Remote Sensing of Environment 171 (2015) 326–336

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Predicting relative species composition within mixed conifer forest pixels using zero-inflated models and Landsat imagery

Shannon L. Savage [a,*], Rick L. Lawrence [a], John R. Squires [b]

[a] Department of Land Resources & Environmental Sciences, Montana State University, PO Box 173120, Bozeman, MT 59717, United States
[b] Rocky Mountain Research Station, USDA Forest Service, 800 E. Beckwith, Missoula, MT 59801, United States

ARTICLE INFO

Article history:
Received 8 May 2015
Received in revised form 28 September 2015
Accepted 22 October 2015
Available online xxxx

Keywords:
Canopy cover mapping
Generalized linear model
Logistic regression
Operational Land Imager
randomForest
Support vector machine
Thematic Mapper
Zero-inflated data

ABSTRACT

Ecological and land management applications would often benefit from maps of relative canopy cover of each species present within a pixel, instead of traditional remote-sensing based maps of either dominant species or percent canopy cover without regard to species composition. Widely used statistical models for remote sensing, such as randomForest (RF), support vector machines (SVM), and generalized linear regression (GLM), are problematic for this purpose as they often fail to properly predict the absence of a target species, especially in areas of high vegetation diversity, due to the relative abundance of absence observations (or zero values) in the reference data used to train predictive models. Experience has shown that RF, SVM, and GLM models trained on such reference data produce biased values of PCC, for example, in forested areas absent the target species, PCC is overestimated, while in forested areas where a target species PCC is abundant, PCC tends to be underestimated. We used zero-inflated regression modeling to reduce such bias and better predict PCC-by-species within each pixel in mixed conifer forests. Zero-inflated regression models use a two-step process to first predict the presence or absence of the target species, and then to predict continuous levels of PCC only where the target species is present. We compared the results of three widely used methods (RF, SVM, and GLM) to nine zero-inflated models for their ability to predict continuous PCC for each of five different conifer species in heterogeneous forests of northwestern Montana using Landsat TM and OLI imagery. Our best zero-inflated models resulted in a mean difference of −3.84% to 2.26%, 95% confidence interval of 6.22% to 13.09%, and RMSE of 11.26% to 22.98%, depending on the species. The success of the zero-inflated model was robust across methods tested. Both the zero-inflated and traditional methods were successful in estimating continuous canopy cover, however, the traditional models showed a substantial bias by *never* correctly predicting the absence of the target species, while the zero-inflated models correctly predicted species absence 57% to 84% of the time, depending on the species. Visual inspection of the predicted maps compared to high-resolution imagery demonstrated that the zero-inflated models also more closely matched the landscape, as traditional models more often incorrectly predicted canopy cover in non-forested areas. Using the zero-inflated process dramatically reduced the bias of the results, allowing end users to make management decisions with increased confidence about where a target species is absent, something not possible with the traditional methods tested.

1. Introduction

Forests are commonly mapped either continuously by percent canopy cover (PCC) regardless of species or categorically by dominant species. Canopy cover is defined as the "proportion of the forest floor covered by the vertical projection of the tree crowns" (Jennings, Brown, & Sheil, 1999). Total PCC is important when studying biophysical factors of vegetation such as leaf area index (LAI) or the total amount of biomass available within an ecosystem and is widely used in forest and agricultural management. Each plant species within a forest, however, has unique ecological functions. Species composition affects ecological functionality and ecological stability within a community in a profound way (Peterson, Allen, & Holling, 1998) and individual species with different ecological roles can play an important part in the health of an ecosystem. Keystone species – both animal and plant – have strong ecological functions and structure the ecosystems in which they exist (Holling, 1992; Peterson et al., 1998). The inherent dynamic structure and nature of vegetation and ecological communities and their response to disturbance regimes make accurate knowledge of species composition within a forest mandatory for any type of management activity (White, 1979). Foresters (or biologists or managers) would greatly benefit from access to accurately mapped PCC-by-species within large landscapes.

Total PCC has been successfully mapped using a variety of multispectral data at different levels of spatial resolution and scales:

* Corresponding author.
E-mail address: shannon.savage@msu.montana.edu (S.L. Savage).

http://dx.doi.org/10.1016/j.rse.2015.10.013
0034-4257/

coarse-resolution imagery such as MODIS (Moderate-Resolution Imaging Spectroradiometer) (Hansen et al., 2003; Tottrup, Rasmussen, Eklundh, & Jonsson, 2007), moderate-resolution imagery such as Landsat (multispectral scanner (MSS), thematic mapper (TM), enhanced thematic mapper plus (ETM+), and Operational Land Imager (OLI)) (Ahmed, Franklin, & Wulder, 2014; Carreiras, Pereira, & Pereira, 2006; Coulston, Jacobs, King, & Elmore, 2013; Homer, Huang, Yang, Wylie, & Coan, 2004) and ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) (Falkowski, Gessler, Morgan, Hudak, & Smith, 2005), and high-resolution imagery such as IKONOS (Johansen & Phinn, 2006), RapidEye (Ozdemir, 2014), and NAIP (National Agriculture Imagery Program) (Coulston et al., 2013). Similarly, a variety of data types and spatial resolutions have been applied in attempts to assess cover for specific habitats. Landsat imagery was used in several successful habitat and crown closure mapping studies: spectral unmixing was used on Landsat TM to derive the spatial distribution of fraction of oak crown closure in Tulare County, California (Pu, Xu, & Gong, 2003); the habitat of an endangered deer species in Myanmar was predicted with analysis of Landsat ETM+ data (Koy, McShea, Leimgruber, Haack, & Aung, 2005); and Landsat ETM+ along with IKONOS and LiDAR imagery was spectrally unmixed to identify shade versus trees in the Black Hills Experimental Forest of western South Dakota (Chen, Vierling, Rowell, & DeFelice, 2004). Airborne laser scanning data was used to compute fractional cover and LAI in Ofenpass Valley of the Swiss National Park with R^2 values of 0.73 and 0.69 respectively (Morsdorf, Kotz, Meier, Itten, & Allgower, 2006). LiDAR data were found acceptably equivalent to field collected forest canopy cover, while ASTER data were not (Smith et al., 2009). In Queensland, Australia, crown cover was estimated with an R^2 of 0.78 and RMSE of 9.25% when LiDAR vegetation returns were compared to field measurements (Lee & Lucas, 2007). Most of these studies, however, either did not use Landsat data or they classified canopy cover into discrete categories of ecosystems instead of predicting a continuous response for individual species.

Predicting continuous PCC-by-species over large areas using moderate-resolution multispectral data such as Landsat has not been accomplished to date. PCC-by-species can be an important factor in studying habitat, since many animal species rely heavily on specific vegetation species for survival. The snowshoe hare (*Lepus americanus*) is an excellent example: it prefers spruce/fir (*Picea engelmannii*/*Abies lasiocarpa*) forests with high horizontal cover for survival from predation by the Canada lynx (*Lynx canadensis*) and other predators (Fuller & Harrison, 2010; Hodges, 2000; Squires, DeCesare, Kolbe, & Ruggiero, 2010). The effects of global climate change on indicator species is also a good example: whitebark pine (*Pinus albicaulis*) is a strong indicator species for climate change, and the loss of this species from subalpine regions in the rocky mountains of the USA has had dramatic ecosystem-wide impacts, including on threatened grizzly bears (Hicke & Logan, 2009; Jewett et al., 2011).

One factor complicating attempts at mapping PCC-by-species is that reference data for PCC-by-species can be inherently *zero-rich* (i.e., many plots will not have the target species present) and will skew the prediction output toward zero. Zero-rich data is common in count data studies, such as those concerning presence/absence or species distribution (Barry & Welsh, 2002; Cunningham & Lindenmayer, 2005). The risk of having many zeroes in the reference data is high when mapping PCC-by-species, especially if said species is one of many in an area of high diversity.

