IDTReeS Competition

Frequently Asked Questions

This is a live document and will be updated frequently. Responses to new important and frequently asked questions will be sent in an email to those who have registered for the competition. If you have a question, please visit this page: https://idtrees.org/questions/

General questions

I registered, but didn't receive confirmation. How do I know if I am registered for the competition?

If you've signed up for the competition using the form on our website, you're all set! There is no other formal registration process. We add your email to a listserv (idtrees@lists.wisc.edu) so you can receive periodic updates during the competition. We don't have a rigid registration structure, and people are free to participate as they like;

- If you registered, you aren't committed to submitting. Think of registering as letting us know you are interested, and opting in to getting email updates.
- If you registered, but no longer want to participate, no problem! You can either stay on our list so you keep getting updates, or email us to ask to be removed from the list.
- If you haven't registered, you are free to download the data and work with it however you choose. If you end up with something great, and want to formally submit your results, you can register at any time before the final submission deadline.

What is the current timeline?

In recognition that for many people, schedules, duties, and priorities have changed over the past few months, we have modified the timeline of this competition. The list below has our current dates for data release and deadlines. Additional updates will be announced on our email list. On June 26 we announced a 1-month extension of the timeline.

- **June 1**: test data released and submission process open for live evaluations
- **August 1**: deadline for participants to submit data for evaluation
- **August 8**: announcement of results/winner and organization of manuscript collection
- **August 15**: deadline for participants to sign up for manuscript submission

Can other sources of data be used in our methods?

The objectives of this competition are 1) to find the best possible method for tree crown delineation and species classification, and 2) to be able to compare methods in a standardized way. The use of external data (meaning data that was not provided in the training data set by the competition organizers) may allow the first objective to be achieved, but hinder the second.

To support the first objective of finding the best possible methods, we allow participants to use external data if it meets a few criteria. First, the data must be freely and publicly available. Second, the data should be available nationally or globally and part of ongoing data collection efforts. For example, the use of Landsat satellite data is acceptable because it is freely and publically available for download (mostly commonly from the USGS), and the data isn't limited to specific regions. The use of spectral reflectance data your research team collected is not

acceptable because it is not freely and publicly available, and it is not data available for other parts of the US NEON study region.

Remote sensing or field data from other NEON sites is allowed, except for the Talladega National Forest (TALL) site. We do not allow the use of any TALL data because this site is a model for how methods will apply when there is no data available, such as high-resolution remote sensing data or field data.

To support the second objective to ensure a standardized comparison across teams, we ask that, if possible, participants provide 2 submissions: one where the external data was used; one where only the provided data was used. Any external data used must be mentioned in the submission description and provided in the submission. Having two submissions will allow us to make comparisons equally across teams and also see the potential of incorporating other data. A discussion of these aspects of the competition will be included in the final publication of this competition.

Will there be a publication from this competition?

Each submitting team will be invited to submit a paper as part of a collection of papers. The collection of papers will be a coordinated submission to a journal with one overview/comparison paper of all participating team submissions, and multiple sub-papers provided by each participating team. Our team will write the comparison paper with a summary of results. The authors of that paper are just the members of the organizing team.

All other participating teams will write separate papers that allows the team to go into more details of the methods and results. Teams can choose to contribute to the paper collection by writing their own papers. In other words, you can participate in the competition, and choose to submit a paper, or not. The overview/comparison paper will cite other papers in the issue.

This structure is similar to what was done in the previous round of the competition. The collection of papers from that competition is published in PeerJ: https://peerj.com/collections/56-remotesensingcomp/

Working with the data

The hyperspectral data has 369 bands, but in the competition instruction, it says "426 bands". What's the range of electromagnetic spectrum of the HSI in the training set?

From the 426 original bands included in the NEON data we removed 27 bands that correspond to the water absorption bands. We removed 27 bands in the range of (1.3445 to 1.429 nm) and

(1.8004 to 1.955 nm). Please email Sarah if you want a list of the NEON bands and wavelengths.

