1

Modelling vegetation understory cover using LiDAR metrics.

3
4
5

Lisa A. Venier[1*], Tom Swystun[1], Marc J. Mazerolle[2], David P. Kreutzweiser[1], Kerrie L. Wainio-Keizer[1], Ken A. McIlwrick[1], Murray E. Woods[3], Xianli Wang[1]

[1] Great Lakes Forestry Centre, Canadian Forest Service, Natural Resources Canada, Saul Ste. Marie, ON, Canada

[2] Department of Biology, Laval University, Quebec, QC, Canada

[3] Ontario Ministry of Natural Resources and Forestry, North Bay, ON, Canada

*Corresponding Author
E-mail: lisa.venier@canada.ca

21

Abstract

Forest understory vegetation is an important feature of wildlife habitat among other things. Predicting and mapping understory is a critical need for forest management and conservation planning, but it has proved difficult. LiDAR has the potential to generate remotely sensed forest understory structure data, yet this potential has to be fully validated. Our objective was to examine the capacity of LiDAR point cloud data to predict forest understory cover. We modeled ground-based observations of understory structure in three vertical strata (0.5 m to < 1.5 m, 1.5 m to < 2.5 m, 2.5 m to < 3.5 m) as a function of a variety of LiDAR metrics using both mixed-effects and Random Forest models. We compared four understory LiDAR metrics designed to control for the spatial heterogeneity of sampling density. The four metrics were highly correlated and they all produced high values of variance explained in mixed-effects models. The top-ranked model used a voxel-based understory metric along with vertical stratum (Akaike weight = 1, explained variance = 87%, SMAPE=15.6%). We found evidence of occlusion of LiDAR pulses in the lowest stratum but no evidence that the occlusion influenced the predictability of understory structure. The Random Forest model results were consistent with those of the mixed-effects models, in that all four understory LiDAR metrics were identified as important, along with vertical stratum. The Random Forest model explained 74.4% of the variance, but had a lower cross-validation error of 12.9%. Based on these results, we conclude that the best approach to predict understory structure is using the mixed-effects model with the voxel-based understory LiDAR metric along with vertical stratum, but that other understory LiDAR metrics (fractional cover, normalized cover and leaf area density) would still be effective in mixed-effects and Random Forest modelling approaches.

Introduction

45

46   Understory vegetation is an important part of the forested ecosystem. It contributes

47   greatly to nutrient cycling (1, 2), wildlife habitat (3-5), fire behaviour (6-8), microclimate (2) and

48   carbon accounting (9). Understory vegetation communities are therefore often considered a

49   good indicator of forest ecological integrity (10, 11) . However, spatial predictions of understory

50   cover or density have been extremely difficult to generate using traditional variables such as

51   topography, overstory and soils (12). Active remote-sensing technology such as LiDAR (light

52   detection and ranging) could potentially address this issue.

53   LiDAR provides an estimate of three-dimensional forest structure including estimates of

54   canopy structure, understory vegetation and terrain. LiDAR is a survey method that measures

55   the distance to a target (in this case, vegetation) by illuminating the vegetation with a laser light

56   pulse, and measuring the reflected pulses with a sensor. These reflected pulses are called LiDAR

57   returns. Three-dimensional representations of the forest are constructed using laser pulse

58   return times. This capacity has conferred large advantages to forest managers, conservationists

59   and researchers in their attempts to manage the forest efficiently and sustainably. LiDAR can

60   generate reliable, robust estimates of many forest structure variables including canopy height

61   and cover (13-15), as well as basal area and tree density (13, 16) and has similar potential for

62   understory structure.

63   To date, relatively few studies have evaluated the potential of LiDAR to describe

64   understory structure by comparing ground-based measures of understory structure and LiDAR

65   data (17-20). In each study, different LiDAR metrics were used with a variety of covariates,

66   analytical approaches, and forest types to test predictions of understory cover or density. There

4

67   is a large discrepancy in the success of the various LiDAR metrics in producing reliable

68   predictions of understory. Our objective in this paper is to evaluate the potential of LiDAR to

69   generate predictions of understory cover by comparing to field measures of understory. To

70   achieve this objective, we examine alternative LiDAR metrics that control for spatial

71   heterogeneity of sampling density, we compare regression and machine learning statistical

72   approaches, and we examine the value of multiple variables in our models.

73   A key challenge of working with LiDAR data is that there is a large amount of spatial

74   heterogeneity in the sampling density over space that occurs in the normal course of

75   generating LiDAR point clouds. This spatial heterogeneity is due to variations in scan angle,

76   flight height, movement of the aircraft during data collection and the degree of overlapping

77   flight lines. Thus, relative measures of vegetation density or cover, where the number of

78   returns in a vertical stratum are scaled relative to some measure of sampling density, should

79   provide better estimates of true understory vegetation cover. A variety of approaches have

80   been used to relativize these measures, for example, dividing the number of returns in a

81   vertical bin by the total number of returns in the column, or by the number of returns in the bin

82   and below the bin (20). We examine four different understory structure metrics based on

83   different approaches to control for sampling density.

84   We explored two statistical approaches for modelling understory vegetation structure

85   as a function of LiDAR data: machine learning and mixed effects regression models. Machine

86   learning, specifically random forest, has been used to model forest inventory variables with a

87   large suite of LiDAR derived predictors (18, 21). Machine learning in this context strives to

88   produce the best prediction of the forest inventory variables. However, machine learning does

89   not produce an ecologically interpretable relationship per se, only estimates of variable

90   importance. Machine learning makes no assumptions about the structure of the data, is ideal

91   for predicting relationships that are non-linear, is insensitive to correlations among variables,

92   and interactions are automatically modeled. However, machine learning is prone to bias

93   associated with incomplete ranges of conditions being sampled. As an alternative, we explored

94   linear mixed-effects regression models. These models make assumptions of homoscedasticity

95   and normality of errors which must be checked but can produce more parsimonious and more

96   interpretable models than machine learning in some instances. In random forest models, large

97   suites of variables are usually included to achieve the best predictive capacity. In the regression

98   models, it is more important to limit the number of variables included to avoid overfitting and

99   strong correlations between explanatory variables.

100   Occlusion has been discussed in the literature as a possible issue limiting LiDAR

101   effectiveness for prediction of understory structure (22, 23), but more recent studies have

102   shown that the potential occlusion may not interfere with generating predictions. Latifi et al.

103   (18) demonstrated that artificially reducing the density of the LiDAR point cloud did not have an

104   appreciable effect on variance explained in models predicting understory structure. In another

105   study, prediction errors of understory vegetation cover were not related with canopy cover

106   (17). However, forest type in some instances can influence the predictive accuracy of models

107   (19). In both of our modelling approaches, we included additional variables beyond the

108   understory LiDAR metrics that may influence the amount of occlusion of the laser pulse,

109   namely, the amount of overstory, the forest type, and the vertical stratum. All three of these

110   variables could reflect the amount of vegetation in the area above the vertical stratum of

111   interest.

112   Our primary objective is to quantify the capacity of LiDAR to estimate understory

113   structure. To achieve this, 1; we compare the effectiveness of four possible understory LiDAR

114   metrics for predicting understory cover that control for sampling density, 2; we examine the

115   influence of potentially important additional explanatory variables on the model which will

116   inform us about the importance of occlusion, and 3; we compare the mixed effects vs random

117   forest approach for generating predictions. Our aim is to generate robust and effective

118   predictions of understory cover that could inform forest management and conservation.

