Part 2 Questions & Answers Sessions

Please type your questions in the Question Box. We will try our best to answer all your questions. If we don't, feel free to email Justin Fain (justin.j.fain@nasa.gov). This document will be shared to the training webpage within one week.

Question 1: I tried downloading some HSL data yesterday. I noted there is HSLL30 and HSLS30. Which one do you choose and how do you apply a data quality mask?
Answer 1: I'm using the HLSS data. Specifically, I am using the data described here. On the lefthand side of the Earthdata Search tool, there is a data quality mask filter.

Question 2: I'm not able to download the .TIF files correctly when I clone the repo. I get this error: Error downloading object: data/rast/HLS_2017_multiband.tif (e52a8ae): Smudge error: Error downloading data/rast/HLS_2017_multiband.tif: batch response: This repository exceeded its LFS budget. The account responsible for the budget should increase it to restore access. Is this a problem on my end or your end?
Answer 2: Looks like that's a problem on our end with so many people trying to access the data simultaneously. I put the data up on Google Drive as well.

Question 3: Is there any code example to do supervised Random Forest classification without downloading the large raster images – like CDSE Statistical API?
Answer 3: If you want to process the data without downloading you will be restricted to a cloud provider like Google Earth Engine.

Question 4: Can the Randomized Forest Model generalize well outside the training data distribution (e.g., model trained in Region A and applied to Region B)?
Answer 4: So long as the spectral characteristics for the classes are similar. If you were to move the domain to a different area, performance may degrade but only testing and validation can tell you to what extent. I have also given more detail in response to other similar questions as it pertains to applications of a trained model outside of the original spatial and temporal domain.

Question 5: How many reference data points are required to train and validate the Random Forest classifier, and what parameters are considered when determining the number of reference points?
Answer 5: There is no exact rule for deciding how many training points you need. Typically, the number of training points is on the order of thousands, but keep in mind that each of the trees in the Random Forest ensemble takes a random sample of the training data and input features, so having too little training data will limit the number of unique permutations of the decision trees. This can lead to overfitting issues if the number of trees is large relative to the number of training data samples. Ideally, you would also spread these training data points across both the “before” and “after” imagery, but that is often impractical due to the constraints of field data collection.

Question 6: How do you create the training sample?
Answer 6: I used the 2017 data as a basemap and created sample points with labels for each of the classes I wanted to represent in the Random Forest model. Keep in mind that the classification demonstrated was intentionally simplified. A true rigorous implementation would have had thousands of training data points spread out over both the 2017 and 2024 images.

Question 7: How can data from different satellite sensors, such as Landsat and Sentinel-2, be harmonized for long-term change detection analysis?
Answer 7: This is an inquiry for the science teams that work directly with this data. You might also be interested in the answer to the question regarding domain adaptation.

Question 8: Can you explain the number of trees in the Random Forest module, and how it affects the result of the classification?
Answer 8: Adding more trees in the n_tree option of the Random Forest setup makes the model more robust to noise, but adding more trees doesn’t always improve the accuracy of a model. In this case, where I didn’t create many training points, there wasn’t any significant difference between 100 and 500 trees because there weren’t enough training points to construct more unique tree permutations in the bagging process.

Question 9: Can we have the Github repo for Random Forest implementation?
Answer 9: https://github.com/NASAARSET/LCLUC_2026. This is the same repo as the previous session.

Question 10: When applying Random Forest models for land cover change detection across multiple years, how can we ensure model stability and avoid bias when spectral conditions differ between dates?
Answer 10: Use bias corrected data and ensure you are using the same sensor for both the past and present imagery. Harmonized products like this HLS data are great for this since they provide a longer period of record. The decision to use an ensemble method like Random Forest also helps to reduce bias through the bagging (bootstrapping) process. Further image correction techniques like normalization can improve classification accuracy, as well as taking training samples from each of the dates for which you have imagery.

