The CCB-ID approach to tree species mapping with airborne imaging spectroscopy

Christopher B Anderson [Corresp. 1, 2]

1 Department of Biology, Stanford University, Stanford, California, United States

2 Center for Conservation Biology, Stanford University, Stanford, California, United States

Corresponding Author: Christopher B Anderson
Email address: cbanders@stanford.edu

Background. Biogeographers assess how species distributions and abundances affect the structure, function, and composition of ecosystems. Yet we face a major challenge: it is difficult to precisely map species across landscapes. Novel Earth observations could overcome this challenge for vegetation mapping. Airborne imaging spectrometers measure plant functional traits at high resolution, and these measurements can be used to identify tree species. Plant traits are highly conserved within species, and highly variable between species, providing the biophysical basis for species mapping. In this paper I describe a trait-based approach to species identification with imaging spectroscopy, the Center for Conservation Biology species identification (CCB-ID) method, which was developed as part of an ecological data science evaluation competition.

Methods. These methods were developed using airborne imaging spectroscopy data from the National Ecological Observatory Network (NEON). CCB-ID classifies tree species using trait-based reflectance variation and decision tree-based machine learning models, approximating a morphological trait and dichotomous key method inspired by botanical classification. First, outliers were removed using a spectral variance threshold. The remaining samples were transformed using principal components analysis and resampled to reduce common species biases. Gradient boosting and random forest classifiers were trained using the transformed and resampled feature data. Prediction probabilities were then calibrated using sigmoid regression, and sample-scale predictions were averaged to the crown scale.

Results. This approach performed well according to the competition metrics, receiving a rank-1 accuracy score of 0.919, and a cross-entropy cost score of 0.447 on the test data. Accuracy and specificity scores were high for all species, but precision and recall scores were variable for rare species. PCA transformation improved accuracy scores compared to models trained using reflectance data, but outlier removal and data resampling exacerbated class imbalance problems.

Discussion. CCB-ID accurately classified tree species using NEON data, reporting the best scores among participants. However, it failed to overcome several well-known species mapping challenges like precisely identifying rare species. Key takeaways include (1) training models to maximize metrics beyond accuracy (e.g. recall) could improve rare species predictions, (2) within-genus trait variation may drive spectral separability, precluding efforts to distinguish between functionally convergent species, (3) outlier removal and data resampling exacerbated class imbalance problems, and should be carefully implemented, (4) PCA transformation greatly improved model results, and (5) feature selection could further improve species classification models. CCB-ID is open source, designed for use with NEON data, and available to support future species mapping efforts.

The CCB-ID approach to tree species mapping with airborne imaging spectroscopy

Christopher B. Anderson[1,2]
[1] Department of Biology, Stanford University, Stanford, CA, USA
[2] Center for Conservation Biology, Stanford University, Stanford, CA, USA

Corresponding Author:
Christopher B. Anderson[1,2]
Email address: cbanders@stanford.edu

Abstract

Background. Biogeographers assess how species distributions and abundances affect the structure, function, and composition of ecosystems. Yet we face a major challenge: it is difficult to precisely map species across landscapes. Novel Earth observations could overcome this challenge for vegetation mapping. Airborne imaging spectrometers measure plant functional traits at high resolution, and these measurements can be used to identify tree species. Plant traits are highly conserved within species, and highly variable between species, providing the biophysical basis for species mapping. In this paper I describe a trait-based approach to species identification with imaging spectroscopy, the Center for Conservation Biology species identification (CCB-ID) method, which was developed as part of an Ecological Data Science Evaluation (ECODSE) competition.

Methods. These methods were developed using airborne imaging spectroscopy data from the National Ecological Observatory Network (NEON). CCB-ID classifies tree species using trait-based reflectance variation and decision tree-based machine learning models, approximating a morphological trait and dichotomous key method inspired by botanical classification. First, outliers were removed using a spectral variance threshold. The remaining samples were transformed using principal components analysis and resampled to reduce common species biases. Gradient boosting and random forest classifiers were trained using the transformed and resampled feature data. Prediction probabilities were then calibrated using sigmoid regression, and sample-scale predictions were averaged to the crown scale.

Results. This approach performed well according to the competition metrics, receiving a rank-1 accuracy score of 0.919, and a cross-entropy cost score of 0.447 on the test data. Accuracy and specificity scores were high for all species, but precision and recall scores were variable for rare species. PCA transformation improved accuracy scores compared to models trained using reflectance data, but outlier removal and data resampling exacerbated class imbalance problems.

Discussion. CCB-ID accurately classified tree species using NEON data, reporting the best scores among participants. However, it failed to overcome several well-known species mapping challenges like precisely identifying rare species. Key takeaways include (1) training models to maximize metrics beyond accuracy (e.g. recall) could improve rare species predictions, (2) within-genus trait variation may drive spectral separability, precluding efforts to distinguish between functionally convergent species, (3) outlier removal and data resampling exacerbated class imbalance problems, and should be carefully implemented, (4) PCA transformation greatly improved model results, and (5) feature selection could further improve species classification models. CCB-ID is open source, designed for use with NEON data, and available to support future species mapping efforts.

1 **Introduction**

2   When you get down to it, biogeographers seek to answer two key questions: where are

3 the species, and why are they where they are? Answering these simple questions has proven

4 remarkably difficult. The former reflects a data gap; we do not have complete or unbiased

5 information on where species occur. This is known as the 'Wallacean shortfall' (Whittaker et al.

6 2005; Bini et al. 2006). Addressing the latter, however, does not necessarily require data; the

7 drivers of species abundances and their spatial distributions can be derived from ecological

8 theory (McGill, 2010). But evaluating these theoretical predictions does require data. Testing

9 generalized theories of species distributions requires continuously-mapped presences and

10 absences for many individuals across large areas. And while field efforts can assess fine-scale

11 distribution patterns, they are often restricted to small extents. Mapping organism-scale species

12 distributions over landscapes could help fill the data gaps that preclude addressing these key

13 biogeographic questions (Anderson 2018). One remote sensing dataset holds the promise to do

14 so for plants: airborne imaging spectroscopy.

15   Airborne imaging spectrometers measure variation in the biophysical properties of soils

16 and vegetation at fine grain sizes across large areas (Goetz et al. 1985). In vegetation mapping,

17 imaging spectroscopy can measure plant structural traits, like leaf area index and leaf angle

18 distribution (Broge & Leblanc, 2001; Asner & Martin, 2008), and plant functional traits, like

19 growth and defense compound concentrations (Kokaly et al. 2009; Asner et al. 2015). These

20 traits tend to be highly conserved within tree species, and highly variable between species (i.e.

21 interspecific trait variation is often much greater than intraspecific trait variation; (Townsend et

22 al. 2007; Asner et al. 2011). This trait conservation provides the conceptual and biophysical basis

23 for species mapping with imaging spectroscopy. Indeed, airborne imaging spectroscopy has been

24 used to map crown-scale species distributions across large extents in several contexts (Fassnacht

25 et al. 2016). These approaches have been applied in temperate (Baldeck et al. 2014) and tropical

26 ecosystems (Hesketh & Sánchez-Azofeifa, 2012), using multiple classification methods (Feret &

27 Asner, 2013) and multiple sensors (Clark, Roberts & Clark, 2005; Colgan et al. 2012; Baldeck et

28 al. 2015). However, this wide range of approaches has not yet identified a canonical best practice

29 for tree species identification.

30   In this paper I describe an approach to tree species classification using airborne imaging

31 spectroscopy data that builds on the above methods to advance the discussion on best practices.