The water absorption bands are regions of the electromagnetic spectrum where energy is absorbed by water molecules in the atmosphere. Because these regions do not measure reflected energy of the surface, they are typically removed in hyperspectral analysis of vegetation.

More information from NEON about hyperspectral data: "*The NEON imaging spectrometer collects data within the 380 nm to 2510 nm portions of the electromagnetic spectrum within bands that are approximately 5 nm in width*"

Why are the images black?

The raster images may appear black when opening a geoTIFF using a normal image viewer. The files can be properly opened and viewed in common raster image formats in R and python. Depending on your default image viewer, they may be blank as a preview.

Here is a simple script to open geoTIFF files in R, with an example of an RGB plot:

```
library(raster)
r <- stack("MLBS_12.tif")
plotRGB(r)
```

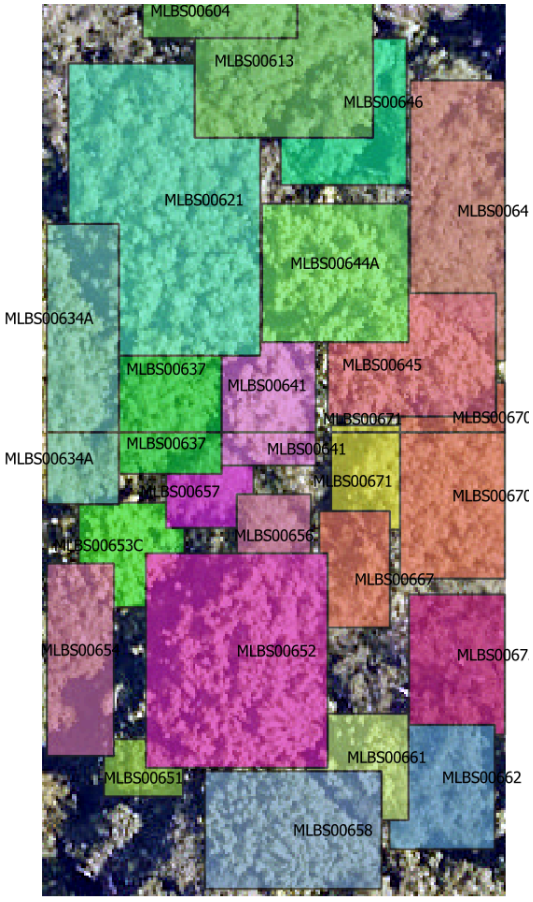
For opening hyperspectral data in python, you can use a script found here.

How do you match the ITC bounding boxes to each image?

The ITCs and the remote sensing data images are matched to each other spatially. That means that the ITCs fall within the geographic extents of the remote sensing data. In a GIS program, opening the ITC vector data and the remote sensing raster data will show that these data geographically overlap. **In version 1 of the training data, the only way to match each ITC to the remote sensing file is by using the geographic attributes of the data.**

In version 2 of the training data, an additional file has been included that specifies which remote sensing data is associated with each ITC. Below is an example of the new table included in the training data. The “id” and “indvdID” columns are unique to each ITC. The “rsFile” specifies which remote sensing file is associated with each ITC. The file extension .tif is recorded, but the same site and number applies to the .las files as well.
In some cases a single tree crown exists in multiple remote sensing files. In these cases, the ITC for the single crown is split where two separate polygons exist for a single “indvdID”. The image below has multiple crowns where this occurs (see “MLBS00637: as an example). In these cases, the “indvdID” will be listed multiple times with different “rsFile” corresponding to each separate part of the ITC polygon.

id	indvdID	rsFile
516	MLBS00621	MLBS_40.tif
519	MLBS00634A	MLBS_39.tif
519	MLBS00634A	MLBS_40.tif
517	MLBS00637	MLBS_39.tif
517	MLBS00637	MLBS_40.tif
518	MLBS00641	MLBS_39.tif
518	MLBS00641	MLBS_40.tif
531	MLBS00644A	MLBS_40.tif
533	MLBS00645	MLBS_40.tif



Left: Example of the table provided to match the ITCs to each remote sensing data file. Right: Example of an area where ITCs are split between two remote sensing files. ITC “MLBS00637” is an example where the ITC spans two remote sensing files and is listed twice in the table.