Zero-rich reference data (data that include a large number of zero observations) can impact statistical models and produce inaccurate results. These data result in models that potentially produce "good" statistics (e.g., low p-values) by under-predicting non-zero observations (i.e., assigning low or zero values to locations with forest cover), a problem when presence of a species is of high interest to the end user, while over-predicting zero observations (assigning values other than zero to locations with no forest cover), a problem when absence of a species is of importance. Zero-inflated regression is one method that has been developed to address this issue of under-prediction of presence data and over-prediction of absence data. It is a two-step process where, first, the presence or absence of a target species is predicted, and then refined to predict a continuous value for "presence" (non-zero) observations (Somers et al., 2012). Zero-inflated regression techniques historically used logistic regression for the first step and a linear model for the second. Existing zero-inflated regression methods, however, do not take advantage of modern machine learning techniques that have been successful with remote sensing imagery, such as randomForest (RF) and support vector machines (SVM) (Mountrakis, Im, & Ogole, 2011; Pal, 2005).

RF, SVM, and linear regression each have been widely used and are well established in remote sensing for predicting vegetation. RF is a decision tree classification and regression method that grows hundreds of decision trees, where each tree is grown using a different bootstrapped (resampled with replacement) random subset of training data, and each split within each tree is based on a different random subset of predictor variables (Breiman, 2001; Lawrence, Wood, & Sheley, 2006). The "forest" of decision trees then votes to assign a class to each input data point for classification (Breiman, 2001; Prasad, Iverson, & Liaw, 2006). In regression mode, the average of the predictions of the "forest" of trees is calculated (Breiman, 2001; Liaw & Wiener, 2002). SVM is a supervised learning mechanism for nonlinear regression and classification problems that identifies hyperplanes (or decision planes) that define decision boundaries based on training data (Zhao, Popescu, Meng, Pang, & Agca, 2011). The decision boundaries measure similarities between objects (kernels), i.e., data features are transformed into multi-dimensional space where the hyperplanes can more easily separate the features into optimal different predictions or classes while minimizing error (StatSoft, Inc., 2013). Linear regression or generalized linear models (referred to hereafter as "GLM") is a straight-forward probabilistic approach to modeling that analyzes the relationship between a predictor variable and one (simple regression) or more (multiple regression) explanatory variables. GLMs can take multiple forms, thus are useful for binary classification (logistic regression) as well as continuous regression.

Our goal was to accurately predict PCC-by-species in highly heterogeneous coniferous forests of northwestern Montana using zero-inflated RF, SVM, and GLM models and Landsat TM and OLI imagery. This would, if successful, identify for each pixel the percentages of each of the conifer species present in each pixel, effectively "unmixing" the forest composition at the pixel level. To our knowledge, this is the first documented attempt to map (1) continuous PCC-by-species in western conifer forests (2) using zero-inflated statistical methods. We focused on five dominant coniferous species within the region. Building on the concept behind zero-inflated regression, we used a two stage model to evaluate whether mixed models might be beneficial, where one algorithm was used to differentiate presence/absence observations, while a different algorithm might best model continuous levels of presence observations. Models were constructed based on forest canopy cover plots measured in the field across the inference area. In addition, we evaluated our model's predictive ability by comparing forest canopy estimates from the model to field data that we measured at independent vegetation plots where canopy was most rigorously quantified.

2. Methods

2.1. Study area

Our study area covers approximately 3.6 million ha in northwestern Montana (Fig. 1). Five United States Forest Service (USFS) forests are included within the study area boundary: Flathead, Helena, Kootenai, Lewis & Clark, and Lolo National Forests. It is a mountainous region ranging from 549 to 3188 m in elevation with a variety of grassland, brushland, and forest types. These complex forests were composed of

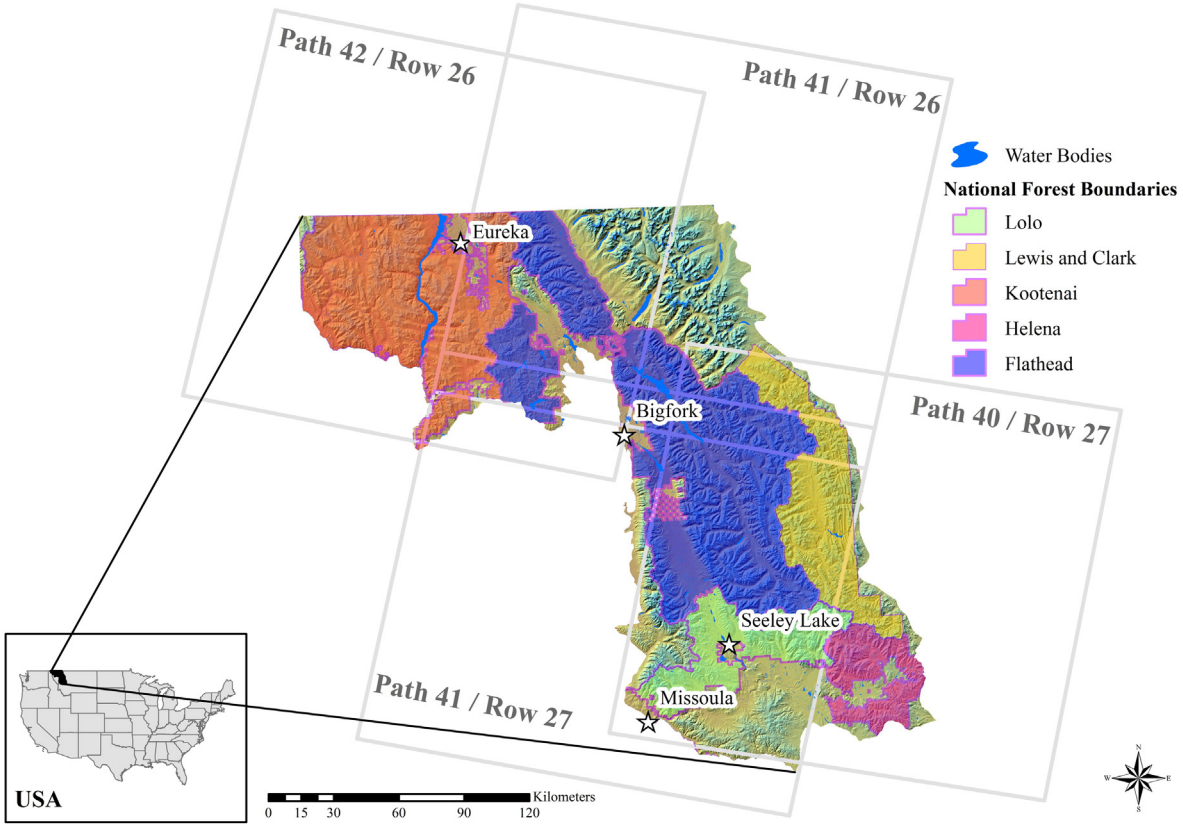


Fig. 1. Location map for the study area within Montana.

mixed conifer species that varied from those dominated by dry ponderosa pine (*Pinus ponderosa*) and Douglas-fir (*Pseudotsuga menziesii*) stands at lower elevations to those dominated by lodgepole pine (*Pinus contorta*), western larch (*Larix occidentalis*), subalpine fir (*A. lasiocarpa*), and Engelmann spruce (*P. engelmannii*) at high-elevation sites. The majority of the study area is covered by 4 Landsat WRS2 scenes (Fig. 1, Table 1).

2.2. Data acquisition

We downloaded mostly cloud-free Landsat TM images from summer and fall 2010 and 2011 from the USGS EROS Center for each of these scenes (Table 1). The images were rectified by the USGS EROS Center in UTM coordinate system, Zones 11 and 12, WGS84 datum. We chose these images to cover the summer and fall seasons to ensure that we had leaf-on and leaf-off for deciduous trees and there was no snow to interfere with the forest canopy mapping. Seven of the images are considered cloud free (less than 1% cloud cover), while the remaining 3 have less than 14% cloud cover.

We downloaded mostly cloud-free Landsat OLI and Thermal Infrared Sensor (TIRS) images from July 2013 from the USGS EROS Center in UTM coordinate system, Zones 11 and 12, WGS84 datum (Table 1). There were no cloud-free OLI images acquired by the satellite during the entire summer and fall of 2013 for our study area. Clouds were masked out of all of these images and filled with radiometrically normalized data from a September image for one scene (through no-change regression normalization (Yuan & Elvidge, 1996)), leaving voids in the other three scenes for which no acceptable fill data was available.

