NEON NIST Data Science Evaluation Report for Baseline

UF DSE Team
November 7, 2017

Overall Performance

Here is the summary of overall performance for all tasks.

- ✓ **Crown Delineation:** 0.0014
- ✓ **Crown Alignment:** 0.4800
- ✓ **Species Classification:** 1.1306 (cross-entropy cost), 0.6667 (rank-1 accuracy)

Task 1 - Crown Delineation

Overall Confusion Matrix

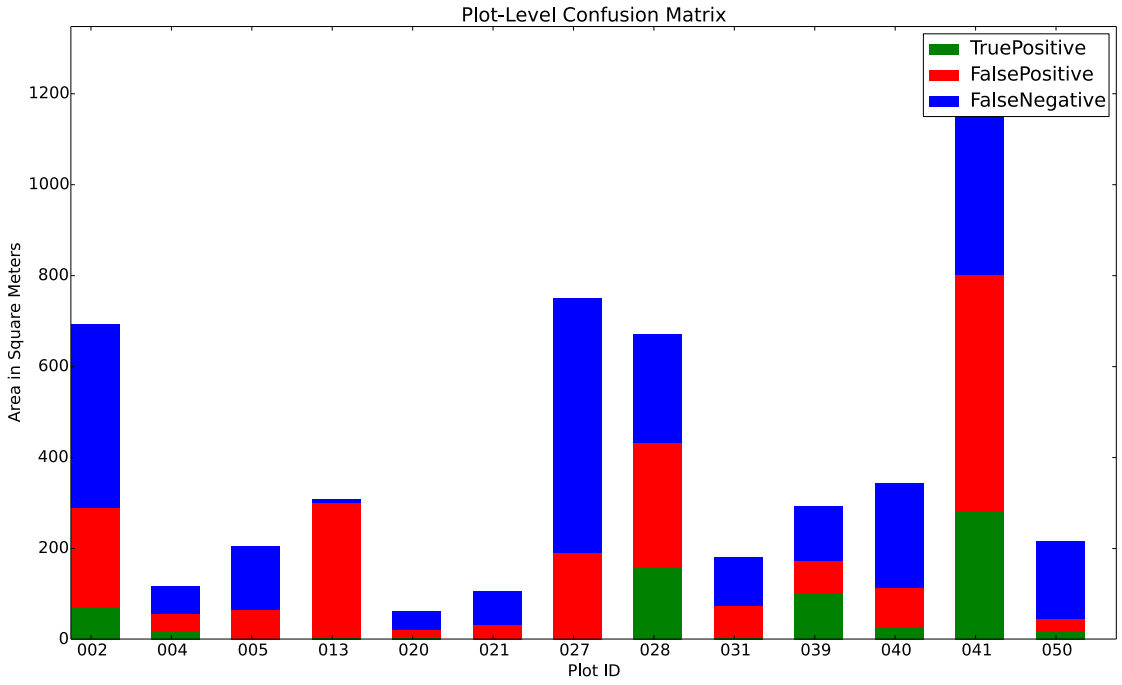
The overall confusion matrix (OCM) measures the area in square meters that is correctly or incorrectly classified as crown or not in the delineation task. The OCM accumulates the counts of area overall all testing plots, as shown in the following table.

Area (Square Meters)	**Positive**	**Negative**
True	684.3	-
False	1916.2	2631.6

In the Table, an area is counted as true positive if it is within any groudtruth crown and it is also classified as part of output crown. The other quantities can be defined similarly.

Plot-Level Confusion Matrix as a Bar Chart

To analyze the performance w.r.t. each plot, we visualize the confusion matrix for each plot, as shown in the following figure. The confusion matrix of each plot is calculated in a similar manner with OCM, the only difference is that the area is accumulated within the plot only, rather than over all testing plots.



Example Delineation

The top-6 best and worest delineations of the system are shown in Table 1 and Table 2, respectively, where Green annotations represent groundtruth polygons, and Red annotations are predicted ones.

