Forest Ecology and Management 272 (2012) 26–34

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

A comparison of selected parametric and imputation methods for estimating snag density and snag quality attributes

B.N.I. Eskelson [a,*], H. Temesgen [a,1], J.C. Hagar [b,2]

[a] Oregon State University, Department of Forest Engineering, Resources and Management, 204 Peavy Hall, Corvallis, OR 97331-5706, United States
[b] US Geological Survey, Forest and Rangeland Ecosystem Science Center, Forestry Sciences Lab, 3200 S.W. Jefferson Way, Corvallis, OR 97331, United States

ARTICLE INFO

Article history:
Available online 3 August 2011

Keywords:
Snag density
Snag decay class
Snag size class
Nearest neighbor imputation
Ordinal regression

ABSTRACT

Snags (standing dead trees) are an essential structural component of forests. Because wildlife use of snags depends on size and decay stage, snag density estimation without any information about snag quality attributes is of little value for wildlife management decision makers. Little work has been done to develop models that allow multivariate estimation of snag density by snag quality class. Using climate, topography, Landsat TM data, stand age and forest type collected for 2356 forested Forest Inventory and Analysis plots in western Washington and western Oregon, we evaluated two multivariate techniques for their abilities to estimate density of snags by three decay classes. The density of live trees and snags in three decay classes (D1: recently dead, little decay; D2: decay, without top, some branches and bark missing; D3: extensive decay, missing bark and most branches) with diameter at breast height (DBH) $\geqslant$ 12.7 cm was estimated using a nonparametric random forest nearest neighbor imputation technique (RF) and a parametric two-stage model (QPORD), for which the number of trees per hectare was estimated with a Quasipoisson model in the first stage and the probability of belonging to a tree status class (live, D1, D2, D3) was estimated with an ordinal regression model in the second stage. The presence of large snags with DBH $\geqslant$ 50 cm was predicted using a logistic regression and RF imputation. Because of the more homogenous conditions on private forest lands, snag density by decay class was predicted with higher accuracies on private forest lands than on public lands, while presence of large snags was more accurately predicted on public lands, owing to the higher prevalence of large snags on public lands. RF outperformed the QPORD model in terms of percent accurate predictions, while QPORD provided smaller root mean square errors in predicting snag density by decay class. The logistic regression model achieved more accurate presence/absence classification of large snags than the RF imputation approach. Adjusting the decision threshold to account for unequal size for presence and absence classes is more straightforward for the logistic regression than for the RF imputation approach. Overall, model accuracies were poor in this study, which can be attributed to the poor predictive quality of the explanatory variables and the large range of forest types and geographic conditions observed in the data.

1. Introduction

Detailed information about the density of standing dead trees (snags) and their quality attributes (e.g., size, decay class) are not only important for carbon storage and fire effects but also essential for managing biodiversity and wildlife habitat. Many wildlife species depend on snags occurring across the landscape in a variety of sizes and decay classes for nesting, roosting, denning, foraging, or shelter (Bull et al., 1997). The potential use of snags by wildlife species depends on snag characteristics such as tree species, diameter, height, decay stage, and proximity to other snags and live trees (Bull et al., 1997).

Size and decay stage are two prominent characteristics of dead wood that influence habitat suitability for individual wildlife species. Many species prefer large snags with diameter at breast height (DBH) greater than 50 cm for nesting, roosting, and foraging (Marcot et al., 2010; Bull, 2002). The number of species provided with suitable habitat, reproductive output of cavity-nesters, security of nest and roost sites from predators and weather, and the longevity of available habitat all increase with snag size (Hagar, 2007). In addition to snag size, decay stage is considered one of the most important attributes that affects snag use by wildlife species associated with dead wood (Vaillancourt et al., 2008) and vertebrates in general (Harmon et al., 1986). The decay stage of

* Corresponding author. Tel.: +1 503 808 2094; fax: +1 503 808 2020.
E-mail addresses: bianca.eskelson@oregonstate.edu (B.N.I. Eskelson), hailemariam.temesgen@oregonstate.edu (H. Temesgen), joan_hagar@usgs.gov (J.C. Hagar).
[1] Tel.: +1 541 737 8549; fax: +1 541 737 4316.
[2] Tel.: +1 541 758 8815; fax: +1 541 758 8806.

0378-1127/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.foreco.2011.06.041

snags influences the ability of cavity-using species to excavate nest sites (Lundquist and Mariani, 1991), and the extent of colonization by arthropods, an important food resource for many species of vertebrate wildlife.

Forest management effects on snag density have been well documented (Graves et al., 2000). Snag density tends to be low in areas of intensive timber harvest and increased human access (Wisdom and Bate, 2008), while large snags are more abundant in unharvested stands (Marcot et al., 2010). Snag density also differs by ownership, ecoregion, and forest type. For Coastal Oregon, Kennedy et al. (2008) and Ohmann et al. (2007) reported more snags on public lands than on private lands. In Oregon and Washington, Ohmann and Waddell (2002) found lowest snag densities in dry habitats east of the Cascade crest of Oregon and Washington and greatest snag densities at high elevations.

To successfully conserve forest structure and biodiversity and manage forests for broad ecological goals including wildlife, information about snag density along with snag quality attributes is necessary. While research has focused on modeling snag density ignoring snag quality attributes (e.g., Frescino et al., 2001; Eskelson et al., 2009a; Pierce et al., 2009), few studies have attempted to model snag density by decay class (Bater et al., 2009) or snag presence by size class (Martinuzzi et al., 2009).

The aim of this study was to (1) apply nearest neighbor imputation and a two-stage parametric model to predict snag density by decay class; (2) predict the presence/absence of large snags (DBH $\geqslant$ 50 cm) in western Oregon and Washington with logistic regression and nearest neighbor imputation; and (3) evaluate the performance of the different approaches on private and public lands in western Washington and western Oregon.

2. Material and methods

2.1. Data

Data from forested Forest Inventory and Analysis (FIA) plots, located in western Oregon (OR) and western Washington (WA) and collected between 2001 and 2008, were used in this study. A detailed description of the FIA inventory data is available in Bechtold and Patterson (2005). Only plots (n = 2356) for which all four subplots belonged to the same condition class were used in this study, which allows using plot-level variables as explanatory variables. The number of live trees per hectare (LTPH) and number of dead trees per hectare (DTPH) by three decay classes (DTPH.D1, DTPH.D2, DTPH.D3) were used as response variables (Table 1). These variables were based on trees with DBH greater or equal to 12.7 cm. Snags that had recently died, showed little decay, and retained bark, branches, and top belonged to decay class 1 (D1). Snags that showed evidence of decay and had lost some bark, branches, and sometimes the top belonged to decay class 2 (D2), and snags with a broken top and extensive decay with missing bark and most branches belonged to decay class 3 (D3).

