Visualizing Land Cover and Land Use Change with NASA Satellite Imagery
February 24 & 26, 2026

Part 1 Question & Answers Session A

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: Which RStudio are you using?
Answer 1: R version 4.5.0 (2025-04-11), I have the colors set to "Solarized Dark" which is why it looks different from yours.

Question 2: What is a spatraster?
Answer 2: Spatial Raster. That is the raster format that the {terra} package uses to represent raster data with geographical coordinates.

Question 3: I can't find the GitHub page, or any other resources. Can I have the links for all necessary documents and resources?
Answer 3: ARSET github: https://github.com/NASAARSET/LCLUC_2026

Question 4: Do you have this script for python? Is there a reason why we are doing classification in R vs using Python?
Answer 4: We do not currently have this in Python. You can do this in Python or any other programming language that has packages/modules for clustering. Scikit-learn is one popular Python package for this sort of task. Google Earth Engine implementation is possible too.

We used R due to personal preference and the availability of packages for the efficient handling of clustering and spatial data.

Question 5: How many clusters are required for 5 land classes?
Answer 5: Five cluster centers to correspond to your 5 classes; but remember that when you are clustering, it is the distance between clusters. If you have 5 known land classes, you pick a center point. We encourage you to experiment with different k values.

Question 6: You mentioned that for K-means classification, it's good to just try a few different values for K. For clustering, you can use approaches like average silhouette width to determine an optimized cluster count. Is there a similar approach for determining a good K here?
Answer 6: You can optimize for K, but some methods may give you conflicting results so it is generally advisable to experiment with different values for K. See this resource for more information.

Question 7: Is there an objective way to figure out when K-mean option is better?
Answer 7: Please see #6.

Question 8: How do you manage a large number of bands if the study area is very large?
Answer 8: If this is a question about the data being large in memory, the {terra} package we are using is capable of doing on-disk processing which takes a bit longer but doesn't need to hold the entire image in memory to do the processing. K-means is computationally cheap, so you're unlikely to run into issues where the compute time required becomes unreasonable. Otherwise cloud computing is an option (such as with hyperspectral data for large areas).

Question 9: What are the advantages of unsupervised and supervised classification in a professional workflow?
Answer 9: Unsupervised methods are computationally cheap, but the clusters are based on spectral profile similarity so they can be inexact and sensitive to changes in the scene. Supervised methods lets you set your classes ahead of time and give you an expanded toolkit for accuracy assessment. If you have something in particular you want to target such as loss of mangroves, it would be easy to run K-means clustering with two centers for very many dates. If you were interested in something more specific like the change in species of tree in a particular patch of forest, you would likely want to choose a supervised method wherein you provide the model with information about what differences (spectral differences between the species) you are interested in detecting.

Question 10: How can I know each class after classification accurately?
Answer 10: You can't, assuming you mean K-mean clusterings. Because the clustering relies solely on the grouping of points in feature space, the clusters aren't inherently

tied to any particular class. It requires you to go back to your original imagery to see what the model picked up on. As we increased K, it picked up on differences which we confirmed with the original imagery. At some point, increasing the value for K will still create more clusters, but those clusters won't be meaningful.

Question 11: What would you recommend as a solution to the heterogeneity likely to be experienced through the application of K-means other than increasing K values?
Answer 11: We recommend supervised as we will cover this in Part 2. K-means is sensitive to heterogeneity since it is always splitting the clusters based on which pixels are most similar to one another. This means it is making the "best" splits possible, but it does not guarantee that these clusters will be meaningful.

Question 12: In the processed imagery, I observed some striping effects. I would like to understand how these image strips and the applied stretching during preprocessing influence the classification results?
Answer 12: Striping and stretching do not affect the result of the clustering. HSL imagery comes in with the digital values as small. If you display them without a stretch it will be shown as very dark. The striping here is primarily caused by the differences in gain between sensors, but the effects are largely exaggerated by the stretch I've applied and are actually much smaller than the variability within an LC class.

Question 13: Does the choice of platform or image processing software impact the classification outcomes?
Answer 13: Yes, in a nuanced way. We set a random seed at the beginning so that the random processes for choosing cluster centers are repeatable. Cluster centers have a bit of randomness to it. As long as you set your seed correctly, your results should match mine, but there are other peculiarities between software, operating system, etc., which could potentially impact which cluster centers are chosen.

Question 14: Could we have code for Python?
Answer 14: We will see if we can produce this in Python for you, but it is not currently in any language other than R.

Question 15: Is there a difference between Ground truth data and training point?

Answer 15: Those terms are used interchangeably. In situ data such as what you would collect with a GPS while doing a field campaign is useful for assessing the accuracy of the model results, or enables you to use supervised classification using that data to train the model.

Question 16: Is there another method for unsupervised classification rather than K-means? If there is, how can we decide what we should use?
Answer 16: Yes, there are a few other unsupervised methods. DBSCAN and ISODATA are common alternatives. As with all things, you should choose the method which best fits your use case, which may be different for different projects.