Table 1
Landsat images used to predict PCC-by-species.

Scene path/row	Acquisition date (July TM)	Acquisition date (September TM)	Acquisition date (July OLI)
42/26	18 July 2011[a]	4 September 2011[a]	23 July 2013
41/26[b]	24 July 2010[a]	29 September 2011	16 July 2013
41/27	24 July 2010	29 September 2011[a]	16 July 2013
40/27[b]	4 July 2011	6 September 2011[a]	9 July 2013

[a] Images considered cloud free (<1% cloud cover).
[b] Scenes provided in UTM Zone 12, while all others were provided in UTM Zone 11.

Several ancillary data sets were acquired for analysis and validation. We acquired a 30-m digital elevation model (DEM) of the study area from the USGS National Elevation Dataset (ned.usgs.gov). We downloaded hydrology data from the USGS National Hydrology Dataset (NHD, nhd.usgs.gov) and used them to mask water bodies from the study area. Recent fire boundaries were provided by the USFS Rocky Mountain Research Station and used to mask fire scars from the study area. Finally, we acquired four-band, one-meter-pixel NAIP images from 2011 from the Montana State Library Natural Resource Information System (nris.mt.gov).

In 2013, we collected field reference data representing percent canopy cover (PCC) by species at 1276 points randomly located within 500 m of roads or trails throughout the study area (Fig. 2). Points were reviewed to be spatially homogeneous so that geometric misregistration would have minimal effect. A moosehorn, a tool for vegetation sampling, allowed field crews to accurately identify the presence or absence of canopy cover by establishing a vertical projection at sample points that was then used to estimate PCC in the field (Fiala, Garman, & Gray, 2006). We established sample points on a 5-m by 5-m grid oriented to the north. We then acquired a moosehorn reading every meter within the 5 × 5 grid (25 readings per field data point), recording if canopy cover existed and what species. The total percent canopy cover for each species was calculated by counting the number of times that a species was listed in the upper canopy within the 25-point grid and multiplying that number by 4. Canopy species recorded at field

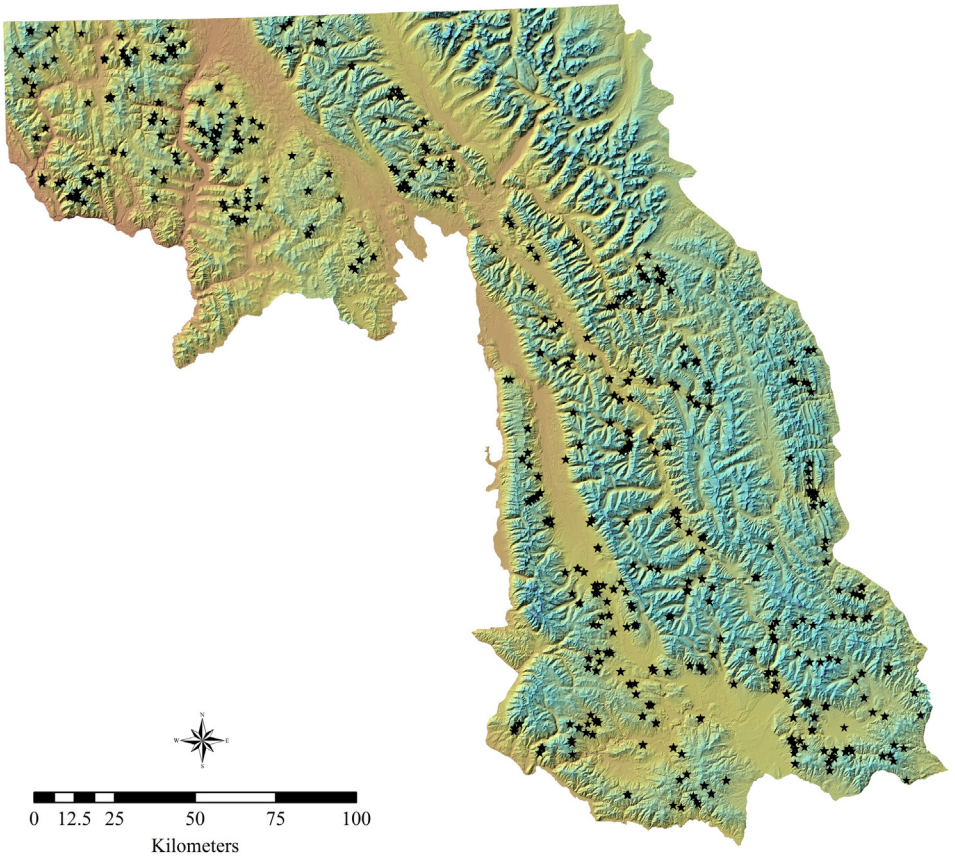


Fig. 2. 2013 field data sites (black stars) randomly located near roads or trails throughout the study area.

plots included: subalpine fir, western larch, lodgepole pine, Engelmann spruce, and Douglas-fir.

2.3. Data processing

We assumed that no change in canopy cover had occurred from 2011 to 2013 (unless from fire or clouds), because the scenes were acquired from similar dates and the two scenes from the center path are from the same date. We applied relative radiometric normalization to adjacent scenes using band-specific regressions based on the grouping around the mean of the difference histogram of the overlapping areas (similar to image regression and no-change regression normalization (Singh, 1989; Yuan & Elvidge, 1996)). The regression was performed on the data within one standard deviation of the mean of the difference histogram for each band for each overlapping area within one month. By removing the data in the histogram's tails - where drastic changes such as fire or cloud cover occurred – we derived the regression equation for normalization from pixels that were most likely unchanged.

We geometrically corrected all Landsat scenes to a master scene as needed, though most images were aligned without detectable error as provided by the USGS EROS Center. The four scenes for each month (July 2011, September 2011, and July 2013) were mosaicked and clipped to the study area boundary. Water bodies and recent fire scars (within 10 years of the image acquisition) were masked from each mosaicked image, since canopy cover is not applicable in those locations.

We derived ancillary datasets from the DEM and NAIP imagery, including slope, aspect, and texture information for the study area. We also used the 1-m NAIP imagery for checking field data for errors (i.e., GPS coordinate mistakes and/or incorrect cover types). The image texture mean and minimum values were derived from the standard deviation of the first principal component of the 1-m NAIP data, then resampled to 30-m pixels for analysis (Brown & Barber, 2012). Finally, the three mosaicked Landsat images were combined with these ancillary datasets to produce a 27-component image for analysis (Table 2).

Table 2
Imagery components used in the analysis.

Number	Component name
1	July 2013 — Blue
2	July 2013 — Green
3	July 2013 — Red
4	July 2013 — NIR
5	July 2013 — MIR1
6	July 2013 — MIR2
7	July 2013 — TIR1
8	July 2013 — TIR2
9	September 2011 — Blue
10	September 2011 — Green
11	September 2011 — Red
12	September 2011 — NIR
13	September 2011 — MIR1
14	September 2011 — TIR
15	September 2011 — MIR2
16	July 2011 — Blue
17	July 2011 — Green
18	July 2011 — Red
19	July 2011 — NIR
20	July 2011 — MIR1
21	July 2011 — TIR
22	July 2011 — MIR2
23	DEM
24	Slope (degrees)
25	Aspect (17 categories)
26	NAIP texture mean
27	NAIP texture minimum

2.4. Zero-inflated models

We utilized the 27 data components created for the forest cover mapping process (Table 2) and reference data extracted using the 2013 field data points for model creation. A zero-inflated model is a two-step process that we applied to each target species on an individual basis, so that each species was analyzed independently through a series of models with a separate output map for each species of percent canopy cover per pixel for that species (Fig. 3). The prediction models were based on combinations of three specific methods: (1) RF, (2) SVM, and (3) GLM. All reference data (input (a) in Fig. 3) are identified first as either zero (absence) data or non-zero (presence) data and given values of 0 or 1. A binary classification (using one of the prediction models listed above; e.g.: GLM in Fig. 3) is then performed on the entire study area to identify whether canopy cover for the species being analyzed is present or absent for each pixel (process (b), Fig. 3) and a binary map is predicted (output (c), Fig. 3). Second, only non-zero (presence) reference data are utilized in a continuous regression model (e.g.: SVM in Fig. 3) to predict PCC where canopy cover is present for the species being analyzed within the entire study area (processes (d) and (f), Fig. 3) and a continuous map is predicted (output (g), Fig. 3). We produced the final study-area-wide zero-inflated prediction map for each species (e.g.: GLM.SVM in Fig. 3) by combining the binary classification map with the continuous prediction map (process (h), Fig. 3). We assigned a value of zero to any pixel that was identified as absence or zero data in the binary map, while the pixels that we identified as presence or non-zero data in the binary map were assigned the PCC value from the continuous prediction (output (i), Fig. 3). By separating out the zero data, the zero-rich nature of the full dataset does not bias the continuous regression model toward zero, nor does the non-zero data result in overestimation of the zero data.