Table 2
Number of plots for which snags with DBH $\geqslant$ 50 cm are present and absent.

	Presence	Absence
WA PUB	308 (61%)	199
WA PRI	175 (42%)	240
OR PUB	536 (61%)	339
OR PRI	175 (31%)	384

The FIA plots were grouped by state and ownership (Table 1). The percentage of plots without snags was higher on private lands (PRI) than on public lands (PUB) for both OR and WA. The range and variability of the response variables was larger on private lands in WA than on private lands in OR, while the range and variability of the response variables on OR public lands exceeded that of WA public lands (Table 1).

Large snags (DBH $\geqslant$ 50 cm) were present on 61% of the public land plots in WA and OR, while large snags only occurred on 42% and 31% of the private land plots in WA and OR, respectively (Table 2). The range and variability of large snags was higher on public lands than on private lands (Table 1).

Climate, remote sensing, and topography data as well as stand age (STDAGE) and three broad forest type groups (Douglas-fir forests, other conifer forests, and hardwood forests) were available as explanatory variables (for details see Table 3). The range and variability of STDAGE was larger on public lands than on private lands (Table 4). Private lands were typically located at lower elevation than public lands (Table 4). Mean % canopy cover (CANOPY) was lower on private lands and more variable than on public lands (Table 4). Climate data were derived from PRISM (http://www.prism.oregonstate.edu) and were interpolated from 800 m normals (1971–2000) to 30 m resolution. The values of the climate, remote sensing, and topography data for the FIA plots were determined by overlaying each FIA plot with nine 30 m pixels, with the center plot falling in the center pixel, and calculating a mean of the nine pixel values.

2.2. Snag density by decay stage

Snag density by decay class was estimated by the non-parametric nearest neighbor (NN) imputation method randomForest (RF) (Crookston and Finley, 2008) and a parametric two-stage model (QPORD). For QPORD the total number of trees per hectare (TPH) was estimated with a Quasipoisson regression model in the first stage and the probability of belonging to either live trees or dead trees in decay classes one through three was estimated with an ordinal regression model (ORD) in the second stage. Because of differences in snag density on public and private lands (Kennedy et al., 2008) and its more accurate prediction in coastal OR than in northeastern WA (Pierce et al., 2009), the models were developed by ownership group (PUB vs. PRI) within state (WA vs. OR). Separate models that provide accurate model predictions for each state/ownership combination will be useful for forest managers.

Table 1
Summary of response variables for public and private forest lands in Washington and Oregon.

	WA PUB (n = 507)			WA PRI (n = 415)			OR PUB (n = 875)			OR PRI (n = 559)		
	Range	Median	%[a]	Range	Median	%	Range	Median	%	Range	Median	%
LTPH	0–1903	416	1	0–1770	449	7	0–1442	349	3	0–1338	389	11
DTPH.D1	0–476	0	60	0–877	0	72	0–684	0	63	0–387	0	71
DTPH.D2	0–372	30	20	0–223	2	48	0–654	17	26	0–268	0	50
DTPH.D3	0–166	5	41	0–104	0	53	0–456	2	48	0–109	0	65
Large snags	0–99	5	39	0–52	0	58	0–104	2	39	0–44	0	69

Note: WA = Washington; OR = Oregon; PUB = public land; PRI = private land; LTPH = live trees/ha; DTPH.D1 = snags/ha in decay class 1; DTPH.D2 = snags/ha in decay class 2; DTPH.D3 = snags/ha in decay class 3; Large snags = snags/ha with DBH $\geqslant$ 50 cm; TPH variables include trees with DBH $\geqslant$ 12.7 cm.

[a] Percent plots with zero observation.

Table 3
Explanatory variables used to model response variables.

Variable	Description
Climate[a]	
ANNPRE	Annual precipitation (natural logarithm, mm)
CONTPRE	Percentage of annual precipitation falling in June–August
SMRPRE	Mean precipitation (natural logarithm, mm) in May–September
CVPRE	Coefficient of variation of mean monthly precipitation of December and July (wettest and driest month)
ANNTMP	Mean annual temperature (°C)
AUGMAXT	Mean maximum temperature in August (°C) (hottest month)
DECMINT	Mean minimum temperature in December (°C) (coldest month)
DIFTMP	AUGMAXT–DECMINT (°C)
SMRTMP	Mean temperature (°C) in May–September
Remote sensing[b]	
TM*x*	Landsat Thematic Mapper™ band, where *x* = 1–5 and 7
R43	Ratio of TM4 to TM3
R54	Ratio of TM5 to TM4
R57	Ratio of TM5 to TM7
BRT	Brightness axis from tasselled cap transformation
GRN	Greenness axis from tasselled cap transformation
WET	Wetness axis from tasselled cap transformation
CANOPY	% canopy cover from NLCD 2001 (Homer et al., 2004)
Topography	
SLOPE	Slope (%), from digital elevation model
ASPECT	Cosine transformation of aspect (°) (Beers et al., 1966)
PRR	Potential relative radiation
STDAGE	Stand age
ForTyp1	Indicator variable Douglas-fir forests
ForTyp2	Indicator variable other conifer forests

[a] Climate data were derived from PRISM (http://www.prism.oregonstate.edu) and were interpolated from 800 m normals (1971–2000).
[b] Remote sensing data were obtained by RSAC (http://www.fs.fed.us/eng/rsac) from GloVis (http://glovis.usgs.gov), except where noted, and a seamless mosaic was compiled for western Oregon and Washington for the year 2006.

For NN approaches the imputed response is a value that was observed for another unit (=reference unit) that is similar to the unit for which the response is imputed (=target unit). Similarity of the units is determined by the explanatory variables. Instead of using a single value of the nearest neighbor, a simple or weighted average of several nearest neighbors is sometimes imputed (e.g., see LeMay and Temesgen, 2005). For a more detailed description of NN imputation methods see Eskelson et al. (2009b). The RF imputation approach employed in this study was implemented in the yaImpute R package (Crookston and Finley, 2008). RF is based on a classification and regression tree method (Breiman, 2001), for which the data and variables are randomly and iteratively sampled to generate a large group of classification and regression trees. Two units are considered similar if they tend to end up in the same terminal nodes in a group of classification and regression trees. The measure of similarity equals one minus the proportion of classification and regression trees where the target unit is in the same terminal node as a reference unit (Crookston and Finley, 2008). The response variables LTPH, DTPH.D1, DTPH.D2, and DTPH.D3 were imputed simultaneously.