32 This approach was developed as a submission to an Ecological Data Science Evaluation

33 competition (ECODSE; https://ecodse.org) sponsored by the National Institute of Standards and

34 Technology (NIST). This competition had participants use airborne imaging spectroscopy data,

35 collected by the National Ecological Observatory Network's Airborne Observation Platform

36 (NEON AOP; Kampe et al. 2010), to identify tree crowns to the species level. The work

37 described was submitted under the team name of the Stanford Center for Conservation Biology

38 (CCB), and has since been formalized under the moniker *CCB-ID* (https://github.com/stanford-

39 ccb/ccb-id). First, I describe the CCB-ID approach to tree species classification using airborne

40 imaging spectroscopy data. Next, I review its successes and shortcomings in the context of this

41 competition. Finally, I highlight key opportunities to improve future imaging spectroscopy-based

42 species classification approaches. The goals of this work are to improve NEON's operational tree

43 species mapping efforts and to reduce barriers for addressing key data gaps in plant

44 biogeography.

45 **Materials & Methods**

46   The CCB-ID approach was inspired by botanical and taxonomic approaches to species

47 classification. In the field, botanists can use plant morphological features and a dichotomous key

48 to identify tree species. These features often include variations in reproductive traits (e.g.

49 flowering bodies, seeds), vascular traits (e.g. types of woody and non-woody tissue), and foliar

50 traits (e.g. waxy or serrated leaves). The dichotomous key approach hierarchically partitions

51 species until each can be identified using a specific combination of traits. Species classification

52 with imaging spectroscopy is rather restricted in comparison; imaging spectrometers can only

53 measure a subset of plant traits. This subset includes growth traits such as leaf chlorophyll and

54 nitrogen content (Lepine et al. 2016), structural traits such as leaf cellulose and water content

55 (Papeş et al. 2010), and defense traits such as leaf phenolic concentrations and lignin content

56 (McManus et al. 2016). Furthermore, the inter- and intraspecific variation in this subset of traits

57 is rarely known *a priori*, precluding the use of a standard dichotomous key (Kichenin et al. 2013;

58 Siefert et al. 2015).

59   Imaging spectroscopy approaches to species classification instead rely on distinguishing

60 species-specific variations in canopy reflectance. However, several confounding factors drive

61 variation in reflectance data, including (1) measurement conditions (e.g. sun and sensor angles),

62 (2) canopy structure (e.g. leaf area index or leaf angle distribution), (3) leaf morphology and

63 physiology (i.e. plant functional traits), and (4) sensor noise (Goetz et al. 1985; Ollinger 2011;

64 Lausch et al. 2016). Measurement conditions and canopy structure tend to drive the majority of

65 variation; up to 79-89% of spectral variance is driven by within-crown variation (Baldeck &

66 Asner, 2014; Yao et al. 2015). Unfortunately, this variation does not help distinguish between

67 species. Interspecific spectral variation is instead driven by functional trait variation (Asner et al.

68 2011; Martin et al. 2018). Disentangling trait-based variation from measurement and structure-

69 based variation is thus central to mapping species with airborne imaging spectroscopy.

70   CCB-ID classifies tree species using trait-based reflectance variation with decision tree-

71   based machine learning models. This approach approximates a morphological trait and

72   dichotomous key model (Godfray 2007), and is described in the following sections. The first

73   section describes the outlier removal and data transformation procedures. The second section

74   describes how the training data were resampled to reduce biases towards common species. The

75   third section describes model selection, training, and probability calibration. The fourth section

76   describes the model performance metrics, and the final section describes two analyses performed

77   post-ECODSE submission.

78   The NEON data included the following products: (1) Woody plant vegetation structure

79   (NEON.DP1.10098), (2) Spectrometer orthorectified surface directional reflectance - flightline

80   (NEON.DP1.30008), (3) Ecosystem structure (NEON.DP3.30015), and (4) High-resolution

81   orthorectified camera imagery (NEON.DP1.30010). These data were provided by the ECODSE

82   group (2017; https://ecodse.org) and are freely available from the NEON website

83   (https://neonscience.org). These analyses used data product (2). All analyses were performed

84   using the Python programming language (Oliphant, 2007; https://python.org) and the following

85   open source packages: NumPy (der Walt, Colbert & Varoquaux, 2011; http://numpy.org), scikit-

86   learn (Pedregosa et al. 2011; http://scikit-learn.org), pandas (McKinney et al. 2010;

87   https://pandas.pydata.org), and matplotlib (Hunter, 2007; https://matplotlib.org). The python

88   scripts used for these analyses have been uploaded to a public GitHub repository

89   (https://github.com/stanford-ccb/ccb-id), including a build script for a Singularity container to

90   ensure computational replicability (Kurtzer, Sochat & Bauer, 2017).

91   *Data preprocessing*

92   The canopy reflectance data were preprocessed using two steps: outlier removal and

93   dimensionality reduction. In the outlier removal step, the reflectance data were spectrally subset,

94   transformed using PCA, then thresholded to isolate spurious values. First, reflectance values

95   from the blue region of the spectrum (0.38-0.49 μm) and from noisy bands (1.35-1.43 μm, 1.80-

96   1.96 μm, and 2.48-2.51 μm) were removed. These bands correspond to wavelengths dominated

97   by atmospheric water vapor, and do not track variations in plant traits (Gao et al. 2009; Asner et

98   al. 2015). This reduced the data from 426 to 345 bands. Next, these spectrally-subset samples

99   were transformed using PCA. The output components were whitened to zero mean and unit

100   variance, and outliers were identified using a three-sigma threshold. Samples with values outside

101   of +/- three standard deviations from the means (i.e. which did not fall within 99.7% of the

102   variation for each component) for the first 20 principal components were excluded from analysis.

103   These samples were expected to contain non-vegetation spectra (e.g. exposed soil), unusually

104   bright or dark spectra, or anomalously noisy spectra (Féret & Asner, 2014). The outlier-removed

105   reflectance profiles for each species are shown in Figure 1.

106   Once the outliers were removed, the remaining spectra were transformed using PCA.

107   This was not performed on the already-transformed data from the outlier removal process, but on

108   the outlier-removed, spectrally-subset reflectance data. PCA transformations are often applied to

109   airborne imaging spectrometer data to handle the high degree of correlation between bands, and

110   these transformations are highly sensitive to input feature variation (Jia & Richards, 1999).

111   Furthermore, transforming reflectance data into principal components can isolate the variation

112   driven by measurement conditions from variation driven by functional traits, which is critical for

113   distinguishing between species. And though trait-based variation drives a small proportion of

114   total reflectance signal, a single trait can be expressed in up to 9 orthogonal components (Asner

115   et al. 2015). After the transformation, the first 100 of 345 possible components were used as

116   feature vectors for the species classification models. This threshold was arbitrary; it was set to

117   capture the majority of biologically-relevant components and to exclude noisy components.

118   *Class imbalance*

119   Class imbalance refers to datasets where the number of samples per class are not evenly

120   distributed among classes. Imbalanced datasets are common in classification contexts, but can

121   lead to problems if left unaddressed. Training classification models with imbalanced data can

122   select for models that overpredict common classes when model performance is based on

123   accuracy metrics. The ECODSE data were imbalanced: after outlier removal, these data

124   contained a total of 6,034 samples from 9 classes (8 identified species, one 'other species' class).

125   The most common species, *Pinus palustris*, contained 4,026 samples (66% of the samples) and

126   the rarest species, *Liquidambar styraciflua*, contained 62 samples (1% of the samples).

127   These data were resampled prior to analysis to reduce the likelihood of overpredicting

128   common species. Resampling was performed by setting a fixed number of samples per class,

129   then undersampling or oversampling each class to that fixed number. This fixed number was set

130   to 400 samples to split the difference of two orders of magnitude between the rarest and the most

131   common classes. This number was arbitrary, but it approximates the number of per-species

132   samples recommended in Baldeck & Asner (2014). To create the final training data, classes with

133   fewer than 400 samples were oversampled with replacement, and classes with more than 400

134   samples were undersampled without replacement. The final training data included 400 samples

135   for each of the 9 classes (3,600 samples total). Each sample contained a feature vector of the

136   principal components derived from the outlier removed, spectrally subset canopy reflectance

137   data.