For the RF and SVM binary analyses (process (b), Fig. 3), we used the models in classification mode, while for the GLM binary analyses, we applied a logistic regression with variable selection based on a stepwise process evaluated using AIC (Akaike Information Criteria). For the RF and SVM continuous analyses (process (f), Fig. 3), we used the models in regression mode, while for the GLM continuous analysis, we used a stepwise generalized linear model based on AIC.

We used a 10-fold cross-validation process to compare models (i.e., each model was executed 10 times while withholding a random 25% of the data each time). We evaluated nine zero-inflated models for each species. The first 3 zero-inflated models were combinations of the same type of model used for both the binary and continuous analyses, for example, we used a binary RF classification to model zero data and a continuous RF-based regression to model non-zero data (zero-inflated randomForest, or RF.RF). The rest of the zero-inflated models were combinations of the three methods and notated as "binary.continuous", e.g., a model using logistic regression to model zero data and SVM to model non-zero data was designated GLM.SVM (Fig. 3). We also evaluated three single-step (non-zero-inflated) models for each species (i.e., we applied only RF, SVM, or GLM continuous regression models) in order to determine whether or not zero-inflated models predict PCC more accurately than traditional methods.

We performed a Wilcoxon's signed rank test on predictions for the combined withheld data for each of the 10-fold cross-validations in order to produce a p-value, the 95% confidence interval (CI), and the mean of the differences. Of the nine zero-inflated models created for each species, we chose the model combination with the smallest average CI (as long as the p-value was greater than 0.05, indicating the difference was not statistically significantly different from zero) as the final model with which to predict the PCC for that species. This allowed us to select the most precise model (smallest CI) having acceptable accuracy (mean difference not significantly different from zero). It is possible that a dataset might have a low mean difference due to a balancing of extreme over- and underestimates, so we calculated a root mean squared error (RMSE, a statistic that penalizes extreme errors (Chai & Draxler, 2014)) to identify the existence of possible extreme errors not indicated by the Wilcoxon's statistics. We also calculated the number of over or underestimates for each model to see if the single-step models were overestimating zeroes, as we expected. For each best/final model, in addition to a predicted map based on the zero-inflated model, we also applied single-step models to the same

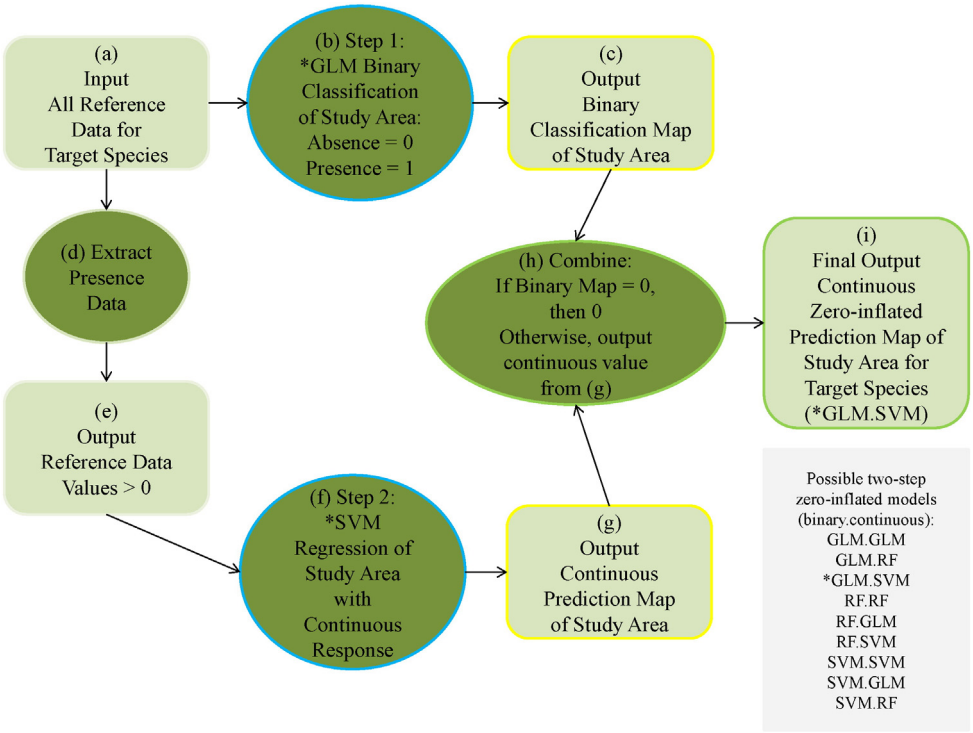


Fig. 3. Flowchart of the two-step zero-inflated process. Light shaded boxes represent data input and output, dark shaded ovals represent processes. Example shown with GLM used for the binary classification (Step 1) and SVM used for the continuous regression (Step 2) (indicated with an asterisk(*)). The full process is performed independently on each target species.

data based on the methods used for binary classification and continuous prediction in the zero-inflated model. Ultimately, we produced up to 3 predicted maps of PCC-by-species for each species (for example, if the best zero-inflated model was a combination of GLM.SVM, we produced three predicted maps: (1) GLM.SVM, (2) single-step GLM, and (3) single-step SVM; or if the best model was SVM.SVM, we produced two predicted maps: (1) SVM.SVM and (2) single-step SVM) and we compared the zero-inflated prediction maps visually to the single-step predicted maps.

2.5. Independent accuracy assessment

In 2014, we quantified a second sample (N = 113 plots) of canopy cover plots to provide an independent assessment of model performance. We established a 20-m by 20-m grid in which technicians sampled canopy by species with moosehorn readings taken every meter (400 readings per field data point). We then calculated PCC for each species by adding up all occurrences of the species within the 400-m^2 grid and dividing that total by 4. The intent of this intensive sampling regime at independent points represented our best attempt to establish estimates of "true" canopy composition by tree species for model comparison. We believed this comparison with independent data provided a valid estimate of model performance as would be expected during actual field application. Sample plots for independent data were randomly established with VMap polygons (a region-wide USFS geospatial database with lifeform and dominant tree canopy cover information; Brown & Barber, 2012) to sample forest patches (>1.6 ha) that were relatively homogeneous for comparison to model outputs. We restricted sample points to within 200 m of open forest roads to facilitate sampling, but plots were located outside of edge-effects from the road surface. We applied the Wilcoxon's signed rank test with p-value, CI, and mean of the differences to the predicted maps and these independent field data points to check the accuracy of our methods.

