A peer-reviewed version of this preprint was published in PeerJ on 8 October 2018.

View the peer-reviewed version (peerj.com/articles/5666), which is the preferred citable publication unless you specifically need to cite this preprint.

Anderson CB. 2018. The CCB-ID approach to tree species mapping with airborne imaging spectroscopy. PeerJ 6:e5666
https://doi.org/10.7717/peerj.5666

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 obviate this challenge. Airborne imaging spectrometers measure plant functional traits at high resolution, and these measurements can be used to identify tree species. Plant traits are often highly conserved within species, and highly variable between species, which provides the biophysical basis for species mapping. In this paper I describe a trait-based approach to species identification with imaging spectroscopy, *CCB-ID*, which was developed as part of a NIST-sponsored ecological data science evaluation (ECODSE).

Methods. These methods were developed using NEON airborne imaging spectroscopy data. 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 traditionally used in botanical classification. First, outliers were removed using a spectral variance threshold. The remaining samples were transformed using principal components analysis and resampled by species 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 imaging spectroscopy data, reporting the best classification 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26972v1 | CC BY 4.0 Open Access | rec: 1 Jun 2018, publ: 1 Jun 2018

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 obviate this challenge. Airborne imaging spectrometers measure plant functional traits at high resolution, and these measurements can be used to identify tree species. Plant traits are often highly conserved within species, and highly variable between species, which provides the biophysical basis for species mapping. In this paper I describe a trait-based approach to species identification with imaging spectroscopy, *CCB-ID*, which was developed as part of a NIST-sponsored ecological data science evaluation (ECODSE).

Methods. These methods were developed using NEON airborne imaging spectroscopy data. 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 traditionally used in botanical classification. First, outliers were removed using a spectral variance threshold. The remaining samples were transformed using principal components analysis and resampled by species 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 imaging spectroscopy data, reporting the best classification 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.

Introduction

When you get down to it, biogeographers seek to answer two key questions: where are the species, and why are they where they are? Answering these simple questions has proven remarkably difficult. The former reflects a data gap; we do not have complete or unbiased information on where species occur. This is known as the 'Wallacean shortfall' (Whittaker et al., 2005; Bini et al., 2006). Addressing the latter, however, does not necessarily require data; the drivers of species abundances and their spatial distributions can be derived from ecological theory (McGill, 2010). But evaluating these theoretical predictions does require data. Testing generalized theories of species distributions requires continuously-mapped presences and absences for many individuals across large areas. And while field efforts can assess fine-scale distribution patterns, they are often restricted to small extents. Mapping organism-scale species distributions over landscapes could help fill the data gaps that preclude addressing these key biogeographic questions. One remote sensing dataset holds the promise to do so: airborne imaging spectroscopy.

Airborne imaging spectrometers measure variation in the biophysical properties of soils and vegetation at fine grain sizes across large areas (Goetz et al., 1985). In vegetation mapping contexts, imaging spectroscopy can map plant structural traits, like leaf area index and leaf angle distribution (Broge & Leblanc, 2001; Asner & Martin, 2008), and plant functional traits, like growth and defense compound concentrations (Kokaly et al., 2009; Asner et al., 2015). These traits tend to be highly conserved within tree species, and highly variable between species (i.e., interspecific trait variation is often much greater than intraspecific trait variation; (Townsend et al., 2007; Asner et al., 2011). This trait conservation provides the conceptual and biophysical basis for species mapping with imaging spectroscopy. Indeed, airborne imaging spectroscopy has been used to map crown-scale species distributions across large extents in several contexts

(Fassnacht et al., 2016). These approaches have been applied in temperate (Baldeck et al., 2014) and tropical ecosystems (Hesketh & Sánchez-Azofeifa, 2012), using multiple classification methods (Feret & Asner, 2013) and multiple sensors (Clark, Roberts & Clark, 2005; Colgan et al., 2012; Baldeck et al., 2015). However, this wide range of approaches has not yet identified a canonical best practice for tree species identification.

In this paper I describe an approach to tree species classification using airborne imaging spectroscopy data that builds on the above methods to advance the discussion on best practices. This approach was developed as a submission to a NIST-sponsored ecological data science evaluation competition (ECODSE; https://ecodse.org). This competition had participants use airborne imaging spectroscopy data, collected by the National Ecological Observatory Network's Airborne Observation Platform (NEON AOP), to identify tree crowns to the species level. The work described below was submitted under the team name of the Stanford Center for Conservation Biology (Stanford CCB), and has since been formalized under the moniker *CCB-ID* (https://github.com/stanford-ccb/ccb-id). First, I describe the CCB-ID approach to tree species classification using airborne imaging spectroscopy data. Next, I review its successes and shortcomings in the context of this competition. Finally, I highlight key opportunities to improve future imaging spectroscopy-based species classification approaches. The goals of this work are to improve NEON's operational tree species mapping efforts and to reduce barriers for addressing key data gaps in biogeography.

Materials & Methods

The CCB-ID approach was inspired by botanical and taxonomic approaches to species classification. In the field, botanists can use plant morphological features and a dichotomous key to identify tree species. These features often include variations in reproductive traits (e.g.,

flowering bodies, seeds), vascular traits (e.g., types of woody and non-woody tissue), and foliar traits (e.g., waxy or serrated leaves). The dichotomous key approach hierarchically partitions species until each can be identified using a specific combination of traits.