138   *Model selection, training, and probability calibration*

139   The CCB-ID approach used two machine learning models: a gradient boosting classifier

140   and a random forest classifier (Friedman, 2001; Breiman, 2001). These models can fit complex,

141   nonlinear relationships between response and feature data, can automatically handle interactions

142   between features, and have built-in mechanisms to reduce overfitting (Mascaro et al. 2014). They

143   were selected because they perform well in species mapping contexts (Elith, Leathwick &

144   Hastie, 2008), in remote sensing contexts (Pal, 2005), and in conjunction with PCA

145   transformations (Rodríguez, Kuncheva & Alonso, 2006). Furthermore, these models are built as

146   ensembles of decision trees, resembling the dichotomous key employed by botanists. Unlike a

147   dichotomous key, these models were trained to learn where to split the data since the trait

148   variation that distinguishes species was not known *a priori*.

149   These models were fit using hyper-parameter tuning and probability calibration

150   procedures. Model hyper-parameters were tuned by selecting the parameters that maximized

151   mean F1 scores in 5-fold cross-validation using an exhaustive grid search. F1 score calculates the

152   weighted average of model precision and recall (see *Model assessment*), and maximizing F1

153   scores during model tuning reduces the likelihood of selecting hyper-parameters that overpredict

154   common classes and underpredict rare classes. The following parameters were tuned for both

155   models: number of estimators, maximum tree depth, minimum number of samples required to

156   split a node, and minimum node impurity split threshold. The learning rate and node split quality

157   criterion were also tuned for the gradient boosting and random forest classifiers, respectively. All

158   samples were used for hyper-parameter tuning, and the best model hyper-parameters (i.e. the

159 hyper-parameters that maximized mean F1 scores in cross validation) were used to fit the final

160 models.

161   Accurately characterizing prediction probabilities is essential for error propagation and

162 for assessing model reliability. Prediction probabilities were calibrated after the final hyper-

163 parameters were selected. Well-calibrated probabilities should scale linearly with the true rate of

164 misclassification (i.e. model predictions should not be under or overconfident). Some ensemble

165 methods, such as random forest, tend to be poorly calibrated. Since they average their predictions

166 from a set of weak learners, which individually have high misclassification rates but gain

167 predictive power post-ensemble, the variance of random forest classifiers can skew high

168 probabilities away from one, and low probabilities away from zero. This results in sigmoid-

169 shaped reliability diagrams (DeGroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005).

170   To reduce these probability biases, prediction probabilities were calibrated using sigmoid

171 regression for both gradient boosting and random forest classifiers. The data were first randomly

172 split into three subsets: model training (50%, or 200 samples per class), probability calibration

173 training (25%, or 100 samples per class), and probability calibration testing (the remaining 25%).

174 Each classifier was fit using the model training subset and the tuned hyper-parameters.

175 Prediction probabilities were calibrated with sigmoid regression using the probability training

176 subset and internal 3-fold cross validation to assess the calibration. Calibrated model

177 performance was assessed using the holdout test data. After these assessments, the final models

178 were fit using the model training data, then calibrated using the full probability training and

179 testing data (i.e. the full 50% of samples not used in initial model training).

180   *Model assessment*

181   During model training, performance was assessed on a per-sample basis using model

182   accuracy and log loss scores. Model accuracy calculates the proportion of correctly classified

183   samples in the test data (Figure 2). High model accuracy scores are desirable. Log loss assesses

184   whether the prediction probabilities were well calibrated, penalising incorrect and uncertain

185   predictions. Low log loss scores indicate that misclassifications occur at rates close to the rates of

186   predicted probabilities. During model testing, performance was assessed using rank-1 accuracy

187   and cross entropy cost (Marconi et al. 2018). Rank-1 accuracy was calculated based on which

188   species ID was predicted with the highest probability. The cross entropy score is similar to the

189   log loss function, but was scaled using an indicator function. These can be interpreted in similar

190   ways to accuracy and log loss; high rank-1 accuracy and low cross entropy scores are desirable

191   (Hastie et al. 2009).

192   Secondary model testing metrics were calculated for each species using the test data.

193   These included model specificity, precision, and recall (Figure 2). These metrics reveal model

194   behavior that accuracy scores may obscure. Specificity assesses model performance on non-

195   target species, penalizing overprediction of the target species (i.e. a high number of false

196   positives). Precision also penalizes overprediction, but assesses the rate of overprediction relative

197   to the rate of true positive predictions. Recall calculates the proportion of true positive

198   predictions to the total number of positive observations per species. Higher values are desirable

199   for each. These metrics were calculated to aid interpretation, but were not used to formally rank

200   model performance.

201   Performance during model training was assessed at the sample scale, meaning the model

202   performance metrics were calculated on every pixel (i.e. sample) in the training data. However,

203   the competition evaluation metrics were calculated using crown-scale prediction probabilities,

204   meaning the model performance metrics were calculated after aggregating each pixel from

205   individual trees to unique crown identities. To address this scale mismatch, prediction

206   probabilities were first calculated for each sample in a crown using both gradient boosting and

207   random forest models. These sample-scale probabilities were then averaged by crown.

208   *Further analyses*

209   Two post-submission analyses were performed to assess how PCA transformations

210   affected model performance. Prior to these analyses, I bootstrapped the original model fits to

211   assess their variance. I then compared these bootstrapped fits to models trained with the

212   spectrally-subset reflectance data instead of the PCA transformed data. Next, I compared models

213   trained using a varying number of principal components. These models were trained using $n_{pcs} \in$

214     $\{10, 20, \ldots, 345\}$ as the input features, with 345 being the maximum number of potential

215   components after spectral subsetting. These comparisons assessed whether the PCA

216   transformations improved model performance, and how changing the amount of spectral

217   variation in the feature data affected performance. These analyses were each bootstrapped 50

218   times.

219   **Results**

220   CCB-ID performed well according to the ECODSE competition metrics, receiving a

221   rank-1 accuracy score of 0.919, and a cross-entropy cost score of 0.447 on the test data. These

222   were the highest rank-1 accuracy and the lowest cross-entropy cost scores among participants.

223   Other methods reported rank-1 accuracy scores from 0.688 to 0.88 and cross entropy scores from

224   0.877 to 1.448 (Marconi et al. 2018). A confusion matrix with the classification results is

225   reported in Table 1. In addition to the high rank-1 accuracy and low cross entropy cost scores,

226   the CCB-ID model performed well according to the secondary crown-scale performance metrics.

227   These secondary metrics calculated a mean accuracy score of 0.979, mean specificity score of

228   0.985, mean precision score of 0.614, and mean recall score of 0.713 across all species. The per-

229   species secondary metrics are summarized in Figure 3. These results were calculated using the

230   categorical classification predictions (i.e. after assigning ones to the species with the highest

231   predicted probability, and zeros to all other species). The probability-based confusion matrix and

232   classification metrics are reported in Table S1 and Figure S1, respectively.

233   During model training, outlier removal excluded 797 samples from analysis. 264 of the

234   797 samples (33%) removed from analysis were from *P. palustris*, while the remaining 533

235   samples (67%) were from non-*P. palustris* species. Outlier removal disproportionately excluded

236   samples from uncommon species; 45% of samples from *Liquidambar styraciflua,* the rarest

237   species, were removed. After outlier removal, the first principal component contained 78% of the

238   explained variance. However, this component did not drive model performance; it ranked 7th

239   and 11th in terms of ranked feature importance scores for the gradient boosting and random

240   forest classifiers. Model accuracy scores, calculated on a sample basis (i.e. not by crown) using

241   the 25% training data holdout, were 0.933 for gradient boosting and 0.956 for random forest.