Question 11: Beyond generating a change matrix, what approaches would you suggest translating land cover change outputs into quantitative environmental risk indicators?
Answer 11: This is going to depend entirely on what changes you are interested in quantifying. You might imagine scenarios where the loss of forest area, changes in coastal vegetation, or signs of aridification could all be considered indicators of unwanted changes in their particular contexts, but there is not one general strategy for translating detected LCLUC into risk indicators. This is why the map of LC change is such an important tool as it allows us to not only observe the magnitude of change, but also to put that change into the correct landscape context.

Question 12: Where do you recommend sourcing high-resolution imagery for creating training data?
Answer 12: Sentinel is relatively high resolution and commercial providers such as Planet also provide imagery as well with a cost. Some governments and other entities have bought extremely high-resolution imagery from these commercial providers and made it available online for free, though these tend to be limited to specific times and regions rather than ongoing global products.

Question 13: Do you use the same training data for 2017 and 2024, or do you collect new training data for subsequent years?
Answer 13: The training data is only related to the 2017 imagery. The Random Forest model is trained on that data and then applied to both years. Keep in mind that the classification demonstrated was intentionally simplified to fit the intermediate level of the training. A true rigorous implementation would have had thousands of training data points spread out over both the 2017 and 2024 images. The process of extracting the

reflectance values, creating a classifier, and applying the same model to both years would still be exactly as demonstrated.

Question 14: Is it possible to get access to a kind of "explainability" of the decision rules (thresholds related to a band...)?
Answer 14: The package {randomForestExplainer} does a good job of giving you information about the model including the importance of variables. There are other methods for creating a representative tree from the randomForest model object, but those features haven't yet been merged with any CRAN library as far as I am aware.

Question 15: Has ARSET done training on this same LCLU using Python before?
Answer 15: There are previous LCLU trainings that ARSET has conducted in the past using other software which may be of interest to you. This particular training has not been conducted in Python, however.

Question 16: Are there seasonal constraints when selecting satellite imagery for land cover comparison? Should images be acquired during the same time of year to minimize phenological and atmospheric variability?
Answer 16: Yes, ideally you should get imagery for both dates that are in the same part of the seasonal/phenological cycle to reduce the effects of seasonality on vegetation, ephemeral wetlands, crop cycles, etc. Your choice of model is also important, but ensemble models like Random Forest are generally tolerant to small variations in spectral characteristics. If atmospheric or sensor effects are degrading the model classification accuracy you can use a method like histogram normalization to better align the images.

Question 17: In Google Earth Engine (GEE), when training a Random Forest model, we usually input the imagery and training samples directly without explicitly extracting reflectance values. Does GEE automatically use the pixel reflectance values from the image, or do we need to extract them first (e.g., using. addBands())?
Answer 17: I am not familiar enough with GEE to give an answer here. I imagine there are other resources for doing Random Forest in GEE.

Question 18: In R, we normally extract reflectance values before training. Could you please clarify how this works in GEE?
Answer 18: See above.

Question 19: How can we determine the optimal number of trees (ntree) for our Random Forest model? Does it depend on the number of training samples, or are there other factors we should consider?
Answer 19: There are papers on this topic that attempt to answer the question in the general case. Usually, we expect that the error rate (out-of-bag estimate of error) rapidly decreases as the number of trees increases and then levels out. Increasing the number of trees beyond the point where the error rate has stabilized will not provide meaningful improvements in the model but will appreciably increase the time it takes to train the model. The optimal number of trees in any case will be dependent on your training data, target location, choice of sensor, etc.

Question 20: I may have misunderstood something, but your binary mask looks like a grey-scaled imagery, why?
Answer 20: The default background for a plot in R is white, so I opted to render the binary mask as two different shades of grey so that the edges were more clearly visible.

Question 21: What is the advantage of doing change detection in R versus in GEE?
Answer 21: It is mostly down to personal preference; R was used for this particular training series but there is no inherent disadvantage or advantage. GEE has the advantage of running on Google Cloud servers, but you might choose R if you would prefer to retain control of your own code and data.