Table 1: The best-6 Delineations

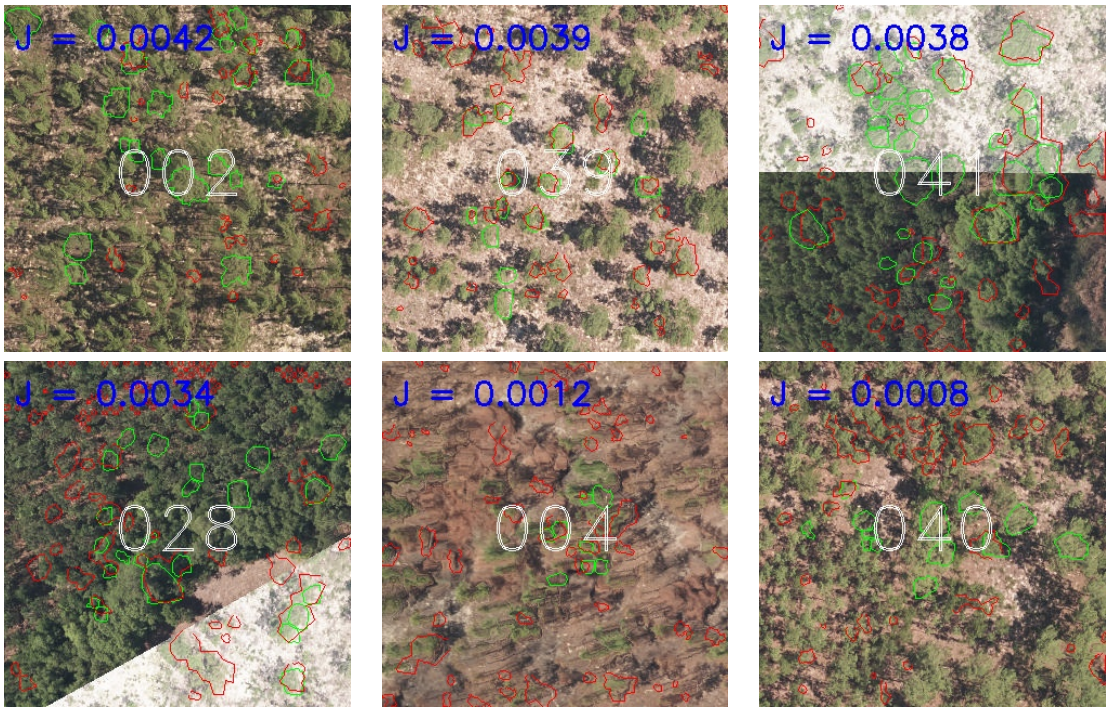
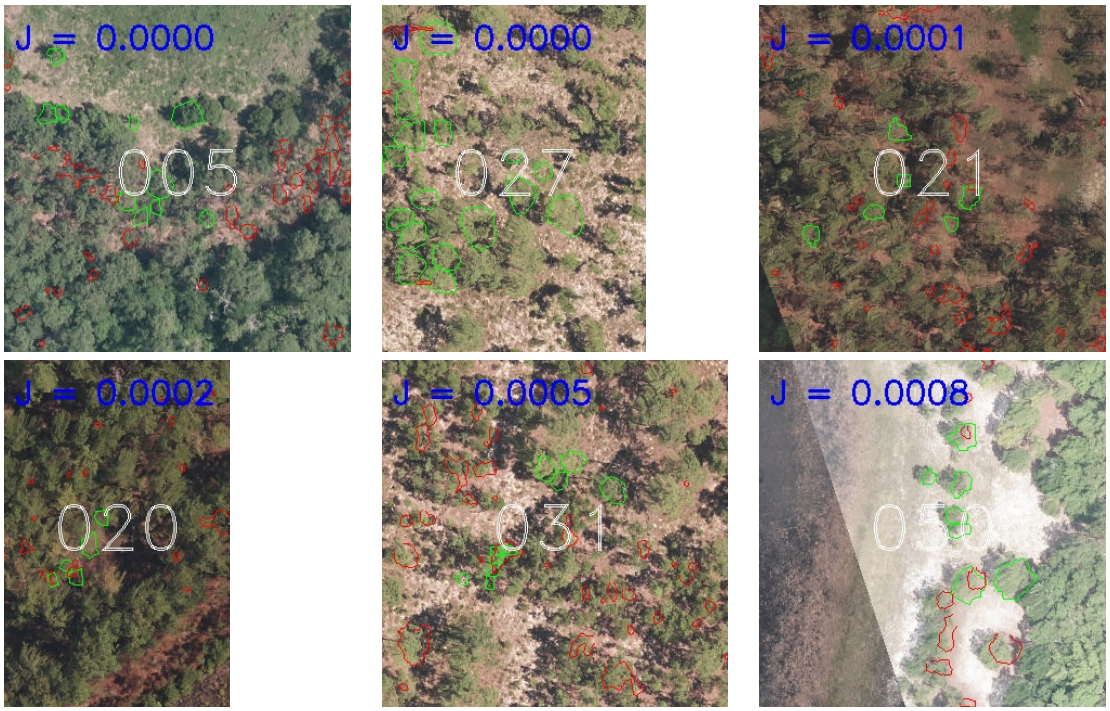