Species classification with imaging spectroscopy is rather restricted, in comparison; imaging spectrometers can only measure a subset of plant traits. Furthermore, the inter- and intraspecific variation in this subset of traits is rarely known *a priori*, precluding the use of a standard dichotomous key. Imaging spectroscopy approaches to species classification instead rely on distinguishing species-specific variations in canopy reflectance signal. However, several confounding factors drive variation in canopy reflectance data, including (1) measurement conditions (e.g., sun and sensor angles), (2) canopy structure (e.g., leaf area index or leaf angle distribution), (3) leaf morphology and physiology (i.e., plant functional traits), and (4) sensor noise. Measurement conditions and canopy structure tend to drive the majority of variation; up to 79-89% of spectral variance is driven by within-crown variation (Baldeck & Asner, 2014; Yao et al., 2015). Unfortunately, this variation does not help distinguish between species. Interspecific spectral variation is instead driven by functional trait variation (Asner et al., 2011). Disentangling trait-based variation from measurement and structure-based variation is thus central to mapping species with airborne imaging spectroscopy.

CCB-ID classifies tree species using trait-based reflectance variation with decision tree-based machine learning models. This approach approximates a morphological trait and dichotomous key model, and is described in the following sections. The first section describes the outlier removal and data transformation procedures. The second section describes how the training data were resampled to reduce biases towards common species. The third section describes model selection, training, and probability calibration. The fourth section describes the

model performance metrics, and the final section describes two analyses performed post-ECODSE submission.

The data from NEON included the following data products: 1) Woody plant vegetation structure (NEON.DP1.10098); 2) Spectrometer orthorectified surface directional reflectance - flightline (NEON.DP1.30008); 3) Ecosystem structure (NEON.DP3.30015); and High-resolution orthorectified camera imagery (NEON.DP1.30010). These data were provided by the ECODSE group (2017; https://ecodse.org). All analyses were performed using the Python programming language (Oliphant, 2007; https://python.org) and the following open source packages: NumPy (der Walt, Colbert & Varoquaux, 2011; http://numpy.org), scikit-learn (Pedregosa et al., 2011; http://scikit-learn.org), pandas (McKinney et al., 2010; https://pandas.pydata.org), and matplotlib (Hunter, 2007; https://matplotlib.org). The python scripts used for these analyses have been uploaded to a public GitHub repository (https://github.com/stanford-ccb/ccb-id), including a build script for a Singularity container to ensure computational replicability (Kurtzer, Sochat & Bauer, 2017).

Data preprocessing

The canopy reflectance data were preprocessed using two steps: outlier removal and dimensionality reduction. In the outlier removal step, the reflectance data were spectrally subset, transformed using PCA, then thresholded to isolate spurious values. First, reflectance values from the blue region of the spectrum (0.38-0.49 μm) and from noisy bands (1.35-1.43 μm, 1.80-1.96 μm, and 2.48-2.51 μm) were removed. These bands correspond to wavelengths dominated by atmospheric water vapor, and do not track variations in plant traits (Gao et al., 2009; Asner et al., 2015). Next, these spectrally-subset samples were transformed using PCA. The output components were whitened to zero mean and unit variance, and outliers were identified using a

three-sigma threshold. Samples with values outside of +/- three standard deviations from the means (i.e., which did not fall within 99.7% of the variation for each component) for the first 20 principal components were excluded from analysis. These samples were expected to contain non-vegetation spectra (e.g., exposed soil), unusually bright or dark spectra, or anomalously noisy spectra (Féret & Asner, 2014). The outlier-removed reflectance profiles for each species are shown in Figure 1.

Once the outliers were removed, the remaining spectra were transformed using PCA. This was not performed on the already-transformed data from the outlier removal process, but on the outlier-removed, spectrally-subset reflectance data. PCA transformations are often applied to airborne imaging spectrometer data to handle the high degree of correlation between bands, and these transformations are highly sensitive to input feature variation (Jia & Richards, 1999). Furthermore, transforming reflectance data into principal components can isolate the variation driven by measurement conditions from variation driven by functional traits, which is critical for distinguishing between species. And though trait-based variation drives a small proportion of total reflectance signal, a single trait can be expressed in up to 9 orthogonal components (Asner et al., 2015). After the transformation, the first 100 components were used as feature vectors for the species classification models. This threshold was arbitrary; it was set to capture the majority of biologically-relevant components and to exclude noisy components.

Class imbalance

Class imbalance refers to datasets where the number of samples per-class are not evenly distributed among classes. Imbalanced data sets are common in classification contexts, but can lead to problems if this imbalance is not addressed. Training classification models with imbalanced data can select for models that overpredict common classes when model

performance is based on metrics like accuracy. The ECODSE data were imbalanced: after outlier removal, these data contained a total of 6,034 samples from 9 classes (8 identified species, one 'other species' class). The most common species, *Pinus palustris*, contained 4,026 samples (66% of the samples) and the rarest species, *Liquidambar styaciflua*, contained 62 samples (1% of the samples).