Question 22: How do you decide which classification method to use?
Answer 22: In the general case, the classification method you chose to use is dependent on your goals. There is a chart in the presentation which lists some examples of supervised and unsupervised classification models, but there are many more to choose from, each with their own advantages and disadvantages.

Question 23: Is it scientifically valid to use training samples collected for one year (e.g., 2016) to classify imagery from another year (e.g., 2025)?
If we only have samples from 2025, how can we use those same sample points to classify imagery from 2016? Is there any proper model or scientific method to handle this type of multi-temporal classification?
Answer 23: Ideally, you would have training data from all of the dates to train your model but given the nature of field campaigns it is often the case that you only have data from one time period. Histogram normalization of the images can improve

classification accuracy by removing small variations due to changes in atmospheric conditions. We try to avoid introducing bias by selecting images taken at the same time of year and assuming that the spectral characteristics of the target classes do not change significantly over time. However, atmospheric effects and sensor differences can violate this assumption. This tends to be less of a concern when the two images are close to one another in time, and a greater concern as the time between the image collections increases. Evaluation and validation of the model results is the best way to determine if your 2025 data can be applied to 2016 (also see the other answers about randomForestExplainer and train/test splits on the training dataset).

Question 24: How do we sign up for ARSET updates for trainings?
Answer 24: Sign up for the ARSET Listserv on the training webpage! To get updates on our latest trainings and receive our quarterly newsletter, please send an email with no subject line to arset-join@lists.nasa.gov and follow the instructions sent in response.

Question 25: Is field-collected training data required for supervised classification?
Answer 25: Field-collected training data is not necessary. In cases where the target site is difficult to access, covers a large area, or other limitations would make field collection infeasible it is more practical to create training data from high-resolution remote sensing data collected at (or nearly at) the same time as the multispectral imagery you plan to use in the model. Keep in mind that a Random Forest model typically requires thousands of points, and collecting a large volume of data in the field might be time or cost prohibitive.

Question 26: What approaches can be used to collect and prepare training data to improve the classification accuracy of visually similar land cover types like palm oil plantations and forests?
Answer 26: You are going to want to collect in-situ data if possible. Hyperspectral imagery in this particular application may be more beneficial since the spectral profiles of oil palms, and the background forests are likely to be very similar. Multispectral imagery may not have sufficient spectral resolution to detect the small differences between the two classes. A segmentation-based approach might be an even better choice since they can detect structural differences in the ways palm oil plantations are organized relative to natural forests (i.e. oil palms are laid out in rows, unlike natural forests).

Question 27: Are there any books you guys recommend going deeper in these subjects?

Answer 27: Here are a few recommendations to get you started.
Land Cover Classification of Remotely Sensed Images: A Textural Approach
Data Classification: Algorithms and Applications
Statistical Rules of Thumb (I keep this one next to my desk at all times)
Random Forests with R

Question 28: If I don't have any training in situ or field data, can I still do LULC? If so, how do I validate the same? I want to track the LULC changes for the past 30 years.
Answer 28: Yes. You can still use contemporary high-resolution remote sensing imagery to aid in the creation of training data in the absence of any field collected data. Finding sufficient data over 30 years may be the bigger challenge, and you will certainly want to consider normalizing the imagery since there is likely to be a high degree of variability across that time period which could degrade the accuracy of your classification.

Question 29: What are the "hyperparameters" of a Random Forest tree, and which are the most sensitive?
Answer 29: The hyperparameters of a model are all of the variables we can change to alter the behavior of the model. In the case of Random Forest, it is the number of features at each split (the mtry argument in the R example) that has the greatest impact on the model, especially in the case where you have many weakly correlated predictors. Tuning mtry adjusts the diversity of the trees in the forest by changing the probability of a feature being included or excluded. I came across this very thorough writeup on the subject which also includes information about variable selection by importance, which is briefly discussed in the “Additional information” section at the end of the Part 2 code document for this training.