119

120

121   **Methods**

122   **Study area**

123   This project was conducted in the Petawawa Research Forest. The research forest

124   covers 9,945 hectares in the Great Lakes-St. Lawrence forest region (45º 58' 46.74" N, 77º 30'

125   22.11" W), Ontario, Canada. The study area is on the Southern end of the Precambrian Shield,

126   on bedrock of granites and gneisses. Forest composition features White Pine (*Pinus strobus*

127   Linnaeus), Red Pine *(Pinus resinosa* Aiton), Red Oak *(Quercus rubra* Linnaeus), Yellow Birch

128   (*Betula alleghaniensis* Britton), Sugar Maple *(Acer saccharum* Marshall), and Red Maple *(Acer*

129   *rubrum* Linnaeus) as dominant species, often in uneven-aged forests. Presently, the Petawawa

130   Research Forest is dominated by healthy but mature and overmature overstory (80-140 years)

131   coupled primarily with low-quality regeneration and understories. For the purpose of the

132   current study, we classified the forest into four types (TYPE) to explore the influence of forest

133   type on the consistency of the relationship between field measured understory vegetation

134   structure and LIDAR metrics. The four forest type classes (TYPE) are Pine, Red Oak, Mixedwood

135   without Pine, and Mixedwood with Pine. These four classes account for approximately 71% of

136   the landbase of the research forest.

137

Field data collection

139   Within the Petawawa Research Forest, plots were selected from a 25 m-resolution

140   rasterized LiDAR database and Forest Resource Inventory data based on aerial photo

141   interpretation. Potential plots were selected based on a stratification by forest type, overstory

142   density and understory density. Initial overstory was measured as the relative number of LiDAR

143   laser pulse returns in overstory (> 4 m), and understory density as the relative number of LiDAR

144   laser pulse returns 4 m or lower. We divided the full range of overstory values into 10 equal

145   bins, and the full range of understory values into 10 equal bins. For each combination of

146   understory by overstory bin we selected five potential plots for each of four forest types, for a

147   total of 2000 plots, 500 in each 10 by 10 matrix, with one matrix per each forest type. This is a

148   rough stratification but helped to fill the statistical space to ensure optimal conditions for

149   model construction. We sampled 437 plots out of the possible 2000, trying to select 1-5 plots

150   from all cells in the matrix.

151   We collected vegetation data on 250 plots in 2015 and on an additional 187 plots in

152   2016. Plots were selected in the field from the list of preselected plots based on accessibility

8

153   and conformity with classified forest type, understory and overstory. At each plot centre, we

154   used an SX Blue II GPS to generate a sub-meter accurate location through averaging a minimum

155   number of 300 points. Our field data collection attempted to generate a field-based point

156   cloud to match the LiDAR based point cloud. We measured forest structure on ground-based

157   plots in nine vertical strata (0-0.5 m, 0.5-1 m, 1-1.5 m, 1.5-2 m, 2-2.5 m, 2.5-3 m, 3-3.5 m, 3.5-

158   4.0 m, > 4 m). From the centre point we created eight radial transects (12 m in length each)

159   starting in a north direction and moving clockwise by 45 degrees for each additional transect.

160   Along each transect, data were collected at each meter for a total of 97 sample locations in

161   each plot, including the centre point (Fig 1). To sample the vegetation structure, observers

162   recorded the presence or absence of vegetation within a 15 cm circle for each of the nine

163   vertical strata. Thus, there were 97 sampling points x 9 strata = 873 presence/absence points

164   collected in each 12 m radius plot volume. The original vertical strata were later grouped into

165   three strata (S1 = 0.5-1.5 m, S2 = 1.5-2.5 m, S3 = 2.5-3.5 m). We excluded points below 0.5 as

166   they are difficult to distinguish from ground points. We excluded points above 3.5m as they

167   were difficult to estimate from the ground. The total number of vegetation presences in each

168   stratum (0-194) were recorded in the FIELD variable for subsequent analysis. This field

169   collection would represent a lower sampling density than the LiDAR data which are at 6 pulses

170   per square meter with up to 8 returns per pulse which resulted in 2.44 returns per m^3

171   compared to the field data with 0.43 returns per m^3. These data are not strictly comparable

172   since the field data represent presence and absence whereas the LiDAR returns represent only

173   presence but give a general impression of relative sampling density.

174

175 **Fig 1. Sampling design for FIELD observations of vegetation structure. r is the radius of the**
176 **measurement area around each point on the transect.**
177

LiDAR acquisition

179   Airborne LiDAR data were collected over the Petawawa Research Forest from August

180 17-20, 2012. The Riegl 680i sensor was carried aboard a Cessna 172 aircraft flown at an

181 average altitude of 750 m. Technical acquisition specifications are provided in Table 1. The data

182 were collected as a full-waveform and provided as a discrete point file (LAS 1.1) for use in this

183 project. Flight overlap was approximately fifty percent.

184
185 **Table 1. Airborne LiDAR acquisition specifications.**
186

Parameter	Value
Pulse repetition rate	150Khz
Frequency	76.67Hz
Scan Angle	± 20 Degrees
FOV	40 Degrees
Line spacing: Cross track	0.6 m
Line spacing: Along track	0.6 m
Line spacing between flight lines	250 m
Laser footprint min:	0.38 m
Laser footprint max	0.42 m
Average point density: All Returns	~ 15 pts/m^2
Average point density: Last Returns	~ 6 pts/m^2

187
188

Data processing and LiDAR variables

190   We developed specific LiDAR understory cover metrics that are expected to capture the

191 vegetation understory density directly. We identified four metrics for our analysis. Three of

192 these metrics are used in the literature: fractional cover (FRAC, modified from Wing et al. (17)),

193 leaf area density (LAD, (24)), and voxel cover (VOX1m, (25)). The fourth metric considered was

194   normalized cover (NORM), because it is an easily interpretable and easily calculated alternative.

195   Fractional cover is calculated by summing the number of LiDAR vegetation returns for each

196   understory vertical stratum and dividing by the sum of understory and ground returns. Leaf

197   area density is calculated as the negative log of the number of returns in a vertical stratum

198   divided by all returns in and below the vertical bin and then divided by a constant. Normalized

199   cover is calculated by dividing all vegetation returns in the understory stratum divided by all

200   first returns. The voxel cover approach filters all returns by estimating presence/absence of

201   returns in each standard voxel (in our case 1 m^3) in the vertical stratum. For example, a 2 m x 5

202   m x 5 m vegetation stratum that contains 50 1-m^3 voxels would have a voxel cover value

203   between 0 and 50, equal to the number of voxels that contain vegetation. Sampling density is

204   extremely heterogeneous due to different factors such as flight line overlap and the pitch and

205   yaw of the plane. The LiDAR metrics provide four alternative ways to scale the number of

206   returns in a vertical bin by sampling density. In addition to these four specific LiDAR understory

207   cover metrics, we calculated a suite of standard LiDAR point cloud metrics such as canopy cover

208   and canopy height (S1 Table).

209

Analysis

211   We used linear mixed effects models to determine the capacity of our four main LiDAR

212   understory cover metrics to predict understory cover recorded in the field (FIELD) in each of the

213   three vertical strata defined above (ST1, ST2, ST3), and to examine the influence of secondary

214   explanatory variables(26). These secondary explanatory variables consisted of forest TYPE

215   (based on overstory composition), STRATUM (vertical 1 m strata, ST1-ST3), and OVERSTORY

216   (Appendix 1). The OVERSTORY variable was a measure of LiDAR vegetation cover in the vertical

217   column above the stratum of interest calculated by classifying canopy cover (CC) into three

218   classes (low, medium, high). We treated the plot as a random effect to account for multiple

219   measurements in each plot. We formulated 16 candidate models consisting of LiDAR variables,

220   with the constraint of maintaining variance inflation factors (VIF) < 10 to avoid issues of

221   multicollinearity (Table 2). For each the four main LiDAR metric, we derived four models: 1) a

222   null model consisting only the LiDAR metric, 2) a model with the LiDAR metric, OVERSTORY and

223   their interaction, 3) a model with the LiDAR metric, TYPE, and their interaction, and 4) a model

224   with the LiDAR metric, STRATUM, and their interactionWe ranked all mixed effects models

225   based on Akaike's information criterion (AIC, (27, 28)) and calculated the R^2 values. We also

226   computed the symmetric mean absolute percentage error (SMAPE), based on 10-fold cross-

227   validation (29) for the top-ranked models, and calculated SMAPE values for each of the 3

228   vertical strata separately. Parameters of the mixed effects models were estimated by maximum

229   likelihood in R with the nlme package(18, 26, 30).