We used a two-stage method (QPORD) for estimating the density of snags by decay stage. In the first stage, we estimated TPH with a Quasipoisson regression model (QP). The QP model is a generalized linear model (GLM) with Poisson-like assumptions that accounts for overdispersion in the data and uses a quasi-likelihood approach (for details see McCullagh and Nelder, 1989, pp. 200–204, p. 323 ff.). The QP model was chosen because overdispersion and strong positive skew were observed in the response. The mean $\mu_i = E(Y_i)$ for the ith observation is related to p explanatory variables through a monotonic, differentiable log link function so that:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p1} x_{pi}) \quad (1)$$

This can be generalized as the vector of mean parameters $\boldsymbol{\mu} = g^{-1}\boldsymbol{\beta}^T\boldsymbol{X}$, where $g^{-1}(.)$ is the exponential function, $\boldsymbol{X}$ is a design matrix of explanatory variables, and $\boldsymbol{\beta}$ is a vector of parameters (regression coefficients). Alternatively, this can be written as $g(\boldsymbol{\mu}) = \boldsymbol{\beta}^T\boldsymbol{X}$, where g(.) is the log link function. The QP models were fit in R (R Development Core Team, 2010) using forward variable selection.

TPH is equal to the sum of all live and dead trees (TPH = LTPH + DTPH.D1 + DTPH.D2 + DTPH.D3). Since LTPH, DTPH.D1, DTPH.D2, and DTPH.D3 can be considered ordinal data with the first class representing live trees and the following three classes representing dead trees with increasing amounts of decay, ordinal regression can be employed in the second stage of the QPORD model to estimate cumulative proportions of the $k = 4$ tree status classes (live, D1, D2, D3) in each FIA plot using the available explanatory variables. We used a proportional odds model with log–log function to predict the cumulative proportions of live trees ($P(k = 1|\mathbf{x}) = \gamma_1 = \pi_1$), live trees and dead trees in D1 combined ($P(k \leqslant 2|\mathbf{x}) = \gamma_2 = \pi_1 + \pi_2$), live trees and dead trees in D1 and D2 combined ($P(k \leqslant 3|\mathbf{x}) = \gamma_3 = \pi_1 + \pi_2 + \pi_3$), and all live and dead trees combined ($P(k \leqslant 4|\mathbf{x}) = \gamma_4 = 1$), using the following equation:

$$\gamma_{ki} = 1 - \exp\left[-\exp(\theta_k - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi}))\right] \quad (2)$$

where γ_{ki} is the cumulative probability of tree status class k occurring on plot i, θ_k is the intercept for class k, and x_{ji} is the value of the jth explanatory variable on plot i and the β_j's are the corresponding p parameter coefficients, respectively (McCullagh and Nelder, 1989, pp. 149–155). The probability π_{ki} of tree status class k occurring on plot i can then be calculated as $\pi_{ki} = \gamma_{ki} - \gamma_{k-1,i}$. Ordinal regression was implemented with the R package 'ordinal' (Christensen, 2010) using forward variable selection in combination with a likelihood ratio test to assess a model's improvement when an explanatory variable was added.

For each tree status class k, TPH could then be estimated for plot i by multiplying μ_i (Eq. (1)) and π_{ki} (Eq. (2)) so that:

$$\text{LTPH}_i = \mu_i * \pi_{1i}$$

$$\text{DTPH}.D1_i = \mu_i * \pi_{2i}$$

$$\text{DTPH}.D2_i = \mu_i * \pi_{3i}$$

$$\text{DTPH}.D3_i = \mu_i * \pi_{4i}$$

2.3. Presence of large snags

The presence of large snags was estimated using the RF imputation approach as well as a logistic regression model. The conditional

Table 4
Summary of three explanatory variables.

	WA PUB (n = 507)			WA PRI (n = 415)			OR PUB (n = 875)			OR PRI (n = 559)		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
STDAGE	130	117	0–500	32	23	0–150	120	108	0–610	36	25	0–150
Elevation (m)	745	449	22–1905	288	290	4–1385	832	435	8–2172	452	302	10–1648
CANOPY	83	23	0–98	73	30	0–100	76	22	0–97	65	33	0–98

Note: WA = Washington, OR = Oregon, PUB = public land, PRI = private land, STDAGE = stand age, DEM = elevation, CANOPY = % canopy cover, SD = standard deviation.

probability that large snags are present on plot i is denoted by $P(Y = 1|\mathbf{x}) = \pi_i$, and the logistic regression model is:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})} \quad (3)$$

where x_{ji} is the value of the jth explanatory variable on plot i and the β_j's with $j = 1, \ldots, p$ are the corresponding p parameter coefficients (Hosmer and Lemeshow, 2000, p. 32). Logistic regression models were fit separately for private and public lands in WA and OR using backwards variable selection based on Akaike's Information Criterion (AIC) in R (R Development Core Team, 2010). In order to classify a predicted probability (continuous between 0 and 1) as either absence or presence of large snags, a decision threshold τ needs to be chosen. The decision threshold τ was chosen to reflect the prevalence of large snags (proportion of plots in which large snags are present), which results in balancing the values of sensitivity (proportion of correctly classified presences) and specificity (proportion of correctly classified absences) (Chen et al., 2006). Hence, $\tau = 0.61$ for public lands in both WA and OR, and $\tau = 0.42$ and 0.31 for private lands in WA and OR, respectively (see Table 2).

The RF imputation method (Crookston and Finley, 2008) was employed to impute a single response: absence or presence of large snags. NN methods such as RF, when used to produce a binary outcome of either 0 or 1 as in the given case, cannot be used in combination with a threshold adjustment that takes prevalence of large snags into account (Chen et al., 2006). In order to apply an adjusted decision threshold τ to the RF imputation approach, we used a simple average of the values observed for the five nearest neighbors to determine presence or absence of large snags. The mean of the five nearest neighbors could result in imputed probabilities of snag presence equal to {0, 0.2, 0.4, 0.6, 0.8, 1}. If the imputed probability of snag presence was greater or equal to τ (for threshold values see Table 2 as described above), presence of large snags was predicted.

2.4. Model validation

The predictive performance of the RF and QPORD models was assessed by using leave-one-out cross-validation. For RF, one observation at a time was used as target unit while the remaining n-1 observations were used as reference units. For the parametric models, one observation at a time was used for the validation process based on the models fitted with the remaining n-1 observations. Based on the cross-validation results, the Spearman correlation coefficient between predicted and observed values and the root mean square error (RMSE) were calculated:

$$\mathrm{RMSE} = \sqrt{\sum_n (\mathrm{predicted} - \mathrm{observed})^2 / n} \quad (4)$$

where n is the sample size. Additionally, scatterplots of observed vs. predicted values were generated to visually show the distribution error (as in Frescino et al., 2001). Predicted values within 30% of the observed value were considered accurate and the percentage of accurate predictions (% accurate predictions) falling within ±30% of the observed values was reported in the scatterplots. The level of ±30% was chosen somewhat arbitrarily for this study and needs to be adjusted based on specific management objectives on a case-by-case basis.