Question 30: Can the same training dataset be used to classify land cover in different years? Does the training process teach the model spectral characteristics of classes, or does it learn location-specific information?
Answer 30: Once the model is trained, it can be applied to a number of years, but the error rate will be higher than if the training data had been drawn from multiple years initially. This effect tends to be greater when the period of time between the images is longer. You can further improve the stability of the model across years by applying image processing techniques to normalize the images. Depending on the area that the model was trained in, interoperability may not be feasible. Tobler’s first law of geography states, “Everything is related to everything else, but near things are more

related than distant things." Consequently, a model trained in one area is likely to retain a decent performance when applied to nearby areas, but unlikely to perform well when applied to distant areas. Testing and validation of the model is the best way to assess the performance of a model when presented with data outside of the training set.

Question 31: Would it be a best practice to apply a filter (3x3 majority) to miss classified pixels? Or would it be best to say any change over 1 Ha is actual change on the ground and not mixed pixels?
Answer 31: This is going to depend on how well matched the spatial resolution of the imagery is to the size of the changes you expect to see. Since we are typically discussing LCLUC on a landscape scale we are expecting to see changes that are large enough to be accurately captured by a single image pixel. In cases where the expected changes are small relative to the resolution of the imagery or the land cover types have similar spectral profiles it might be necessary to explore advanced signal processing techniques like spectral unmixing and/or establish a lower bound to the area of change you will consider to be actual change.

Question 32: If I want to perform a LULC classification on two dates, 2015 and 2025, and I only have field data for 2025, what should I do regarding 2015? For example, I want to perform a supervised classification using Random Forest.
Answer 32: See the answers to other similar questions for a more detailed discussion. In short, while this is not the ideal situation it is a common problem. There are normalization techniques that can help and you might try to create training data using high resolution imagery collected in 2015. Only testing and validation can give you a true assessment of how well the model performs on data outside of the training set.

Question 33: What methods can be applied to improve classification accuracy between spectrally similar land-cover classes, such as bare land and built-up/settlement areas?
Answer 33: Moving to hyperspectral data, adding more narrow bands may allow you to pick up on differences. Note that when your model includes a very large number of dimensions it is generally a good idea to perform some level of parameter tuning and/or feature selection to improve performance.

Question 34: Isn't ten training points per class a very low number for training data?
Answer 34: Yes. Though I only created ~10 points for each class for this demonstration, you will see that papers with the Random Forest model often have

upwards of 1000 points. The important takeaway is that the training data should accurately capture the full variability of spectral profiles that belong to a particular class.

Question 35: Will Random Forest provide a classified image with higher accuracy compared to other classification algorithms like Maximum Likelihood Classification, Minimum Distance, while we are using the same training set?
Answer 35: Random Forest belongs to the ensemble classification methods group, meaning that it synthesizes the results of hundreds of individually simple decision trees. This gives it some characteristics which are well suited to land cover classification such as the speed of classification on large images and its robustness against overfitting. Adaboost and XGboost are two other popular ensemble classification methods that have some additional features. These methods are usually preferred in land cover classification over methods like MLC, SVM, etc.

Question 36: Which methods can be used to obtain training samples for classifying past satellite imagery when contemporary ground truth data are unavailable?
Answer 36: Using high resolution imagery should be sufficient to build training data for Random Forest. See the answers to other similar questions for a more detailed discussion of the options and limitations when training data is not available.

Question 37: Does increasing the geographic extent of a study area (e.g., 10 counties) necessitate a proportional increase in training data to ensure accurate land cover classification?
Answer 37: One main concern when building training data, it needs to represent the full variance for that class. Since there is variation in classes, the whole range needs to be represented. For Random Forest, you are likely going to need to collect thousands of training data points, so it might make sense to do a stratified sample to ensure that the variability of each class across the entire geographic region is captured.