230
231   **Table 2: Mixed effects model explaining understory cover recorded in the field (FIELD)** : TYPE
232   = forest type based on overstory composition, STRATUM = vertical 1 m strata, S1-S3, and
233   OVERSTORY = a measure of LiDAR vegetation cover in the vertical column above the stratum of
234   interest calculated by classifying canopy cover (CC) into three classes (low, medium, high), see
235   Appendix 1. The plot was treated as a random effect in each model.
236

Model Name	Model fixed effects structure	Biological interpretation
FRAC null	FRAC	Relationship between FRAC and FIELD is constant
FRAC * STRATUM	FRAC + STRATUM + FRAC*STRATUM	Relationship between FRAC and FIELD differs among STRATUM
FRAC * OVERSTORY	FRAC + OVERSTORY + FRAC*OVERSTORY	Relationship between FRAC and FIELD differs among OVERSTORY

12

FRAC * TYPE	FRAC + TYPE + FRAC*TYPE	Relationship between FRAC and FIELD differs among TYPE
NORM null	NORM	Relationship between NORM and FIELD is constant
NORM * STRATUM	NORM + STRATUM + NORM*STRATUM	Relationship between NORM and FIELD differs among STRATUM
NORM * OVERSTORY	NORM + OVERSTORY + FRAC*OVERSTORY	Relationship between NORM and FIELD differs among OVERSTORY
NORM * TYPE	NORM + TYPE + FRAC*TYPE	Relationship between NORM and FIELD differs among TYPE
VOX1m null	VOX1m	Relationship between VOX1m and FIELD is constant
VOX1m * STRATUM	VOX1m +STRATUM + VOX1m*STRATUM	Relationship between VOX1m and FIELD differs among STRATUM
VOX1m * OVERSTORY	VOX1m + OVERSTORY + VOX1m*OVERSTORY	Relationship between VOX1m and FIELD differs among OVERSTORY
VOX1m * TYPE	VOX1m + TYPE + VOX1m*TYPE	Relationship between VOX1m and FIELD differs among TYPE
LAD (null)	LAD	Relationship between LAD and FIELD is constant
LAD * STRATUM	LAD + STRATUM + LAD*STRATUM	Relationship between LAD and FIELD differs among STRATUM
LAD * OVERSTORY	LAD + OVERSTORY + LAD*OVERSTORY	Relationship between LAD and FIELD differs among OVERSTORY
LAD * TYPE	LAD + TYPE + LAD*TYPE	Relationship between LAD and FIELD differs among TYPE

237
238
239

240

241   We used random forest with the same FIELD response variable as in the mixed-effects models

242   described above. Because random forests are non-parametric and do not yield a log-likelihood,

243   we ran a stepwise procedure with 341 LiDAR derived variables (which includes overstory

244   estimates) (S1 Table), plus secondary variables forest TYPE (from Forest Resource Inventory),

245   and STRATUM. We used mean decrease in accuracy to rank variable importance (31). At each

246   iteration, we removed the 20% least influential variables and compared the explained variance.

247   Models were built using the randomForest package in R (31). We examined the importance of

248   variables in the suite of random forest models. Similar to the mixed effects models above, we

249   quantified model performance with the percent variance explained and SMAPE based on 10-

250   fold cross-validation. Finally, we compared the prediction performance of the mixed effects and

251   random forest approaches.

252

253   ## Results

254   ### Relationship among LiDAR metrics

255   The FIELD measure of understory cover was strongly correlated with all of the four main

256   LiDAR metrics we investigated (Fig 2a-d). However, the FRAC and VOX1m metrics appeared to

257   be the most linearly related to the FIELD measure (Fig 2a-d). Nonetheless, the four understory

258   vegetation metrics were all highly correlated with one another (Table 3).

259   **Fig 2. Scatterplot of FIELD (measured density) against the LiDAR metrics**, a) fractional cover
260   (FRAC), b) normalized cover (NORM), c) leaf area density (LAD), and d) voxel cover (VOX1m),
261   including Pearson product-moment correlation coefficients.
262

263 **Table 3. Pearson product-moment correlations between pairs of understory cover LiDAR**
264 **metrics included in analysis (n = 1310).**

265

Correlation	r	Lower 95% CL	Upper 95% CL
FRAC vs NORM	0.77	0.751	0.794
FRAC vs VOX1m	0.84	0.819	0.852
FRAC vs LAD	0.77	0.744	0.789
NORM vs LAD	0.81	0.79	0.827
NORM vs VOX1m	0.92	0.911	0.927
VOX1m vs LAD	0.79	0.767	0.808

266

Mixed-effects models

268   The model consisting of the voxel-based cover estimate (VOX1m) with STRATUM and

269   their interaction was the most parsimonious among all sixteen models considered (Table 4).

270   This model had all the support (Akaike weight = 1, Table 4, Fig 3). This model also had the

271   highest conditional R^2 (along with the FRAC + STRATUM + interaction model, although all

272   sixteen models had high R^2 values (0.71-0.87). For each the four LiDAR metrics we considered,

273   we observed the same pattern: the addition of STRATUM and the interaction to the null models

274   resulted in consistently better model performance in terms of delta AIC and R^2. The addition of

275   OVERSTORY or TYPE resulted in much less model improvement than the addition of STRATUM.

276   The model with most support did not include forest type or overstory, which is important since

277   forest type was derived from forest inventory data and cannot be extracted from LiDAR point

278   clouds.

279

280   **Table 4. R^2 and AIC values for sixteen candidate linear mixed-effects models.** Note that
281   marginal R^2 denotes the percent variance explained by the fixed effects, whereas the
282   conditional R^2 includes both fixed effects and random effects. Delta AIC is the difference

283   between each model relative to the most parsimonious model and Akaike weight indicates the

284   percent support of a given model. .

285

Model	Marginal R^2	Conditional R^2	AIC	Delta AIC	Akaike weight
VOX1m * STRATUM	0.62	0.87	11868.87	0	1
FRAC * STRATUM	0.65	0.87	11901.00	32.13	0
LAD * STRATUM	0.56	0.82	11998.29	129.42	0
NORM * STRATUM	0.52	0.83	12099.16	230.29	0
VOX1m * OVERSTORY	0.60	0.82	12348.32	479.45	0
LAD * OVERSTORY	0.51	0.73	12384.88	516.01	0
VOX1m * TYPE	0.60	0.82	12384.88	516.01	0
VOX1m null	0.60	0.82	12385.78	516.91	0
LAD * TYPE	0.51	0.72	12396.42	527.55	0
LAD null	0.50	0.71	12407.11	538.24	0
NORM * OVERSTORY	0.53	0.75	12450.66	581.79	0
NORM * TYPE	0.51	0.75	12563.97	695.1	0
NORM null	0.49	0.75	12568.4	699.53	0
FRAC * OVERSTORY	0.58	0.77	12585.04	716.17	0
FRAC * TYPE	0.57	0.75	12613.19	744.32	0
FRAC null	0.56	0.75	12617.05	748.18	0

286

287   **Figure 3: Predicted versus observed scatterplot.** Predictions of FIELD generated from mixed-

288   effects model consisting of VOX1m + STRATUM + interaction.

