition, the diagnostic plots (Fig. 4) were used to compare the models and to identify potential model misspecifications. Differences, d_k, were calculated as predicted proportions minus observed proportions (eq. 8), so that positive values indicate overestimations and negative values indicate underestimation. All three regression models underestimated zero counts and showed a good fit for counts of four and greater (Fig. 4). The ZANB model had the lowest w value (w = 0.0244) followed by the NB model (w = 0.0252).

The diagnostic plot for RF imputation with the untransformed data as *Y* variables showed that the RF imputation highly overestimated the zero cavity tree counts and highly underestimated the number of stands with one cavity tree. MSN imputation with the untransformed and log-transformed data resulted in small underestimations of zero and three cavity tree counts and small overestimations of the number of stands with two and four cavity trees. The w values indicated that the differences were closest to the reference line $d_k = 0$ for the logarithmic MSN model (w = 0.0126) followed by the MSN model with the untransformed data as *Y* variables (w = 0.0128) (Fig. 4).

For estimating snag abundance, the NB model outperformed the ZINB and ZANB models according to the χ^2 goodness-of-fit statistic. The critical χ^2 statistic was 98.48 for the probability of a greater value = 0.05 and 77 degrees of freedom. For both the ZINB and ZANB models, the χ^2 goodness-of-fit statistic greatly exceeded the critical χ^2 statistic, suggesting that the model did not adequately characterize snag abundance. The ZANB model resulted in a smaller w value (w = 0.0371) than the NB model (w = 0.0424). The smallest w value for the ZANB model indicates generally small d_k values and a good model fit. The diagnostic plots (Fig. 5) show that the ZANB model perfectly estimates the number of stands with zero snags, while the NB model slightly overestimates the number of stands with zero snags. The ZINB model greatly overestimated the number of stands with zero snags and greatly underestimated the number of stands with one snag, which caused the large w and χ^2 values.

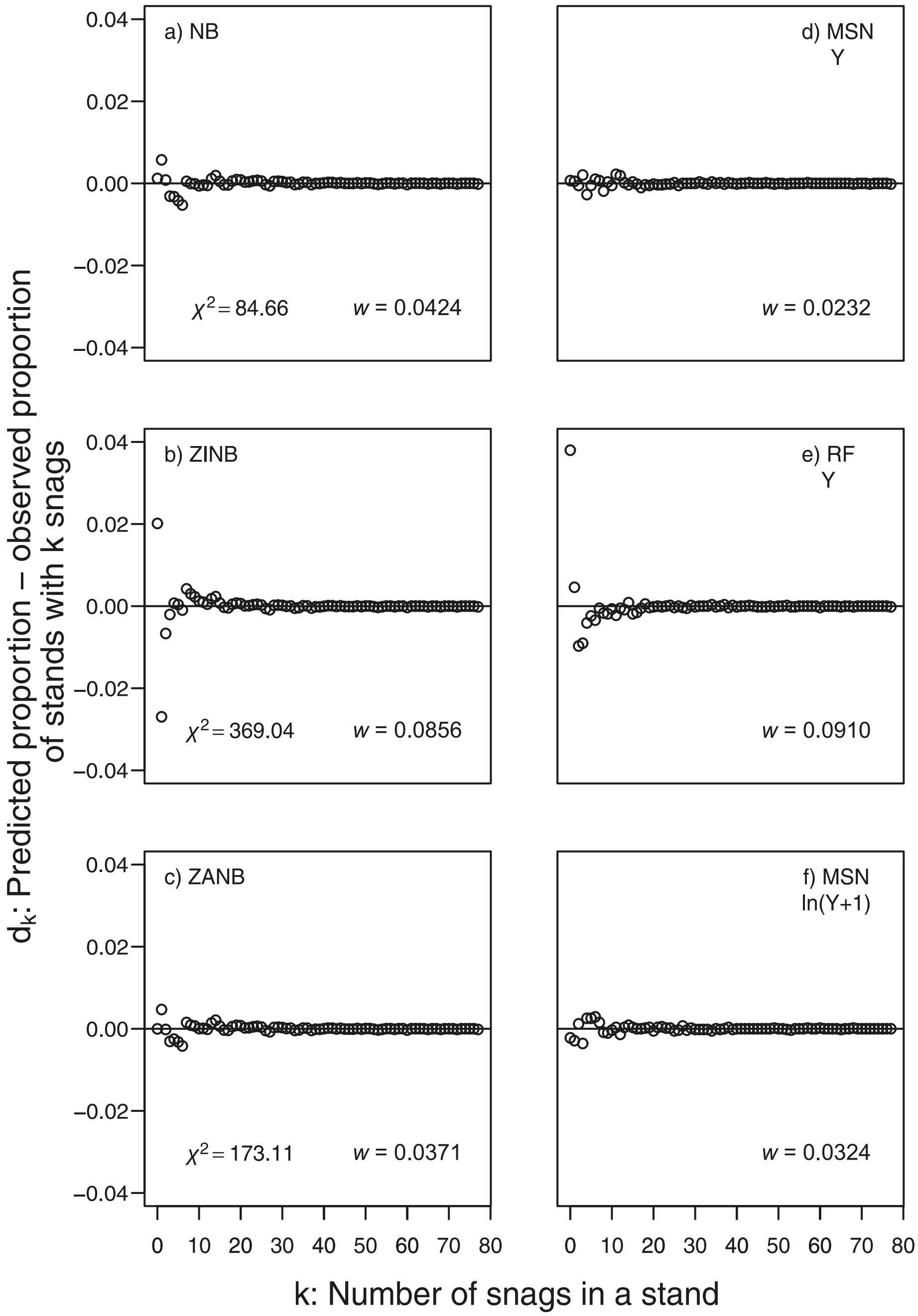
Fig. 4. Diagnostic plots for cavity tree abundance for the negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB), and three nearest neighbor (NN) imputation methods. χ^2 is the χ^2 statistic for the NB, ZINB, and ZANB models. *w* is the sum of the absolute differences (d_k). *Y* indicates that untransformed data were used; ln(*Y* + 1) indicates that the data were log tranformed.