The accuracy of predicted presence/absence of large snags was assessed with confusion matrices. The confusion matrices were used to calculate overall accuracy, sensitivity, specificity, and the kappa statistic (κ) (see e.g., Allouche et al., 2006). κ is a measure of improvement of the model over random predictions (Cohen, 1960).

3. Results

3.1. Snag density by decay stage

For both states and both ownership types, CANOPY, STDAGE, TM7, WET and the forest type indicator variables were significant explanatory variables in the QP model for estimating TPH. Other significant variables that were selected in three of the four QP models were R54, R57, and TM5. For the OR private land model, no climate variables were significant, while the other three models included both temperature and precipitation variables among the significant explanatory variables (Table 5).

Explanatory variables that were significant for the four ORD models that estimated the proportion of live trees and dead trees by decay stage were STDAGE, R43, R54 and the forest type indicator variables. BRT, GRN, and TM5 were included in three of the four ORD models. All ORD models included at least one climate variable among the explanatory variables (Table 5).

QPORD models outperformed the RF imputation approach in terms of RMSE for all response variables on public and private lands in both OR and WA (Table 6). QPORD resulted in larger Spearman correlation coefficients than RF for LTPH in both states on private and public lands as well as for DTPH in all three decay classes on private lands in WA and OR. For DTPH across the three decay classes RF outperformed QPORD in terms of the correlation coefficient on public lands in WA and OR (Table 7).

The scatterplots of observed versus predicted values depicted the poor fit of both the QPORD models and the RF imputation approach. QPORD models overpredicted LTPH, DTPH.D1, DTPH.D2, and DTPH.D3 for small observed values and highly underpredicted large observed values (Figs. 1 and 2). On the contrary, the RF imputation approach resulted in large over- and underpredictions of LTPH, DTPH.D1, DTPH.D2, and DTPH.D3 across the whole range of observed values in OR (Fig. 1) and WA (Fig. 2).

For LTPH the QPORD models provided larger % accurate predictions for OR (PUB = 41%, PRI = 45%) and WA (PUB = 49%, PRI = 41%)

Table 5
Explanatory variables selected for the Quasipoisson (QP) and ordinal (ORD) regression models on public and private lands in western OR and WA.

	QP	ORD
WA PUB	TPH = fn (CANOPY, STDAGE, GRN, WET, TM5, TM7, R54, R57, AUGMAXT, CVPRE, PRR, ForTyp1, ForTyp2)	γ_j = fn (STDAGE, BRT, GRN, TM5, R43, R54, R57, DECMINT, ForTyp1, ForTyp2)
WA PRI	TPH = fn (CANOPY, STDAGE, WET, TM7, R57, AUGMAXT, DIFTMP, SMRPRE, ForTyp1, ForTyp2)	γ_j = fn (STDAGE, BRT, WET, TM2, TM5, R43, R54, SMRPRE, ForTyp1, ForTyp2)
OR PUB	TPH = fn (CANOPY, STDAGE, ANNTMP, CVPRE, SMRTMP, GRN, WET, TM5, TM7, R54, R57, ANNTMP, CVPRE, SMRTMP, PRR, ForTyp1, ForTyp2)	γ_j = fn (STDAGE, GRN, TM5, R43, R54, R57, SMRPRE, ANNPRE, AUGMAXT. ForTyp1, ForTyp2)
OR PRI	TPH = fn (CANOPY, STDAGE, WET, TM2, TM5, TM7, R54, ForTyp1, ForTyp2)	γ_j = fn (STDAGE, BRT, GRN, TM2, TM3, R43, R54, DECMINT, ForTyp1, ForTyp2)

Table 6
Root mean square error between the observed values and the model predictions.

	WA PUB		WA PRI		OR PUB		OR PRI	
	QPORD	RF	QPORD	RF	QPORD	RF	QPORD	RF
LTPH	237	296	250	310	218	254	213	265
DTPH.D1	28	33	48	51	47	52	25	33
DTPH.D2	57	64	34	40	70	80	31	38
DTPH.D3	25	28	16	20	29	31	14	17

Table 7
Spearman correlation coefficient between the observed values and the model predictions.

	WA PUB		WA PRI		OR PUB		OR PRI	
	QPORD	RF	QPORD	RF	QPORD	RF	QPORD	RF
LTPH	0.52	0.34	0.65	0.53	0.48	0.44	0.64	0.56
DTPH.D1	0.11	0.05	0.25	0.05	0.15	0.17	0.25	0.08
DTPH.D2	0.09	0.28	0.40	0.23	0.16	0.21	0.37	0.26
DTPH.D3	0.05	0.28	0.26	0.07	0.02	0.14	0.26	0.08

than RF, which only achieved a maximum of 41% accurate predictions for OR private lands and only 38% accurate predictions for OR public lands as well as for WA private and public lands (Figs. 1 and 2). In both WA and OR and both private and public lands, the RF imputation approach resulted in % accurate prediction values between 21% and 59% for DTPH across all decay classes, while the QPORD models only achieved% accurate prediction values between 6% and 18% for DTPH across the three decay classes. The RF imputation approach achieved higher % accurate prediction values on private lands than on public lands (Figs. 1 and 2).

3.2. Presence of large snags

The accuracy of the large snag presence/absence classification based on the logistic regression model was superior to the accuracy of the RF classification in terms of overall accuracy and κ for both WA and OR on private and public lands (Table 8). For the logistic regression model, overall accuracy and κ ranged from 64% to 70% and from 0.28 to 0.41, respectively, with the highest and the lowest κ values on public and private lands in WA, respectively. The RF approach only achieved overall accuracy and κ values ranging from 50% to 66% and −0.06 to 0.30, respectively. RF provided less