289

290   In all of the mixed effects models, the four LiDAR metrics had positive slopes (Fig 4,

291   Table 5, for example). In our best model, the intercept of the lowest STRATUM was higher than

292   in the upper strata (Fig 4). Although the model included the interaction between STRATUM and

293   voxel cover, there was no evidence of different slopes of LiDAR among strata (Fig 4, Table 5).

294   Symmetric mean absolute percentage (SMAPE) errors for the top-ranked mixed effects model

295   was 0.156, but these values varied when investigating each stratum separately (Table 6).

296   Contrary to expectations, the SMAPE value was lowest for the lowest strata (0.107) and

297   greatest for the highest strata (0.190). There were 437 observations for each stratum.

298

299 **Figure 4: Predictions of FIELD for each of three strata based on the mixed-effects model**
300 **consisting of VOX1m + STRATUM + interaction.** Dashed lines around solid lines denote 95%
301 confidence intervals around predictions.
302

303 **Table 5. Estimates of the best supported mixed-effects model consisting of VOX1m +**
304 **STRATUM + interaction and a random effect of plot.**
305

	Estimate	Lower 95% CL	Upper 95% CL
intercept	64.35	60.25	68.46
LIDAR	0.03	0.29	0.32
STRATUM.ST2	-21.94	-25.96	-17.98
STRATUM.ST3	-29.38	-33.48	-25.28
LIDAR*STRATUM.ST2	-0.016	-0.039	0.008
LIDAR*STRATUM.ST3	-0.010	-0.037	0.017

306
307

308 **Table 6. Ten-fold cross-validation results from top linear mixed-effects model and the**
309 **selected random forest model, based on symmetric mean absolute percentage error**
310 **(SMAPE).** Note that average values of SMAPE are given for predictions of all STRATUM levels,
311 but also for predictions specific to STRATUM levels.
312

Model		SMAPE mean	SMAPE sd (n=10)
VOX1m * STRATUM	predictions of all STRATUM levels	0.156	0.014
	predictions of STRATUM 1	0.107	0.016
	predictions of STRATUM 2	0.170	0.024
	predictions of STRATUM 3	0.190	0.020
Random forest (59 predictors)		0.129	0.015

313
314
315

Random forest models

316

317   We examined the percent variance explained and the number of variables included to

318   choose a final random forest model. The base model with all 341 LiDAR-derived variables,

319   forest TYPE, and STRATUM explained 74.8% of the variance, but the final model with only 59

320   predictors had a very similar variance explained (74.4%) (Fig 5, Table 7). The 10-fold cross-

321   validation on this reduced model showed an overall mean error rate of 0.129 (Table 6).

322

323   **Fig 5. Predicted versus observed scatterplot for Random Forest model with 59 predictors.**
324

325   Some variables appeared more often than others among the 18 random forest models

326   considered. These variables consisted of STRATUM, GAP (the inverse of LAD), and LAD. In

327   addition, most or all of the LiDAR understory vegetation cover metrics (VOX1m, FRAC, NORM)

328   were represented in the top 10 variables of most of the 18 potential models (S2 Table). Crown

329   closure (CC), an estimate of overstory, was also often among the top 10 most important

330   variables within the models considered. Forest TYPE never occurred among the top 10 variables

331   (S2 Table).

332
333   **Table 7. Random forest models: mean squared residuals and percent variance explained.**
334

Number of Predictors in model	Mean Squared Residuals	Percent variance Explained
341 (Base model)	484	74.8
276	485	74.7
223	485	74.8
180	484	74.7
145	476	75.2
116	486	74.7
93	481	75.0

74	492	74.3
59	490	74.4
47	513	73.3
37	508	73.5
29	531	72.4
22	553	71.2
17	528	72.5
13	558	70.9
10	580	69.8
7	569	70.4
5	632	67.1

335

Discussion

337   In this study, our primary objective was to quantify the capacity of LiDAR to estimate

338 understory structure so that it can be predicted across a landscape. To address this objective,

339 first we compared the effectiveness of four possible understory LiDAR metrics (fractional cover,

340 leaf area density, voxel cover, and normalized cover) for predicting understory cover. Each of

341 these metrics used some measure of the number or presence of LiDAR returns in an understory

342 vertical stratum and standardized these measures with an estimate of sampling density. All four

343 LiDAR metrics were effective at predicting the amount of structure in an understory stratum,

344 but the best metric based on mixed effects modelling was the voxel-based cover estimate

345 (VOX1m) with the addition of STRATUM with a conditional R^2 of 0.87. The voxel-based

346 approach is relatively easy to calculate and provides a direct measure of the amount of

347 understory structure.

348   We anticipated that other variables could influence the predictions of understory. We

349 identified three potentially important variables that might influence occlusion of understory

350 structure: overstory, forest type and stratum. Increased overstory can reduce the ability of

351   LiDAR to predict understory structure due to occlusion (22, 23). For LiDAR to detect the

352   understory structure, LiDAR pulses must reach and be reflected by understory vegetation. A

353   greater vegetation interception above the area of interest will result in fewer pulses returning

354   from the understory. Both forest type and stratum will also influence the amount of vegetation

355   in the area above the area of interest and therefore potentially alter the relationship of field

356   measured and LiDAR measured understory.

357   Correlations between the three secondary explanatory variables (STRATUM, forest

358   TYPE, and OVERSTORY) made it impossible to include all variables in a single model. Our best

359   model included STRATUM, where we found that the lowest stratum had the highest intercept.

360   This is consistent with occlusion in that we have more vegetation in ST1 than ST2 and ST3 for a

361   given value of VOX1m. This is consistent with the idea that fewer laser pulses are reaching the

362   lower stratum. The relationship between the field observed structure and VOX1m did not vary

363   with STRATUM. Surprisingly, we found that the error in the predicted relationship was greatest

364   in the highest STRATUM and lowest in the lowest STRATUM suggesting that there was no

365   reduction in predictability associated with potential occlusion. This may be due to the

366   possibility that the understory vegetation in the lower stratum is easier to estimate on the

367   ground and therefore there is less noise in the relationship between the field and the LiDAR

368   measures in the lower stratum. Either way, we conclude that our LiDAR sampling intensity was

369   sufficient in our forest system to capture the understory structure regardless of the density of

370   vegetation above the area of interest and the related potential for occlusion.

371   There is some discrepancy in the literature on the effect of occlusion. Latifi et al. (19)

372   found that thinning LiDAR data by artificially reducing the sampling density did not impact the

373   effectiveness of models to predict understory. Their original data had a high point density of 30-

374   40 points per m^2 and a maximum of 11 returns. Data were thinned to two different levels but

375   Latifi et al. (19) do not report on the final point density after thinning. Our data are at roughly

376   11.69 vegetation returns per m^2, with about 0.55 vegetation returns per m^3 in the 0.5-4m

377   understory stratum. Obviously, the effectiveness of LiDAR to capture understory structure will

378   eventually be undermined by a sufficient reduction in sampling density, but this limit does not

379   seem to have been reached in the Petawawa research forest. Gonzalez-Ferreiro et al. (32)

380   showed that reducing pulse density from 8 pulses per m^2 to 0.5 pulses per m^2, did not decrease

381   model precision in estimating stand variables. Wing et al. (17) found no trends between

382   understory vegetation cover prediction error and canopy cover, lending support to the idea

383   that under some natural overstory conditions and common LiDAR sampling densities, occlusion

384   is not an issue for predicting understory with LiDAR. In contrast, Ruiz et al. (33) reported an

385   effect of LiDAR sampling density on model R^2 values but only at levels below around 5

386   points/m^2. It is unclear how this number translates into pulses reaching the understory. The

387   lack of influence of forest type on understory cover predictions enables predicting understory

388   from LiDAR alone without relying on traditional forest resource inventory data.