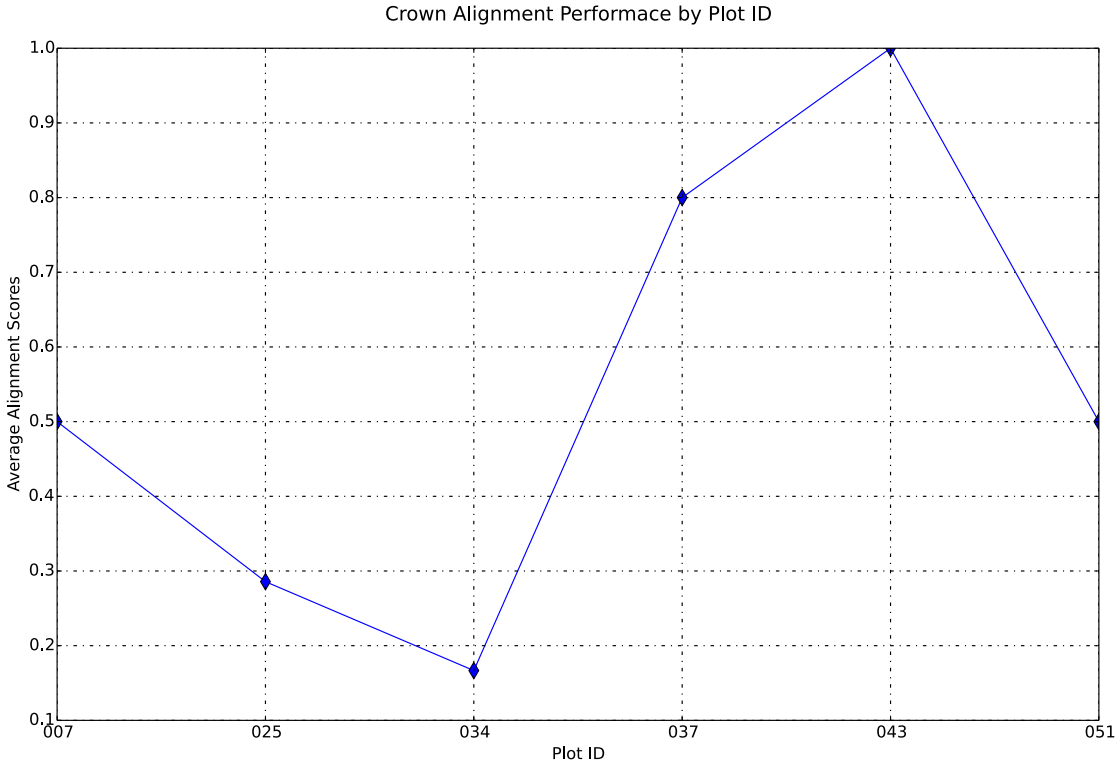
Table 2: The worst-6 Delineations



Task 2 - Crown Alignment

Plot-Level Crown Alignment Performance

While the overall alignment performance gives a good measure for effectiveness of a system, it is also useful to analyze how a system perform w.r.t. each testing plot. Being able to locate those badly performed plots help with finding the source of potential shortcomings, which can be used to further improve the system. The plot-level crown alignment performance is shown in the following figure.