Question 17: Do you evaluate with other software or plugin like MOLUSE (QGIS)?
Answer 17: There are ways to do this classification with a GUI interface. QGIS is a good option, but we didn't explore any of the plugins for this particular training.

Question 18: What methods of accuracy assessment are available for unsupervised classification?
Answer 18: K means, which we covered today in Part 1, relies on the spectral similarity among classes. Without ground truth points, we can't get a direct metric for accuracy assessment. However, the strength of the clustering can be thought of as the average distance from the points within a cluster to the cluster center versus the distance of the most extreme points in a cluster to the other clusters. This would be a metric of how tightly the clusters are grouped and how distinct they are from each other. If your clustering is good (good centers) it should be tightly grouped around those centers and with no points lying close to the decision boundary between clusters. In Part 2, we will introduce training data and talk a bit about accuracy assessment.

Question 19: During the R session, when you visualized the RGB bands, there was obvious banding on the image. Can you suggest ways of cleaning the data?
Answer 19: See the answer to Question 12. The banding/striping of the imagery is greatly exaggerated by the stretch I applied to increase the contrast and doesn't pose a problem for the clustering or the supervised methods we see in Part 2.

Question 20: Is there any advantages of using R and R studio over python and GEE for LULC? Can I use Google Earth Engine instead of R or RStudio?

Answer 20: Yes, there are no particular advantages of using R other than this was my personal preference. GEE gives you the advantage of using Google servers for data processing, but there may be a fee associated with using cloud resources.

Question 21: What techniques can we use to reduce server load without degrading LCLUC accuracy?
Answer 21: You have a few approaches available to you and some might be more or less appropriate depending on your compute environment, data, and other constraints. Dimensionality reduction techniques can limit the number of bands you need to store and process without significantly degrading the classification accuracy. Some models are more computationally efficient than others, with the caveat that there is not any single model which is the best for all situations. You can explore parallelization and multiprocessing or look into cloud-optimized file formats, or many other similar techniques to reduce server load. The best solution is likely to be a combination of different approaches which fit your particular situation.

Question 22: I loved that you used R in this tutorial! Have you ever tried that in new positron version of R?
Answer 22: Not yet, though it looks promising.

Question 23: What is the address to the training page?
Answer 23: Visualizing Land Cover and Land Use Change with NASA Satellite Imagery | NASA Earthdata

Question 24: Is it possible to run R in Vs studio code rather than R studio? I feel comfortable with the Vs studio code IDE.
Answer 24: Yes! I used the standard Rstudio IDE, but anything that can run R is fine to use. VS Code, VSCodium, Sublime Text, Jupyter, etc. are all great options if you are more comfortable working in that environment.

Question 25: For climate mitigation and carbon or nutrient credit projects, small misclassification between vegetation, wetland, and barren land can significantly impact credit quantification. What best practices should practitioners follow to ensure classification outputs are defensible for policy, carbon accounting, and regulatory reporting?

Answer 25: If a small misclassification will have a large effect, we suggest using supervised so you have more opportunities to go back and check your work. You start with a metric that you can assess the accuracy of and quantify the impacts of misclassification.

Question 26: Is there any problem in solving the homework in python?
Answer 26: Homework will be posted on Thursday on the training webpage. I don't believe the HW has questions that will need you to run code, but the R code and HLS data are provided for you to test if you're interested.

Question 27: Are there some location-specific code in the R script provided via your GitHub? Or would it be possible to test the workflow with the same imagery and bands, but from a different location?
Answer 27: In Part 1 (unsupervised), you can apply this same method to any time or place for which you have imagery. In Part 2, I have created training data specific to this imagery. You'd need to create your own training dataset for your area of interest to use the supervised methods.

Question 28: Is K-mean clustering heavily affected by cloud coverage?
Answer 28: K-means will detect the bright reflective clouds and sort cloudy pixels into their own category. This can be good if you want to use K-means for cloud detection, but might pose a problem if you aren't interested in the clouds.

Question 29: If we use too many clusters (Ks) could it become too messy for the code to work properly?
Answer 29: Yes, you will hit a point of diminishing returns. It is best to use as few clusters that you can get away with to make sure your spectral clustering means something. Also, see the answer to Question 6 for some extra information about choosing a value for K.

Question 30: Is the data set you used available or do we pick our own on Earthdata?
Answer 30: Yes, you can find the data for this training (both parts) on the ARSET GitHub. However, you could download similar data for your area of interest and apply the same k-means clustering procedure. Since K-means doesn't need any training data we don't have to worry about making training/ground truth points.

Question 32: You have selected the random seed value at the very start. Does the selection of it affect anything with image processing? In classification, there is K=5. What if we set it to 2 or 8 values? How will the K mean cluster work?
Answer 32: Your code will select the cluster centers at random. In the code we explored what would result of you change the # of K. You can increase or decrease the value of K but you may not get meaningful results if you pick too many.