Question 38: Which satellite dataset is best for LULC assessment from 1980–2015? What do you recommend and why?
Answer 38: Ideally, you would need to use the same sensor across the time scale, but I can’t provide any definitive recommendations for a particular place. Perhaps it would be acceptable to train a series of models for each sensor, though only testing and validation would be able to confirm if that approach works for your location and use case.

Question 39: I'm working on mapping agricultural land use in Benin. We're using QFIELD on tablets to collect field data. Do you have any other methodologies to suggest?
Answer 39: QFIELD is perfectly capable of doing the field collection. Photos with associated GPS metadata are also useful for this application.

Question 40: Will single pixels of training data provide a better classification result compared to polygons of pixels of training data?
Answer 40: If it is possible to draw a polygon such that it only contains pixels of one class it is always going to be more time efficient to draw polygons rather than creating and labeling thousands of individual point samples. However, this approach is often not possible in highly homogeneous landscapes. You aren't limited to only one vector type when creating your training data and can use a combination of points, lines, and polygons as needed to fit your landscape.

Question 41: To what extent can domain adaptation methods mitigate spectral and distributional shifts when transferring LCLUC classification models across distinct ecological regions or multi-sensor datasets? Any stable method(s)?
Answer 41: There are a number of possible techniques, but since your question specifically targets stability I think that you would be interested in correlation alignment (CORAL), Maximum Mean Discrepancy, Self-Training, and this recent paper. You might also consider exploring AugMix as a potential solution. However, domain generalization is probably a better fit for operational applications since we'd prefer to have a model that performs well against unknown future shifts rather than a reactive adaptation strategy. Unfortunately, deeper discussion is outside the realm of something we would be able to cover in an intermediate level training. If there is sufficient interest, we might conduct an advanced level training on these topics in the future.

Question 42: I have a difficult time understanding how to best collect ground truth data to classify a Sentinel image given the 10m spatial resolution. Does the training data have to be uniform 10m areas? Potential classes in my scene would be water, bare soil, mangrove vegetation, and salt marsh vegetation. The latter three are likely to exist within any given 10m area. I appreciate any advice you may have!
Answer 42: In relation to mixed pixel effects, it can be an issue and using a different model may help for your use case. You may have to explore some advanced

approaches like spectral unmixing or look into the possibility of using higher spatial resolution data if available.

Question 43: To what extent can domain adaptation methods mitigate spectral and distributional shifts when transferring LCLUC classification models across distinct ecological regions or multi-sensor datasets? Any Stable method(s)?
Answer 43: Refer to Q41.

Question 44: For satellite-based LULC classification, which supervised model (Random Forest, XGBoost, SVM, or KNN) generally provides the highest classification accuracy, and why?
Answer 44: XGBoost tends to be highly performing in classification tasks and is also computationally efficient. XGBoost is similar to Random Forest, but with more complexity on the backend. The particular boosting method, regularization, and depth-first search make it an exceptionally performant classification model. Since it is also relatively fast and computationally cheap, XGBoost is very easy to iteratively test and tune which helps to achieve maximum accuracy. Here is a great article with more details.

Question 45: Could you briefly reflect on using Random Forest for classifying LiDAR data, if you have experience on that (e.g. for urban areas or identifying changes in forest biomass, success, pitfalls…)?
Answer 45: Refer to our previous LiDAR training conducted in November of 2025.

Question 46: I need a quality image for my study area for the year 2010. Where can I download it?
Answer 46: Depending on your area, Earthdata Search should be sufficient. Sentinel and HLS can provide high quality data.

Question 47: What is the best vector type for training data (polygon or point)?
Answer 47: You can be precise with point data, but polygon data can be a more efficient way of collecting training data.

Question 48: What do I do if a particular class has several variable spectral signatures? Will it provide better results if we classify each spectral subclass and later merge them into a single class?
Answer 48: Yes, classifying each subclass and combining them later is likely to give better results. Random Forest learns an envelope which describes the spectral

characteristics of a class. If your two subclasses have substantially different spectral signatures, the envelope will be larger than necessary and prone to overprediction.