242   Log loss scores, calculated prior to probability calibration, were 0.19 for gradient boosting, and

243   0.47 for random forest. After probability calibration, log loss scores were 0.24 for gradient

244   boosting and 0.16 for random forest. The per-class secondary metrics reported a mean specificity

245   score of 0.987, mean precision score of 0.908, and mean recall score of 0.907 across all species.

246   The post-submission analyses found PCA transformations improved model accuracy.

247   Models fit using the original methods calculated mean bootstrapped accuracy scores of 0.944 (*s*

248   = 0.009) for gradient boosting and 0.955 (*s* = 0.008) for random forest. Models fit using the

249   spectrally-subset reflectance data as features calculated mean accuracy scores of 0.883 ($s =$

250   0.012) for gradient boosting and 0.877 ($s = 0.011$) for random forest, and mean log loss scores of

251   0.46 ($s = 0.03$) for gradient boosting and 0.48 ($s = 0.03$) for random forest. Mean model

252   accuracies declined and mean log loss scores increased after including more than 20 components

253   as features for the models fit using varying numbers of principal components (Figure 4).

254   **Discussion**

255   The CCB-ID approach accurately classified tree species using NEON imaging

256   spectroscopy data, reporting the highest rank-1 accuracy score and lowest cross-entropy cost

257   score among ECOSDE participants. These scores compare favorably to other imaging

258   spectroscopy-based species classification efforts (Fassnacht et al. 2016). These crown-scale test

259   results highlight the potential to develop species mapping methods that approximate botanical

260   and taxonomic approaches to classification. However, this method failed to overcome several

261   well-known species mapping challenges, like precisely identifying rare species. Below I discuss

262   some key takeaways and suggest opportunities to improve future imaging spectroscopy-based

263   species classification approaches.

264   The high per-species accuracy scores indicate a high proportion of correctly classified

265   crowns in the test data. However, accuracy can be a misleading metric in imbalanced contexts.

266   Since seven of the nine classes had six or fewer crowns in the test data (out of 126 total test

267   crowns), classification metrics weighted by the true negative rate (i.e. accuracy and specificity)

268   were expected to be high if the majority class were correctly predicted. Metrics weighted instead

269   by the true positive rate (i.e. precision and recall) showed much higher variation across rare

270   species, as a single misclassification greatly alters these metrics when there are few observed

271   crowns (Figure 3). Due to the small sample size, it is difficult to assess if these patterns portend

272   problems at larger scales. For example, there were two observed *Acer rubrum* crowns in the test

273   data, yet only one was correctly predicted. Was the misclassified crown an anomaly? Or will this

274   low precision persist across the landscape, predicting *Acer rubrum* occurrences at half its actual

275   frequency? The latter seems unlikely, in this case; the low cross entropy and log loss scores

276   suggest misclassified crowns were appropriately uncertain in assigning the wrong label (Table

277   S1). However, since airborne species mapping is employed to address large-scale ecological

278   patterns where precision is key (e.g. in biogeography, macroecology, and biogeochemistry), we

279   should be assessing classification performance on more than one or two crowns per species.

280   Model performance between and within taxonomic groups revealed some notable

281   patterns. *Quercus* and *Pinus* individuals (i.e. Oaks and Pines) accounted for 120 of the 126 test

282   crowns and there was high fidelity between them. Only one *Quercus* crown was misclassified as

283   *Pinus*, and two *Pinus* crowns were misclassified as *Quercus*. From a botanical perspective, this

284   makes sense; these genera exhibit very different growth forms (i.e. different canopy structures

285   and foliar traits), and should thus be easy to distinguish in reflectance data. However, within-

286   genus model performance varied between *Quercus* and *Pinus*. *Quercus* crowns were never

287   misclassified as other *Quercus* species, yet there were several within-*Pinus* misclassifications.

288   This may be because *Quercus* species tightly conserve their canopy structures and foliar traits

289   (Cavender-Bares et al. 2016), while *Pinus* species may express trait plasticity. *Pinus* species

290   maintain similar growth forms (i.e. their needles grow in whorls bunched through the canopy),

291   limiting opportunities to distinguish species-specific structural variation. Furthermore, they are

292   distributed across the varying climates of the southern, eastern, and central United States,

293   suggesting some degree of niche plasticity. If this plasticity is expressed in each species'

294   functional traits, then convergence among species may then preclude trait-based classification

295   efforts. Quantifying the extent to which foliar traits are conserved within and between species

296   and genera will be essential for assessing the potential for imaging spectroscopy to map

297   community composition across large extents (Violle et al. 2012; Siefert et al. 2015).

298   The post-submission analyses revealed further notable patterns. First, PCA

299   transformation increased mean model accuracy scores compared to the spectrally-subset

300   reflectance data. I suspect this is because the models could focus on the spectral variation driven

301   by biologically meaningful components instead of searching for that signal in the reflectance

302   spectrum where the majority of variation is driven by abiotic factors. The low feature importance

303   scores of the first principal component support this interpretation. The first component in

304   reflectance data is typically driven by brightness (i.e. not a driver of interspecific variation) and

305   contained 78% of the explained reflectance variance, but ranked low in feature importance for

306   both models. This preprocessing transformation approximates the 'rotation forest' approach

307   developed by Rodríguez, Kuncheva & Alonso (2006), who found PCA preprocessing improved

308   tree-based ensemble models in several contexts. They suggested retaining all components to

309   maintain the original dimensionality of the input data. However, the analysis that varied the

310   number of feature components showed model accuracy decreased when including more than the

311   first 20 components (Figure 4). This suggests that using all components could overfit to noise.

312   Performing feature selection on transformed data may help overcome this. Feature selection has

313   been applied to reflectance data to find the spectral features that track functional trait variation

314   (Feilhauer, Asner & Martin, 2015), and I believe it could help identify trait-based components

315   that discriminate between species. Furthermore, other transformation methods may be more

316   appropriate than PCA; principal components serve only as proxies for functional traits in this

317   context. I expect transforming reflectance data directly into trait features, further extending the

318   analogy of a taxonomic approach to classification, could improve species mapping efforts,

319   improve model interpretability, and further develop the biophysical basis for species mapping

320   with imaging spectroscopy.

321   Despite the successes of CCB-ID, there were a few missteps in model design and

322   implementation. For example, outlier removal and resampling were employed to reduce class

323   imbalance problems but may instead have exacerbated them. First, the PCA-based outlier

324   removal excluded samples based on deviation from the mean of each component. However,

325   since the transformations were calculated using imbalanced data, the majority of the variance

326   was driven by variation in the most common class. This means outlier removal excluded samples

327   that deviated too far from the mean-centered variance weighted by *Pinus palustris*. Indeed, 533

328   of the 797 samples excluded from analysis (67%) were from non-*P. palustris* species (which

329   comprised only 37% of the full dataset). This removed up to 45% of samples from the rarest

330   species (*Liquidambar styraciflua*), reducing the spectral variance these models should be trained

331   to identify. This suggests outlier removal should either be skipped or implemented using other

332   methods (e.g. using spectral mixture analysis to identify samples with high soil fractions) to

333   reduce imbalance problems for rare species.