These data were resampled prior to analysis to reduce the likelihood of overpredicting common species. Resampling was performed by setting a fixed number of samples per class, then undersampling or oversampling each class to that fixed number. This fixed number was set to 400 samples to split the difference of two orders of magnitude between the rarest and the most common classes. This number was arbitrary, but it approximates the number of per-species samples recommended in Baldeck & Asner (2014). To create the final training data, classes with fewer than 400 samples were oversampled with replacement, and classes with more than 400 samples were undersampled without replacement. The final training data included 400 samples for each of the 9 classes (3,600 samples total), where each sample contained a feature vector of the principal components derived from the outlier removed, spectrally subset canopy reflectance data.

Model selection, training, and probability calibration

Th CCB-ID approach used two machine learning models: a gradient boosting classifier (GBC) and a random forest classifier (RFC; (Friedman, 2001; Breiman, 2001). These models can fit complex, nonlinear relationships between response and feature data, can automatically handle interactions between features, and have built-in mechanisms to reduce overfitting. They were selected because they perform well in species mapping contexts (Elith, Leathwick & Hastie, 2008), in remote sensing contexts (Pal, 2005), and in conjunction with PCA transformations

(Rodríguez, Kuncheva & Alonso, 2006). Furthermore, these models are built as ensembles of decision trees, resembling the dichotomous key employed by botanists. Unlike a dichotomous key, these models were trained to learn where to split the data since the trait variation that distinguishes species was not known *a priori*.

These models were fit using hyper-parameter tuning and probability calibration procedures. Model hyper-parameters were tuned by selecting the parameters that maximized mean F1 scores in 5-fold cross-validation using an exhaustive grid search. The following parameters were tuned for both models: number of estimators, maximum tree depth, minimum number of samples required to split a node, and minimum node impurity split threshold. The learning rate and node split quality criterion were also tuned for GBC and RFC, respectively. All samples were used for hyper-parameter tuning, and the best model hyper-parameters (i.e., the hyper-parameters that maximized mean F1 scores in cross validation) were used to fit the final models.

Accurately characterizing prediction probabilities is essential for error propagation and for assessing model reliability, and prediction probabilities were calibrated once the final hyper-parameters were selected. Well-calibrated probabilities should scale linearly with the true rate of misclassification (i.e., should not be under or overconfident in predictions). Some ensemble methods, like RFC, tend to be poorly calibrated; their variance can skew high probabilities away from one, and low probabilities away from zero, since they average their predictions from a set of weak learners, which individually have high misclassification rates but gain predictive power post-ensemble. This results in sigmoid-shaped reliability diagrams (DeGroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005).

To reduce these probability biases, prediction probabilities were calibrated using sigmoid regression for both RFC and GBC. The data were first randomly split into three subsets: model training (50%, or 200 samples per class), probability calibration training (25%, or 100 samples per class), and model testing (the remaining 25%). Each classifier was fit using the model training subset and the tuned hyper-parameters. Prediction probabilities were calibrated with sigmoid regression, using the probability training subset and internal 3-fold cross validation to assess the calibration. Calibrated model performance was assessed using the holdout test set. After these assessments, the final models were fit using the model training data, then calibrated using the full probability training and testing data (i.e., the full 50% of samples not used in initial model training).

Model assessment

During model training, performance was assessed on a per-sample basis using model accuracy and log loss scores. Model accuracy calculates the proportion of correctly classified samples in the test data (Figure 2). High model accuracy scores are desirable. Log loss assesses whether the prediction probabilities were well calibrated, penalising incorrect and uncertain predictions. Low log loss scores indicate that misclassifications occur at rates close to the rates predicted by the reported probabilities. During model testing, performance was assessed using rank-1 accuracy and cross entropy cost (Marconi et al., 2018). Rank-1 accuracy was calculated based on which species ID was predicted with the highest probability. The cross entropy score is similar to the log loss function, but was scaled using an indicator function. These can be interpreted in similar ways to accuracy and log loss; high rank-1 accuracy and low cross entropy scores are desirable.

Secondary model testing metrics were calculated for each species using the test data. These included model specificity, precision, and recall (Figure 2). These metrics reveal model behavior that accuracy scores may obscure. Specificity assesses model performance on non-target species, penalizing overprediction of the target species (i.e., a high number of false positives). Precision also penalizes overprediction, but assesses the rate of overprediction relative to the rate of true positive predictions. Recall calculates the proportion of true positive predictions to the total number of positive observations per species. Higher values are desirable for each. These metrics were calculated to aid interpretation, but were not used to formally rank model performance.

Performance during model training was assessed at the sample scale, but the competition evaluation metrics were calculated using crown-scale prediction probabilities. To address this scale mismatch, prediction probabilities were first calculated for each sample in a crown using both GBC and RFC models. These sample-scale probabilities were then averaged by crown.