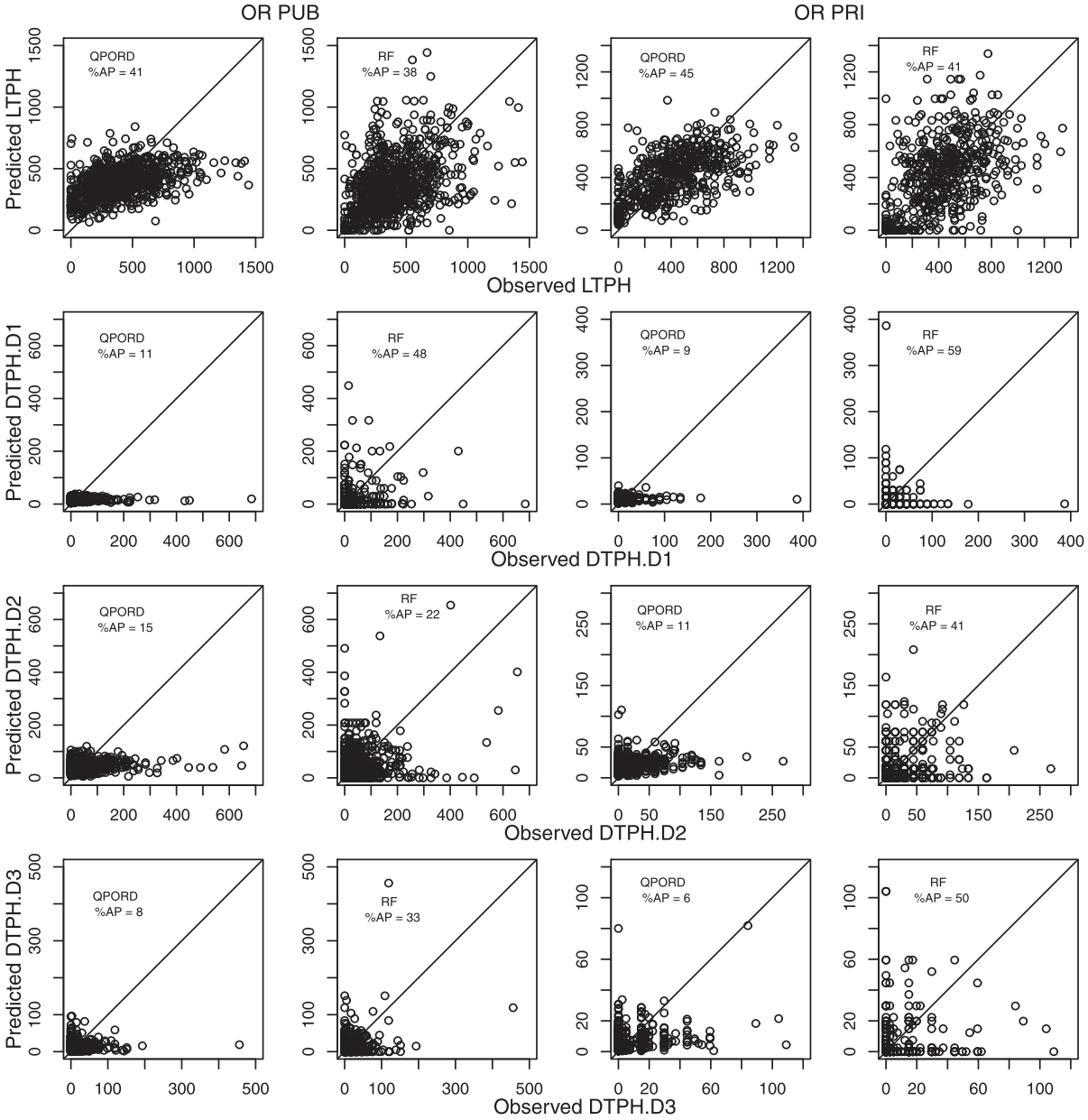

Fig. 1. Scatterplots of observed vs. predicted snag densities by decay class on public (PUB) and private (PRI) lands in Oregon (OR), including % accurate predictions (% AP). The solid line represents perfect correlation between observed and predicted values.

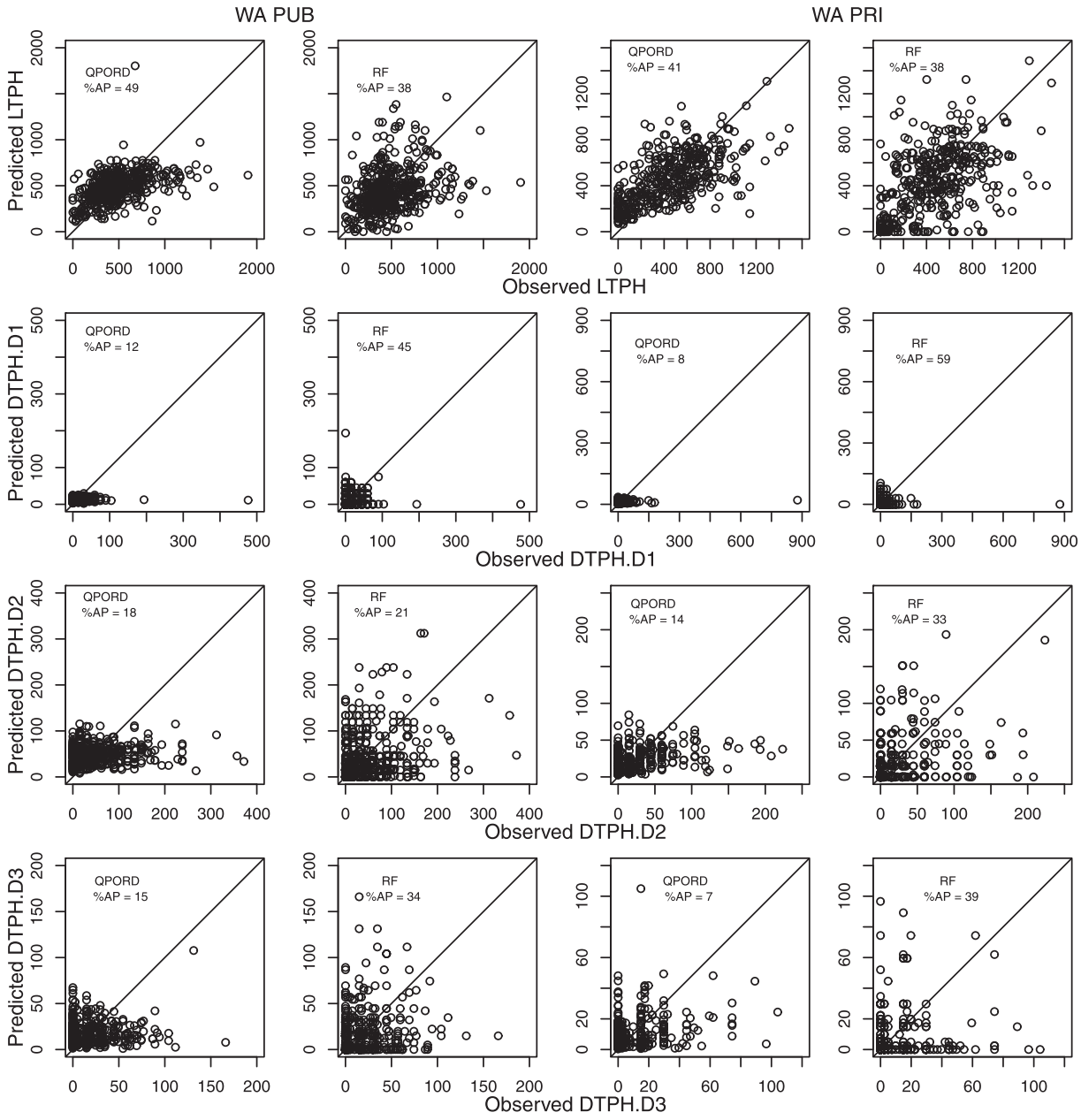

Fig. 2. Scatterplots of observed vs. predicted snag densities by decay class on public (PUB) and private (PRI) lands in Washington (WA), including % accurate predictions (% AP). The solid line represents perfect correlation between observed and predicted values.

accurate classifications of presence/absence of large snags on private lands than on public lands (Table 8).