3. Results

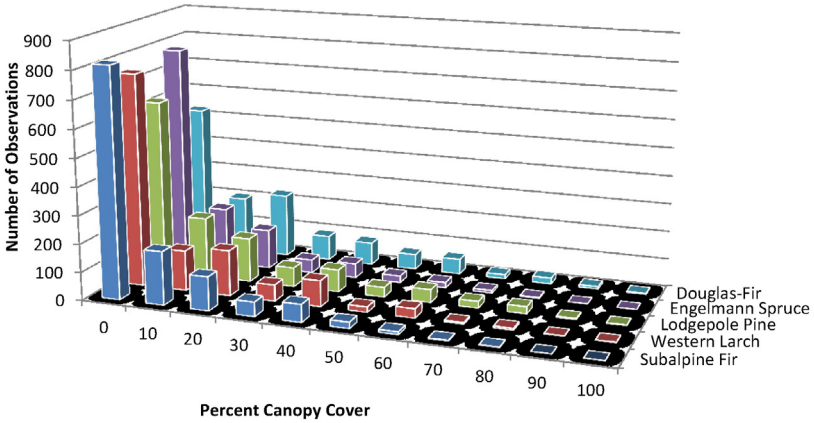
3.1. Prediction accuracies from 10-fold cross-validation

The forests of northwestern Montana are diverse and highly heterogeneous at fine (sub-Landsat pixel) scales, which makes accurately mapping PCC-by-species with Landsat data a difficult process — and up to this point, rarely, if ever, done (and documented). Our intensive field sampling in 2013 (N = 1276) showed that the forests in our study area are composed of no less than 18 different tree species with a wide range of PCC. On average, our five species of interest (subalpine fir, western larch, lodgepole pine, Engelmann spruce, and Douglas-fir) had PCC values ranging from 18.58% for subalpine fir to 25.02% for Douglas-fir, while the average total PCC of the field sites was 60.81%. Subalpine fir, western larch, and Engelmann spruce were observed in less than half of the field points, while lodgepole pine and Douglas-fir were observed in just over half of the field points, demonstrating the zero-inflated nature of these reference data. Maximum values of the field observations of PCC ranged from 72% for subalpine fir to 100% for lodgepole pine. After accounting for the zeroes in these observations, we see most observations have low values and very few have the higher values (Fig. 4).

We evaluated 3 different statistical methods for our zero-inflated models: RF, SVM, and GLM. We chose the best/final PCC-by-species model for each of our species of interest by identifying the model with the smallest average confidence interval in the 10-fold cross-validation results. All of the final models included SVM for continuous modeling of the non-zero data. All three statistical methods were used, depending on species, for the final binary classification of absence, or zero, data (See Appendix, Table S1).

None of the single-step models were able to predict absence – not a single observation of zero was accurately predicted as zero – and in many cases these models had a larger mean difference than the zero-inflated models (Table 3). The zero-inflated models, on the other hand, accurately predicted absence on average 74% of the time (57% to 84% of the time depending on species). In all cases, the single-step models overestimated more often than underestimated. The zero-inflated models over- or underestimated PCC equally: subalpine fir and Engelmann spruce were underestimated, while western larch, lodgepole pine, and Douglas-fir were overestimated though not as strongly as the single-step models.

Overall mean differences (predicted minus observed) for the best/final zero-inflated models ranged from −1.81% to −0.54% with CI widths ranging from 4.81% to 6.33% and RMSE values ranging from 11.07% to 18.99.% (See Appendix, Table S2). The zero-inflated models predicting PCC-by-species for subalpine fir and Douglas-fir had the best results overall based on the CI width. The p-value of nearly all of the zero-inflated models (with the exception of many of the



	0	10	20	30	40	50	60	70	80	90	100
Subalpine Fir	817	190	120	51	63	22	10	2	1	0	0
Western Larch	756	143	162	58	92	19	33	5	6	1	1
Lodgepole Pine	622	214	157	68	79	35	46	24	25	4	2
Engelmann Spruce	780	203	140	47	51	24	19	9	3	0	0
Douglas-Fir	525	200	227	88	80	54	54	15	21	9	3

Fig. 4. 2013 field sample observations of percent canopy cover for five conifer species of interest (N = 1276).

Table 3
Comparison of over- and underestimates and accurately predicted zero observations for the final zero-inflated model and the associated single-step model(s) for each species. Each validation consisted of 318 observations.

Species	Model	Mean difference	# Overestimates	# Underestimates	Total # zeroes accurately predicted/observed
Subalpine fir	GLM.SVM	−1.18	65	83	169/201
	GLM*	0.10	206	112	0/201
	SVM	−2.75	184	134	0/201
Western larch	SVM.SVM	−0.54	89	84	145/188
	SVM	−3.46	171	147	0/188
Lodgepole pine	SVM.SVM	−1.81	113	105	100/155
	SVM*	−5.81	168	150	0/155
Engelmann spruce	GLM.SVM	−1.14	79	82	157/192
	GLM*	−0.12	198	120	0/192
	SVM	−2.70	180	138	0/192
Douglas-fir	RF.SVM	−0.56	132	113	73/129
	RF*	0.37	207	111	0/129
	SVM	−3.44	165	154	0/129

* Models with a p-value less than 0.05 (i.e., the difference between predicted and observed was statistically significantly different from zero).

Douglas-fir models) was greater than 0.05, indicating that the model differences were not statistically significantly different from zero and supporting the use of these models for prediction of PCC-by-species. While each final model was identified based on CI width, it is important to note that there were no major differences in model performance among the zero-inflated models (See Appendix, Table S1) (less than 2% difference in CI width and mean differences, and from just over 1% to just under 3% differences in RMSE). Also, there is no clear evidence that zero-inflated GLM methods (i.e., non-machine learning) performed consistently worse than those that used RF or SVM, since even the worst models (those with the widest CI for each species) included RF or SVM combined with GLM. All of the single-step RF models and all but one of the single-step GLM models had p-values less than 0.05 with very small CI widths and mean differences. The single-step SVM models had p-values greater than 0.05, but demonstrably larger mean differences than the zero-inflated models. The RMSE values of the single-step models were only slightly larger than the zero-inflated models of western larch, lodgepole pine, and Engelmann spruce.

3.2. Independent accuracy assessment

The CI widths and mean differences from the independent accuracy assessment were all less than 13% and 4% respectively, although in all cases larger than the results of the 10-fold cross-validation (Table 4). The CI widths from the independent accuracy assessment ranged from 6.22% for Engelmann spruce to 13.09% for lodgepole pine and mean differences ranged from 2.26% for Engelmann spruce to −3.84% for lodgepole pine. Western larch, lodgepole pine, and Douglas-fir all had p-values greater than 0.05, while subalpine fir and Engelmann spruce had p-values that indicate that statistically there was a difference between predicted and observed for these species. RMSEs were smaller for the 10-fold cross-validation for all species, except for Engelmann spruce, where the RMSE differed by merely half a percent.

Table 4
Comparison of independent accuracy assessment (IAA) and 10-fold cross-validation (C-V) error rates. The IAA compares percent canopy cover at 113 test plots to the final zero-inflated model predictions for each species.

Species	Model	Mean differences	CI lower/upper	RMSE
Subalpine fir	GLM.SVM (IAA)*	−2.86	−9.02/−0.38	12.29
	GLM.SVM (C-V)	−1.18	−4.53/0.91	11.07
Western larch	SVM.SVM (IAA)	−1.27	−3.75/3.81	16.74
	SVM.SVM (C-V)	−0.54	−3.01/2.97	13.81
Lodgepole pine	SVM.SVM (IAA)	−3.84	−9.74/3.35	22.98
	SVM.SVM (C-V)	−1.81	−3.92/2.41	18.99
Engelmann spruce	GLM.SVM (IAA)*	2.26	1.85/8.07	11.26
	GLM.SVM (C-V)	−1.14	−3.52/1.92	11.76
Douglas-fir	RF.SVM (IAA)	−3.02	−5.83/2.28	18.69
	RF.SVM (C-V)	−0.56	−1.94/2.87	16.85

* Models with a p-value less than 0.05 (i.e., the difference between predicted and observed was statistically significantly different from zero).

4. Discussion

Our objective was to develop a method that combined zero-inflated models and Landsat imagery to accurately map PCC-by-species in coniferous forests of northwestern Montana. We found that our zero-inflated models were successful in estimating canopy cover by species. This novel application of zero-inflated regression along with RF, SVM, and GLM resulted in accurate predictions of PCC-by-species (all mean differences between predicted and observed <4%). Meanwhile, the single-step (or more traditional) RF, SVM, and GLM models showed a bias issue that indicated they were more error-prone, especially when the end user is most interested in presence data. In other words, the single-step models did not successfully predict absence, but rather predicted low values where the species did not actually exist. We demonstrated that with proper reference data, Landsat imagery and zero-inflated methods provided accurate and useful predictive models of forest canopy (with mean differences of −3.84% to 2.26%) despite the inherent zero-richness of species-specific canopy cover data within highly heterogeneous coniferous forests (Fig. 5).