Technically, K-means will find better centers by trying to maximize the variance between classes and minimizing the sum of the square of the distances between each point in a cluster and the cluster center. The first random set of points chosen isn't always the best set of centers. I cut that particular detail from the presentation for the sake of simplicity.

Question 33: Is the K-means algorithm sensitive to the initialization of the first centers?
Answer 33: Yes, that is why we set the random seed at the beginning. K nearest neighbor uses optimization techniques. Part 2 will cover this a bit.

Question 34: I have two classifications 2 years apart and the classification shows change in wetlands (from one type to another) but that wouldn't be the case in two years. How can you better understand real change? Should you look at classification probability? Will this type of question be covered in the next session
Answer 34: If the wetlands are shifting from one type to another, they are likely to still be more similar to each other than they are to the background, so K-means might not catch the subtle changes in spectral profile. Use a supervised classification instead where your targets will be the 2 different types of wetlands. That way the model can be trained to focus on detecting the changes in wetland type and deemphasize the other changes in the scene which you'd rather ignore.

Question 35: Would you please let me know how did you identify the right RGB band ids
Answer 35: Searching the internet for information about the data and reading the user guide(s) are generally good ideas for any sort of data you're not familiar with. When you work with these datasets frequently you begin to remember details like the ordering of band numbers. In the case of HLS in particular, the RGB bands are 4, 3, and 2 (rather

than 3, 2, and 1 as you might expect) because band 1 is a narrow blue wavelength band for detecting coastal aerosols.

Question 36: Which distance is appropriate and effective (Euclidean or Manhattan) in K means clustering?
Answer 36: K-means uses Euclidian distance. I would guess that there is already a paper which explores the implications of manhattan distance in K-means clustering, but I am not familiar with one at the moment.

Question 37: When you use K-means classification twice in a row for the same number of K set, is it always going to return the same 'clusters' for the same image?
Answer 37: Yes, which is why we set the random seed at the beginning. This provides the random number generator with a common starting point so we can have repeatable clustering outcomes.

Question 38: What is the true value of classification if I should go back to the original image to figure out the classes?
Answer 38: Going back to the original image is far faster than labeling each pixel by hand. You can use the original image to cross-reference with the clustering outputs to determine which differences the K-means clustering algorithm used to make the splits. Remember also that K-means is using all of the available bands, so it might even pick up on subtle differences that are only detectable in wavelengths we can't see or correlations across more bands than we could visualize simultaneously.

Question 39: How do you interpret the output of a K-means clustering algorithm in the context of land cover classification, and what techniques can be used to validate the clustering results with ground truth data?
Answer 39: Unsupervised classification (K-means clustering) does not use training data which is an advantage and disadvantage. In Part 2, we go over a bit of the information about how we can assess supervised models, because the training data gives us a "ground truth" to compare our classification to. You could compare the results of K-means clustering to known labeled training data where your accuracy metric would be the strength of correlation between the label and the predicted cluster. For example, a good result would see all of the training points labeled "grass" sorted into the same cluster, none of the "grass" points appearing in other clusters, and no points of other

classes in the same cluster as the “grass” points. More complicated models allow you to tune the model parameters and assess the precision and recall in detail.

Question 41: How should we use K-means clustering for more than 20 crop types based on their NDVI?
Answer 41: I would not suggest this. Twenty crop types may be too much and the spectral signatures will be too similar to each other. NDVI is a simple band ratio. We suggest using the actual imagery to detect the subtle differences between crops types and a *supervised* model which allows you to take direct control of which classes (crop types in this case) the model uses.

Question 42: What would be the direct address to earthdata download/search data you showed us?
Answer 42: I can’t provide a direct link to my Earthdata project, but this is the link for the HLS data I’m using:
https://www.earthdata.nasa.gov/data/catalog/lpcloud-hlss30-2.0

Question 43: Is there a way to create a graph with the mean spectral profile for each center that was generated from the K-means classification?
Answer 43: Yes. The resulting object returned from the kmeans function can be combined with the original raster imagery to associate each pixel’s spectral profile with its predicted cluster. Then it is just a matter of taking the average reflectance in each band for each cluster. We also need to use the factor function with the levels argument to ensure that the bands graph in the same order as they appear in the HLS data, but we could just as easily sort the bands in order of wavelength.

Here is the code I came up with using our K-means run where K=2 as an example:

```
c(km2, hls[1]) %>%
  values() %>%
  as.data.frame() %>%
  pivot_longer(cols = -lyr1) %>%
  group_by(lyr1, name) %>%
  summarize(avg=mean(value)) %>%
  mutate(cluster=factor(lyr1),
         band=factor(name, levels=c("Coastal_Aerosol", "Blue", "Green", "Red",
                                    "Red_Edge1", "Red_Edge2", "Red_Edge3",
                              "NIR_Broad", "Water_Vapor", "Cirrus",
                              "SWIR1", "SWIR2", "NIR_Narrow"),
                 ordered=T)) %>%
  select(band, avg, cluster) %>%
  ggplot() +
  geom_line(mapping=aes(x=band, y=avg, color=cluster, group=cluster))
```

Question 44: When cloning the repo to local, the data/raster folder is empty (in my case, at least). Perhaps due to size of the rasters? Manually downloading the files work as an alternative though.
Answer 44: I had to upload them using the Git Large File Storage. That may cause the clone operation to fail on the larger data files. Download the raster data from the GitHub repository in your internet browser instead if you run into trouble.