Task 3 - Species Classification

Overview of Additional Metrics

In addition to the cross entropy cost and rank-1 accuracy, we also present the following evaluation metrics for further analysis:

Classification Accuracy: 0.8817

Average F1 score: 0.1108

Average Specificity: 0.5445

Classification Accuracy

The accuracy of a classifier is defined as the number of true predictions made by the classifier divided by the total number of predictions. In our evaluation, accuracy is computed per class and averaged across all classes to give an average accuracy score. For class k, its accuracy is defined as

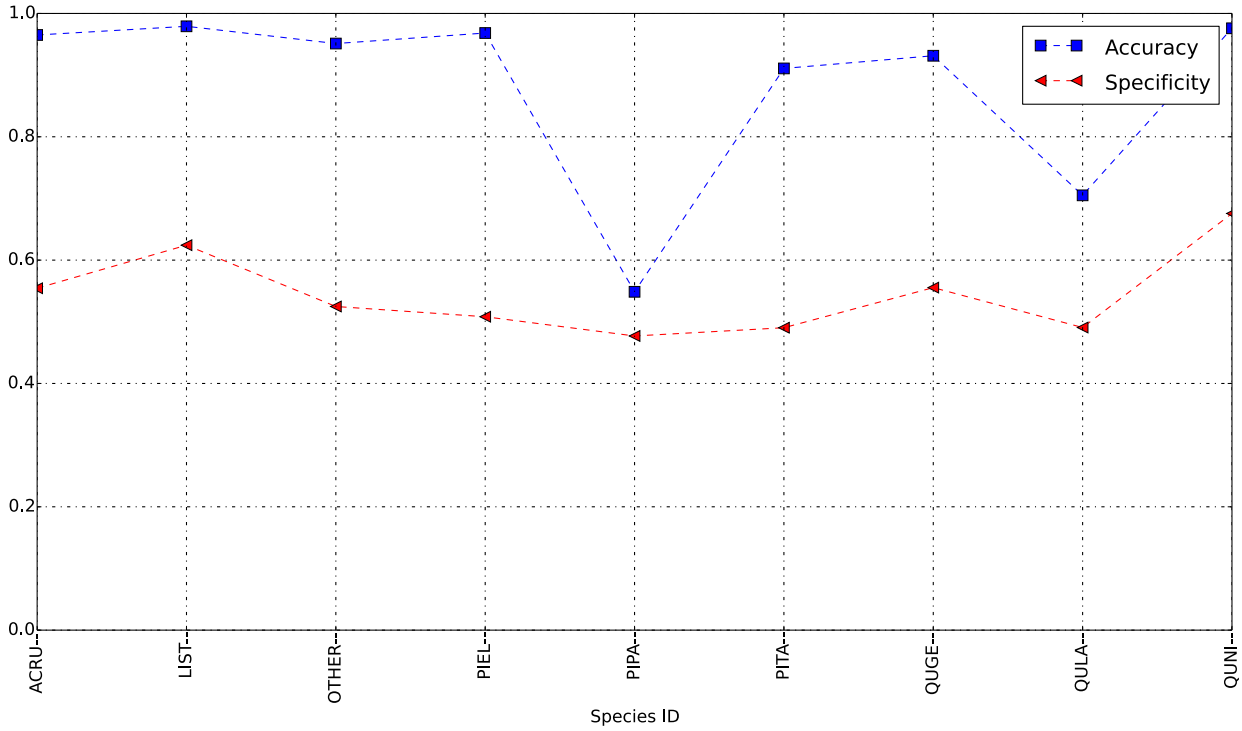
$$\text{Accuracy}_k = \frac{\text{TP}_k + \text{TN}_k}{\text{TP}_k + \text{TN}_k + \text{FP}_k + \text{FN}_k} \tag{1}$$

where:

- TP_k is the **True Positive** for class k, i.e. number of samples of class k that are classified as class k.
- TN_k is the **True Negative**, i.e. number of samples not belonging to and not classified as class k.
- FP_k is the **False Positive**, i.e. number of samples not belonging to class k but classified as class k.
- FN_k is the **False Negative**, i.e. number of samples of class k but classified as not being class k.