334   Data resampling further exacerbated class imbalance. By setting the resampling threshold

335   an order of magnitude above the least sampled class, the rarest species were oversampled nearly

336   tenfold in model training. This oversampling inflated per-class model performance metrics by

337   double-counting (or more) correctly classified samples for oversampled species. These metrics

338   were further inflated as a result of how the train/test data were split. The split was performed

339   after resampling, meaning the train/test data for oversampled species were likely not

340   independent. This invalidated their use as true test data, overestimating performance during

341   model training. This is unequivocally bad practice; I call this "user error." Undersampling the

342   common species was also detrimental. Excluding samples from common species meant the

343   models were exposed to less intraspecific spectral variation during training. This is a key source

344   of variation the models should recognize. Excluding this spectral variation made it more difficult

345   for the models to distinguish inter and intraspecific variation. Assigning sample weights (e.g.

346   proportional to the number of samples per class) and using actually independent holdout data

347   could overcome these issues. These will be implemented in future versions of CCB-ID.

348   However, these need not be the only updates to this method; CCB-ID is an open source, freely

349   available project (https://github.com/stanford-ccb/ccb-id). I invite you to to use it and improve it.

350   **Conclusions**

351   It was not always possible to classify tree species from airplanes; now it is. Airborne

352   imaging spectrometers can identify trees at crown scales across large areas, and these data are

353   now publicly available through NEON. However, there is currently no canonical imaging

354   spectroscopy-based species mapping approach, limiting opportunities to explore key patterns in

355   biogeography. CCB-ID was developed to identify best practices for species classification in this

356   context, and to further the conversation on how to implement these practices. CCB-ID performed

357   well within the scope of the ECODSE competition, reporting the highest rank-1 accuracy and

358   lowest cross entropy scores among participants. Yet further testing is necessary to identify

359   whether this method can scale to other regions (e.g. to high diversity forests). I hope CCB-ID

360   will be used to improve future species mapping efforts to pursue answers to biogeography's great

361   mysteries of where the species are, and why they are there.

362   **Acknowledgements**

363   I would like to thank Gretchen Daily for her continued advisement, support, and

364   inspiration. Thanks to the organizers of the NSF NEON workshop on mapping species, foliar

365   chemistry and soil properties with spectroscopy, including Nancy Glenn, Nathan Leisso, Jessica

366   Mitchell, Yi Qi, and Dar Roberts. Thanks to two anonymous reviewers for their insightful

367   comments. Thanks to Phil Brodrick for being good at models, and even better at explaining

368   them. Finally, thanks to Jeff Smith for comments on this manuscript, and for fruitful

369   conversations on hyperspectral image mixing.

References

Anderson CB. 2018. Biodiversity monitoring, earth observations and the ecology of scale. *Ecology letters* 4:438. DOI: 10.1111/ele.13106.

Asner GP., Martin RE. 2008. Spectral and chemical analysis of tropical forests: Scaling from leaf to canopy levels. *Remote sensing of environment* 112:3958–3970. DOI: 10.1016/j.rse.2008.07.003.

Asner GP., Martin RE., Anderson CB., Knapp DE. 2015. Quantifying forest canopy traits: Imaging spectroscopy versus field survey. *Remote sensing of environment* 158:15–27. DOI: 10.1016/j.rse.2014.11.011.

Asner GP., Martin RE., Knapp DE., Tupayachi R., Anderson C., Carranza L., Martinez P., Houcheime M., Sinca F., Weiss P. 2011. Spectroscopy of canopy chemicals in humid tropical forests. *Remote sensing of environment* 115:3587–3598. DOI: 10.1016/j.rse.2011.08.020.

Baldeck CA., Asner GP. 2014. Improving Remote Species Identification through Efficient Training Data Collection. *Remote Sensing* 6:2682–2698. DOI: 10.3390/rs6042682.

Baldeck CA., Asner GP., Martin RE., Anderson CB., Knapp DE., Kellner JR., Wright SJ. 2015. Operational Tree Species Mapping in a Diverse Tropical Forest with Airborne Imaging Spectroscopy. *PloS one* 10:e0118403. DOI: 10.1371/journal.pone.0118403.

Baldeck CA., Colgan MS., Féret JB., Levick SR., Martin RE., Asner GP. 2014. Landscape-scale variation in plant community composition of an African savanna from airborne species mapping. *Ecological applications: a publication of the Ecological Society of America* 24:84–93.

Bini LM., Diniz-Filho JAF., Rangel TFLVB., Bastos RP., Pinto MP. 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. *Diversity & distributions* 12:475–482. DOI: 10.1111/j.1366-9516.2006.00286.x.

Breiman L. 2001. Random Forests. *Machine learning* 45:5–32. DOI: 10.1023/A:1010933404324.

Broge NH., Leblanc E. 2001. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. *Remote sensing of environment* 76:156–172. DOI: 10.1016/S0034-4257(00)00197-8.

Cavender-Bares J., Meireles JE., Couture JJ., Kaproth MA., Kingdon CC., Singh A., Serbin SP., Center A., Zuniga E., Pilz G., Townsend PA. 2016. Associations of Leaf Spectra with Genetic and Phylogenetic Variation in Oaks: Prospects for Remote Detection of Biodiversity. *Remote Sensing* 8:221. DOI: 10.3390/rs8030221.

Clark ML., Roberts DA., Clark DB. 2005. Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote sensing of environment* 96:375–398. DOI: 10.1016/j.rse.2005.03.009.

Colgan MS., Baldeck CA., Féret J-B., Asner GP. 2012. Mapping Savanna Tree Species at Ecosystem Scales Using Support Vector Machine Classification and BRDF Correction on Airborne Hyperspectral and LiDAR Data. *Remote Sensing* 4:3462–3480. DOI: 10.3390/rs4113462.

DeGroot MH., Fienberg SE. 1983. The Comparison and Evaluation of Forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)* 32:12–22. DOI: 10.2307/2987588.

ECODSE group. 2017. *ECODSE competition training set*. DOI: 10.5281/zenodo.1206101.

Elith J., Leathwick JR., Hastie T. 2008. A working guide to boosted regression trees. *The Journal of animal ecology* 77:802–813. DOI: 10.1111/j.1365-2656.2008.01390.x.

Fassnacht FE., Latifi H., Sterénczak K., Modzelewska A., Lefsky M., Waser LT., Straub C., Ghosh A. 2016. Review of studies on tree species classification from remotely sensed data. *Remote sensing of environment* 186:64–87. DOI: 10.1016/j.rse.2016.08.013.

Feilhauer H., Asner GP., Martin RE. 2015. Multi-method ensemble selection of spectral bands related to leaf biochemistry. *Remote sensing of environment* 164:57–65. DOI: 10.1016/j.rse.2015.03.033.

Feret JB., Asner GP. 2013. Tree Species Discrimination in Tropical Forests Using Airborne Imaging Spectroscopy. *IEEE transactions on geoscience and remote sensing: a publication of the IEEE Geoscience and Remote Sensing Society* 51:73–84. DOI: 10.1109/TGRS.2012.2199323.

Féret J-B., Asner GP. 2014. Mapping tropical forest canopy diversity using high-fidelity imaging spectroscopy. *Ecological applications: a publication of the Ecological Society of America* 24:1289–1296.

Friedman JH. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of statistics* 29:1189–1232.

Gao B-C., Montes MJ., Davis CO., Goetz AFH. 2009. Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean. *Remote sensing of environment* 113:S17–S24. DOI: 10.1016/j.rse.2007.12.015.

Godfray HCJ Jr. 2007. Linnaeus in the information age. *Nature* 446:259–260. DOI: 10.1038/446259a.

Goetz AF., Vane G., Solomon JE., Rock BN. 1985. Imaging spectrometry for Earth remote sensing. *Science* 228:1147–1153. DOI: 10.1126/science.228.4704.1147.

Hastie T., Tibshirani R., Friedman J. 2009. Unsupervised Learning. In: Hastie T, Tibshirani R, Friedman J eds. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, 485–585. DOI: 10.1007/978-0-387-84858-7_14.