389   The comparisons of mixed effects and random forest models revealed some obvious

390   alignment. All four of the LiDAR metrics considered (fractional cover, leaf area density,

391   normalized cover, and voxel cover) produced models with high R^2 values. All four of these

392   variables also had very high variable importance in the random forest models. The stratum

393   variable appeared often in the top random forest models and was also important in the top-

394   ranked mixed-effects model (VOX1m * STRATUM). The random forest model had a high

395   variance explained (75%), but not as high as the best mixed effects model that included the

396   voxel-based measure of cover (87%). Our selected random forest model had 59 explanatory

397   variables, whereas the best mixed effects model had two explanatory variables and their

398   interaction, as well as a random effect of plot. Based on our results, generating landscape-wide

399   predictions using the mixed-effects model should be simpler and more efficient than with the

400   random forest model. For these reasons, although the random forest model is effective, we

401   recommend the mixed effects model as the better choice for predicting understory vegetation

402   structure with LiDAR.

403   Direct evaluations of LiDAR metrics to capture understory cover are relatively rare.

404   Studies have shown good agreement between field and LiDAR measures of forest stand

405   biomass (34, 35), but biomass is likely driven primarily by tree biomass rather than understory.

406   Asner et al. (36) explored structural transformation of rain forests due to invasive plants and

407   used LiDAR to estimate structural changes in the understory. However, Asner et al. (36) did not

408   report quantitative comparisons of field and LiDAR measures. Martinuzzi et al. (37) produced

409   classification accuracies of 83% in predicting the presence of shrubs, but not their abundance.

410   Wing et al. (17) compared understory vegetation cover and airborne LiDAR estimates with the

411   addition of a filter for intensity values in an interior ponderosa pine forest. Their models had R^2

412   values from 0.7 to 0.8 and accuracies of ± 22%. Our models achieved slightly higher R^2 with

413   slightly lower error rates without the use of the intensity filter, suggesting that the latter filter

414   may not always be necessary to generate good estimates. As well, the intensity filter is affected

415   by a number of factors such as elevation and the nature of the object intercepted that are

416   difficult to normalize, so we prefer models that do not require intensity filters. Latifi et al. (19)

417   also made a direct comparison of ground-based vs LiDAR estimates of understory cover in

418   temperate mixed stands, and found strong relationships in the top canopy and the herbal layer

419   with lower predictive power in the intermediate stand layers. Their shrub layer regression

420   model had a relatively low R^2 value of 37%. In a later study, Latifi et al. (18) showed an R^2 of

421   80% for the shrub layer based on thinned LiDAR point clouds and new analytical methods.

422   Campbell et al. (20) also compared field and LiDAR measures of understory directly in

423   mixedwood forests and generated an R^2 of 0.44 based on a relative point density similar to

424   metrics that we used here. It is unclear why there is so much variation in the ability of LiDAR to

425   predict understory structure but it suggests that we should be somewhat cautious in assuming

426   that individual LiDAR metrics are capturing the understory structure. It is important to note

427   that some of the error in prediction in our models is likely the result of the lag between the

428   LiDAR acquisition (2012) and the field data acquisition (2016-2017). This lag is likely to result in

429   the most error in the youngest stands where changes in herb and shrub growth are likely to be

430   greatest. The majority of stands included in the analysis are mature forest, and even with this

431   source of error our ability to predict was good.

432   Despite the limited work directly evaluating LiDAR measures of understory vegetation

433   structure, many studies have explored the use of LiDAR to capture wildlife habitat structure

434   some of which is related to understory (38-42). One of the most commonly reported

435   relationships is between vegetation structural diversity or understory density and wildlife

436   diversity (5, 43-45). In addition, vegetation understory structure explained bird species

437   composition in a number of studies (5, 46, 47). Melin et al. (48) found that a LiDAR metric

438   similar to fractional cover to estimate shrub density below 5 m was a good predictor of grouse

439   brood occurrence in Finland, consistent with expectations based on known habitat preferences

440   of the species. However, they did not test the assumption that the LiDAR metric effectively

441   estimates vegetation density below 5 m. All of these studies do however, provide indirect

442   evidence for the effectiveness of LiDAR estimates to predict understory cover or density.

443

Conclusions

445   Based on the highest variance explained, the fewest number of explanatory variables,

446   and ease of interpretation and application, we would recommend using the mixed-effects

447   model consisting of voxel-based cover estimate, stratum, and their interaction to generate

448   spatial estimates of understory cover. Nonetheless, all four LiDAR metrics that we considered

449   and both analytical approaches (mixed effects models, random forests) produced predictions

450   suitable for many ecological and forest planning applications. This information could improve

451   spatially-explicit mapping of wildlife habitat, fire behaviour, or forest ecosystem dynamics.

452   Measuring understory cover *in situ* is not difficult, but many applications require maps or

453   spatial estimates of attributes for forest management and conservation applications over large

454   areas. LiDAR remote sensing is the most efficient approach to generating these spatial

455   estimates of forest attributes. Our results fully support the indirect evidence provided from

456   wildlife studies that LiDAR can predict understory vegetation structure even in the presence of

457   a mature tree canopy. With error percentages of around 15%, these spatial predictions will

458   introduce some uncertainty into predictions, which should be factored into decision-making.

459

Acknowledgements

461   Nicholas Coops and Piotr Tompalski provided guidance and training on LiDAR data

462   handling. Peter Arbour and staff at the Petawawa Research Forest provided logistic support.

463

References

465   1.   Yarie J. The role of understory vegetation in the nutrient cycle of forested ecosystems in
466   Mountain Hemlock Biogeoclimatic Zone. Ecology. 1980;61:1498-514.
467   2.   Nilsson M-C, D.A. W. Understory vegetation as a forest ecosystem driver: evidence from the
468   Northern Swedish boreal forest. Front Ecol Environ. 2005;3:421-8.
469   3.   MacArthur RH, MacArthur JW. On bird species diversity. Ecology. 1961;42:594-8.
470   4.   Venier LA, Pearce JL. Boreal forest landbirds in relation to forest composition, structure, and
471   landscape: Implications for forest management. Can J For Res. 2007;37(7):1214-26.
472   5.   Lesak AA, Radeloff VC, Hawbaker TJ, Pidgeon AM, Gobakken T, Contrucci K. Modeling forest
473   songbird species richness using LiDAR-derived measures of forest structure. Remote Sens Environ.
474   2011;115:2823-35.
475   6.   Bessie WC, Johnson EA. The relative importance of fuels and weather on fire behavior in
476   subalpine forests. Ecology. 1995;76:747-62.
477   7.   Call PT, Albini FA. Aerial and surface fuel consumption in crown fires. Int J Wildland Fire.
478   1997;7:259-64.
479   8.   Hély C, Bergeron Y, Flannigan MD. Effects of stand composition on fire hazard in mixed-wood
480   Canadian boreal forest. J Veg Sci. 2000;11:813-24.
481   9.   Roxburgh SH, Karunaratne SB, Paul KI, Lucas RM, Armston D, Sun J. A revised above-ground
482   maximum biomass layer for the Australian continent. For Ecol Manage. 2019;432:264-75.
483   10.   Kerns BK, Ohmann JL. Evaluation and predicition of shrub cover in coastal Oregon forests (USA).
484   Ecol Indic. 2004;4:83-98.
485   11.   Suchar VA, Crookston NL. Understory cover and biomass indices predictions for forest
486   ecosystems of Northwestern United States. Ecol Indic. 2010;10:602-9.
487   12.   Eskelson BNI, Madsen L, Hagar JC, Temesgen H. Estimating riparian understory vegetation cover
488   with beta regression and copula models. Forest Sci. 2011;57:212-21.
489   13.   Lim K, Treitz P, Baldwin K, Morrison I, Green J. LiDAR remote sensing of biophysical properties of
490   tolerant northern hardwood forests. Can J Remote Sens. 2003;29:658-78.
491   14.   Naesset E. Practical large-scale forest stand inventory using a small-footprint airborne scanning
492   laser. Scand J Forest Res. 2004;19:164-79.
493   15.   Thomas V, Treitz P, McCaughey JH, Morrison I. Mapping stand-level forest biophysical variables
494   for a mixedwood boreal forest using lidar: an examination of scanning density. Can J For Res.
495   2006;36:34-47.
496   16.   Woods M, Lim K, Treitz P. Predicting forest stand variables from LiDAR data in the Great Lakes-
497   St. Lawrence Forest of Ontario. Forest Chron. 2008;84:827-39.