Among the NN imputation methods, RF imputation resulted in the largest d_k values (w = 0.0910), highly overestimating the number of stands with zero and one snag counts. MSN imputation with the untransformed data had the smallest d_k values, resulting in the smallest w value of 0.0232. Both MSN approaches had smaller w values than the three NB regression models (Fig. 5).

The NN imputation methods had larger numbers of prediction errors between –0.5 and 0.5 than the NB regression models for the cavity tree counts. The RF imputation using the untransformed data as *Y* variables had the largest number of prediction errors (3455), and the ZINB model had the smallest number of prediction errors (2798) within this range. The prediction errors of MSN, RF, and logarithmic MSN fell between –13 and 12, –12 and 12, and –11 and 11, respectively, covering almost the whole range of possible values between –13 and 13 and resulting in mean square prediction error (MSPE) values that were about twice as large as the MSPE observed for the NB regression models (Fig. 6). The NB and ZINB models had the smallest MSPE value of 1.35. None of the three NB regression models resulted in overpredictions larger than three counts. However, each model had one underprediction that exceeded 12 counts (Fig. 6).

Fig. 5. Diagnostic plots for snag abundance for the negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB), and three nearest neighbor (NN) imputation methods. χ^2 is the χ^2 statistic for the NB, ZINB, and ZANB models. w is the sum of the absolute differences (d_k). Y indicates that untransformed data were used; $\ln(Y + 1)$ indicates that the data were log tranformed.