Hesketh M., Sánchez-Azofeifa GA. 2012. The effect of seasonal spectral variation on species classification in the Panamanian tropical forest. *Remote sensing of environment* 118:73–82. DOI: 10.1016/j.rse.2011.11.005.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* 9:90–95. DOI: 10.1109/MCSE.2007.55.

Jia X., Richards JA. 1999. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. *IEEE transactions on geoscience and remote sensing: a publication of the IEEE Geoscience and Remote Sensing Society* 37:538–542. DOI: 10.1109/36.739109.

Kampe TU., Johnson BR., Kuester MA., Keller M. 2010. NEON: the first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure. *Journal of Applied Remote Sensing* 4:043510. DOI: 10.1117/1.3361375.

Kichenin E., Wardle DA., Peltzer DA., Morse CW., Freschet GT. 2013. Contrasting effects of plant inter-and intraspecific variation on community-level trait measures along an environmental gradient. *Functional ecology* 27:1254–1261.

Kokaly RF., Asner GP., Ollinger SV., Martin ME., Wessman CA. 2009. Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote sensing of environment* 113:S78–S91. DOI: 10.1016/j.rse.2008.10.018.

Kurtzer GM., Sochat V., Bauer MW. 2017. Singularity: Scientific containers for mobility of compute. *PloS one* 12:e0177459. DOI: 10.1371/journal.pone.0177459.

Lausch A., Bannehr L., Beckmann M., Boehm C., Feilhauer H., Hacker JM., Heurich M., Jung A., Klenke R., Neumann C., Pause M., Rocchini D., Schaepman ME., Schmidtlein S., Schulz K., Selsam P., Settele J., Skidmore AK., Cord AF. 2016. Linking Earth Observation and taxonomic, structural and functional biodiversity: Local to ecosystem perspectives. *Ecological indicators* 70:317–339. DOI: 10.1016/j.ecolind.2016.06.022.

Lepine LC., Ollinger SV., Ouimette AP., Martin ME. 2016. Examining spectral reflectance features related to foliar nitrogen in forests: Implications for broad-scale nitrogen mapping. *Remote sensing of environment* 173:174–186. DOI: 10.1016/j.rse.2015.11.028.

Marconi S., Graves SJ., Gong D., Nia MS., Le Bras M., Dorr BJ., Fontana P., Gearhart J., Greenberg C., Harris DJ., Kumar SA., Nishant A., Prarabdh J., Rege SU., Bohlman SA., White EP., Wang DZ. 2018. *A data science challenge for converting airborne remote sensing data into ecological information*. PeerJ Preprints. DOI: 10.7287/peerj.preprints.26966v1.

Martin RE., Chadwick KD., Brodrick PG., Carranza-Jimenez L., Vaughn NR., Asner GP. 2018. An Approach for Foliar Trait Retrieval from Airborne Imaging Spectroscopy of Tropical Forests. *Remote Sensing* 10:199. DOI: 10.3390/rs10020199.

Mascaro J., Asner GP., Knapp DE., Kennedy-Bowdoin T., Martin RE., Anderson C., Higgins M., Chadwick KD. 2014. A tale of two "forests": random forest machine learning aids tropical forest carbon mapping. *PloS one* 9:e85993. DOI: 10.1371/journal.pone.0085993.

McGill BJ. 2010. Towards a unification of unified theories of biodiversity. *Ecology letters* 13:627–642. DOI: 10.1111/j.1461-0248.2010.01449.x.

McKinney W., Others. 2010. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. Austin, TX, 51–56.

McManus KM., Asner GP., Martin RE., Dexter KG., Kress WJ., Field CB. 2016. Phylogenetic Structure of Foliar Spectral Traits in Tropical Forest Canopies. *Remote Sensing* 8:196. DOI: 10.3390/rs8030196.

Niculescu-Mizil A., Caruana R. 2005. Predicting Good Probabilities with Supervised Learning. In: *Proceedings of the 22Nd International Conference on Machine Learning*. ICML '05. New York, NY, USA: ACM, 625–632. DOI: 10.1145/1102351.1102430.

Oliphant TE. 2007. Python for Scientific Computing. *Computing in Science Engineering* 9:10–20. DOI: 10.1109/MCSE.2007.58.

Ollinger SV. 2011. Sources of variability in canopy reflectance and the convergent properties of plants. *The New phytologist* 189:375–394. DOI: 10.1111/j.1469-8137.2010.03536.x.

Pal M. 2005. Random forest classifier for remote sensing classification. *International journal of remote sensing* 26:217–222. DOI: 10.1080/01431160412331269698.

Papeş M., Tupayachi R., Martínez P., Peterson AT., Powell GVN. 2010. Using hyperspectral satellite imagery for regional inventories: a test with tropical emergent trees in the Amazon Basin. *Journal of vegetation science*: 21:342–354. DOI: 10.1111/j.1654-1103.2009.01147.x.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher

M., Perrot M., Duchesnay É. 2011. Scikit-learn: Machine Learning in Python. *Journal of machine learning research: JMLR* 12:2825–2830.

Rodríguez JJ., Kuncheva LI., Alonso CJ. 2006. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence* 28:1619–1630. DOI: 10.1109/TPAMI.2006.211.

Siefert A., Violle C., Chalmandrier L., Albert CH., Taudiere A., Fajardo A., Aarssen LW., Baraloto C., Carlucci MB., Cianciaruso MV., Others. 2015. A global meta-analysis of the relative extent of intraspecific trait variation in plant communities. *Ecology letters* 18:1406–1419.

Townsend AR., Cleveland CC., Asner GP., Bustamante MMC. 2007. Controls over foliar N:P ratios in tropical rain forests. *Ecology* 88:107–118. DOI: 10.1890/0012-9658(2007)88[107:COFNRI]2.0.CO;2.

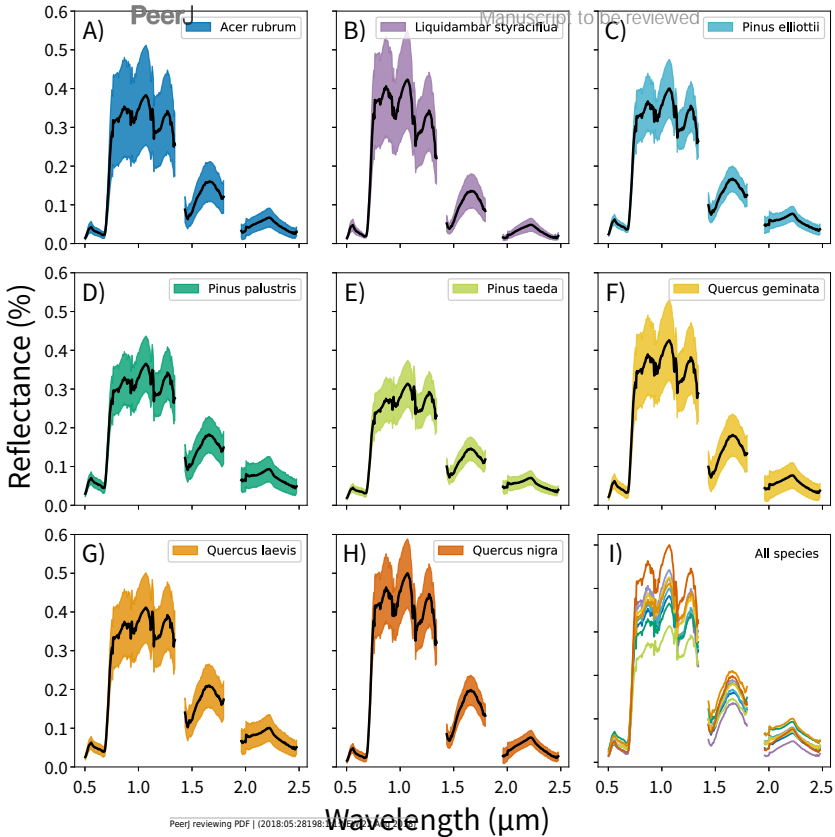
Violle C., Enquist BJ., McGill BJ., Jiang L., Albert CH., Hulshof C., Jung V., Messier J. 2012. The return of the variance: intraspecific variability in community ecology. *Trends in ecology & evolution* 27:244–252. DOI: 10.1016/j.tree.2011.11.014.

der Walt S van., Colbert SC., Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* 13:22–30. DOI: 10.1109/MCSE.2011.37.