25

498 17. Wing BM, Ritchie MW, Boston K, Cohen WB, Gitelman A, Olsen MJ. Prediction of understory
499 vegetation cover with airborne LiDAR in an interior ponderosa pine forest. Remote Sens Environ.
500 2012;124:730-41.
501 18. Latifi H, Hill S, Schumann B, Heurich M, Dech S. Multi-model estimation of understory shrub,
502 herb and moss cover in temperate forest stands by laser scanner data. Forestry. 2017;90:496-514.
503 19. Latifi H, Heurich M, Hartig F, Müller J, Krzystek P, Jehl H, et al. Estimating over- and understorey
504 canopy density of temperate mixed stands by airborne LiDAR data. Forestry. 2016;89:69-81.
505 20. Campbell MJ, Dennison PEH, A., Parham LM, Butler BW. Quantifying understory vegetation
506 density using small-footprint airborne lidar. Remote Sens Environ. 2018;215:330-42.
507 21. Penner M, Pitt DG, Woods ME. Parametric vs. nonparametric LiDAR models for operational
508 forest inventory in boreal Ontario. Can J Remote Sens. 2013;39(5):426-43.
509 22. Hill RA, Broughton RK. Mapping the understory of deciduous woodland from leaf-on and leaf-off
510 airborne LiDAR data: a case study in lowland Britain. ISPRS J Photogramm. 2009;64:223-33.
511 23. Morsdorf F, Marell A, Koetz B, Cassagne N, Pimont F, Rigolot E, et al. Discrimination of
512 vegetation strata in a multi-layered Mediterranean forest ecosystem using height and intensity
513 information derived from airborne laser scanning. Remote Sens Environ. 2010;114:1404-15.
514 24. Bouvier M, Durrieu S, Rounier RA. Generalizing predictive models of forest inventory attributes
515 usin an area-based approach with airborne LiDAR data. . Remote Sens Environ. 2015;156:322-34.
516 25. Kim E, Lee W-K, Yoon M, Lee J-Y, Son Y, Abu Salim K. Estimation of Voxel-Based Above-Ground
517 Biomass Using Airborne LiDAR Data in an Intact Tropical Rain Forest, Brunei. Forests. 2016;7(11):259.
518 26. Pinheiro J, Bates D, DebRoy S, Sarkar D, RCoreTeam. nlme: Linear and Nonlinear Mixed Effects
519 Models. R package version 3.1-137
520 ed2018.
521 27. Burnham KP, Anderson DR. Model Selection and Multimodel Inference: a practical information-
522 theoretic approach. 2 ed. New York: Springer-Verlag; 2002.
523 28. Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear
524 mixed effects models. Methods Ecol Evol. 2013;4:133-42.
525 29. Gneiting T. Making and evaluating point forecasts. Journal of American Statistical Association.
526 2011;106:746-62.
527 30. RCoreTeam. R: A language and environment for statistical computing. . R Foundation for
528 Statistical Computing. Vienna, Austria2018.
529 31. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2:18-22.
530 32. Gonzalez-Ferreiro E, Dieguez-Aranda, Miranda D. Estimation of stand variables in Pinus radiata
531 D. Don plantations using different LiDAR pulse densities. Forestry. 2012;85:281-92.
532 33. Ruiz LA, Hermosilla T, Mauro F, Godino M. Analysis of the influence of plot size and LiDAR
533 density on forest structure attribute estimates. Forests 2014;5:936-51.
534 34. Hyde P, Dubayah R, Walker W, Blair JB, Hofton M, Hunsaker C. Mapping forest structure for
535 wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM+, Quickbird) synergy. Remote Sens
536 Environ. 2006;102:63-73.
537 35. Hyde P, Dubayah RP, B., Blair JB, Hofton M, Hunsaker C, Knox R, et al. Mapping forest structure
538 for wildlife habitat analsyis using waveform lidar: validaton of montane ecosystems. Remote Sens
539 Environ. 2005;96:427-37.
540 36. Asner GP, Hughes RF, Vitousek PM, Knapp DE, Kennedy-Bowdoinm T, Boardmann J, et al.
541 Invasive plants transform the three-dimensional structure of rain forests. PNAS. 2008;105:4519-23.
542 37. Martinuzzi S, Vierling LA, Gould WA, Falkowski MJ, Evans JS, Hudak AT, et al. Prediction of
543 understory vegetation cover with airborne LiDAR in an interior ponderosa pine forest. Remote Sens
544 Environ. 2009;124:730-41.

545 38.   Lefsky MA, Cohen WB, Parker GG, Harding DJ. LiDAR remote sensing for ecosystem studies.
546 BioScience. 2002;52:19-30.
547 39.   Bradbury RB, Hill RA, Mason DC, Hinsley SA, Wilson JD, Balzter H, et al. Modelling relationships
548 between birds and vegetation structure using airborne LiDAR data: A review with case studies from
549 agricultural and woodland environments. Ibis. 2005;147(3):443-52.
550 40.   Vierling KT, Vierling LA, Gould WA, Martinuzzi S, Clawges RM. Lidar: Shedding new light on
551 habitat characterization and modeling. Front Ecol Environ. 2008;6(2):90-8.
552 41.   Davies AB, Asner GP. Advances in animal ecology from 3D-LiDAR ecosystem mapping. Trends in
553 Ecology and Evolution 2014;29:681-91.
554 42.   Rechsteiner C, Zellweger F, Gerber A, Breiner FT, Bollmann K. Remotely sensed forest habitat
555 structures improve regional species conservation. Remote Sens Ecol Conserv. 2017;3:247-58.
556 43.   Vogeler JC, Hudak AT, Vierling LA, Evans J, Green P, Vierling KT. Terrain and vegetation structural
557 influences on local avian species richness in two mixed-conifer forests. Remote Sens Environ.
558 2014;147:13-22.
559 44.   Coops NC, Tompaski P, Nijland W, Rickbeil GJM, Nielsen SE, Bater CW, et al. A forest structure
560 habitat index based on airborne laser scanning. Ecol Indic. 2016;67:346-57.
561 45.   Clawges R, Vierling K, Vierling L, Rowell E. The use of airborne lidar to assess avian species
562 diversity, density, and occurrence in a pine/aspen forest. Remote Sens Environ. 2008;112(5):2064-73.
563 46.   Müller J, Stadler J, Brandl R. Composition versus physiognomy of vegetation as predictors of bird
564 assemblages: The role of lidar. Remote Sens Environ. 2010;114(3):490-5.
565 47.   Vierling LA, Vierling KT, Adam P, Hudak AT. Using satellite and airborne LiDAR to model
566 woodpecker habitat occupancy at the landscape scale. PLoS ONE. 2013;8(12).
567 48.   Melin M, Mehtätalo L, Miettinen J, Tossavainen S, Packalen P. Forest structure as a determinant
568 of grouse brood occurrence: an analysis linking LiDAR data with presence/absence field data. For Ecol
569 Manage. 2016;380:202-11.
570

Supporting information

S1 Table. Definitions of all variables included in at least one of the mixed-effects or Random

Forest models.

S2 Table. Frequency of explanatory variables among the 18 random forest models run with

341 to 7 variables.