Do all ITCs represent full tree crowns?

The training data includes all ITCs that exist in the canopy. That is, the crowns of all trees that appear in the remote sensing data are associated with an ITC bounding box. For trees that fall in the middle of the plot, the ITC bounding boxes encompass the full spatial extent of the tree

crown. For trees that fall on the edges of the plot, the ITC bounding boxes only encompass a portion of the tree crown since some of the tree exists outside of the spatial extent of the box.

The training data provided is accurate and represents a real issue in generating ITCs from remote sensing data. The testing data will not have any partial ITCs. In other words, you will only be evaluated on the delineation of complete tree crowns, or ITCs that exist fully within the extent of the plot.

What is the spatial resolution of the RGB data?

The spatial resolution (or pixel size) of the RGB data is 0.1 meters. Each 20 x 20 meter RGB .tif file is 200 by 200 pixels. The original instructions document incorrectly said the spatial resolution was 25 cm.

There is no overlap of species in the OSBS and MLBS sites. Is there a way that we can self-evaluate our ability to do cross-site predictions on the training data?

The test data contains information for a third site. No training data was provided for this site, so evaluation on this site is to assess the ability of methods to apply to novel sites. This third site contains some species that exist in the MLBS and OSBS sites, and some species that do not exist in either. To evaluate the ability of your methods to transfer to untrained sites, we recommend a few approaches:

1. Split the data within a site. The training data is geographically spread across an individual NEON site. You could split the site in a way to have some overlap and some differences in the species between the subsites. The limitation of this approach is that the remote sensing data is similar for the full site, so it would not show how your methods work with remote sensing data collected on a different day under different environmental conditions.
2. Use different years of remote sensing data. This approach would allow you to see how your methods trained remote sending data collected from one year apply to remote sensing data collected for the same area but a different time. All ITC data was created with reference to the remote sensing data provided. For MLBS, the data were provided for 2018, and for OSBS the data were provided for 2019. Data from previous years is available to download from the NEON data portal (https://data.neonscience.org/data-products/explore), and also be accessed with the API (https://data.neonscience.org/data-api) and in R (https://www.neonscience.org/neonDataStackR).

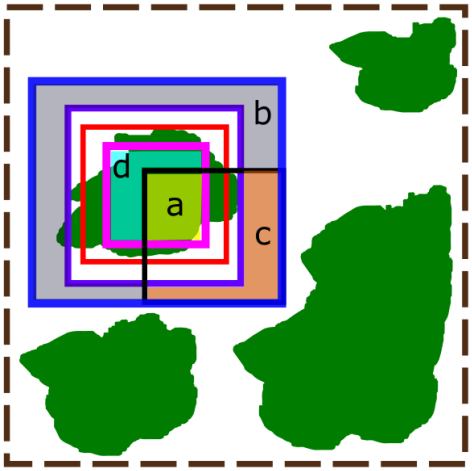
Evaluation

Can we get the evaluation code?

The evaluation code can be found on Zenodo, and the most updated version on GitHub.

What are the inner and outer halos?

To account for small uncertainties and inaccuracies in the size of the ITC delineations, when we compute the Rand index, we will use a set of “halos” around each truth location during scoring. There is an inner halo, outer halo, and edge halo that are used in the Rand Index. The inner and outer halos are set to be 10 cm within (1 pixel in the RGB imagery data) and outside of the ground truth, and the edge halo is computed so that the area between the edge and outer halo is the same size as that of the inner halo. Predicted bounding boxes that have no overlap with a ground truth bounding box will be given a score of 0. The figure below provides more details about the halos, the ground truth ITC, and the submitted ITC.