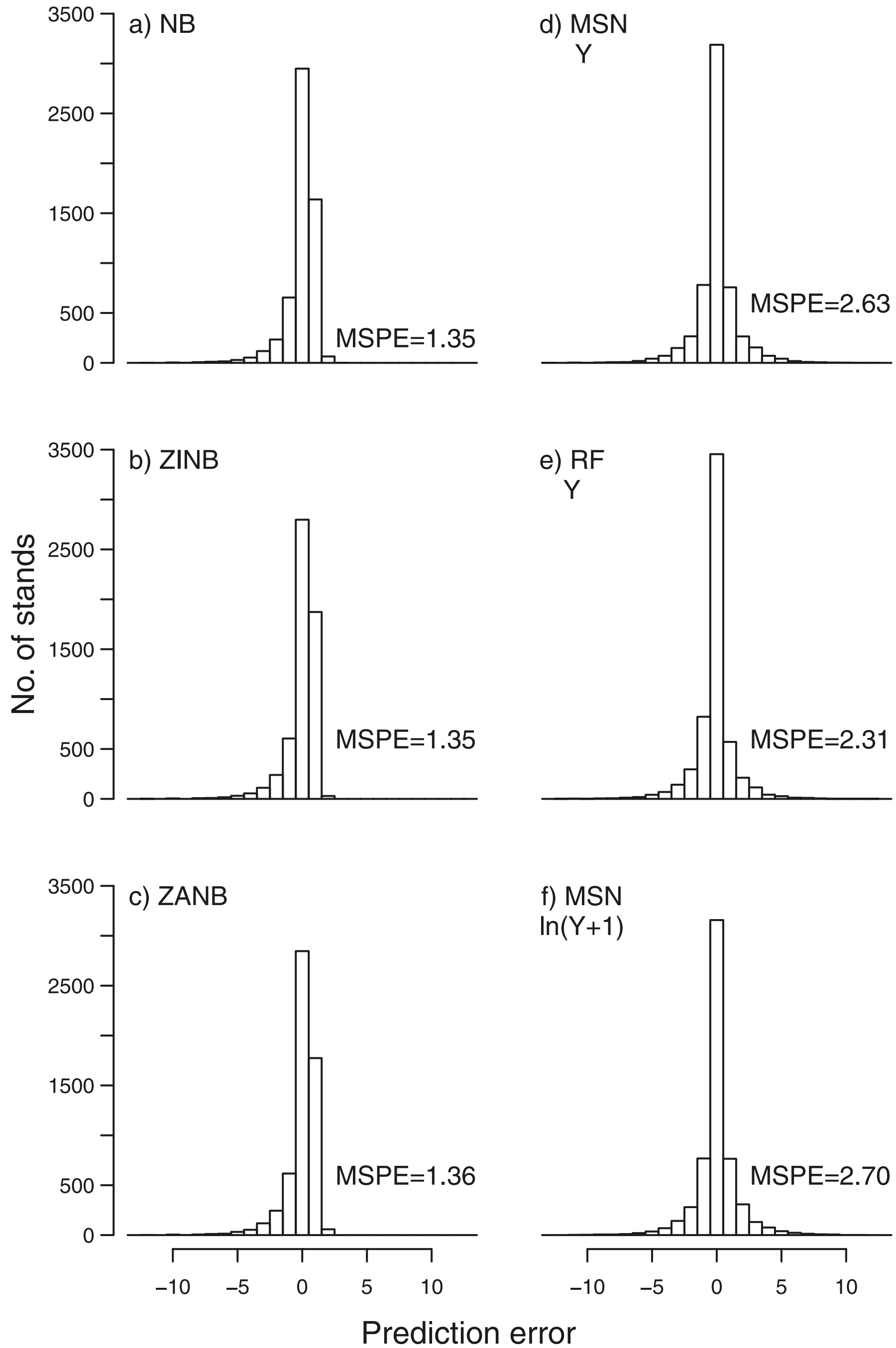


As for the cavity tree abundance on a per-stand level, the NB regression models resulted in smaller MSPE values for the snag abundance with the ZINB model having the smallest MSPE value (28.08) (Fig. 7). None of the NB regression models overpredicted snag counts by more than 17 counts, but for all three models the largest underprediction was around 71 counts. Most of the prediction errors were between 1 and 3 for the NB regression models, whereas most of the prediction errors of the NN imputation methods were between –1 and 1. RF imputation performed best among the NN imputation methods in terms of MSPE (Fig. 7). The prediction errors for the RF imputation ranged between –76 and 56. For the MSN imputation using the untransformed data as Y variables the range was –71 to 62, and for the logarithmic MSN imputation the range was –72 to 76.

Discussion

The results of the NN imputation methods indicate that RF generally performed better than MSN in terms of relative RMSE for predicting cavity tree and snag abundance. The RF method was employed in this study because it produces results that are generally superior to those of other NN imputation methods for predicting basal area and tree density

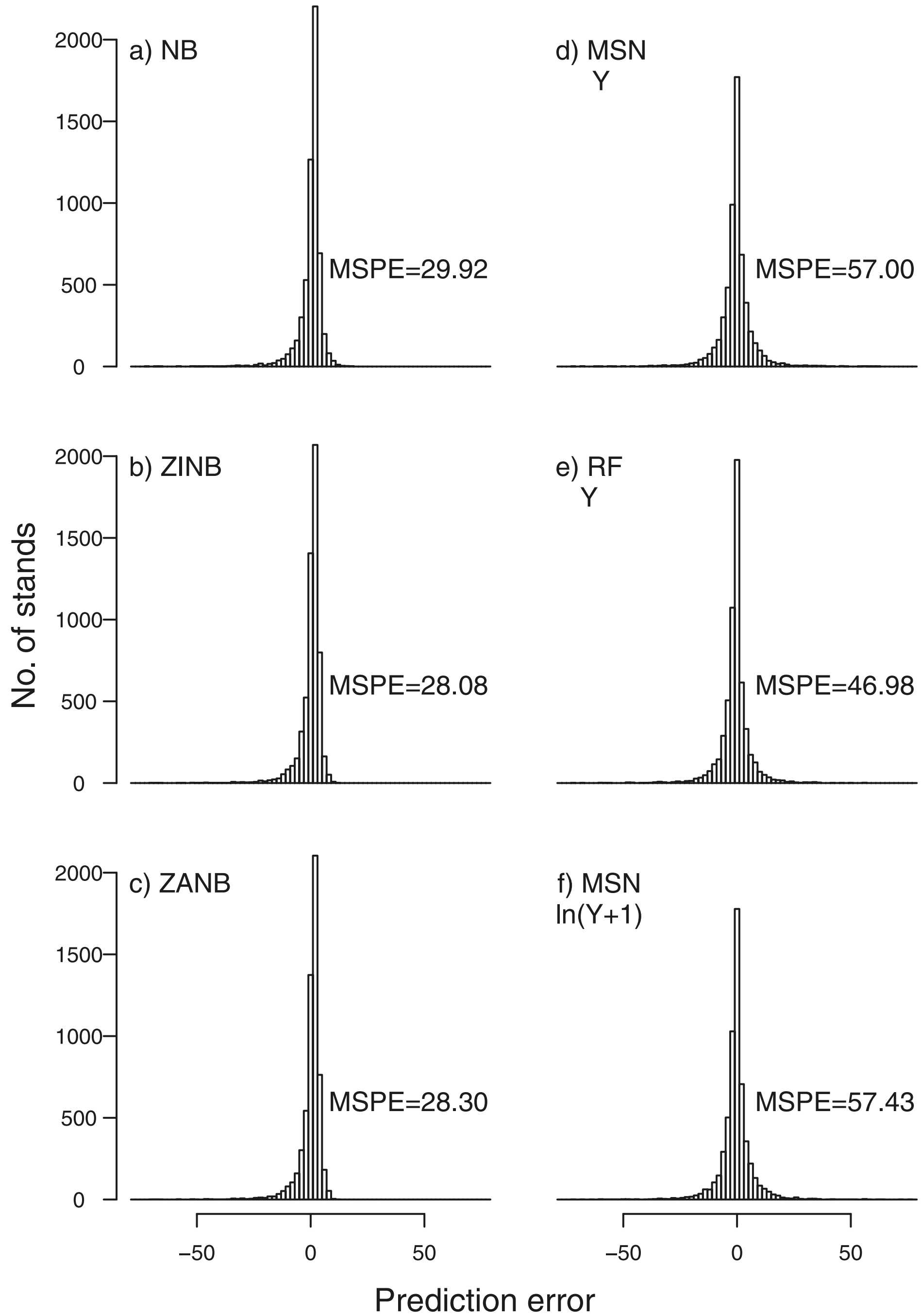
Fig. 6. Frequency plots of prediction errors of cavity tree abundance for the negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB), and three nearest neighbor (NN) imputation methods. MSPE is the mean square prediction error. *Y* indicates that untransformed data were used; ln(*Y* + 1) indicates that the data were log tranformed.



by species (Hudak et al. 2008). The results of this study confirm the conclusion of Hudak et al. (2008) that RF imputation represents an alternative to traditional NN imputation methods.