For both modeling approaches, overall accuracy and κ were smaller when the decision threshold was set to $\tau = 0.5$ than for the adjusted τ values. Additionally, sensitivity was far greater than specificity on public lands for both modeling approaches while specificity was far greater than sensitivity on private lands. For the adjusted τ values, sensitivity and specificity values were more balanced with values fairly close to each other. For RF, specificity was still larger than sensitivity on private lands and sensitivity was greater than specificity on public lands. However, this trend was reversed for the logistic regression model, which resulted in larger specificity values on public lands and larger sensitivity values on private lands (Table 8). Sensitivity was lowest for WA private lands, which had large proportion of plots without large snags and the smallest number of observations (Table 8).

4. Discussion

4.1. Snag density by decay stage

The results of our study indicate that STDAGE was the most important explanatory variable in all QP and ORD models, since TPH tends to decrease with stand age and the proportion of DTPH is larger in old forests than young and pole-sapling stands (Bater et al., 2009) and hence, increases with STDAGE. CANOPY, however, was only significant in the QP models, but did not explain much variability in the ORD models for estimating the proportion of LTPH and DTPH by decay class. This is due to CANOPY being able to capture openings in the canopy that are important in determining the density of standing trees, but CANOPY fails to describe the vertical variability of the canopy that Bater et al. (2009) showed to be important in estimating the proportions of standing dead trees.

Table 8
Accuracy statistics for the different models of the presence/absence of large snags (DBH ⩾ 50 cm).

		Logistic Regression Observed		
		Present	Absent	Sum
WA PUB				
Predicted	Present	194	38	232
	Absent	114	161	275
	Sum	308	199	507
		OA = 70%; SN = 63; SP = 81; κ = 0.41		
WA PRI				
Predicted	Present	118	93	211
	Absent	57	147	204
	Sum	175	240	415
		OA = 64%; SN = 67; SP = 61; κ = 0.28		
OR PUB				
Predicted	Present	335	93	428
	Absent	201	246	447
	Sum	536	339	875
		OA = 66%; SN = 62; SP = 73; κ = 0.33		
OR PRI				
Predicted	Present	121	119	240
	Absent	54	265	319
	Sum	175	384	559
		OA = 69%; SN = 69; SP = 69; κ = 0.35		
		RF Observed		
		Present	Absent	Sum
WA PUB				
Predicted	Present	200	71	271
	Absent	108	128	236
	Sum	308	199	507
		OA= 65%; SN = 65; SP = 64; κ = 0.28		
WA PRI				
Predicted	Present	53	86	139
	Absent	122	154	276
	Sum	175	240	415
		OA = 50%; SN = 30; SP = 64; κ = −0.06		
OR PUB				
Predicted	Present	363	126	489
	Absent	173	213	386
	Sum	536	339	875
		OA = 66%; SN = 68; SP = 63; κ = 0.30		
OR PRI				
Predicted	Present	73	86	159
	Absent	102	298	400
	Sum	175	384	559
		OA = 66%; SN = 42; SP = 78; κ = 0.20		

Note: OA = overall accuracy; SN = sensitivity; SP = specificity; κ = Kappa statistic.

Although Ohmann and Waddell (2002) had reported large snag densities at high elevations, we did not include elevation in our models since elevation is a proxy for temperature and moisture gradients that should be captured by the available climate variables. Additionally, the occurrence of greater snag densities at higher elevations may be explained by ownership, since federal lands are predominately located in the Cascades, Olympics, and Klamath mountain ranges at higher elevation than private lands.

The differences in RMSE, correlation coefficient, and % accurate prediction values between the QPORD and RF approaches demonstrate the importance of choosing appropriate measures to validate model performance. Many studies only present a single model validation criterion which may be misleading, especially if visualization of the modeling results is omitted. Based on the coefficient of correlation between observed and predicted values, QPORD clearly outperformed RF on private lands in WA and OR for all response variables. However, the correlation coefficient provides no insights on the actual accuracy of model predictions, but only indicates the degree of correlation between observed and predicted values. The RMSE values suggested that the QPORD models outperform the RF imputation in predicting LTPH and DTPH across the three decay classes (Table 6). However, the scatterplots (Figs. 1 and 2) show that the QPORD models resulted in smaller RMSE values because fairly small overpredictions were consistently made for small observed values and very large underpredictions were made for large observed values of the four response variables. In the given case, this resulted in smaller RMSE values than those obtained from the RF approach, although RF resulted in higher % accurate prediction for DTPH in all three decay classes, suggesting that the RF approach was superior at least for predicting DTPH.D1, DTPH.D2, and DTPH.D3. Typical for all NN imputation methods, RF provided large over- and underpredictions across the whole range of the response variables but also a large number of accurate predictions which explains why RF outperforms QPORD in terms of % accurate prediction. It is noteworthy that the distribution of predicted values using NN imputation methods like the RF approach is more like the observed distribution in terms of variability than the distribution of predicted values generated by parametric regressions like QPORD, which tend to exhibit much smaller variability than the distribution of observed values. The ultimate decision to use one method over the other depends on the evaluation criteria and the intended use of the model results.

The QPORD model provided poor results in this study for two reasons: (1) the QP model tended to underpredict TPH for large observed TPH and overpredict TPH for small observed TPH. This trend carried over to the next stage of the model for which TPH was multiplied by the estimated proportions of tree status class; (2) The ORD model highly underestimated the proportions of DTPH.D1, DTPH.D2, and DTPH.D3 when large proportions were observed. Multiplying low predicted proportions with low predicted TPH, resulted in extremely low density predictions of DTPH by decay class.