The per-class classification accuracy is shown in Figure 1.

Figure 1: Accuracy and Specificity Scores (Per-Class).



Specificity

Specificity for a class k refers to the number of samples correctly rejected for a class k in proprtion to the total number of non-class k samples in the data set. This measurement is used to evaluate the efficiency of classifier in ruling out samples as not belonging to a certain class. A higher specificity means that the classifier can better reject data points not belonging to a class of species. The average specificity score is calculated as arithmetic mean of specificity scores for all classes. The per-class specificity scores are illustrated in Figure 1.

Precision, Recall, and F1

Precision is defined as number of true positive samples divided by the total number of samples predicted as positive by the classifier. For class k, it is

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{FP}_k + \text{TP}_k}. \tag{2}$$

Similarly, recall is the ratio of true positive samples that are correctly identified by the classifier. Mathematically, it is defined as

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{FN}_k + \text{TP}_k}. \tag{3}$$

The per-class F1 score is thus twice of the harmonic mean of precision and recall, and average F1 score is calculated as arithmetic mean of F1 scores for all classes. The per-class precision, recall and f1 are illustrated in Figure 2.

Figure 2: F1 Score

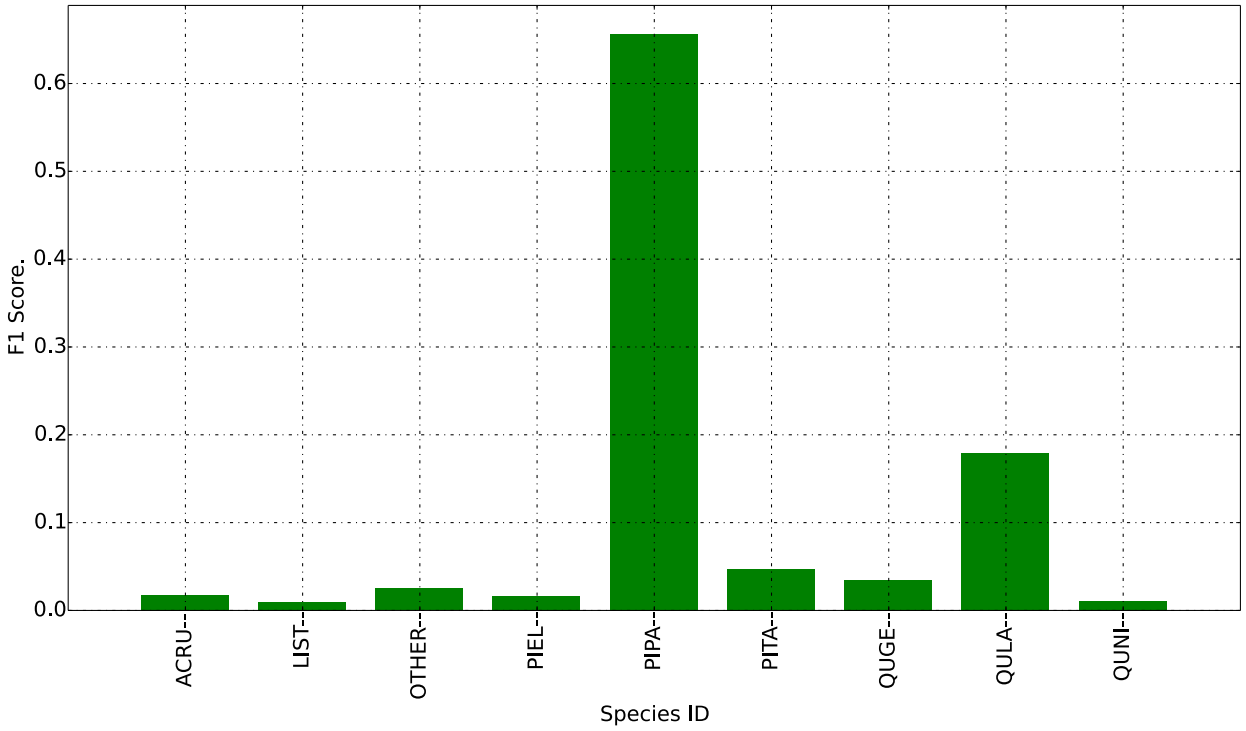


Figure 3: Precision

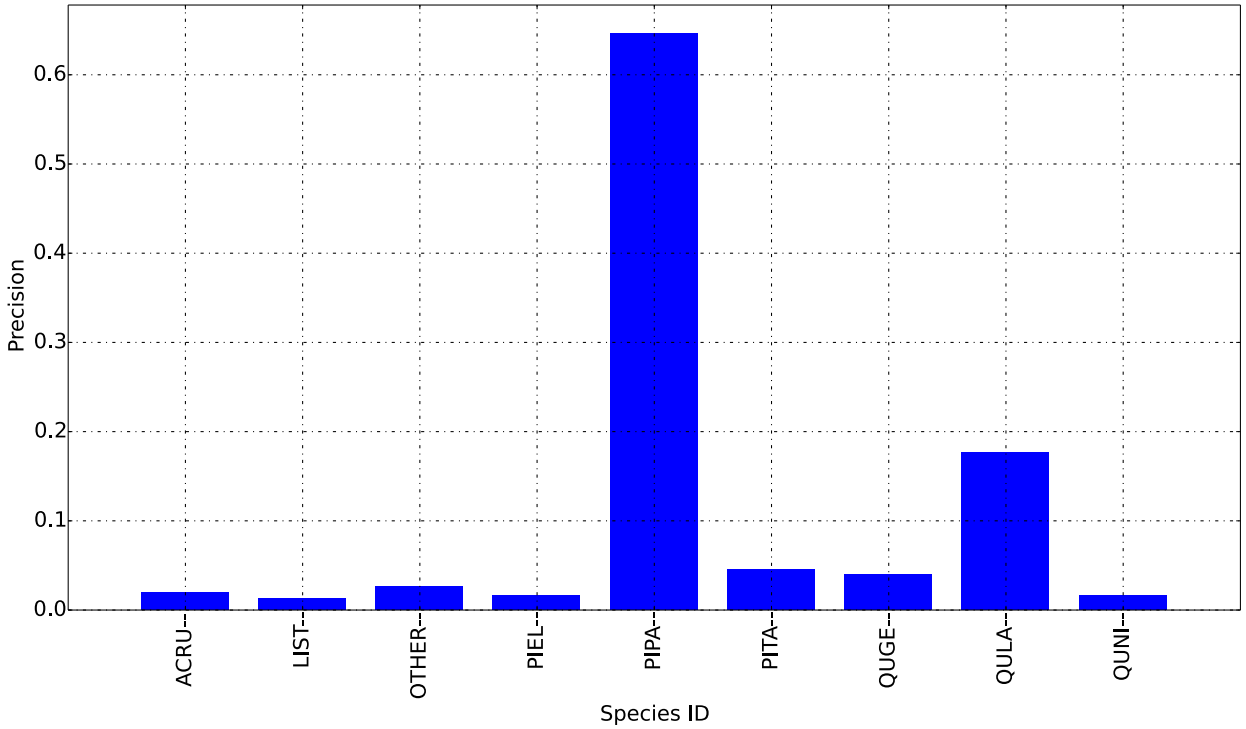