Transformations on the *Y* variables were tested with the hope of improving the relationships between the *X* and *Y* variable sets. An improved relationship between *X* and *Y* variables could have had a positive impact on the canonical correlation analysis that is used in MSN imputation to determine the weight matrix, **W**. Temesgen et al. (2008) imputed cavity trees per hectare using MSN imputation and found that using the square-root transformation of cavity trees per hectare as the *Y* variable improved the imputation results. The authors considered that the assumption about linear correlations was better met with the square-root transformation (Temesgen et al. 2008). The ranges of cavity tree abundance (0–13) and snag abundance (0–77) were very small in this study because actual tree counts rather than the expanded tree per hectare values were used as *Y* variables. None of the transformations substantially improved the results of the MSN and RF imputation in terms of bias and RMSE, likely because of the small ranges of the *Y* variables.

Fig. 7. Frequency plots of prediction errors of snag abundance for the negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB), and three nearest neighbor (NN) imputation methods. MSPE is the mean square prediction error. *Y* indicates that untransformed data were used; ln(*Y* + 1) indicates that the data were log tranformed.



Poisson, ZIP, and ZAP models were fit to the data but did not provide adequate results because of the overdispersion that is present in the data. The expected values for zero as well as large cavity tree and snag counts were much too low for the Poisson model and were only somewhat improved by the ZIP and ZAP models (results not shown). Hence, this study focused on the application of NB regression models that allow for overdispersion. For predicting overall cavity tree and snag abundance the NB and ZANB models both fit reasonably well. The NB model resulted in the lowest χ^2 goodness-of-fit statistics, whereas the ZANB model resulted in the smallest *w* values. Since many of the observations have zero or less than five counts, the reliability of the results provided by the χ^2 goodness-of-fit statistic is questionable. This dubiety was confirmed by the contradictory results of the χ^2 goodness-of-fit statistic and *w* values for the ZANB model. For predicting cavity tree abundance on a per-stand basis, the NB and ZINB models slightly outperformed the ZANB model in terms of the MSPE value. For predicting snag abundance on a per-stand basis, the ZINB model slightly outperformed the NB and ZANB models in terms of MSPE. However, the differences in the MSPE values among the three models were only minimal. Both the ZINB and ZANB models require the estimation of more parameters than the NB model. Therefore, the NB model should be preferred for predicting cavity tree and snag abundance because it respects the principle of model parsimony. The empirical distributions of cavity tree and snag abundance (Fig. 1) are rather smooth and display a negative exponential trend. There is no clear evidence of a mixture of two distributions (zero point mass and count distribution) in the given case, and the NB regression model seems to be flexible enough to fit the empirical distributions. However, if cavity tree and snag abundance exhibited empirical distributions that clearly showed evidence of two distributions (for an example see Fig. 1 in Cunningham and Lindenmayer (2005)), the ZINB and ZANB models would probably produce better results than the NB model. If mechanisms could be identified that separate the conditions associated with zero cavity trees and snags from conditions associated with positive counts of cavity trees and snags, ZANB models would provide the advantage of modeling these two aspects separately (Welsh et al. 1996). However, in the given case, this property of the ZANB model is an unnecessary complication to the application. In addition, the difficulty of assigning biological meaning to the components of the ZINB and ZANB models raises the issue of overfitting (Affleck 2006) and interpretation of these models.

Cavity trees and snags are rare in forest ecosystems and thus more difficult to predict than many other forest attributes (Temesgen et al. 2008). The prediction of cavity tree and snag abundance is complicated by the fact that generally little association is found with environmental factors and stand-level attributes such as forest type, slope, aspect, and site index (Fan et al. 2003*b*). No strong relationships were found between cavity tree abundance or snag abundance and the continuous variables available in this study (Table 3). This finding is due to the fact that random processes such as fire, wind, and insect outbreaks play a major role in creating snags and cavities, resulting in a large variability in cavity tree and snag abundance (Carey 1983). Hence, it will probably always be difficult to predict cavity tree and snag abundance from stand-level variables that are readily available from forest cover maps or remotely sensed data.

Snag abundance generally increases with successional development (Ohmann and Waddell 2002). Fan et al. (2005) found that mean cavity tree abundance increased with increasing stand-size class, expressed as seedling–sapling, pole, sawtimber, or old growth. In this study, mean stand age was the only explanatory variable that was used to represent stand development stage. Including other variables such as stand-size class in the set of explanatory variables may improve the prediction of cavity tree and snag abundance. Timber harvest and human access can have substantial effects, decreasing snag abundance in areas of intensive timber harvest and increased human access (Wisdom and Bate 2008). Explanatory variables that represent harvest history and degree of human access could potentially improve the prediction of cavity tree and snag abundance.

Formal tests that allow comparing parametric models such as the NB regression models with NN imputation methods do not exist. The diagnostic plots proposed by Lambert (1992) to detect model misspecifications in zero-inflated models as well as the goodness-of-fit statistic *w* introduced by Fortin and DeBlois (2007) provided efficient and convenient ways to show differences between observed and expected counts that could not only be used to detect model misspecification in the NB, ZINB, and ZANB models but also to compare the results of these models with those provided by the NN imputation methods. All NB regression models and NN imputation methods resulted in large differences between observed and expected counts for counts of five and smaller. However, the differences between observed and expected counts decreased faster with increasing counts for the NB regression models than for the NN imputation models. This finding suggests that the NB regression models, in particular the NB model, should be preferred over the NN imputation methods to predict overall cavity tree and snag abundance.