The five species we focused on in this study were from five different genera rather than five different species within a genus. We believe it is likely that distinguishing species within a genera (e.g., distinguishing among true firs) generally would be more difficult, although there are likely exceptions, such as where species are spatially distinguishable (for example, a species might be distinguishable because it is constrained by elevation). We were, however, unable to evaluate this question with the data from our study.

Zero-inflated models that incorporated RF worked well, as expected from the literature (Coulston et al., 2013; Fassnacht et al., 2014; Ho et al., 2014). SVM- and GLM-based models, however, worked as well, and in many cases better (SVM was consistently used in all 5 of the best species-specific zero-inflated models), indicating that the zero-inflated modeling approach is robust across the methods tested. We expected the most accurate results to be from models that utilized RF and SVM based on literature review of machine learning algorithms, however, the models that utilized GLM performed nearly as well as the RF and SVM. Although the GLM.SVM model was established as best for subalpine fir and Engelmann spruce in the 10-fold cross-validation, the statistics were not markedly better than those of the RF.SVM or SVM.SVM models, as they were less than a quarter-percent smaller in the CI width (See Appendix, Table S1.A and S1.D). In fact, the statistics for nearly all the models were similar enough to make choosing a best model decidedly data dependent. Based on our results, a user might

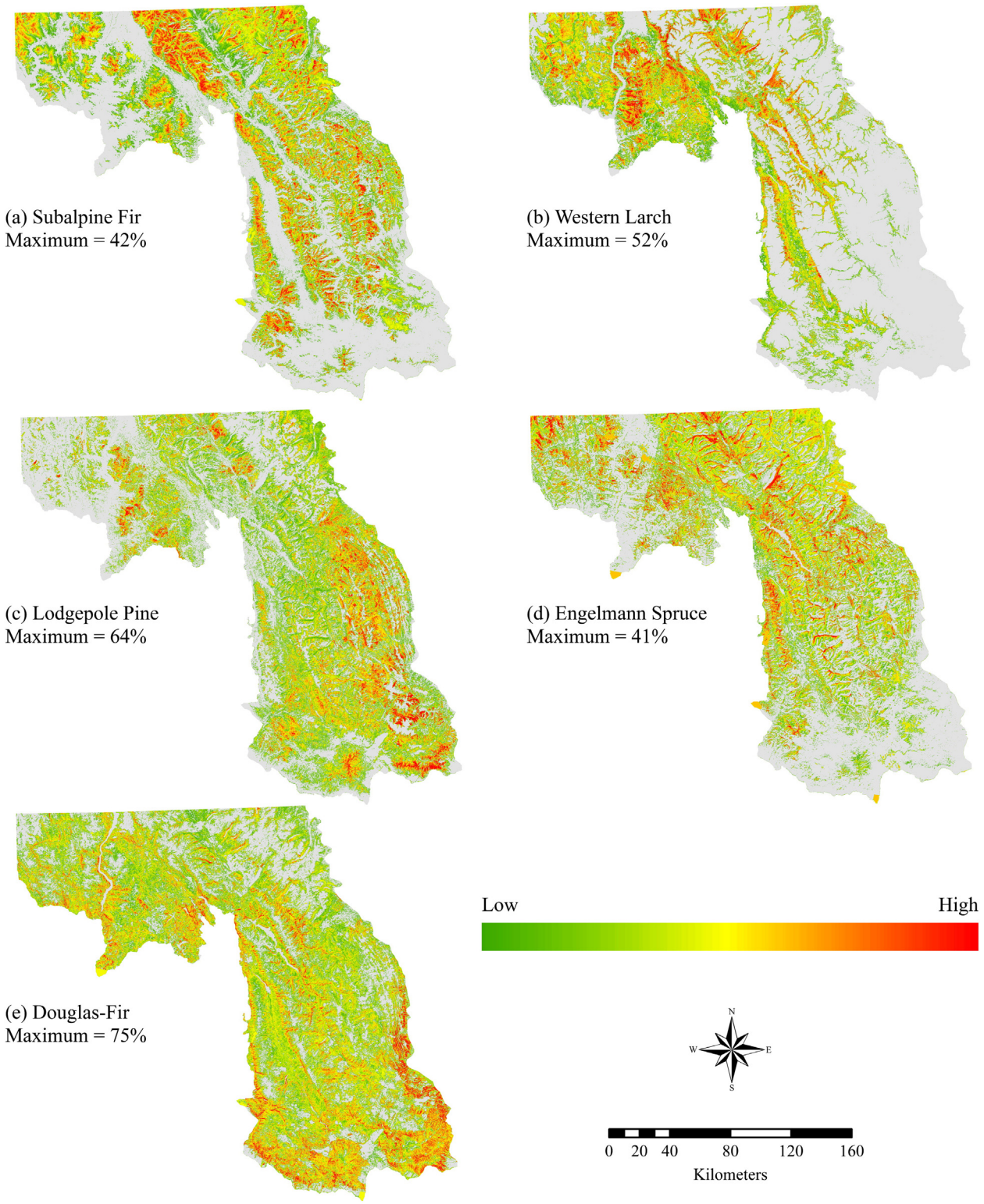


Fig. 5. Percent canopy cover by species. Areas in gray represent 0% canopy cover. Areas in red represent the highest PCC values. Ranges of PCC differ for each species: (a) subalpine fir — 0 to 42%, (b) western larch — 0 to 52%, (c) lodgepole pine — 0 to 64%, (d) Engelmann spruce — 0 to 41%, and (e) Douglas-fir — 0 to 75%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

choose to either (1) select any method and expect satisfactory results or (2) be advised to compare all methods to achieve optimal results.

From a forest management standpoint, the 10-fold cross-validation statistics suggest that the model combinations chosen for our zero-inflated modeling produced accurate and precise predictions of PCC-by-species for all five species, as all of the models had a p-value greater than 0.05. Additionally, the mean differences were quite small and the spread of the CIs around the mean show high precision and were, therefore, potentially within the bounds of acceptability for land managers and biologists. The RMSE values, however, were higher than the mean differences and indicated there were some large over- and underestimations, meaning care should be taken when using the resulting

maps for pixel-level analyses. These models successfully predicted PCC-by-species out of the zero-rich reference data (at an alpha of 0.05, the predicted versus observed is not significantly different from zero), i.e., where there is forest on the ground, the models were on average underestimating PCC-by-species by 0.54% to 1.81% and we are 95% certain that the true mean error with respect to the population is within 2.41% to 3.17% of that estimate.

Our accuracy assessments demonstrate that zero-inflated methods have distinct advantages over single-step methods to develop accurate and useful models of PCC-by-species. The 10-fold cross-validation