The poor performance of the ORD model can be attributed to the lack of explanatory variables in our data set that described the disturbance history as well as the lack of explanatory variables that explained the vertical variability in the forest canopy. Canopies become more structurally variable with the presence of snags (Bater et al., 2009; Martinuzzi et al., 2009). Since our set of explanatory variables did not include a remote-sensing variable that described the vertical variability in canopy structure, we were not able to predict the proportions of tree status class well, and hence, our model predictions were of lesser quality than the results presented by Bater et al. (2009) who used LiDAR-derived metrics of variation in canopy height. It can be assumed that, once the predictions of the ORD model are improved, the accuracy of the QPORD model will improve for all four response variables. Given that the QPORD model already provides predictions of LTPH that are superior to those of the RF approach in terms of % accurate prediction, it can be expected that QPORD may also provide comparable predictions for DTPH.D1, DTPH.D2, and DTPH.D3, once the ORD model provides improved predictions of the proportions of tree status class.

The predictive abilities of RF and QPORD rely on the assumption that a strong relationship exists between snag density and the explanatory variables. Weak relationships were present in the data of this study, and both RF and QPORD may be more effective if a stronger relationship was present.

The results of our study indicate that density of snags by decay class is easier to model and predict on private lands than on public lands, which can be attributed to private forest lands typically being dominated by even aged stands with similar stand origin that exhibit less variability than forests managed for multiple resources on public lands (Hansen et al., 1991).

4.2. Presence of large snags

Explanatory variables that describe the variation in canopy height, topography as well as disturbance history and forest succession are important for predicting the distribution of diameter classes of snags (Martinuzzi et al., 2009). The accuracy of our large snags presence/absence classification was not very high compared to the accuracy reported by Martinuzzi et al. (2009) who achieved overall accuracy $\geqslant$72% and $\kappa \geqslant 0.43$, which are similar to the maximum values that we achieved with our models. These discrepancies may be explained by a variety of differences in the data. STDAGE, which represents forest succession, was the most important explanatory variable in our models. Without the use of LiDAR-derived metrics, we lacked remote-sensing explanatory variables that described the vertical canopy variation in our data. Metrics describing vertical forest structure may be the key to predicting snag density by size or decay class with high accuracies (Bater et al., 2009; Martinuzzi et al., 2009), which cannot be achieved with explanatory variables derived from Landsat TM data. Even though we had some information about topography (elevation, SLOPE, and ASPECT) available in our data, these variables were not nearly as powerful explanatory variables as the LiDAR-derived metrics describing landform employed in Martinuzzi et al. (2009). Information about disturbance history, which has a major effect on snag abundance and snag characteristics, may also improve model performance.

The large variability in geographic range and forest types in our data set, that covered all of western WA and western OR, may also have contributed to fairly low accuracies of our classification models. We attempted to predict the presence and absence of large snags (DBH $\geqslant$ 50 cm), which is inherently difficult because snags of that size have low prevalence, especially in heavily managed and harvested forests. The results of our study suggest that it is more difficult to predict presence/absence of large snags on private lands, which are heavily managed, than on public lands. This may be explained by the lower prevalence of large snags on private lands than on public lands (Table 2). The RF and logistic regression models for WA private lands had the lowest accuracy, which may be due to the data set having the smallest number of observations ($n = 415$) in combination with small prevalence of large snags. The $\kappa = -0.06$ value for the RF imputation approach indicates that RF did not achieve higher accuracy than a classification by chance could have achieved for private lands in WA.

Sensitivity and specificity can depend on the prevalence of large snags, because classification methods tend to favor the larger class, when the size of presence and absence classes is unequal (Chen et al., 2006). When the decision threshold $\tau = 0.5$, we found that sensitivity was high on public lands where prevalence of large snags was high, but low on private lands where prevalence of large snags was small, thus favoring the larger presence class. Similarly, specificity was large when prevalence was low, but low when prevalence was high. This indicates that when prevalence is low, the models tend to falsely predict absence of large snags where large snags are actually present, and when prevalence is high the models tend to falsely predict large snags being present although they are not. Similar results on specificity and sensitivity were reported by Martinuzzi et al. (2009) who found lower sensitivity for snag classes $\geqslant$25 cm and $\geqslant$30 cm than for snag class $\geqslant$15 cm, which can be explained by the lower prevalence of snag classes with larger diameters. By adjusting the threshold value τ so that it accounts for the prevalence of large snags, we were able to counteract the tendency of the classification methods to favor the larger classes which resulted in similar sensitivity and specificity values. Although this did not necessarily increase overall accuracy, κ values increased by balancing sensitivity and specificity. Adjusting τ is fairly easy for logistic regression models that predict probabilities. If three or more nearest neighbors are used in the RF approach, τ can also be adjusted (see Chen et al., 2006).

The logistic regression model provided higher accuracies for the classification of large snag presence/absence than the RF approach. This may suggest that the parametric regression model may be superior over the non-parametric RF approach. While NN imputation methods such as the RF approach have many advantages for predicting multivariate responses (LeMay and Temesgen, 2005; Eskelson et al., 2009b), it can be assumed that a parametric regression model tends to outperform non-parametric NN imputation methods when only one response variable is of interest, as in the case of presence/absence of large snags.

4.3. Management implications

Management decisions must consider the availability of snags by size and decay classes, in order to ensure habitat availability and suitability. Large snags ($\geqslant$50 cm DBH) are critical habitat elements for many wildlife species. Hence, monitoring their presence or absence is fundamental to habitat management. Similarly, monitoring and managing decay stages of snags is necessary to provide the range of habitat functions associated with dead wood. Models that allow the prediction of existing snag density by size and decay class will help forest managers and planners evaluate wildlife habitat suitability.

This study shows the importance of the accurate presentation and validation of modeling results, since some validation measures can be misleading. We recommend that forest managers critically examine the reported model validation criteria and if possible visualize the results when they choose a model to predict DTPH by decay class. Relying on a single model validation criterion, may be misleading.

Higher accuracies in predictions of snag density by decay class were achieved on private lands which show less variability than public forest lands. Managers of public forest lands may be able to improve the performance of the models by fitting models on small subsets of homogenous data (e.g., fit separate models for different forest types).

The models presented in our study achieved low accuracies, which was attributed to the poor predictive abilities of the available explanatory variables and the highly variable nature of dead wood populations. LiDAR-derived metrics describing the vertical variability in forest canopy structure have been shown to be valuable explanatory variables for predicting snag size class distributions and proportions of snag density classes. Hence, the incorporation of LiDAR-derived metrics into snag density models promises to improve model accuracies in the future when LiDAR-derived metrics become more readily available. Including measures of disturbance history may also improve model performance.