Question 45: Should I use the provided git code in git's codespace or download and use in RStudio?
Answer 45: Running locally is recommended. GitHub charges some amount for running on their servers. Everything you see here is designed to run easily on the typical consumer-grade computer.

Question 46: Can you give a quick pipeline for the same work you have shown for K-means clustering based on NDVI?
Answer 46: See Q41. NDVI is a single value. It will only give you groups of similar NDVI values and may not give you meaningful information. We suggest using the original imagery.

Question 47: Wouldn't a high dimensional analysis, say if 3 or more bands are utilized for the K-means, contribute to much finer separation of classes? If so, shouldn't this be preferred over the first?
Answer 47: We are using all bands in the data. You may be asking about doing a PCA (principal components analysis) which is not something K-means does, but other methods do have dimensionality reduction. XGBoost, for example, does a sort of dimensionality reduction inherently. If there is audience interest I would be open to the idea of doing an advanced training to dive deeper into these sorts of statistical questions.

Question 48: Is there any benefit to using Band 8 of Landsat to sharpen images?
Answer 48: Possibly, but we are using all the HLS data bands.

Question 49: If you know what to expect in an area (water, crop, forest), can I pick 3 K-mean clusters?
Answer 49: Yes, that is the method. These all will be distinct LC types in the model. However, if there is anything else in the image, it will get classified into one of those three classes. Grass pastures, for example, would likely end up being combined with crop or forest since the k-means clustering would only “see” that those pixels have spectral characteristics similar to other forms of vegetation.

Question 50: Is it possible to run the K-mean clustering on a scene that has water masked already (to focus on classifying land only)?
Answer 50: Yes. If you have all the water taken out, you will not have any issues with K-means trying to classify water as a cluster. The only caveat is that you will need to make sure that the masked water pixels have their value set to N/A (as opposed to setting the reflectance values to 0) or K-means will still classify all of the masked pixels as a separate class.

Question 51: What if we don’t need high frequency imaginery, can we use this with only sentinel data (for better resolution purposes)?
Answer 51: Yes. Sentinel has great resolutions. The HLS data gives you the benefit of greater historical data to detect change since it includes data from Landsat which has been collecting imagery for much longer than Sentinel. The same logic applies to high resolution commercial data which might be cost-prohibitive to access, and airborne data which might be very high resolution but is only available for a few dates when the sensor was flown.

Question 52: K-means tends to group clusters with similar spectral characteristics but at times, some features have a similar response but they're actually different features on ground. How do we fix or prevent that?
Answer 52: Since K-means operates solely on the spectral similarity, you are limited to using imagery that actively captures the differences between the two similar classes. Supervised classification will let you set the classes ahead of time, which may give the model a better chance of catching subtle differences, such as distinguishing different types of crops or different types of trees.

Question 53: Is it possible to use the Deepness plugin in QGIS for LCLUC?
Answer 53: It appears that the Deepness plugin is capable of classification, but it uses a neural network method which isn't covered in this training.

Question 54: Do we need GPU to train the classification while replicating on the project?
Answer 54: No, all this was done using a consumer-grade laptop. If you are trying to do something larger-scale or using something like a neural network for classification you might run into compute limitations, but the models and data I am demonstrating here don't require a GPU.

Question 55: I would like to develop a study of years prior to 1980. Where can I download the images?
Answer 55: Landsat has data of a longer period of record; AVHRR as well. Earthdata provides a temporal filter which will help you find data. Prior to 1980, the data could be sparse. Also, keep in mind that comparison between different sensors is a complicated process. You will ideally have imagery from the same sensor for all of the dates you are interested in, but satellites don't remain operating forever.

Question 56: Is it possible to classify into vegetation, water, and mining areas using K-means?
Answer 56: Likely yes. Vegetation and water are spectrally distinct. While mining is a land use category, meaning we can't observe it directly, it is probably associated with a particular land cover type. Mining will not involve vegetation or water, so you should be able to identify bare ground (as a result of mining). Just know that there are reasons other than mining for bare ground to appear in an image, such as plowed but unplanted croplands.

Question 57: I am interested to know how we can do this on GEE. Can you share a quick idea of workflow?
Answer 57: A lot of this imagery is available in GEE's catalog. I am not up to date on my JS coding for GEE, but I found this resource which might be helpful.

Question 58: Do you apply dimensionality reduction like PCA before clustering multispectral satellite data?