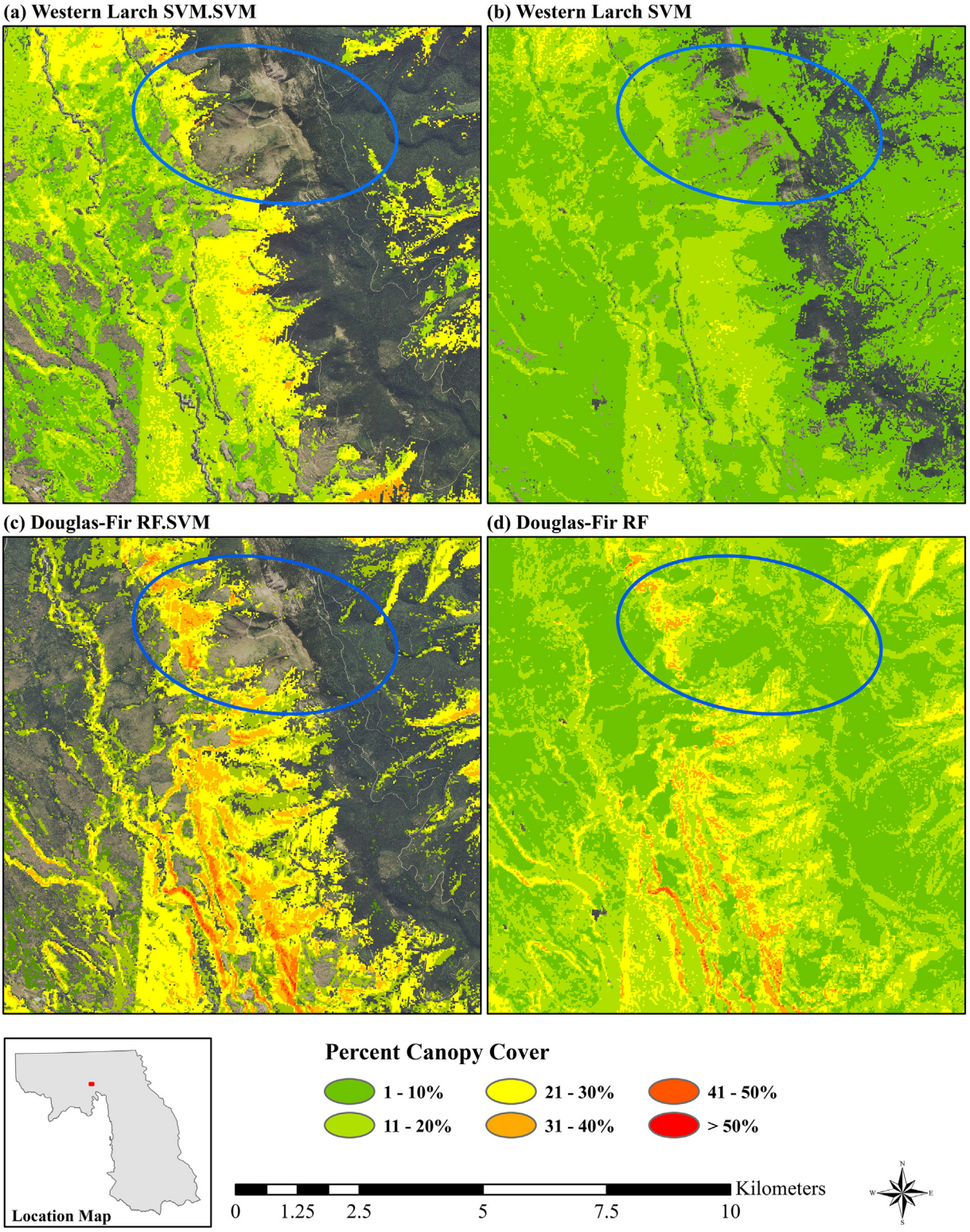


Fig. 6. Examples of two western larch prediction maps: (a) zero-inflated SVM.SVM and (b) single-step SVM; and two Douglas-fir prediction models: (c) zero-inflated RF.SVM and (d) single-step RF. Maps are displayed over NAIP imagery. The blue ovals indicate areas of non-forest. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

showed that the single-step models had very similar and in some cases slightly better results than the zero-inflated models for the complete dataset when evaluated solely on p-values and CIs. The independent accuracy assessment showed that with small CI widths and mean differences, the best/final zero-inflated models were successful at predicting PCC-by-species. Where there is forest on the ground, these models were on average underestimating PCC for subalpine fir by 2.86%, for western larch by 1.27%, for lodgepole pine by 3.84%, and for Douglas-fir by 3.02%. The models were overestimating PCC for Engelmann spruce by 2.26%. We are 95% certain that the true mean error with respect to each population is within 3.11% to 6.54% of those estimates. We believe our independent accuracy assessment best describes what land managers should expect when using these models to estimate the canopy composition of western conifer forests.

The greatest shortcoming of the single-step models was that they had substantially more errors than the zero-inflated models, and in all cases the majority of the errors were overestimates of PCC (Table 3). The single-step models were erroneously predicting PCC values (presence) where there was, in fact, absence. Single-step models successfully predicted zero values (absence) exactly *zero times* out of 865 chances in the 10-fold cross-validation process (Table 3), while the zero-inflated models successfully predicted zeroes on average 74% of the time. This suggests the single-step models were attempting to balance out the overestimations of zero data with underestimations of non-zero data. This is obvious under visual inspection of the produced PCC maps (Fig. 6b and d). Firstly, the single-step model maps identified the majority of the area as having PCC greater than zero, as opposed to the zero-inflated model maps that showed a smaller area with PCC greater than zero. Secondly, the single-step model maps identified the majority of the pixels as having low PCC with relatively few pixels having high PCC, as opposed to the zero-inflated model maps that had more pixels identified as higher PCC. Visual inspection of the PCC maps displayed over NAIP imagery (Fig. 6) demonstrated that the zero-inflated models more closely matched the landscape. The single-step maps incorrectly indicated trees (non-zero PCC) in non-forested areas, including mountain tops and valley floors (blue ovals in Fig. 6) and the Douglas-fir single-step map incorrectly showed Douglas-fir across nearly the entire landscape (Fig. 6d).

The ranges of predicted PCC for each species are noticeably smaller than the actual ranges from the 2013 field sites (Fig. 4) of PCC for these species. While the zero-inflated models have successfully predicted locations with zero PCC (Table 3), their predictions of non-zero data tended toward the mean and underrepresented extremely high values, as is evidenced by the relatively high RMSE values in the accuracy assessment. We expected the zero-inflated model to account for the low and zero values – this is the primary reason we attempted zero-inflated modeling – however the process was imperfect and did not effectively account for the higher values of PCC. One reason for this is that there were very few instances of high values in the 2013 reference data (Fig. 4) so there were relatively few observations available for training for those high values. Additionally, there is a well-known occurrence of "regression to the mean", whereby variability in predictions is reduced/the slope of the true relationship is underestimated because of possible random error (both small-scale and systematic) in the data (Kuchler, Ecker, Feldmeyer-Christe, Graf, & Waser, 2007; Pierce, Ohmann, Wimberly, Gregory, & Fried, 2009; Smart & Scott, 2004).

Use of single-step SVM models, which often had the highest accuracy based on CI width (and p-value greater than zero) (See Appendix, Table S2), was therefore problematic because of the consistent overestimation of zeroes that resulted along with the frequent underestimation of non-zero data. It might be better in the case of habitat mapping, for example, to miss a small amount of vegetation when identifying a particular habitat or impacts to that habitat, than to have high levels of false positives (Fielding & Bell, 1997). However, of utmost importance was to produce accurate maps, and our zero-inflated models, while possibly missing critical habitat, were less biased than the single-step models and produced better maps (visually) with high accuracy and precision. Thus many of the zero-inflated models we produced that had wider CIs but substantially fewer overestimates than the single-step models appeared to, in fact, produce better prediction maps for our task of mapping canopy cover for individual species.

5. Conclusions

Mapping PCC-by-species as a continuous response within mixed forest pixels has rarely been attempted using remote sensing, and Landsat imagery in particular; however, we demonstrated that it can be done with zero-inflated modeling, as evidenced in the results of the independent accuracy assessment (mean differences of −3.84% to 2.26%), as well as through visual inspection of the maps. We demonstrated that zero-inflated methods were useful for predicting continuous PCC-by-species in highly heterogeneous conifer forests. While both the zero-inflated and single-step methods were successful in estimating continuous canopy cover, the single-step models showed a substantial bias by *never* (0/865 observations) correctly predicting the absence of the target species, indicating a considerable bias. Conversely, the zero-inflated models correctly predicted species absence 74% of the time (644/865 observations). Using the zero-inflated process dramatically reduced the bias of the results, allowing end users to make management decisions with greater confidence about where a target species is absent, something not possible with traditional/single-step methods. Further manipulation of zero-inflated models through different data sources (such as hyperspectral data), field data collection methods (more and better reference data, particularly for multiple species within one genus), and/or prediction models (novel regression or classification methods) might lead to even higher accuracies. The application of zero-inflated prediction models to zero-rich vegetation data might be appropriate to many fields of study, especially those that have a continuous response and are patchy in nature across the landscape, for example, rare/threatened/endangered plant species and wildlife habitat.

Acknowledgments

This work was supported by USDA Forest Service, Joint Venture Agreement 12-CS-11221635-176 and many scientists at the Rocky Mountain Research Station in Missoula, Montana. We thank the many field technicians for collecting the reference data and Lucretia Olson for her untiring and invaluable assistance throughout the project. Additionally, we are grateful to the reviewers and the journal editor for their thorough and constructive comments that helped improve our manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.rse.2015.10.013.

References

Ahmed, O. S., Franklin, S. E., & Wulder, M. A. (2014). Integration of LiDAR and Landsat data to estimate forest canopy cover in coastal British Columbia. *Photogrammetric Engineering & Remote Sensing, 80*, 953–961.