An illustration of how crown delinations are scored. The **green** polygons represent individual tree crowns, the **red box** is the ground truth delineation, and the **brown dashed box** is the full 20 x 20 meter plot. Within the ground truth delineation is a **pink box** and this is the inner halo. Outside of the ground truth delineation is a **purple box** and this is the outer halo. The most outer **blue box** is the edge halo. The **black box** is a sample delineation. The true positives are shown in the yellow highlighted region (marked with an “a”). This is the intersection of the detection and the inner halo. The true negatives are the blue highlighted region (marked with a “b”). This is the difference between the edge halo and the outer halo and the detection. The false positives are in the orange highlighted region (marked with a “c”). This region shows the portion of the detection that is encompassed in the negative region between the edge and outer halo. The false negatives are in the teal highlighted region (marked with a “d”). This region shows the difference between the inner halo and what was detected in the outer halo.

How has the evaluation code been updated?

Task 1 evaluation. There were two updates to the evaluation code for task 1 in early July that affected the output scores. The first corrected an issue with the plotting. The output images of the ground truth and delineated crown are now correct. The second corrected the order arguments were called in one of the functions. This correction results in slight differences to the IoU and Rand Index scores.

Task 2 classification. No changes were made to the classification scoring. However, other changes have been made to the evaluation code to account for differences in submissions across teams and the differences in taxonID lists in the test and evaluation data.

How will task 2 (classification) be evaluated?

The evaluation code for tasks can be found on a GitHub repository. For task 2, we use the logistic loss or cross-entropy loss function from sklearn. *We recognize that the number of samples in each class is not equal. This is a real and important issue in ecological classification tasks, and one reason why we are so motivated by this competition!* Because we know that this imbalance impacts evaluation scores, we report both the macro and micro F1 metrics. The average macro F1 weights each class evenly, while the average micro F1 is sensitive to the different number of samples in each class. The teams with the top scores for both metrics will be released.

What is the proper submission format for each task?

Please follow the format in the sample data in the evaluation code (see the eval/submission folder).

For task 1 (delineation), the preferred submission format is a .csv file with the geometry included in the WKT format. Please have separate files for each site (TALL, MLBS, OSBS). We can also accept the Esri “shapefile” format and will convert to WKT format using the R “sf” package.

For task 2 (classification), please submit a .csv file with the attributes of “indvdID”, “taxonID”, and “probability”. It is best to include a prediction for all taxonIDs in the training data, even if the probability is zero.

Known issues with Version 1 of the training data

Version 2 of the training data has been added to the Zenodo repository. The issues below have been corrected.

Why is there a shift in the bounding boxes and the images? Is there a way to correct that? Will the test data have the same issue?

Below is an example of the ITC bounding boxes overlaid on the RGB remote sensing data.The bounding boxes are in the correct geographic location relative to the remote sensing data. However, it is common that on 1 or 2 edges of the image, the bounding boxes extend outside the **geographic extent** of the image data. This is a difference in the geographic extent of the bounding boxes and the image data, and not an error due to a geographic shift.

While not an error in the data, we recognize that the difference in extent may make working with the data challenging. In Version 2 of the training data the ITC bounding boxes are clipped to fall

within the extent of the remote sensing data. This will make the apparent size of the ITCs smaller and will not change the geographic location of the ITCs.

MLBS_7 has 1 bounding box - is that correct?

Yes, that is correct. Version 1 of the training data included a MLBS_7 plot. There was only 1 ITC that was included for that plot. Version 2 of the data has that single ITC removed.

There are ITCs in the training data without attribute information.

There are some ITCs that do not appear in the field data. These ITCs do not have corresponding field data and can be used for training and evaluating delineation methods.