Frequency histograms of the prediction errors were used to visualize and compare the predictions of cavity tree and snag counts per stand of the NB regression models and the NN imputation methods. The NB regression models resulted in a few large underpredictions but no large overpredictions of cavity tree and snag counts. Hence, the NB regression models tend to be more reliable in their predictions of cavity tree and snag abundance than NN imputation methods, which result both in large over- and under-predictions. For management applications that take into account wildlife habitat it is better to base actions on models that are reliable with respect to overpredictions of cavity tree and snag abundance.

All cavity trees and snags were considered equally valuable in this study. Neither cavity size and cavity location on the tree nor decay stage and size of snags, which are important criteria in evaluating habitat quality for certain wildlife species, were taken into account. Hence, no inferences can be made about the quality or potential use of cavity trees and snags for wildlife species, even though the results of this study provided reasonable estimates of cavity tree and snag abundance. For sound forest management purposes, it will be necessary to use methods that allow the estimation of cavity tree and snag abundance while simultaneously providing information on the size and location of cavities as well as the decay stage and size of snags. Because of their multivariate nature, NN imputation methods could be used to simultaneously impute the abundance of cavity trees and snags as well as their quality attributes. Their multivariate nature could be a major advantage of the NN imputation methods over the NB regression models. Other quality attributes that are important for certain wildlife species are tree species, percent bark cover, and presence of a broken top (Spiering and Knight 2005). Information on snag dynamics, such as longevity and the rates at which their quality changes, is also required to be able to fully take snags into account in forest management (Aakala et al. 2008).

If the threshold number of cavity trees and snags necessary for a specific wildlife species to successfully inhabit a stand were known, the analysis could be confined to these conditions. The NB regression models and NN imputation methods provided good estimates for the number of stands with four or more cavity trees and 15 or more snags. If smaller counts than that have minimum utility in wildlife habitat assessment, it would not be necessary to attempt to improve the prediction of low counts.

Conclusions

NB and ZANB models provided reasonable results for predicting overall cavity tree and snag abundance as well as for predicting cavity tree and snag abundance per stand. The ZINB model did not provide very good predictions of overall cavity tree and snag abundance but was comparable to the NB and ZANB models when cavity tree and snag abundance were predicted by stand. In the given case, the NB model should be preferred to the ZINB and ZANB models because it respects the principle of model parsimony and is easier to apply and simpler to interpret. However, ZINB and ZANB models could potentially provide better results than the NB model if the distributions of cavity tree and snag abundance showed evidence of two distributions where some of the processes could be related to the point mass at zero.

MSN imputation outperformed the NB regression models for predicting overall cavity tree and snag abundance, while RF imputation performed by far the worst for this objective.

For predicting cavity tree and snag abundance per stand, NB regression models performed better than NN imputation methods. For this objective, NB regression models should be preferred to NN imputation methods, since they do not result in large overpredictions of the cavity tree and snag counts and hence provide more reliable results.

Acknowledgements

We gratefully acknowledge the support provided by the Forest Inventory and Analysis program, Pacific Northwest Research Station, United States Forest Service. We also thank Dr. Achim Zeileis for his support with the pscl R package, Nick Crookston for his support with the yaImpute package, Dr. Mary Christman for her insights on negative binomial regression models, Dr. Vicente Monleon for his assistance in creating graphs in R, and Dr. Dan Schafer, the Associate Editor, and three anonymous reviewers for their constructive comments.

References

Aakala, T., Kuuluvainen, T., Gauthier, S., and De Grandpré, L. 2008. Standing dead trees and their decay-class dynamics in the northeastern boreal old-growth forests of Quebec. For. Ecol. Manage. **255**: 410–420. doi:10.1016/j.foreco.2007.09.008.

Affleck, D.L.R. 2006. Poisson mixture models for regression analysis of stand-level mortality. Can. J. For. Res. **36**: 2994–3006. doi:10.1139/X06-189.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *In* Second International Symposium on Information Theory. *Edited by* B.N. Petrov and F. Csaki. Akademiai Kaido, Budapest. pp. 267–281.

Akaike, H. 1974. A new look at the statistical identification model. IEEE Trans. Automat. Contr. **19**: 716–723. doi:10.1109/TAC.1974.1100705.

Allen, A.W., and Corn, J.G. 1990. Relationships between live tree diameter and cavity abundance in a Missouri oak–hickory forest. North. J. Appl. For. **7**: 179–183.

Bate, L.J., Garton, E.O., and Wisdom, M.J. 1999. Estimating snag and large tree densities and distributions on a landscape for wildlife management. USDA For. Serv. Gen. Tech. Rep. PNW-GTR-425.

Bechtold, W.A., and Patterson, P.L. (*Editors*). 2005. The enhanced forest inventory and analysis program — national sampling design and estimation procedures. USDA For. Serv. Gen. Tech. Rep. SRS-80.

Bechtold, W.A., and Scott, C.T. 2005. The forest inventory and analysis plot design. *In* The enhanced forest inventory and analysis program — national sampling design and estimation procedures. *Edited by* W.A. Bechtold and P.L. Patterson. USDA For. Serv. Gen. Tech. Rep. SRS-80. pp. 37–52.