Barry, S. C., & Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling, 157*, 179–188.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 4–32.

Brown, S., & Barber, J. (2012). The region 1 existing vegetation mapping program (VMap) flathead national forest overview; version 12. *Region one vegetation classification, mapping, inventory and analysis report 12–34*. Missoula, MT: USDA Forest Service.

Carreiras, J. M. B., Pereira, J. M. C., & Pereira, J. S. (2006). Estimation of tree canopy cover in evergreen oak woodlands using remote sensing. *Forest Ecology and Management, 223*, 45–53.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? —Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*, 1247–1250.

Chen, X., Vierling, L., Rowell, E., & DeFelice, T. (2004). Using LiDAR and effective LAI data to evaluate IKONOS and Landsat 7 ETM+ vegetation cover estimates in a ponderosa pine forest. *Remote Sensing of Environment*, *91*, 14–26.

Coulston, J. W., Jacobs, D. M., King, C. R., & Elmore, I. C. (2013). The influence of multi-season imagery on models of canopy cover: A case study. *Photogrammetric Engineering & Remote Sensing*, *79*, 469–477.

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

Falkowski, M. J., Gessler, P. E., Morgan, P., Hudak, A. T., & Smith, A. M. S. (2005). Characterizing and mapping forest fire fuels using ASTER imagery and gradient modeling. *Forest Ecology and Management*, *217*, 129–146.

Fassnacht, F. E., Hartig, F., Latifi, H., Berger, C., Hernandez, J., Corvalan, P., & Koch, B. (2014). Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*, *154*, 102–114.

Fiala, A. C. S., Garman, S. L., & Gray, A. N. (2006). Comparison of five canopy cover estimation techniques in the western Oregon cascades. *Forest Ecology and Management*, *232*, 188–197.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*, 38–49.

Fuller, A. K., & Harrison, D. J. (2010). Movement paths reveal scale-dependent habitat decisions by Canada lynx. *Journal of Mammalogy*, *91*, 1269–1279.

Hansen, M. C., DeFries, R. S., Townshend, J. R. G., Carroll, M., Dimiceli, C., & Sohlberg, R. A. (2003). Global percent tree cover at spatial resolution of 500 meters: First results of the MODIS vegetation continuous fields algorithm. *Earth Interactions*, *7*, 1–15.

Hicke, J. H., & Logan, J. (2009). Mapping whitebark pine mortality caused by a mountain pine beetle outbreak with high spatial resolution satellite imagery. *International Journal of Remote Sensing*, *30*, 4427–4441.

Ho, H. C., Knudby, A., Sirovyak, P., Xu, Y., Hodul, M., & Henderson, S. B. (2014). Mapping maximum urban air temperature on hot summer days. *Remote Sensing of Environment*, *154*, 38–45.

Hodges, K. E. (2000). Ecology of snowshoe hares in southern boreal and montane forests. In L. F. Ruggiero, K. B. Aubry, S. W. Buskirk, G. M. Koehler, C. J. Krebs, K. S. McKelvey, & J. R. Squires (Eds.), *Ecology and conservation of lynx in the United States* (pp. 163–206). Boulder, CO: University Press of Colorado (480 pp.).

Holling, C. S. (1992). Cross-scale morphology, geometry and dynamics of ecosystems. *Ecological Monographs*, *62*, 447–502.

Homer, C., Huang, C., Yang, L., Wylie, B., & Coan, M. (2004). Development of a 2001 national land-cover database for the United States. *Photogrammetric Engineering & Remote Sensing*, *70*, 829–840.

Jennings, S. B., Brown, N. D., & Sheil, D. (1999). Assessing forest canopies and understory illumination: Canopy closure, canopy cover and other measures. *Forestry*, *72*, 59–73.

Jewett, J., Lawrence, R., Marshall, L., Gessler, P., Powell, S., & Savage, S. (2011). Spatiotemporal relationships between climate and whitebark pine mortality in the Greater Yellowstone Ecosystem. *Forest Science*, *57*, 320–335.

Johansen, K., & Phinn, S. (2006). Mapping structural parameters and species composition of riparian vegetation using IKONOS and Landsat ETM+ data in Australian tropical savannahs. *Photogrammetric Engineering & Remote Sensing*, *72*, 71–80.

Koy, K., McShea, W. J., Leimgruber, P., Haack, B. N., & Aung, M. (2005). Percentage canopy cover — Using Landsat imagery to delineate habitat for Myanmar's endangered Eld's deer (*Cervus eldi*). *Animal Conservation*, *8*, 289–296.

Kuchler, M., Ecker, K., Feldmeyer-Christe, E., Graf, U., & Waser, L. T. (2007). Predictive models of mire habitats: Bias in detection of change. *WETLANDS: Monitoring, modelling and management* (pp. 91–101). London: Taylor and Francis Group.

Lawrence, R. L., Wood, S., & Sheley, R. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*, *100*, 356–362.

Lee, A. C., & Lucas, R. M. (2007). A LiDAR-derived canopy density model for tree stem and crown mapping in Australian forests. *Remote Sensing of Environment*, *111*, 493–518.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2/3*, 18–22.

Morsdorf, F., Kotz, B., Meier, E., Itten, K. I., & Allgower, B. (2006). Estimation of LAI and fractional cover from small footprint airborne laser scanning data based on gap fraction. *Remote Sensing of Environment*, *104*, 50–61.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*, 247–259.

Ozdemir, I. (2014). Linear transformation to minimize the effects of variability in understory to estimate percent tree canopy cover using RapidEye data. *GIScience & Remote Sensing*, *51*, 288–300.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*, 217–222.

Peterson, G., Allen, C. R., & Holling, C. S. (1998). Ecological resilience, biodiversity, and scale. *Ecosystems*, *1*, 6–18.

Pierce, K. B., Jr., 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. *Canadian Journal of Forest Research*, *39*, 1901–1916.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). New classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*, 181–199.

Pu, R., Xu, B., & Gong, P. (2003). Oakwood crown closure estimation by unmixing Landsat TM data. *International Journal of Remote Sensing*, *24*, 4433–4445.

Singh, A. (1989). Review article: Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, *10*, 989–1003.

Smart, S. M., & Scott, W. A. (2004). Bias in Ellenberg indicator values: Problems with detection of the effect of vegetation type. *Journal of Vegetation Science*, *15*, 843–846.

Smith, A. M. S., Falkowski, M. J., Hudak, A. T., Evans, J. S., Robinson, A. P., & Steele, C. M. (2009). A cross-comparison of field, spectral, and lidar estimates of forest canopy cover. *Canadian Journal of Remote Sensing*, *35*, 447–459.

Somers, E. C., Zhao, W., Lewis, E. E., Wang, L., Wing, J. J., Sundaram, B., ... Kaplan, M. J. (2012). Type I interferons are associated with subclinical markers of cardiovascular disease in a cohort of systemic lupus erythematosus patients. *PloS One*, *7*, e37000.

Squires, J. R., DeCesare, N. J., Kolbe, J. A., & Ruggiero, L. F. (2010). Seasonal resource selection of Canada lynx in managed forests of the Northern Rocky Mountains. *Journal of Wildlife Management*, *74*, 1648–1660.

StatSoft, Inc. (2013). *Electronic statistics textbook.* Tulsa, OK: StatSoft (WEB: http://www.statsoft.com/textbook/, last accessed 15 April 2014).

Tottrup, C., Rasmussen, M. S., Eklundh, L., & Jonsson, P. (2007). Mapping the fractional forest cover across the highlands of Southeast Asia using MODIS data and regression tree modelling. *International Journal of Remote Sensing*, *28*, 23–46.

White, P. S. (1979). Pattern, process, and natural disturbance in vegetation. *Botanical Review*, *45*, 229–299.

Yuan, D., & Elvidge, C. D. (1996). Comparison of relative radiometric normalization techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, *51*, 117–126.

Zhao, K., Popescu, S., Meng, X., Pang, Y., & Agca, M. (2011). Characterizing forest canopy structure with LiDAR composite metrics and machine learning. *Remote Sensing of Environment*, *115*, 1978–1996.