Question 49: Is every training point that you use only associated with a latitude/longitude coordinate or are you selecting 2-D polygons?
Answer 49: In this case, points were used (lat/long), though you can create training data with any combination of points, lines, and polygons.

Question 50: How might you compare the results from supervised vs unsupervised methods? Is this something worthwhile or typically looked at?
Answer 50: Early on in a study, the smallest and least complicated model is preferred. From there you can move up in complexity if your study warrants it. I am not able to think of a useful scientific conclusion you could draw from comparing the results of a supervised and unsupervised classification beyond assessing which model provides the best classification accuracy.

Question 51: Can Synthetic Aperture Radar (SAR) data be integrated with optical imagery to improve classification accuracy?
Answer 51: You can add any information associated with a raster. The connection between SAR and optical data is somewhat complicated so check existing literature on the topic before proceeding. Also keep in mind that SAR does also tend to have a lower spatial resolution which may be a source of additional complications.

Question 52: Can you recommend a super resolution model for enhancing the resolution of the images?
Answer 52: Unfortunately, not. The issue with super resolution models in this case is that they attempt to create data where no data originally existed and this additional “ghost” data is correlated with the base data.

Question 53: What proportion of the training data points do you reserve as testing points?
Answer 53: 70/30 can be a good reference for a sufficient split. With Random Forest, the out-of-bag estimate of error serves as a kind of cross validation, so to some degree the very structure of the Random Forest algorithm performs a train/test split with each iteration.

Part 2 Questions & Answers Session B

Please type your questions in the Question Box. We will try our best to answer all your questions. If we don't, feel free to email Justin Fain (justin.j.fain@nasa.gov). This document will be shared to the training webpage within one week.

Question 1: I'm not able to download the .TIF files correctly when I clone the repo. I get this error: Error downloading object: data/rast/HLS_2017_multiband.tif (e52a8ae): Smudge error: Error downloading data/rast/HLS_2017_multiband.tif: batch response: This repository exceeded its LFS budget. The account responsible for the budget should increase it to restore access. Is this a problem on my end or your end?
Answer 1: Looks like that's a problem on our end with so many people trying to access the data simultaneously. I put the data up on Google Drive as well.

Question 2: What is the difference between this method and Google Earth Engine (GEE)? Also, is Land Use/Land Cover (LULC) data for my [country/region] available only for the year 2017?
Answer 2: This method should be equivalent to the GEE implementation. There may be LCLUC data for your study area already, or you can make your own landcover map from multispectral or hyperspectral data using the process we showed today.

Question 3: With regard to forests, what classification method should I use, in addition to reflectance values, if I want to add elevation, aspect, slope, or other values such as tree type (e.g., evergreen/deciduous)? Would there be too many variables to consider?

Answer 3: That doesn't seem like an unreasonable number of variables if those have impact on the land cover type in your study area. You simply need to stack the extra data with your multispectral bands using the rast and sprc functions from {terra} before you extract that information at each training point. Do note that when working with a very high number of additional variables it might improve the Random Forest model performance to perform some degree of dimensionality reduction based on the variable importance. This is discussed in more detail in my answers to other similar questions.

Question 4: Can we do random classification in Google Earth Engine?
Answer 4: Yes. Here is an example of a GEE implementation of Random Forest for mangrove mapping.

Question 5: How were the locations on the training dataset generated? Was it done in points on QGIS?
Answer 5: Yes. I used the 2017 imagery as a base map and created the points using QGIS. That made it easy to add labels for each of the classes and visualize my training points to ensure I was capturing the variability in each land cover class, such as sampling both the clear and turbid water in the image.

Question 6: What is the most efficient way to obtain a training data set for a Random Forest model?
Answer 6: If you can't get data through field work on the dates for which you have RS imagery, you can create your training data using the imagery itself. Polygons are efficient for homogenous areas since they allow you to get samples for many points (pixels in the image) at once.