Answer 58: Please see the answer to Question 47. In short, we aren't going to do any dimensionality reduction in these examples, but it might be a good idea in some scenarios.

Question 59: Would you recommend K-means over supervised classification methods for large-scale land cover monitoring? Why?
Answer 59: For large scale and relatively simple classification (things with distinct spectral profiles like water versus vegetation) K-means clustering is a great option because it is computationally inexpensive. The downside is that the clusters aren't correlated to any specific land cover type, so while the clustering is fast and easy, the results require some care and consideration to turn them into meaningful analysis.

Question 60: In high-resolution imagery, do spatial relationships between pixels get considered, or is clustering purely spectral?
Answer 60: Purely spectral. We project all the pixels into feature space which only represents the reflectance values in each band without retaining any of the spatial/geographic information. There are spatially-aware clustering methods but we did not cover that here. See my answer to Question 72 for some examples.

Question 61: Is there any prebuilt model to identify LCLUC from satellite imagery into different important classes like settlement, crop area, waterbody, kiln, etc.?
Answer 61: Yes. NASA and other orgs have produced these. There are landcover products that can be downloaded assuming it suits your applications. If you have specific classes in mind which aren't already mapped by the existing products you will likely have to make your own LC product. See the links in Questions 62 for some examples of existing products.

Question 62: Can we classify the land cover as vegetation, built up, and bare land?
Answer 62: This is a great candidate for supervised classification. Most LC products that have already been produced have these types. For CONUS, the NLCD is a great resource, but there are also global products as well as products specific to other countries or geographical regions. See this paper for an example.

Question 63: Does the code discussed here take care of any cloud cover influence that might affect the classification if I am using a Landsat image?

Answer 63: EarthData Search allows you to filter by cloud cover percentage. I specifically chose HLS imagery with low cloud cover, but there are other products which use averaging and smoothing methods to address cloud cover. Just note that when using those cloud-removal techniques you aren't looking at a single date, but a variable range of dates combined into one image. That's fine for most cases, but might not be appropriate for rapid changes on a short time scale.

Question 64: How do you mitigate the *curse* of dimensionality when clustering high-band multispectral or hyperspectral imagery?
Answer 64: I love this question because it really gets at the core of the question of how ML models store and represent data in higher dimensions. You can always look to apply techniques for dimensionality reduction like PCA, but in practice it's often even better to have someone on your team with domain expertise related to the thing you are interested in studying to help you target which parts of the spectrum (bands) are going to show the highest between-class distance and within-class similarity. In my experience, data with a lot of bands becomes a storage/compute issue well before the dimensionality causes a degradation in clustering performance.

Question 65: Do you use any cluster validity indices beyond silhouette score for geospatial datasets, considering spatial dependence violates independence assumptions?
Answer 65: There are spatial aware clustering methods. K-means is not one of those. See my answer to Question 72 for some links you might find helpful, but know that my answer isn't comprehensive.

Question 66: I had a problem with the GUI on my project for LCLUC. How can I build a GUI for LCLUC classification in QGIS in a correct way?
Answer 66: I am not familiar with building GUIs in QGIS. In R, you can create a minimal API pretty easily using the {plumber} package and build a more complicated UI with {shiny}. I tend to work on backend systems so I am sorry I can't provide more guidance on the frontend development work.

Question 67: When monitoring land use change over time, do you recluster each time step independently or use centroid initialization from previous years for temporal consistency?

Answer 67: Because K-means is always starting fresh and does not depend upon what has come before it is best applied to a single time step. In Part 2, we will address Land cover change using models we can train and apply repeatedly.

Question 68: Will inclusion of indices enhance classification accuracy?
Answer 68: Possibly, but unlikely. They tend to be simple band combinations. The variability should already be captured in the original imagery. If you have information in band A and information in band B, an index such as A+B isn't providing any more descriptive information than the individual bands. If you had other data, such as surface temperature, soil moisture, etc. then that might improve the classification since those values aren't already included in the multispectral imagery.

Question 69: K-means identifies each cluster differently each time you run it (I mean number for each cluster). How do you compare clusters in different years if they get different numbers? Just eyeing it?
Answer 69: Part 2 will look at the change. K-means is great for large areas and single years. I intentionally did *not* use K-means for comparison between years because the clusters on two images aren't directly comparable.

Question 70: In large-scale global land cover mapping, how do you ensure scalability and computational efficiency when clustering petabyte-scale satellite datasets?
(Follow-up to Q.70) Beyond language optimization, do you leverage distributed computing frameworks like Spark or cloud-native geospatial platforms to scale clustering?
Answer 70: Petabyte scale data is likely to come with petabyte scale compute infrastructure. The R terra package includes some options for parallel computation but at some point you might need to consider switching to a different coding language that gives you even more control over the parallelization. Minimal representation of the data is also very important, but may not be possible in cases where the differences between spectral profiles between classes is very small.
Follow up: Yes, I have worked with distributed computing frameworks in the past for doing work at large scale so it's absolutely possible. Using COG and ZARR formats can help with very large imagery data as well. I intended this training to teach the fundamentals of the *how* and *why* of LCLUC monitoring rather than provide scalable implementations of the classification models, so all of the data I use is intentionally

kept small. I would be super interested in developing a training at the advanced level to explore how you could scale this up, if there is an audience interested in that kind of deep dive!