Further analyses

Two post-submission analyses were performed to assess how PCA transformations affected model performance. Prior to these analyses, I bootstrapped the original model fits to assess their variance. I then compared these bootstrapped fits to models trained with the spectrally-subset reflectance data instead of the PCA transformed data. Next, I compared models trained using a varying number of principal components. These models were trained using *npcs* $\in \{10, 20, \ldots, 345\}$ as the input features, with 345 being the maximum number of potential components after spectral subsetting. These comparisons assessed whether the PCA transformations improved model performance, and how changing the amount of spectral

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

Results

CCB-ID performed well according to the ECODSE competition metrics, receiving a rank-1 accuracy score of 0.919, and a cross-entropy cost score of 0.447 on the test data. These were the highest rank-1 accuracy and the lowest cross-entropy cost scores among participants. A confusion matrix with the classification results is reported in Table 1. In addition to the high rank-1 accuracy and low cross entropy cost scores, the CCB-ID model performed well according to the secondary crown-scale performance metrics. These secondary metrics calculated a mean accuracy score of 0.979, mean specificity score of 0.985 , mean precision score of 0.614, and mean recall score of 0.713 across all species. The per-species secondary metrics are summarized in Figure 3. These results were calculated using the categorical classification predictions (i.e., after assigning ones to the species with the highest probabilities, and zeros to all other species). The probability-based confusion matrix and classification metrics are reported in Table S1 and Figure S1, respectively.

During model training, outlier removal excluded 797 samples from analysis. After removal, the first principal component contained 78% of the explained variance. However, this component did not drive model performance; it ranked 7th and 11th in terms of ranked feature importance scores for GBC and RFC. Model accuracy scores, calculated on the 25% training data holdout, were 0.933 for GBC and 0.956 for RFC. Log loss scores, calculated prior to probability calibration, were 0.19 for GBC, and 0.47 for RFC. After probability calibration, log loss scores were 0.24 for GBC and 0.16 for RFC. The per-class secondary metrics reported a

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

The post-submission analyses found PCA transformations improved model accuracy. Models fit using the original methods calculated mean bootstrapped accuracy scores of 0.944 (s = 0.009) for GBC and 0.955 (s = 0.008) for RFC. Models fit using the spectrally-subset reflectance data as features calculated mean accuracy scores of 0.883 (s = 0.012) for GBC and 0.877 (s = 0.011) for RFC, and mean log loss scores of 0.46 (s = 0.03) for GBC and 0.48 (s = 0.03) for RFC. For the models fit using varying numbers of principal components, mean model accuracies declined and mean log loss scores increased after including more than 20 components as features (Figure 4).

Discussion

The CCB-ID approach accurately classified tree species using NEON imaging spectroscopy data, reporting the highest rank-1 accuracy score and lowest cross-entropy cost score among ECOSDE participants. These scores compare favorably to other imaging spectroscopy-based species classification efforts (Fassnacht et al., 2016). The crown-scale test results highlight the potential to develop species mapping methods that approximate botanical and taxonomic approaches to classification. However, this method failed to overcome several well-known species mapping challenges, like precisely identifying rare species. Below, I discuss some key takeaways, and suggest opportunities to improve future imaging spectroscopy-based species classification approaches.

The high per-species accuracy scores indicate a high proportion of correctly classified crowns in the test data. However, accuracy can be a misleading metric in imbalanced contexts. Since seven of the nine classes had six or fewer crowns in the test data (out of 126 total test

crowns), classification metrics weighted by the true negative rate (i.e., accuracy and specificity) were expected to be high if the majority class were correctly predicted. Metrics weighted instead by the true positive rate (i.e., precision and recall) showed much higher variation across rare species, as a single misclassification greatly alters these metrics when there are few observed crowns (Figure 3). Due to the small sample size, it is difficult to assess if these patterns portend problems at larger scales. For example, there were two observed *Acer rubrum* crowns in the test data, yet only one was correctly predicted. Was the misclassified crown an anomaly? Or will this low precision persist across the landscape, predicting *Acer rubrum* occurrences at half its actual frequency? The latter seems unlikely, in this case; the low cross entropy and log loss scores suggest misclassified crowns were appropriately uncertain in assigning the wrong label (Table S1). However, since airborne species mapping is employed to address large-scale ecological patterns where precision is key (e.g., in biogeography, macroecology, and biogeochemistry), we should be assessing classification performance on more than one or two crowns per species.

Model performance between and within taxonomic groups revealed some notable patterns. *Quercus* and *Pinus* individuals (i.e., Oaks and Pines) accounted for 120 of the 126 test crowns and there was high fidelity between them; only one *Quercus* crown was misclassified as *Pinus*, and two *Pinus* crowns were misclassified as *Quercus*. From a botanical perspective, this makes sense; these genera exhibit very different growth forms (i.e., different canopy structures and foliar traits), and should thus be easy to distinguish in reflectance data. However, *Quercus* and *Pinus* showed different within genus patterns. *Quercus* crowns were never misclassified as other *Quercus* species, yet there were several within-*Pinus* misclassifications. This may be because *Quercus* species tightly conserve their canopy structures and foliar traits (Cavender-Bares et al., 2016), while *Pinus* species may express trait plasticity. *Pinus* species maintain

similar growth forms (i.e., their needles grow in whorls bunched through the canopy), limiting opportunities to identify species-specific structural variation. Furthermore, they are distributed across the varying climates of the southern, eastern, and central United States, suggesting some degree of niche plasticity. If this plasticity is expressed in each species' functional traits, then convergence among species may then preclude trait-based classification efforts. Quantifying the extent to which foliar traits are conserved within and between species and genera will be essential for assessing the potential for imaging spectroscopy to map community composition across large extents.