Brand, G.J., Nelson, M.D., Wendt, D.G., and Nimerfro, K.K. 2000. The hexagon/panel system for selecting FIA plots under an annual inventory. *In* Proceedings of the First Annual Forest Inventory and Analysis Symposium, 2–3 November 1999, San Antonio, Texas. *Edited by* R.E. McRoberts, G.A. Reams, and P.C. Van Deusen. USDA For. Serv. Gen. Tech. Rep. NC-213. pp. 8–10.

Breiman, L. 2001. Random forests. Mach. Learn. **45**(1): 5–32. doi:10.1023/A:1010933404324.

Cameron, A.C., and Trivedi, P.K. 1998. Regression analysis of count data. Cambridge University Press, Cambridge, UK.

Cameron, A.C., and Trivedi, P.K. 2005. Microeconometrics: methods and applications. Cambridge University Press, Cambridge, UK.

Carey, A.B. 1983. Cavities in trees in hardwood forests. *In* Snag habitat management symposium. USDA For. Serv. Proc. RMRS-P-25. pp. 167–184.

Crookston, N.L., and Finley, A.O. 2008. yaImpute: An R package for kNN Imputation. J. Stat. Softw. **23**(10): 1–16.

Cunningham, R.B., and Lindenmayer, D.B. 2005. Modeling count data of rare species: some statistical issues. Ecology, **86**(5): 1135–1142.

DeLong, S.C., Sutherland, G.D., Daniels, L.D., Heemskerk, B.H., and Storaunet, K.O. 2008. Temporal dynamics of snags and development of snag habitats in wet spruce–fir stands in east-central British Columbia. For. Ecol. Manage. **255**: 3613–3620. doi:10.1016/j.foreco.2008.02.043.

Fan, Z., Larsen, D.R., Shifley, S.R., and Thompson, F.R. 2003*a*. Estimating cavity tree abundance by stand age and basal area, Missouri, USA. For. Ecol. Manage. **179**: 231–242. doi:10.1016/S0378-1127(02)00522-4.

Fan, Z., Shifley, S.R., Spetich, M.A., Thompson, F.R., and Larsen, D.R. 2003*b*. Distribution of cavity trees in Midwestern old-growth and second-growth forests. Can. J. For. Res. **33**: 1481–1494. doi:10.1139/x03-068.

Fan, Z., Shifley, S.R., Thompson, F.R., and Larsen, D.R. 2004. Simulated cavity tree dynamics under alternative timber harvest regimes. For. Ecol. Manage. **193**: 399–412. doi:10.1016/j.foreco.2004.02.008.

Fan, Z., Shifley, S.R., Spetich, M.A., Thompson, F.R., and Larsen, D.R. 2005. Abundance and size distribution of cavity trees in second-growth and old-growth central hardwood forests. North. J. Appl. For. **22**(3): 162–169.

Fortin, M., and DeBlois, J. 2007. Modeling tree recruitment with zero-inflated models: the example of hardwood stands in southern Québec, Canada. For. Sci. **53**(4): 529–539.

Ganey, J.L., and Vojta, S.C. 2005. Changes in snag populations in northern Arizona mixed-conifer and Ponderosa pine forests, 1997–2002. For. Sci. **51**(5): 396–405.

Goodburn, J.M., and Lorimer, C.G. 1998. Cavity trees and coarse woody debris in old-growth and managed northern hardwood forests in Wisconsin and Michigan. Can. J. For. Res. **28**: 427–438. doi:10.1139/cjfr-28-3-427.

Hall, D.B. 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics, **56**: 1030–1039. doi:10.1111/j.0006-341X.2000.01030.x. PMID:11129458.

Harmon, M.E., Franklin, J.F., Swanson, F.J., Sollins, P., Gregory, S.V., Lattin, J.D., Anderson, N.H., Cline, S.P., Aumen, N.G., Sedell, J.R., Lienkaemper, G.W., Cromack, K., and Cummins, K.W. 1986. Ecology of coarse woody debris in temperate ecosystems. Adv. Ecol. Res. **15**: 133–302. doi:10.1016/S0065-2504(08)60121-X.

Hilbe, J.M. 2007. Negative binomial regression. Cambridge University Press, Cambridge, UK.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., and Falkowski, M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sens. Environ. **112**(5): 2232–2245. doi:10.1016/j.rse.2007.10.009. Corrigendum: Remote Sens. Environ. 2009. **113**(1): 289–290.

Jensen, R.G., Kabrick, J.M., and Zenner, E.K. 2002. Tree cavity estimation and verification in the Missouri Ozarks. *In* Proceedings of the Second Missouris Ozark Forest Ecosystem Project Symposium: Post-treatment Results of the Landscape Experiment. *Edited by* S.R. Shifley and J.M. Kabrick. USDA For. Serv. Gen. Tech. Rep. NC-227. pp. 114–129.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, **34**(1): 1–14. doi:10.2307/1269547.

LeMay, V., and Temesgen, H. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. For. Sci. **51**(2): 109–199.

Moeur, M., and Stage, A.R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For. Sci. **41**(2): 337–359.

Nilsson, S.G., Niklasson, M., Hedin, J., Aronsson, G., Gutowski, J.M., Linder, P., Ljungberg, H., Mikusinski, G., and Ranius, T. 2002. Densities of large living and dead trees in old-growth temperate and boreal forests. For. Ecol. Manage. **161**: 189–204. doi:10.1016/S0378-1127(01)00480-7.