Question 71: We are facing trouble classifying urban areas accurately using optical data. Are there any suggestions for how we could increase classification accuracy for urban areas?
Answer 71: Optical data (visible spectrum). Urban built up areas will show a higher reflectance in longer wavelengths, but if you are limited to the visible spectrum you might have some difficulty separating urban areas and bare soil, which is a common source of error in LC classification. There are global products that have been processed already, and most of them include urban areas as a LC class, so what you're looking for might already exist.

Question 72: Do you have some references for spatial based clustering?
Answer 72: There are a number of available methods which include the spatial arrangement of pixels in the clustering/classification. I am personally aware of work by Nowosad and Stepinski on the subject, but there are other approaches as well.

1. https://www.researchgate.net/publication/326138711_Spatial_association_between_regionalizations_using_the_information-theoretical_V-measure
2. https://www.researchgate.net/publication/349739046_Pattern-based_identification_and_mapping_of_landscape_types_using_multi-thematic_data

Part 1 Question & 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: Is it possible to use another programming language? Is it something tied to homework? Can we submit in another language?
Answer 1: You won't be asked to run any code in the homework. However, it is possible to do all of the things you see in this training in other programming languages. Python is a popular option, or Javascript in the case of Google Earth Engine (GEE) code.

Question 2: Where can I access the R codes for today's presentation?
Answer 2: Github https://github.com/NASAARSET/LCLUC_2026

Question 3: Is there any particular or special reason for using R as compared to say Python?
Answer 3: No, you can use Python for this same process. The scikit-learn module is a popular choice. Using R in this demonstration is just a matter of personal preference.

Question 4: From a practical perspective; what are the limitations of the K-means clustering method? What are alternative unsupervised classification methods?
Answer 4: Refer to the slides for a more comprehensive list of methods. DBSCAN and ISODATA are two alternative methods that come to mind, but there are many others. K-means is an unsupervised method which means you are dependent on the assumption that there are distinct spectral differences *between* classes and that the things that fall *within each class* are spectrally similar to one another.

Question 5: How sensitive is the satellite imagery and classification technology in 2026? Can we distinguish between perennial and annual crops or between two different types of crops using satellite imagery?
Answer 5: Yes, possibly. There is very high resolution data available from commercial vendors that would help in determining plant types, but the differences in their spectral

response curves would still be broadly similar so you'd also need high spectral resolution (hyperspectral) data to capture the small differences. From a computation perspective, there are significantly more complex models and frequent improvements in the efficiency of machine learning which make the classification technology landscape much more powerful in 2026 than it was even a few years ago. With the large-scale adoption of cloud computing and efficiency improvements on consumer hardware, these large and complex models are increasingly within reach of more scientists, researchers, and technicians. It's an exciting field which has seen and will likely continue to see innovations.

Question 6: Please explain what supervised and unsupervised machine learning is.
Answer 6: Supervised machine learning is using training data to show what you can expect within the scene while unsupervised machine learning relies only on the data from the scene. I hope that Part 2 will give you a better understanding of the differences between supervised and unsupervised classification. We only covered unsupervised classification in Part 1, so don't worry too much if the differences aren't yet clear.

Question 7: Will you send an email with the link for the assignment?
Answer 7: The homework, presentation slides and presentation recordings will be posted on the training website after Part 2: www.earthdata.nasa.gov/learn/trainings/visualizing-land-cover-land-use-change-nasa-satellite-imagery?utm_source=social&utm_medium=ext&utm_campaign=LCLUC2026.

Question 8: Can you show slowly how to download data from https://search.earthdata.nasa.gov/? Also, how do I interact with the file? Do I just load it up on RStudio?
Answer 8: Refer to the presentation to see how you can download data from Earthdata Search. The rast function in R from the package {terra} is the function for bringing imagery data into R in a format that we can interact with via our code. The "rast" name of the function is a shortening of the word "raster" which is another word for the pixel-based file format that is typical of imagery data. The resulting object created by using the rast function is called a SpatRaster which is a shortening of the phrase "spatial raster" which means that the image also has location data (coordinates).

Question 9: In the earthdata search tool, what dataset are you using for this training? Are there preferential datasets based on location, function, etc.?
Answer 9: The data used in the presentation was the Harmonized Landsat Sentinel (HLS) data. The HLS data is a good choice for LCLUC analysis because it has a long historical period of record, which allows us to look at long-term changes. Your choice of data will vary given the location, time, and application area.