Whittaker RJ., Araújo MB., Jepson P., Ladle RJ., Watson JEM., Willis KJ. 2005. Conservation Biogeography: assessment and prospect: Conservation Biogeography. *Diversity and Distributions* 11:3–23. DOI: 10.1111/j.1366-9516.2005.00143.x.

Yao W., van Leeuwen M., Romanczyk P., Kelbe D., van Aardt J. 2015. Assessing the impact of sub-pixel vegetation structure on imaging spectroscopy via simulation. In: *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXI*. International Society for Optics and Photonics, 94721K. DOI: 10.1117/12.2176992.

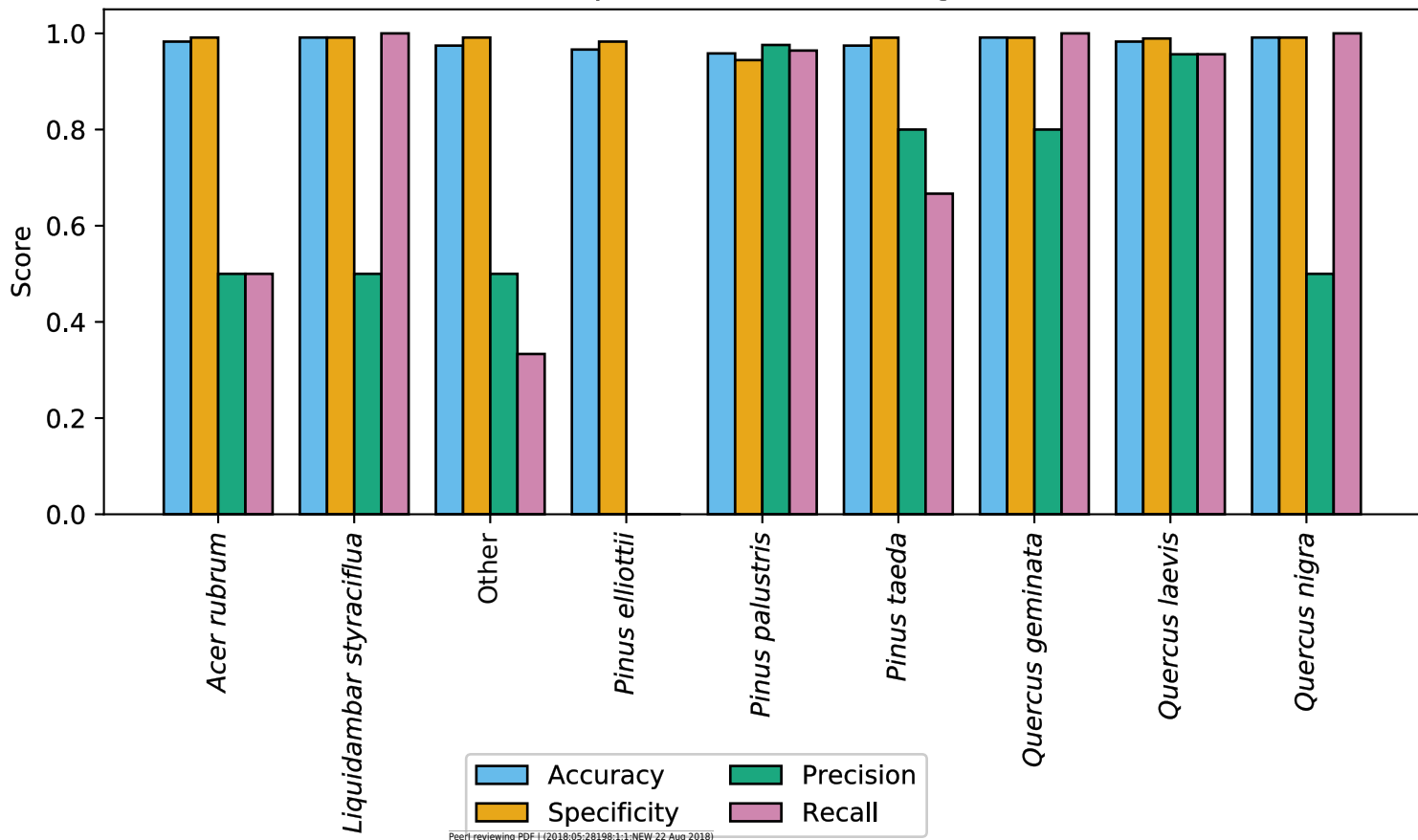
Per-species canopy reflectance profiles

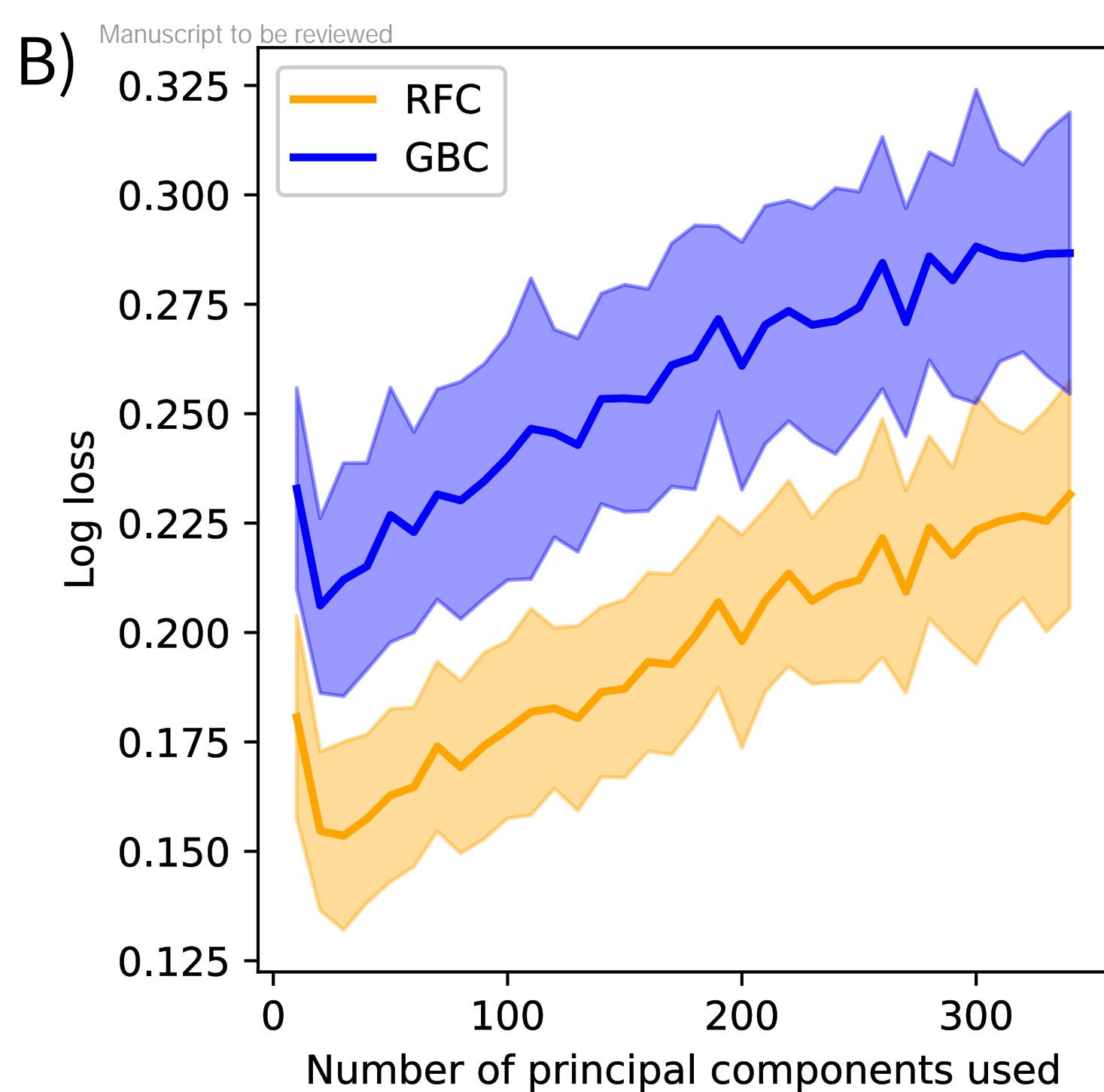
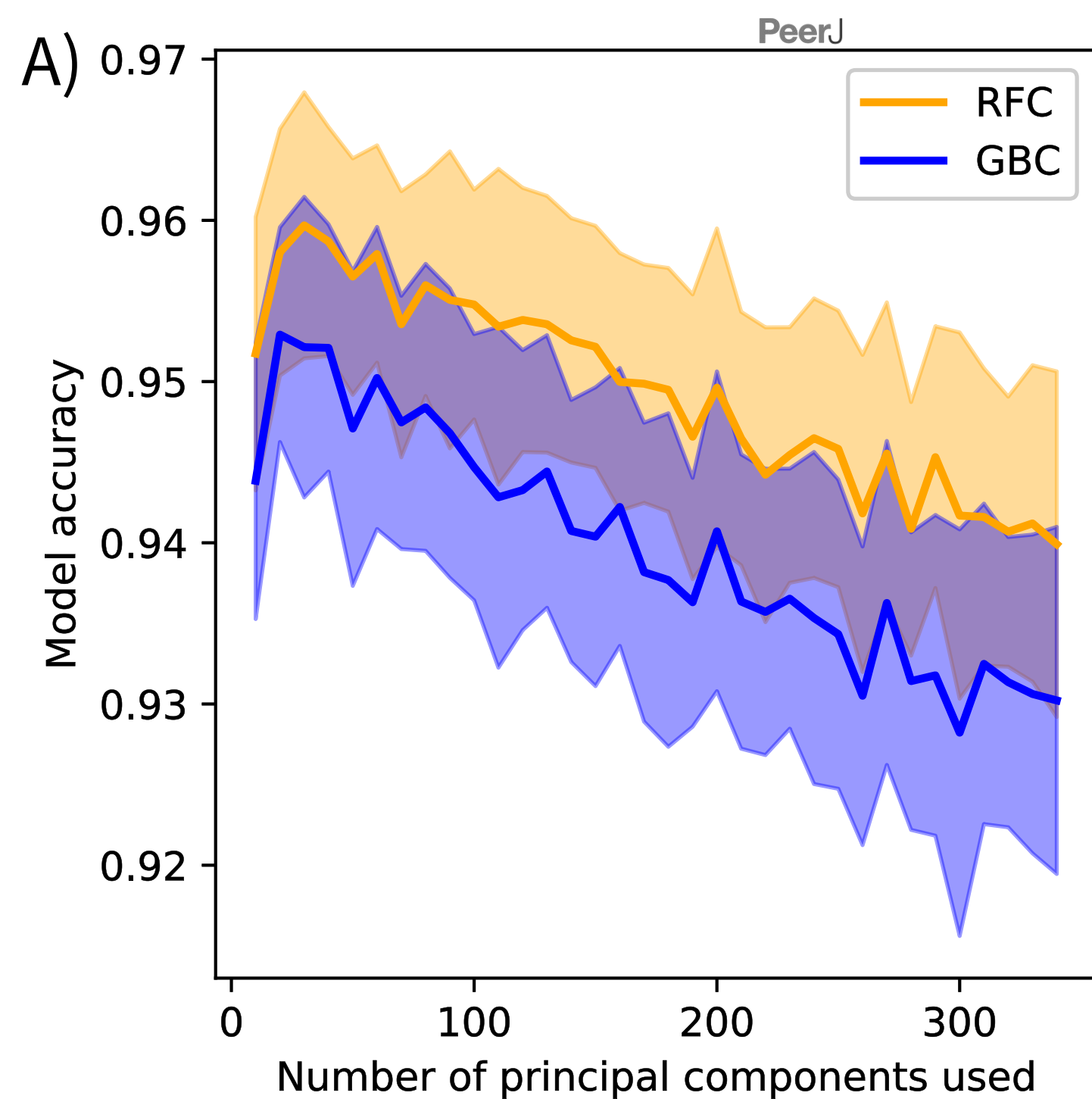
(A-H) Canopy reflectance profiles for the eight tree species analyzed, with mean reflectance values in black and +/- 1 standard deviation values in color. (I) Mean reflectance values for all species, with each color corresponding to the individual species panels. Though the mean reflectance signals show high interspecific variation, the high intraspecific variation complicates classification efforts.

Model performance metrics

Visual representation of the classification model metrics calculated on a per-species basis. A confusion matrix was computed for each species, and each metric was calculated in a one-vs.-all fashion.

CCB-ID model performance

Per-species secondary performance metrics from the test data. These metrics were calculated using the binary confusion matrix reported in Table 1. Metrics weighted by the true negative rate (i.e., accuracy and specificity) were high for all species since the models correctly predicted the most common species, *Pinus palustris*. However, metrics weighted by the true positive rate (i.e., precision and recall) were much more variable