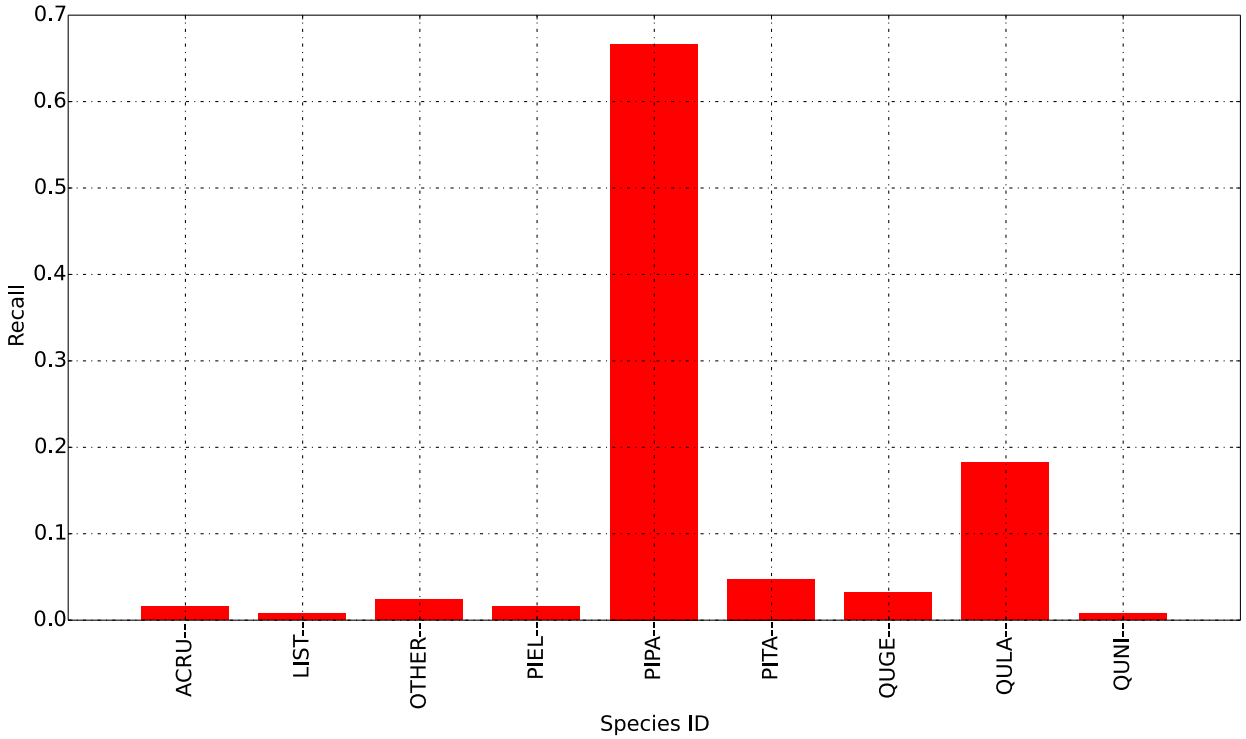


Figure 4: Recall



Confusion Matrix

Finally, we attach the classification confusion matrix where element (i, j) is the sum of probabilities for testing samples from class i predicted as class j.

Species ID	ACRU	LIST	OTHER	PIEL	PIPA	PITA	QUGE	QULA	QUNI
ACRU	**0.04**	0.02	0.06	0.04	1.65	0.12	0.08	0.45	0.02
LIST	0.03	**0.01**	0.04	0.03	1.10	0.08	0.05	0.30	0.01
OTHER	0.05	0.03	**0.08**	0.05	2.20	0.16	0.10	0.60	0.03
PIEL	0.03	0.02	0.05	**0.03**	1.38	0.10	0.07	0.38	0.02
PIPA	1.29	0.65	1.94	1.29	**54.26**	3.88	2.58	14.86	0.65
PITA	0.09	0.05	0.14	0.09	3.86	**0.28**	0.18	1.06	0.05
QUGE	0.08	0.04	0.12	0.08	3.30	0.24	**0.16**	0.90	0.04
QULA	0.35	0.18	0.53	0.35	14.87	1.06	0.71	**4.07**	0.18
QUNI	0.03	0.02	0.05	0.03	1.38	0.10	0.07	0.38	**0.02**