Ohmann, J.L., and Waddell, K.L. 2002. Regional patterns of dead wood in forested habitats of Oregon and Washington. USDA For. Serv. Gen. Tech. Rep. PSW-GTR-181.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/ [accessed September 2008].

Rawlings, J.O., Pantula, S.G., and Dickey, D.A. 1998. Applied regression analysis: a research tool. 2nd ed. Springer, New York.

Russell, R.E., Saab, V.A., Dudley, J.G., and Rotella, J.J. 2006. Snag longevity in relation to wildfire and postfire salvage logging. For. Ecol. Manage. **232**: 179–187. doi:10.1016/j.foreco.2006.05.068.

Salas, C., Stage, A.R., and Robinson, A.P. 2008. Modeling effects of overstory density and competing vegetation on tree height growth. For. Sci. **54**(1): 107–122.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. **6**: 461–464. doi:10.1214/aos/1176344136.

Shorohova, E., and Tetioukhin, S. 2004. Natural disturbances and the amount of large trees, deciduous trees and coarse woody debris in the forests of Novgorod region, Russia. Ecol. Bull. **51**: 137–147.

Spiering, D.J., and Knight, R.L. 2005. Snag density and use by cavity-nesting birds in managed stands of the Black Hills National Forest. For. Ecol. Manage. **214**: 40–52. doi:10.1016/j.foreco.2005.03.054.

Sustainable Forestry Initiative. 2002. The Sustainable Forestry Initiative Standard, 2002–2004 SFI Standard. American Forest and Paper Association. Washington, D.C.

Temesgen, H., Barrett, T.M., and Latta, G. 2008. Estimating cavity tree abundance using nearest neighbor imputation methods for western Oregon and Washington forests. Silva Fenn. **42**(3): 337–354.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F., and Lindenmayer, D.B. 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol. Model. **88**: 297–308. doi:10.1016/0304-3800(95)00113-1.

Wisdom, M.J., and Bate, L.J. 2008. Snag density varies with intensity of timber harvest and human access. For. Ecol. Manage. **255**: 2085–2093. doi:10.1016/j.foreco.2007.12.027.

Zeileis, A., Kleiber, C., and Jackman, S. 2008. Regression models for count data in R. J. Stat. Softw. **27**(8): 1–25.

Appendix A

Table A1. Summary of fitted negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB) regression models for cavity tree abundance: coefficient estimates from count and zero models with standard errors in parentheses and significance levels (****, $p < 0.0001$; ***, $p < 0.001$; **, $p < 0.01$; *, $p < 0.05$).

	NB		ZINB		ZANB	
	Coefficient estimate	Significance	Coefficient estimate	Significance	Coefficient estimate	Significance
Count model						
Intercept	–2.0890 (–0.2635)	****	0.0749 (0.2011)		–1.0361 (0.3945)	***
Mean stand age (years)	–0.0048 (0.0010)	****				
% conifer	1.6130 (0.4744)	****	–0.6946 (0.1539)	****		
Height class midpoint (m)	0.0574 (0.0091)	****	0.0183 (0.0031)	****	0.0141 (0.0042)	****
Elevation (m)	–0.0002 (0.0001)	****	–0.0001 (<0.0001)	**	–0.0002 (0.0001)	**
Slope (%/100)	–0.5669 (0.3675)		0.2485 (0.1388)	*	0.4713 (0.1941)	**

Table A1 (*concluded*).

	NB		ZINB		ZANB	
	Coefficient estimate	Significance	Coefficient estimate	Significance	Coefficient estimate	Significance
Aspect (°)	–0.0100 (0.0131)					
Slope × cosine(aspect)	0.1457 (0.0850)	*				
Slope × ln(elevation)	0.1120 (0.0536)	**				
Forest type: fir – spruce – mountain hemlock	0.2112 (0.0970)	**	0.1329 (0.1020)		0.1062 (0.1523)	
Forest type: other conifers	–0.1752 (0.0746)	**	–0.1341 (0.0812)	*	–0.0488 (0.1223)	
Forest type: hardwoods	–0.2197 (0.1108)	**	0.0421 (0.1271)		0.5291 (0.1340)	****
Owner: other federal	–0.3578 (0.0940)	****	–0.3802 (0.0924)	****	–0.1994 (0.1318)	
Owner: state and local government	–0.09505 (0.1107)		–0.1745 (0.1079)		–0.2592 (0.1745)	
Owner: private	–0.1874 (0.0690)	***	–0.1992 (0.0688)	***	–0.2540 (0.1135)	**
Site class midpoint ($m^3 \cdot ha^{-1} \cdot year^{-1}$)	0.1261 (0.0349)	****				
(Mean stand age)^2	0.0001 (<0.0001)	****				
(% conifer)^2	–1.4630 (0.3752)	****				
(Height class midpoint)^2	–0.0005 (0.0002)	****				
(Site class midpoint)^2	–0.0036 (0.0016)	**				
% conifer × site class midpoint	–0.0690 (0.0176)	****				
Zero model						
Intercept			1.9750 (0.3528)	****	–1.3910 (0.1556)	****
Mean stand age			–0.0141 (0.0326)	****	0.0010 (0.0004)	**
% conifer					–0.4588 (0.1428)	***
Height class midpoint (m)			–0.1132 (0.0183)	****	0.0450 (0.0032)	****
Slope (%/100)			–0.0313 (0.5501)			
Aspect (°)			–0.0141 (0.0495)			
Slope × cosine(aspect)			–0.8520 (0.3620)	**		
Forest type: fir – spruce – mountain hemlock			–0.4665 (0.4506)		0.2280 (0.1052)	**
Forest type: other conifers			0.3420 (0.2791)		–0.2235 (0.0815)	***
Forest type: hardwoods			–1.0973 (0.3347)	***	0.3999 (0.1200)	****
Owner: other federal					–0.4898 (0.1076)	****
Owner: state and local government					–0.0213 (0.1252)	
Owner: private					–0.1388 (0.0725)	*
No. of estimated parameters	22		21		21	
AIC	11 450		11 413		11 463	
BIC	11 597		11 553		11 604	

Note: AIC, Akaike's information criterion; BIC, Schwarz's Bayesian information criterion.