The post-submission analyses revealed further notable patterns. First, PCA transformation increased mean model accuracy scores compared to the spectrally-subset reflectance data. I suspect this is because the models could focus on the spectral variation driven by more biologically meaningful components instead of searching for that signal in the reflectance spectrum where the majority of variation is driven by abiotic factors. The low feature importance scores of the first principal component support this interpretation. The first component in reflectance data is typically driven by brightness (i.e., not a driver of interspecific variation) and contained 78% of the explained reflectance variance, but ranked low in feature importance for both models. This preprocessing transformation approximates the 'rotation forest' approach developed by Rodríguez, Kuncheva & Alonso (2006). They found PCA transformation improved tree-based ensemble models in several contexts, and they suggested retaining all components to maintain the dimensionality of the input features. However, the analysis that varied the number of feature components showed model accuracy decreased when including more than 20 components (Figure 4). This suggests that using all components could overfit to noise. Performing feature selection on transformed data may help overcome this. Feature

297 selection has been applied to reflectance data to find the spectral features that track functional
298 trait variation (Feilhauer, Asner & Martin, 2015), and I believe it could help identify trait-based
299 components that discriminate between species. Furthermore, other transformation methods may
300 be more appropriate than PCA; principal components serve only as proxies to functional traits in
301 this context. Transforming reflectance data directly into trait-based features could improve
302 species classification efforts, improve model interpretability, and further develop the biophysical
303 basis for species mapping with imaging spectroscopy.

304 Despite the successes of CCB-ID, there were a few missteps in model design and
305 implementation. For example, outlier removal and resampling were employed to obviate class
306 imbalance problems but may instead have exacerbated them. First, the PCA-based outlier
307 removal excluded samples based on deviation from the mean of each component. However,
308 since the transformations were calculated using imbalanced data, the majority of the variance
309 was driven by variation in the most common class. This means outlier removal excluded samples
310 that deviated too far from the mean-centered variance weighted by *Pinus palustris*. Indeed, 533
311 of the 797 samples excluded from analysis (67%) were from non-*P. palustris* species (which
312 comprised only 37% of the full dataset). This removed up to 45% of samples from the rarest
313 species (*Liquidambar styraciflua*), reducing the spectral variance these models should be trained
314 on to better identify rare species. This suggests, for rare species, outlier removal should either be
315 skipped or implemented using other methods (e.g., using spectral mixture analysis to identify
316 samples with high soil fractions).

317 Data resampling further exacerbated the class imbalance problem. By setting the
318 resampling threshold an order of magnitude above the least sampled class, the rarest species
319 were oversampled nearly tenfold in model training. This inflated per-class model performance

metrics by double-counting (or more) correctly classified samples for oversampled species. These metrics were further inflated as a result of how the train/test data were split. The split was performed after resampling, meaning the train/test data for oversampled species were likely not independent. This invalidated their use as true test data, and further overestimated performance during model training. This is unequivocally bad practice; I call this "user error." Undersampling the common species was also detrimental. Excluding samples from common species meant the models were exposed to less intraspecific spectral variation during training, a key source of variance the models should recognize. This made it more difficult for the models to distinguish inter and intraspecific variation. Assigning sample weights (e.g., proportional to the number of samples per class) and using actually independent holdout data could obviate these issues. These will be implemented in future versions of CCB-ID. However, these need not be the only updates to this method; CCB-ID is an open source, freely available project (https://github.com/stanford-ccb/ccb-id), and I invite you to to use it and improve it.

Conclusions

It wasn't always possible to classify tree species from airplanes; now it is. Airborne imaging spectrometers can identify trees at crown scales across large areas, and these data are now publicly available through NEON. However, there is currently no canonical imaging spectroscopy-based species mapping approach, limiting opportunities to explore key patterns in biogeography. CCB-ID was developed to identify best practices for species classification in this context, and to further the conversation on how to implement these practices. CCB-ID performed well within the scope of the ECODSE competition, reporting the highest rank-1 accuracy and lowest cross entropy scores among participants, but further testing is necessary to identify whether this method can scale to other regions (e.g., to high diversity forests). I hope CCB-ID

will be used to improve future species mapping efforts, and to pursue answers to biogeography's great mysteries of where the species are, and why they are there.

Acknowledgements

I would like to thank Gretchen Daily for her continued advisement, support, and inspiration. Thanks to the organizers of the NSF NEON workshop on mapping species, foliar chemistry and soil properties with spectroscopy, including Nancy Glenn, Nathan Leisso, Jessica Mitchell, Yi Qi, and Dar Roberts. Thanks to Phil Brodrick for being good at models, and even better at explaining them. Finally, I would like to thank Jeff Smith for comments on this manuscript, and for fruitful conversations on hyperspectral image mixing.