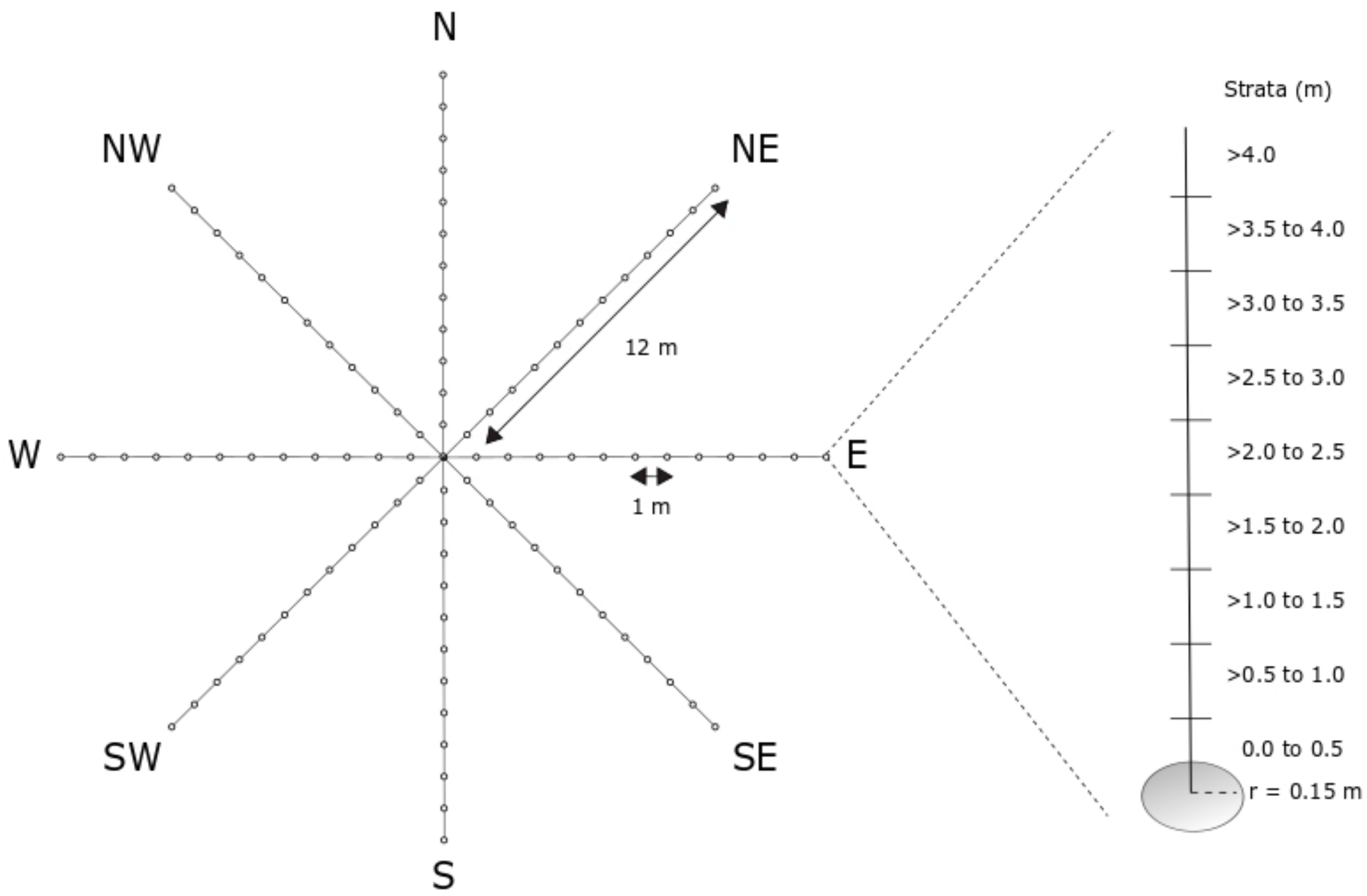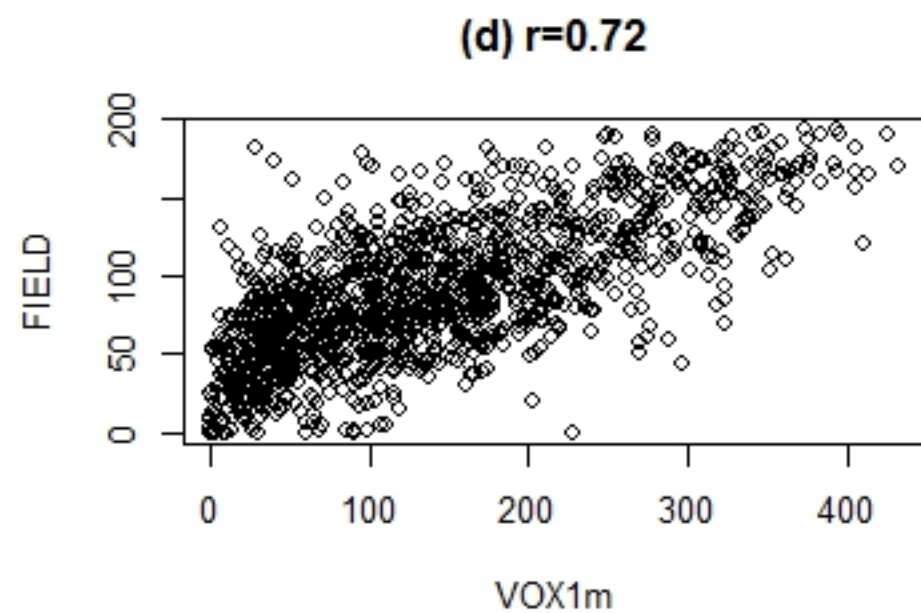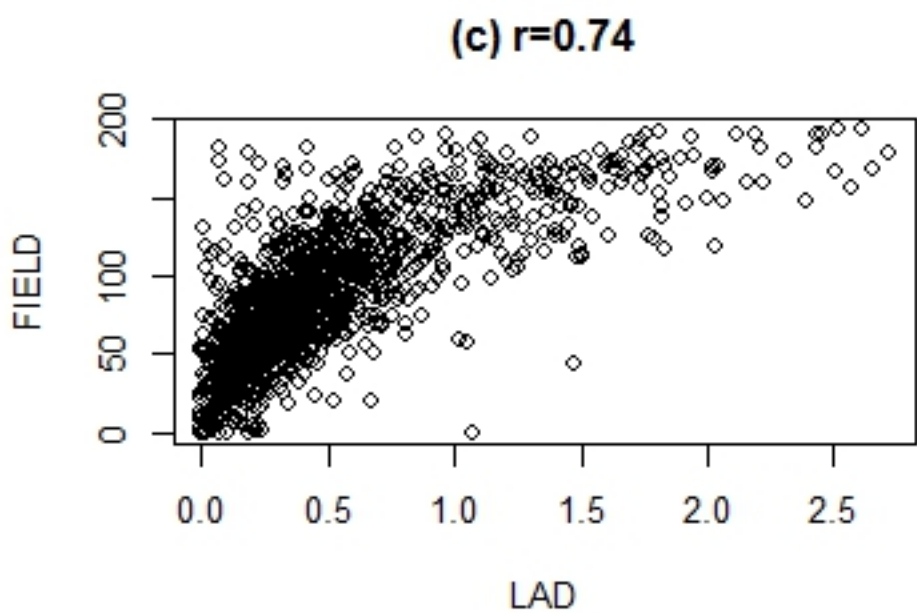**S3 Data. Data used in analyses for manuscript. Variable definitions are found in S1.**

N

NW   NE

12 m

W   E

1 m

SW   SE

S

Strata (m)