Question 7: For the label class data that you used in the training, which year was it taken? Did it below to the condition of y17 or y24?
Answer 7: The training data was created using the 2017 imagery, but the 2017 and 2024 imagery were both taken in roughly the same season, so we are operating under the assumption that the spectral profiles for each land cover class are similar between the years. For optimal model performance we would have preferred to have thousands of training data samples distributed across both years, but that wasn't necessary to demonstrate the process of building and applying a classification model.

Question 8: What about metrics? How can we evaluate the classification performance?

Answer 8: See the additional information at the end of the Part 2 document where I cover some of these topics using the {randomForestExplainer} package. The variable importance and out-of-bag estimate of error are great places to begin when evaluating a Random Forest model.

Question 9: When using the R Random Forest tool, does R randomly create the training and validation data from the entire dataset? Or is this a manual process?
Answer 9: There are evaluation metrics that are inherent to the way that Random Forest builds its model, such as the out-of-bag estimate of error. The bootstrapping process takes a randomized subset of the training data and features to create each tree in the ensemble, which is essentially equivalent to cross-validation. A 70/30 split is a good baseline for splitting the training data for use in a more thorough validation process.

Question 10: What would be the best way to do a multi-year analysis of land cover (same area in several different times months/years)?
Answer 10: As mentioned previously, direct comparison has to match the spectral signatures within the image. You can train the model to match seasonality in the phenological cycle but know that a model trained in one season is not likely to give good results when applied to another season. For example, a model trained during a warm and dry summer will have no training data available for snow and ice and so fail in the winter. There are also more advanced techniques for domain generalization that can assist in expanding the range of contexts in which a model is applicable, but those are beyond the scope of this intermediate training.

Question 11: I am not sure if this is within the scope of this webinar, but can you talk about the relationship between study scale and resolution of images? For example, if one wants to examine the changes in vegetation in the home range of an animal species that lives in a forest in the DRC in the last 25 years compared with changes in the vegetation in the African continent during the same period require different resolutions.
Answer 11: As you move to lower spatial resolution imagery, such as a continental scale, you are going to get more mixed pixel effects, or pixels that have multiple land cover types. Using a commonly aligned grid or fixed scale imagery will solve this problem but could cause issues when it comes to the storage and processing of large datasets. For Random Forest in particular, you will likely want to have labeled training data samples on the order of thousands of points, so the number of pixels in the image should be at least an order of magnitude greater. As the number of pixels in the image

approaches the number of samples necessary you get closer to the case where you're hand-labeling each pixel, at which point the model is not saving you any time or effort.

Question 12: Is it possible to have the R script used for the demo?
Answer 12: Github: https://github.com/NASAARSET/LCLUC_2026

Question 13: Given that working with spatial data in R can be computationally intensive, what would you recommend for this type of analysis? Specifically, what spatial resolution and approximate raster size (e.g., number of pixels) would be reasonable to handle on a standard personal computer?
Answer 13: The {terra} package has the capability of doing on-disk processing which significantly reduces the memory demand when processing large images at the cost of processing speed. Other configurable GDAL options can help to reduce the size and computational overhead. I have a standard consumer-grade laptop both for work and for my personal use and both are capable of processing rasters exceeding 10GB without causing any significant freezing/crashing issues, though the newer model will finish the processing in a much shorter time than my older personal laptop.

Question 14: My supervisor suggested me to use Landsat data and Digital Elevation Model (DEM). Is it possible to layout them together for LULC analysis?
Answer 14: Stack your raster before doing your extraction so you have all of the bands and the DEM value. The documentation for the rast function from the {terra} package will give you examples of how to read and concatenate raster data. You can always run ?rast in R to open the help documentation for the rast function (the same shortcut works with any other function as well).

Question 15: How does spatial and temporal resolution of NASA satellite imagery affect the accuracy of land use change detection?
Answer 15: The spatial resolution for legacy products run into the issue of mixed pixel effects, so it is important to understand the scale of your imagery and study area. Temporal resolution and constraints such as cloud cover can also inhibit the ability to have accurate detections.