References

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., Stereńczak 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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

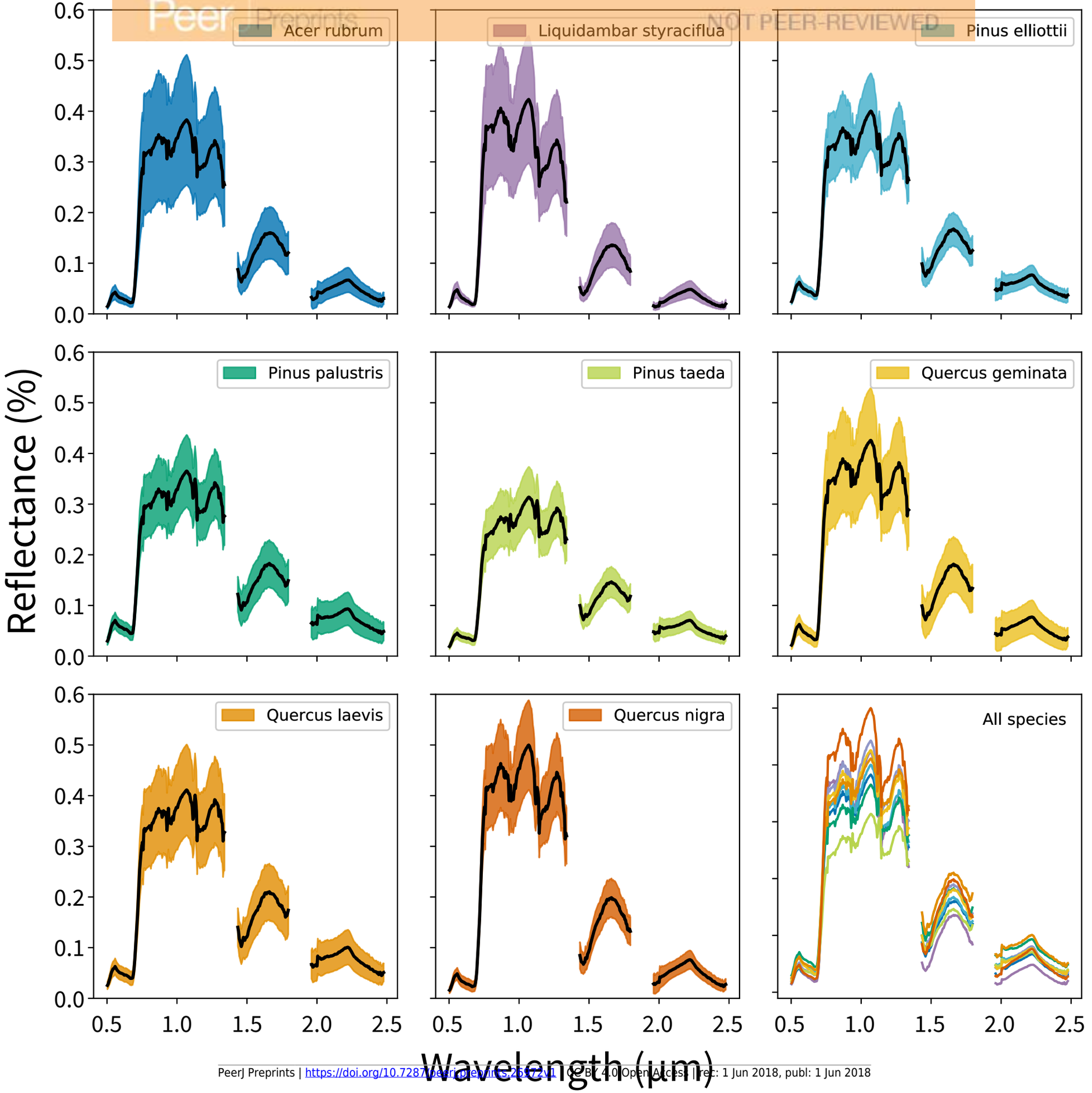
Whittaker RJ., Araújo MB., Jepson P., Ladle RJ., Watson JEM., Willis KJ. 2005. Conservation Biogeography: assessment and prospect: Conservation Biogeography. *Diversity and Distributions* 11:3–23. DOI: 10.1111/j.1366-9516.2005.00143.x.

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.

Figure 1(on next page)

Per-species canopy reflectance profiles

Canopy reflectance profiles for the eight tree species analyzed, with mean reflectance values in black and +/- 1 standard deviation values in color. The bottom right panel shows the mean reflectance values for each 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.