>4.0

>3.5 to 4.0

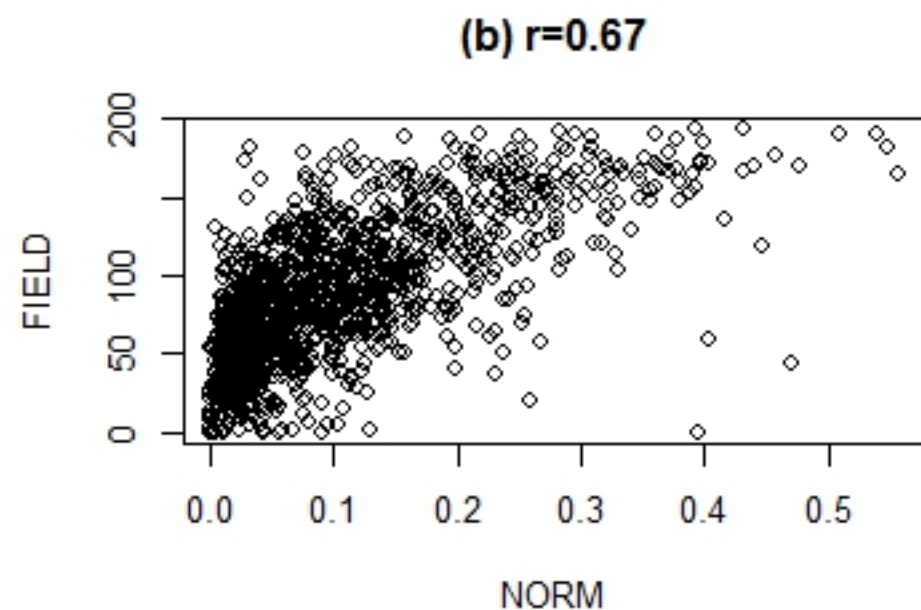>3.0 to 3.5

>2.5 to 3.0

>2.0 to 2.5
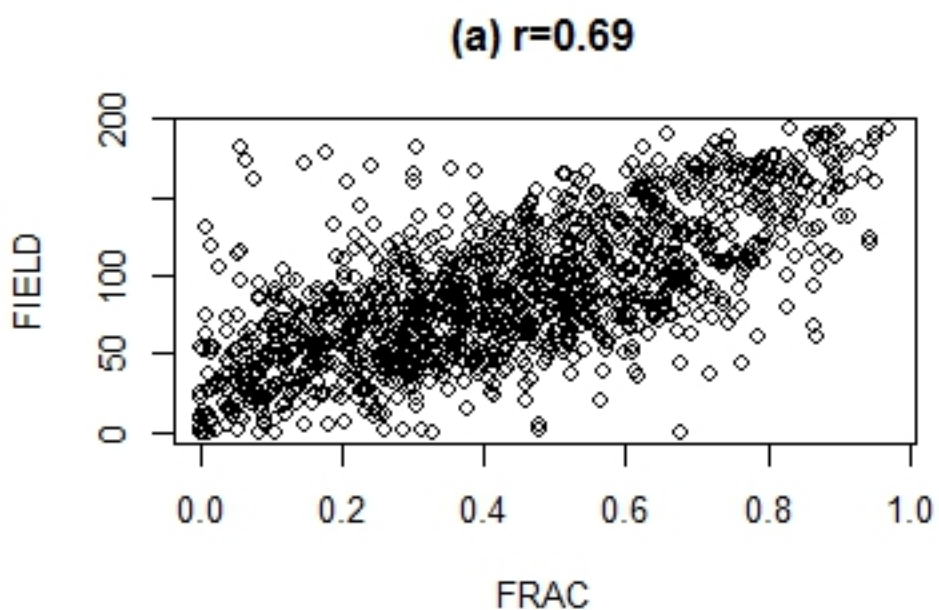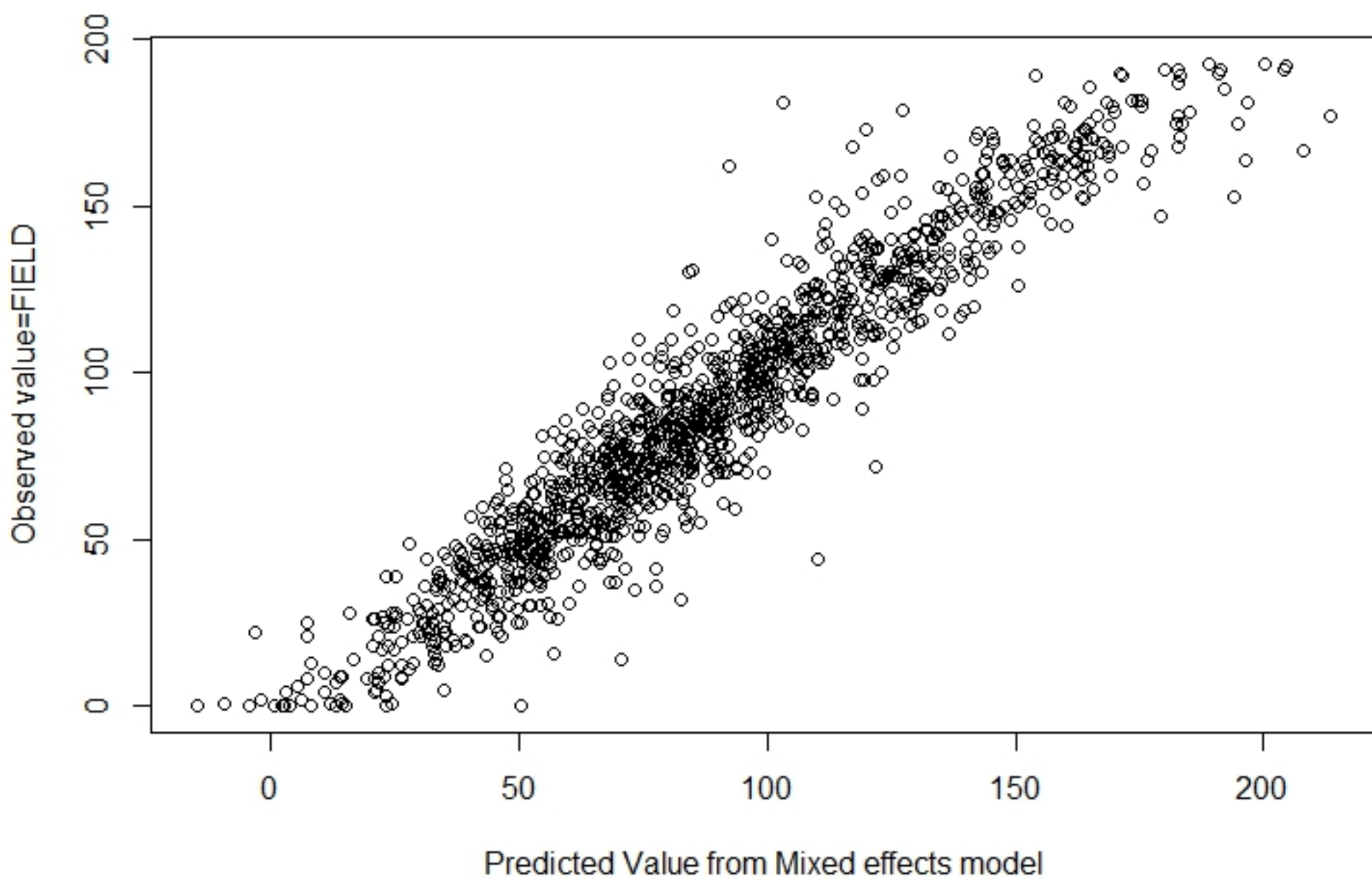
>1.5 to 2.0

>1.0 to 1.5

>0.5 to 1.0

0.0 to 0.5

r = 0.15 m

Figure 1

Figure 2

Figure 3

Figure 5

Figure 4