Since large snags are rare in highly managed forests, which are prevalent on private lands, presence/absence of large snags is more difficult to model on private lands than on public lands. Forest managers of private lands should be aware that the presence classification tends to be low while the absence classification tends to be high, when prevalence of large snags is low. Forest managers of public lands should be aware that prevalence is higher on public lands and hence absence classification tends to be low. By adjusting the decision threshold, we were able to balance sensitivity and specificity. When choosing the decision threshold, managers can either favor sensitivity or specificity depending on whether falsely predicting large snags where they are absent or falsely predicting absence of large snags where they are present has a larger impact on management decisions. For example falsely predicting absence of large snags where they are present may result in managers investing in

artificially creating snags although it may not be needed. The decision threshold can more easily be adjusted in logistic regression models than in NN imputation methods such as RF.

Models to predict snag density by decay class and presence of snags by size class will become increasingly important for making management decisions with regards to wildlife habitat. Managers need to be aware of the advantages, disadvantages and trends inherent to some of the modeling approaches, when they use model results in their decision making.

Acknowledgments

We thank Dr. Janet Ohmann, Travis Woolley, and two anonymous reviewers for their helpful comments on an earlier version of the manuscript. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US Government.

References

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232.

Bater, C.W., Coops, N.C., Gergel, S.E., LeMay, V., Collins, D., 2009. Estimation of standing dead tree class distributions in northwest coastal forests using lidar remote sensing. Can. J. For. Res. 39, 1080–1091.

Bechtold, W.A., Patterson, P.L. (eds.), 2005. The enhanced Forest Inventory and Analysis program—national sampling design and estimation procedures. USDA For. Serv. Gen. Tech. Rep. SRS-80.

Beers, T.W., Dress, P.E., Wensel, L.C., 1966. Aspect transformation in site productivity research. J. Forestry 64, 691–692.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Bull, E.L., 2002. The value of coarse woody debris to vertebrates in the Pacific Northwest. In: Proceedings of the Symposium on the Ecology and Management of Dead Wood in Western Forests. USDA Forest Service, Reno, NV, pp. 171–178 (PSW-GTR-181).

Bull, E.L., Parks, C.G., Torgersen, T.R., 1997. Trees and Logs Important to Wildlife in the Interior Columbia River Basin. Gen. Tech. Rep. PNW-GTR-391. US Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR, 55p.

Chen, J.J., Tsai, C.-A., Moon, H., Ahn, H., Young, J.J., Chen, C.J., 2006. Decision threshold adjustment in class prediction. SAR QAR Environ. Res. 17 (3), 337–352.

Christensen, R.H.B., 2010. Ordinal–regression models for ordinal data. R package version 2010.07-23. <http://www.cran.r-project.org/package=ordinal/>.

Cohen, J., 1960. A coefficient of agreement of nominal scales. Educ. Psychol. Meas. 20, 37–46.

Crookston, N.L., Finley, A.O., 2008. YaImpute: an R package for kNN imputation. J. Stat. Softw. 23 (10), 1–16.

Eskelson, B.N.I., Temesgen, H., Barrett, T.M., 2009a. Estimating cavity tree and snag abundance using negative binomial regression models and nearest neighbor imputation methods. Can. J. For. Res. 39, 1749–1765.

Eskelson, B.N.I., Temesgen, H., LeMay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T., 2009b. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand. J. For. Res. 24 (3), 235–246.

Frescino, T.S., Edwards Jr., T.C., Moisen, G.G., 2001. Modelling spatially explicit forest structural variables using generalized additive models. J. Veg. Sci. 12 (1), 15–26.

Graves, A.T., Fajvan, M.A., Miller, G.W., 2000. The effects of thinning intensity on snag and cavity tree abundance in an Appalachian hardwood stand. Can. J. For. Res. 30, 1214–1220.

Hagar, J.C., 2007. Key elements of stand structure for wildlife in production forests west of the Cascade Mountains. In: Harrington, T.B., Nicholas, G.E. (Eds.), Managing for Wildlife Habitat in Westside Production Forests. PNW-GTR-695. U.S. Forest Service, Pacific Northwest Research Station, Portland, OR, pp. 35–48.

Hansen, A.J., Spies, T.A., Swanson, F.J., Ohmann, J.L., 1991. Conserving biodiversity in managed forests. Bioscience 41 (6), 382–392.

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., Cummins, K.W., 1986. Ecology of coarse woody debris in temperate ecosystems. Adv. Ecol. Res. 15, 133–302.

Homer, C.C., Huang, L., Yang, B., Wylie, Coan, M., 2004. Development of a 2001 National Landcover Database for the United States. Photog. Eng. Rem. Sens. 70 (7), 829–840.

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd ed. John Wiley and Sons, New York, 375p.

Kennedy, R.S., Spies, T.A., Gregory, M.J., 2008. Relationships of dead wood patterns with biophysical characteristics and ownership according to scale in Coastal Oregon. USA. Landsc. Ecol. 23, 55–68.

LeMay, V., Temesgen, H., 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. For. Sci. 51 (2), 109–119.

Lundquist, R.W., Mariani, J.M., 1991. Nesting habitat and abundance of snag-dependent birds in the southern Washington Cascade range. US For. Serv. Gen. Tech. Rep. PNW-GTR-285, pp. 221–238.

Marcot, B.G., Ohmann, J.L., Mellen-McLean, K., Waddell, K.L., 2010. Synthesis of regional wildlife and vegetation field studies to guide management of standing and down dead trees. For. Sci. 56 (4), 391–404.

Martinuzzi, S., Vierling, L.A., Gould, W.A., Falkowski, M.J., Evans, J.S., Hudak, A.T., Vierling, K.T., 2009. Mapping snags and understory shrubs for LiDAR-based assessment of wildlife habitat suitability. Rem. Sens. Environ. 113, 2533–2546.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman and Hall/CRC, Boca Raton, 511p.

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

Ohmann, J.L., Gregory, M.J., Spies, T.A., 2007. Influence of environment, disturbance, and ownership on forest vegetation of Coastal Oregon. Ecol. Appl. 17 (1), 18–33.

Pierce Jr., K.B., Ohmann, J.L., Wimberly, M.C., Gregory, M.J., Fried, J.S., 2009. Mapping wildland fuels and forest structure for land management: a comparison of nearest neighbor imputation and other methods. Can. J. For. Res. 39, 1901–1916.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Url: <http://www.R-project.org>.

Vaillancourt, M.A., Drapeau, P., Gauthier, S., Robert, M., 2008. Availability of standing trees for large cavity-nesting birds in the eastern boreal forest of Québec, Canada. Forest Ecol. Manage. 255, 2272–2285.

Wisdom, M.J., Bate, L.J., 2008. Snag density varies with intensity of timber harvest and human access. Forest Ecol. Manage. 255, 2085–2093.