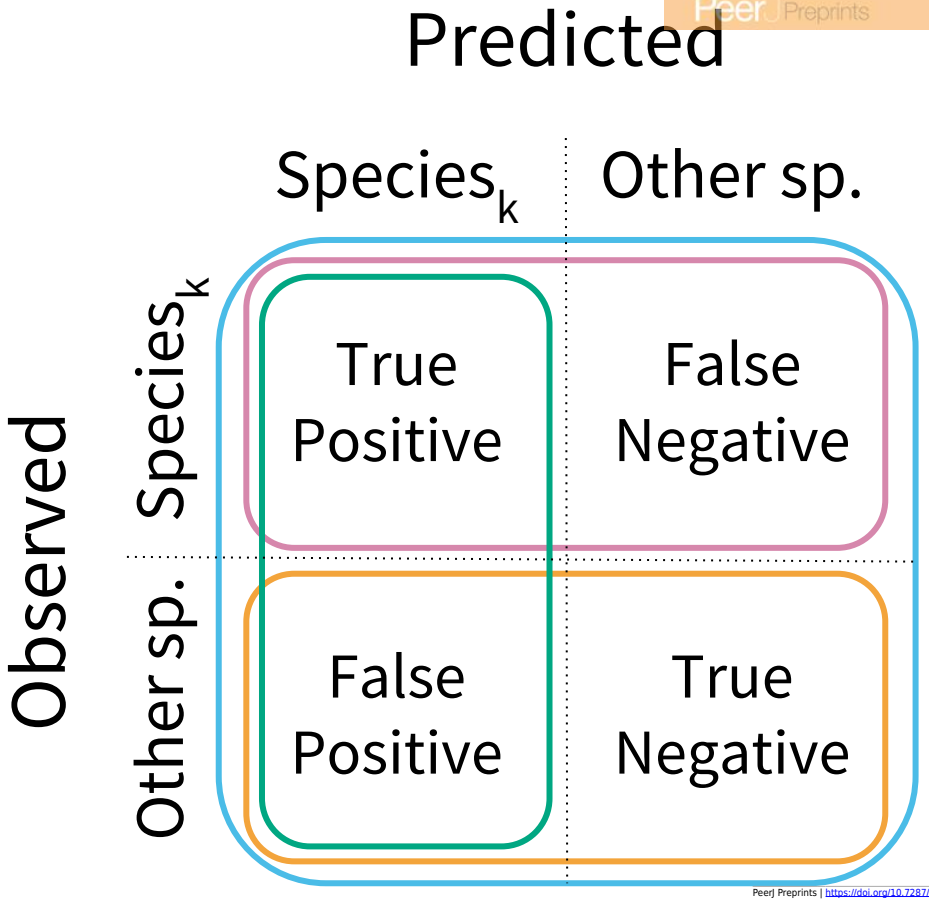


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26972v1 | CC BY 4.0 Open Access | rec: 1 Jun 2018, publ: 1 Jun 2018

Figure 2(on next page)

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.



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

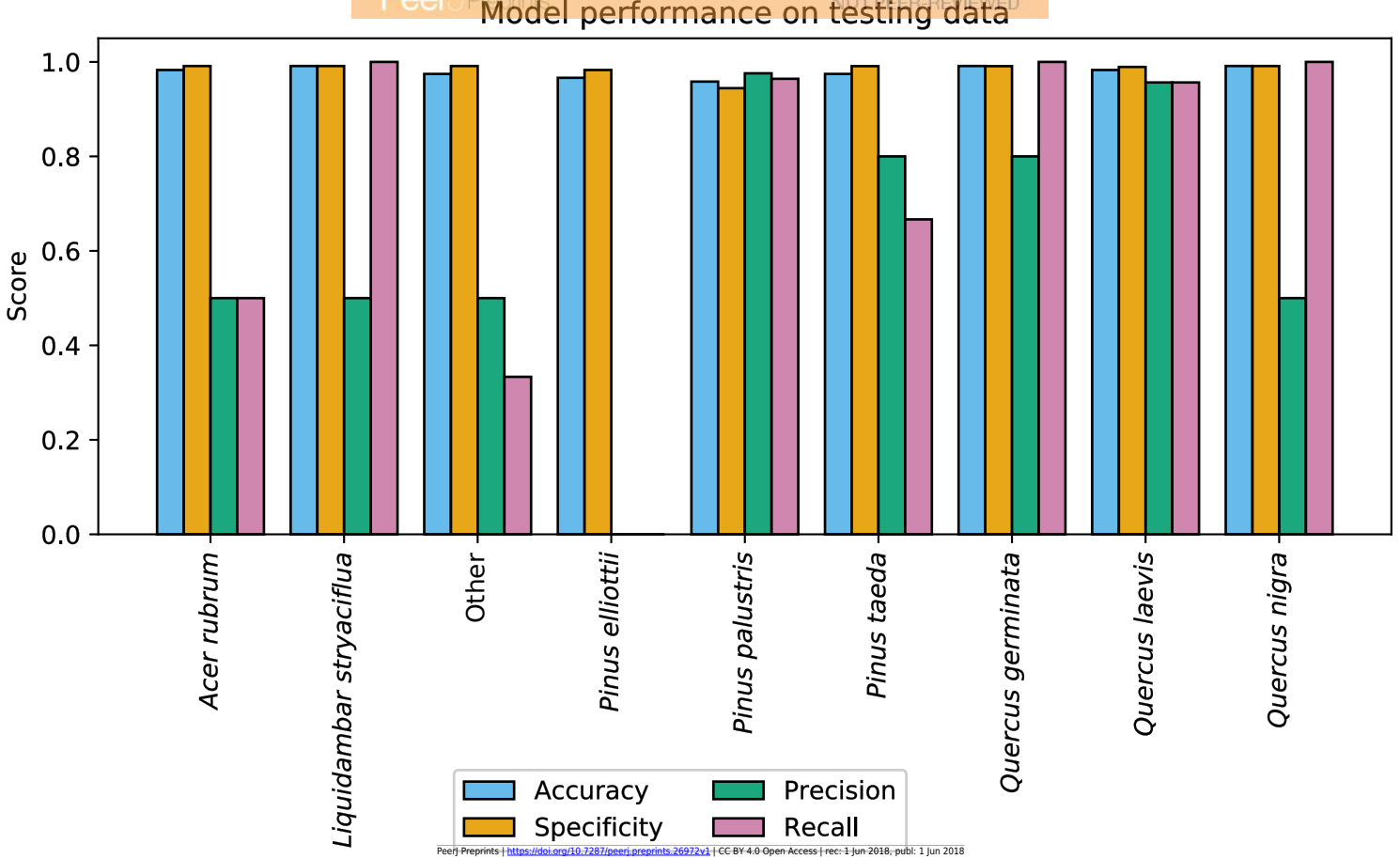
$$\text{Recall} = \frac{TP}{TP + FN}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26972v1 | CC BY 4.0 Open Access | rec: 1 Jun 2018, publ: 1 Jun 2018

Figure 3(on next page)

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.