since there were only 1 to 6 observed crowns for 7 of the 9 species (*P. palustris* and *Quercus laevis* had 84 and 23 crowns, respectively). This penalized misclassifications of rare species. These metrics were re-calculated using the per-crown prediction probabilities, and can be found in Figure S1.

Model performance on testing data

Spectral variance and model performance

The effects of increasing spectral variance on model performance through altering the number of principal component features. These plots show the mean (solid) and standard deviation (shaded) of (A) model accuracy and (B) log loss scores for each classification method. Scores were calculated on holdout data from the training set, not the competition test data. These results suggest that using all available spectral variance (i.e., all principal components) may decrease model performance. Using feature selection to identify components that track variation in plant traits may prevent overfitting to noisy features. In each panel, RFC stands for random forest classifier and GBC stands for gradient boosting classifier.

Confusion matrix of classification results

Binary classification results of the CCB-ID model on the competition test data. These metrics were calculated using the independent crown data.

	Species ID	Acer rubrum	Liquidambar styraciflua	Other	Pinus elliottii	Pinus palustris	Pinus taeda	Quercus geminata	Quercus laevis	Quercus nigra
					Predicted					
	Acer rubrum	**1**	0	0	0	0	0	0	0	1
	Liquidambar styraciflua	0	**1**	0	0	0	0	0	0	0
	Other	1	1	**1**	0	0	0	0	0	0
Observed	Pinus elliottii	0	0	0	**0**	1	1	0	0	0
	Pinus palustris	0	0	0	2	**81**	0	0	1	0
	Pinus taeda	0	0	1	0	0	**4**	1	0	0
	Quercus geminata	0	0	0	0	0	0	**4**	0	0
	Quercus laevis	0	0	0	0	1	0	0	**22**	0
	Quercus nigra	0	0	0	0	0	0	0	0	**1**