Question 10: I have access to ArcGIS Pro. Can I accomplish the same K-means unsupervised classification in that tool?
Answer 10: It appears that ArcGIS Pro includes K-means clustering as part of the spatial analysis toolkit.
https://doc.arcgis.com/en/insights/latest/analyze/find-k-means-clusters.htm

Question 11: How does self-coding compare to the built-in tools for classification in programs such as QGIS or SNAP Toolbox?
Answer 11: Most of those tools use the GDAL library. Writing the code for ourselves gives us more control over the small details, which becomes more relevant with more complex models that have more parameters to optimize. Also, when you write code, there are fewer opportunities for the software to hide important parts of the process behind abstraction. This means that we can closely analyze the process and diagnose any problems that arise. Possibly of interest, in QGIS you can see the underlying GDAL/OGR function calls in the processing history or processing log menus. Every GUI software has to translate your clicks into code at some point, but some software is more transparent about that process than others.

Question 12: If your area of interest only has vegetation and no water or large bare soil patches, could K-means clustering be effective? Given the high level of spectral similarity overall, would the difference then be relatively enough to use K-means classification?
Answer 12: Yes. As K-means is not given any data prior to classification, it is sensitive to changes in the rest of the image. We will go into further detail during Part 2 of this training series.

Question 13: If supervised and unsupervised classification methods allow us to map past and present land cover changes, how might combining Random Forest models with harmonized Landsat Sentinel data enable us to anticipate future

ecological tipping points such as irreversible deforestation or wetland collapse and what responsibilities do scientists and policymakers bear in acting on such forecasts?
Answer 13: We will discuss Random Forest in Part 2 of this training series, but this exact question is beyond the scope of this training series. There is great work being done in this field.

Question 14: Based on Question 5 about the sensitivity of satellite imagery, are there any practices that can extend this sensitivity using in situ resources through algorithmic connections? As resolution of satellite imagery can be a major impediment to smaller scale features, it would be good to know if these resources can be extended to these scales in some way.
Answer 14: K-means is an unsupervised model and as such, does not have any background data prior to classification. There will always be a constraint in any classification based on a variety of factors. Using in situ data either as part of a supervised model or for accuracy assessment in unsupervised models can help to improve the classification and/or highlight sources of error. Even non-spatial data collected in the field can help to improve classification, such as detailed spectral information about a particular group of plant species assisting in selecting data which will best capture the important differences.

Question 15: I am getting an error with the raster code. It looks like it has the file pathway for your computer. How would I update that call to reflect my directory?
Answer 15: Once you download the data, you can change the path to make it work for your application. By default, the data should appear in the /data/rast directory. If you cannot download the data from Github clone, you may need to download it from Github directly.

Question 16: Can you show the system that you use for searching tools?
Answer 16: If this in reference to Earthdata Search, refer to the presentation for that information.

Question 17: Is there a way to run the GitHub R code on Google Colab?
Answer 17: Yes. The documents provided are compatible.

Question 18: Do I have to process images before uploading to R studio?

Answer 18: As referenced in the presentation, the only preprocessing I did for the data was to combine all of the individual HLS bands into a single file for the sake of convenience. Some imagery data is available as a single multi-band file (each file has many bands), while other data will be offered only as multiple single-band files (each band is its own file).

Question 19: Do you know if K-means clustering is available as an add-in in QGIS or ArcGIS?
Answer 19: As referenced previously, there are tools available, but exercise caution when utilizing them.

Question 20: Can you explain more about deciding centers in K-means?
Answer 20: Centers are decided from existing points in the deciding space. These points are random. We will go into further detail during Part 2.

Question 21: Wouldn't Sentinel 1 be better? Sentinel 2 could have cloud cover affecting the image.
Answer 21: Harmonized Landsat Sentinel data was used and the Earthdata filter to limit cloud cover percentage ensures that we only download cloud-free imagery. Use data that works for your particular use case and utilize the filtering options on Earthdata.

Question 22: Given harmonized Landsat Sentinel time series and Random Forest classification outputs, how does NASA quantify and communicate change-detection confidence so that transient spectral effects (phenology, clouds, atmospheric scattering, sensor drift) and classification uncertainty are not mistaken for real land cover change? What formal protocols or decision thresholds prevent automated change products from being used to trigger policy, enforcement, or funding actions without independent ground or high-res?
Answer 22: As we cannot provide any formal recommendations, this training serves more as a basic understanding of classifications.

Question 23: Can I access and submit the homework by any other means? I don't agree with Google's Terms of Use.
Answer 23: At this time, we use a Google form for all homework submissions. Our apologies.

Question 24: What exactly does "Harmonized" mean in an HLS data set?
Answer 24: Landsat is a legacy product and Sentinel is relatively new. The two data sets coming together is the “harmonization.” You can check out this link for more information about the project.

Question 25: Is this K-means model the same for cities and rural areas (since it's just based on an optical pixel, anything blue/dark could potentially be water)? Wouldn’t different models be better? If they're different, how would you align labels?
Answer 25: K-means is best suited for individual scenes. We didn’t use K-means for multiple scenes and we will cover that in Part 2.