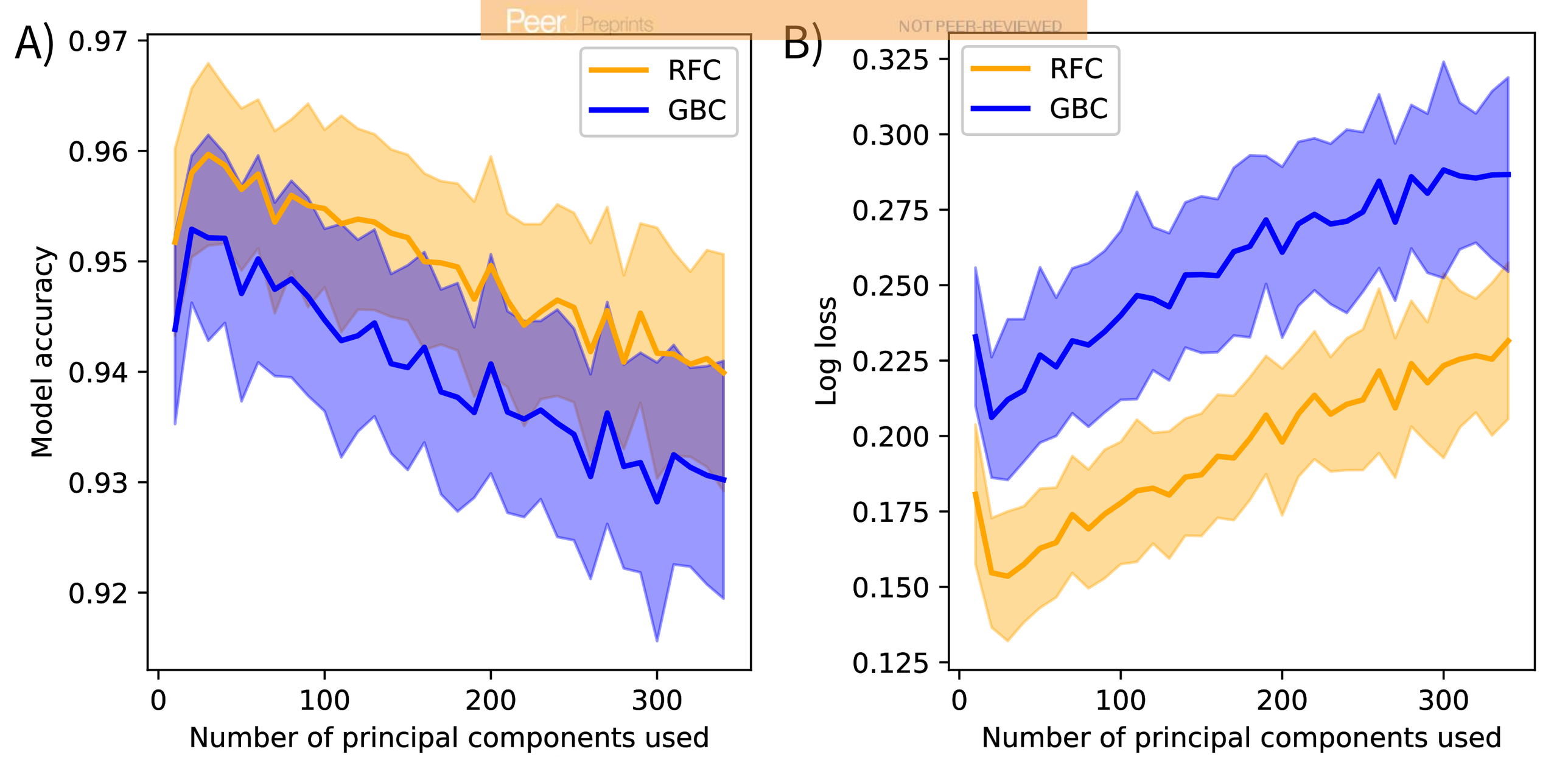


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26972v1 | CC BY 4.0 Open Access | rec: 1 Jun 2018, publ: 1 Jun 2018

Figure 4(on next page)

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.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26972v1 | CC BY 4.0 Open Access | rec: 1 Jun 2018, publ: 1 Jun 2018

Table 1(on next page)

Confusion matrix of classification results

Binary classification results of the CCB-ID model on the competition test data.

					Predicted					
	Species ID	Acer rubrum	Liquidambar stryaciflua	Other	Pinus elliottii	Pinus palustris	Pinus taeda	Quercus germinata	Quercus laevis	Quercus nigra
	Acer rubrum	**1**	0	0	0	0	0	0	0	1
	Liquidambar stryaciflua	0	**1**	0	0	0	0	0	0	0
	Other	1	1	**1**	0	0	0	0	0	0
Actual	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 germinata	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**