Table A2. Summary of fitted negative binomial (NB), zero-inflated NB (ZINB), zero-altered NB (ZANB) regression models for snag abundance: coefficient estimates from count and zero models with standard errors in parentheses and significance levels (****, $p < 0.0001$; ***, $p < 0.001$; **, $p < 0.01$; *, $p < 0.05$).

	NB		ZINB		ZANB	
	Coefficient estimate	Significance	Coefficient estimate	Significance	Coefficient estimate	Significance
Count model						
Intercept	1.2050 (0.1207)	****	1.9224 (0.1390)	****	1.8150 (0.1601)	****
Mean stand age (years)	–0.0019 (0.0003)	****	–0.0058 (0.0009)	****	–0.0077 (0.0012)	****
% conifer	–0.4704 (0.0862)	****	–0.7467 (0.1202)	****	–0.8670 (0.1423)	****
Height class midpoint (m)	0.0356 (0.0024)	****	–0.0029 (0.0072)		–0.0035 (0.0088)	
Elevation (m)	0.0001 (<0.0001)	****			0.0001 (<0.0001)	
Slope (%/100)	0.0872 (0.0828)		–0.0102 (0.8278)		0.0756 (0.3434)	
Aspect (°)	0.0219 (0.0092)	**	0.0254 (0.0090)	***	0.0204 (0.0111)	*
Slope × cosine(aspect)	0.1531 (0.0624)	**	0.1422 (0.0585)	**		
Slope × ln(elevation)					–4.7350 (2.2280)	**
Slope × ln($elevation^2$)					2.3670 (1.1130)	**
Forest type: fir – spruce – mountain hemlock	0.4957 (0.0676)	****	0.4312 (0.0598)	****	0.4157 (0.0790)	****
Forest type: other conifers	–0.2019 (0.0523)	****	–0.1383 (0.0460)	***	–0.0854 (0.0600)	
Forest type: hardwoods	–0.0634 (0.0737)		0.0329 (0.0813)		0.0183 (0.1012)	
Owner: other federal	–0.4395 (0.0609)	****	–0.2513 (0.0620)	****	–0.2715 (0.0770)	****
Owner: state and local government	–0.3099 (0.0790)	****	–0.2174 (0.0747)	***	–0.2093 (0.0935)	**
Owner: private	–0.6858 (0.0471)	****	–0.5988 (0.0456)	****	–0.6286 (0.0599)	****
Site class midpoint ($m^3 \cdot ha^{-1} \cdot year^{-1}$)	–0.0137 (0.0055)	**	–0.0026 (0.0056)			
Mean stand age × % conifer			0.0041 (0.0010)	****	0.0061 (0.0013)	****
% conifer × height class midpoint			0.0250 (0.0077)	***	0.0254 (0.0690)	***
Zero model						
Intercept			0.4584 (0.4342)		0.0938 (0.1900)	
Mean stand age (years)			0.0190 (0.0035)	****		
% conifer			–0.0097 (0.4706)		–0.3350 (0.1330)	**
Height class midpoint (m)			–0.1052 (0.0389)	***	0.0737 (0.0040)	****
Elevation (m)			–0.0003 (0.0001)	**	0.0001 (0.0001)	**
Slope (%/100)			–1.2142 (0.3354)	****	0.4784 (0.1449)	****
Aspect (°)			0.0873 (0.0337)	***	0.0020 (0.0159)	
Slope × cosine(aspect)					0.2885 (0.1133)	**
Forest type: fir – spruce – mountain hemlock					0.6048 (0.1385)	****

Table A2 (*concluded*).

	NB		ZINB		ZANB	
	Coefficient estimate	Significance	Coefficient estimate	Significance	Coefficient estimate	Significance
Forest type: other conifers					–0.3117 (0.0927)	****
Forest type: hardwoods					0.0921 (0.1202)	
Owner: other federal			0.3730 (0.2091)	*	–0.3909 (0.1028)	****
Owner: state and local government			–0.0243 (0.3440)		–0.1261 (0.1452)	
Owner: private			0.0104 (0.1641)		–0.4413 (0.0860)	****
Site class midpoint ($m^3 \cdot ha^{-1} \cdot year^{-1}$)			–0.1092 (0.0370)	***	–0.0453 (0.0087)	****
Mean stand age × % conifer			–0.0125 (0.0050)	**		
% conifer × height class midpoint			–0.0029 (0.0004)	****		
% conifer × site class midpoint			0.1430 (0.0438)	***		
Height class midpoint × site class midpoint			0.0143 (0.0038)	****		
No. of estimated parameters	16		32		32	
AIC	26 205		25 726		25 632	
BIC	26 311		25 940		25 846	

Note: AIC, Akaike's information criterion; BIC, Schwarz's Bayesian information criterion.