Question 25: What method would you use to validate the classified raster for a large area of interest, such as for a province or state? How can you produce metrics for error in classification? Is it possible to create boundaries for each class that will enable areas to be calculated for each class?
Answer 25: We will cover ground truth points in Part 2 which is good for larger scale applications. With that data, we can calculate model performance metrics for accuracy assessment as well as calculating the magnitude and direction of change. You can calculate total land area for each class by multiplying the per-class pixel count by the pixel area. If you need individual polygons representing groups of pixels of the same class you are going to want to look into segmentation.

Question 26: Water is generally delineated successfully. How can we use satellite imagery to identify changes in soil moisture (irrigation or precipitation)? Does noise from vegetation affect reflectance in this case?
Answer 26: There are methods using radar for soil moisture, but soil moisture is also more complicated than typical land cover classifications with the exception of standing water. The HLS data (and similar kinds of remote sensing data) can only see surface features.

Question 27: Would it make sense to:
1. Calculate a series of indices (NDVI, NDWI, SAVI, RNB, etc.) for two different dates.
2. Group each of the indices for each date into a single spatrast object.
3. Perform the k-means process in a similar way to the example.

I would be considering using this to monitor forests and forest fire risks. I am unsure whether using the original bands would produce a better classification.
Answer 27: K-means operates on a closeness of bands within the feature space. The variability that would be captured by an index such as NDVI should already be captured by the information in the original bands. For your particular application, try supervised classification with the original bands.

Question 28: Can K-Means be used to determine the Normalized Difference Vegetation Index (NDVI) is a remote sensing metric used to assess live green vegetation by measuring the difference between near-infrared (reflected by plants) and red light (absorbed by plants)?
Answer 28: K-means may not be the best tool for this application. Refer to our previous QGIS training for more information about the theory and practice behind spectral indices such as NDVI.

Question 29: What method would you use to validate the classified raster for a large area of interest, such as for a province or state? How can you produce metrics for error in classification? Is it possible to create boundaries for each class that will enable areas to be calculated for each class?
Answer 29: Refer to the previous questions.

Question 30: Given a Random Forest model trained on imagery from one date, how do you quantify and mitigate temporal domain shift when applying it to imagery from another date with different phenology, sensor calibration, and land management practices? Specifically, which validation strategies, transfer‑learning or domain‑adaptation techniques, and uncertainty propagation methods do you recommend to ensure detected changes reflect true LCLU transitions rather than model?
Answer 30: We will look into ground truth data in more detail during Part 2 of this training series. I've also included some extra information at the end of the document for Part 2 which goes into more detail about explaining and evaluating RF models.

Question 31: Lately, I wanted to use some sentinel data, but the cloud coverage made it unable to do so. Any way to deal with that?

Answer 31: Yes and no. In reference to HLS data, cloud cover can be an issue. Filtering by maximum cloud coverage in Earthdata Search can help. Unfortunately, sometimes there isn't any cloud-free data for the precise time and place you are interested in studying.

Question 32: I cannot install package tidyterra. Do you have any guidance?
Answer 32: Not without seeing the specific error message. Perhaps setting your CRAN mirror to the cloud version or downloading via the r-universe. More information can be found on the tidyterra page.

Question 33: In remote sensing, when working with multi temporal satellite imagery to monitor deforestation across a large region, what preprocessing steps are necessary to ensure consistency between datasets? How can you quantify the accuracy of change detection results, and what statistical methods might be applied to validate the findings?
Answer 33: In this case, you are looking for L2 or L3 data that have been pre-processed. The harmonization process itself is very complicated and is not recommended unless you are familiar with the satellite data.

Question 34: A few years ago you gave a workshop that used semi-automatic plug-in on QGIS to do LCLU analysis. It was very labor intensive. Now, do you think that AI has advanced that unsupervised analysis enough (I am not saying that K-means method uses AI) so that it is very efficient, rapid, and accurate at classifying land classes that such manually-intensive analysis has become obsolete?
Answer 34: No. The standard of science is reproducibility. The why and how of classification is an important indicator of producing consistent outcomes. The current state of AI doesn't allow us to produce consistent and explainable results, though there is some promising work in the field of explainable AI which would give us the ability to break down and analyze the "thought process" behind AI outputs.

Question 35: Are there methods to determine the error rate in classification using K-means, or are results verified only through the image result?
Answer 35: K-means as covered in the previous questions is unsupervised. An error rate assessment can be done with some ground truth data.

Question 36: I am getting a specific error with the rast function: "Error in h(simpleError(msg, call)) :
error in evaluating the argument 'x' in selecting a method for function 'RGB': [rast] cannot open this file as a SpatRaster: /Users/name/Documents/ARSET/LCLUC_2026/data/rast/HLS_2017_multiband.tif"
Answer 36: Check your file path and make sure it is correct. You can also try to open the file using another program like QGIS to confirm that it isn’t corrupted.