Part 4: Questions & Answers

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 Erika Podest (erika.podest@jpl.nasa.gov) or Sean McCartney (sean.mccartney@nasa.gov).

Question 1: Can you get training data from high resolution satellite data? If yes, how does a non-agricultural person who has not identified crops before identifying them from imagery?
Answer 1: Yes, you can. If you do not have in situ training data, the safest identification you can do by simply viewing satellite images is to identify whether an area contains a crop or not, but you can not securely identify the type of crop.

Question 2: What is the difference between overall Accuracy and the F-Score in accuracy assessment?
Answer 2: F-Score is calculated per class. This means that we can see how well our algorithm performs for each class and not only overall. The overall accuracy can be very misleading, since some classes can have accuracy that is extremely low - indicating potential issues in our training data or in their size, while other classes may have very high accuracy.

Question 3: Is there a relationship between the minimum number of samples and the number of trees in RF classification training?
Answer 3: Not necessarily. There is a relationship between the number of features in your dataset and the number of trees. The more dimensions in your dataset, the more trees should be used. If your training dataset is too small (meaning the number of samples), it makes no sense to have a high number of trees because the samples will repeat in the random sampling.

Question 4: I want to perform a segmentation after an object based classification using an NDVI temporal stack. How can one perform this using Sentinel-2 data?
Answer 4: There are segmentation tools available in SNAP within the Orfeo Toolbox plugin. You can test them.

Question 5: For the classification algorithms presented, when we use for example Entropy instead of Gini (which increases the computational time), does it necessarily mean better results?
Answer 5: No. Based on the literature and testing, the results tend to be very similar.

Question 6: The Alaska Satellite Facility (ASF) offers Vertex On-Demand Radiometric Terrain Corrected Sentinel-1 data. Should we use ASF Vertex RTC processed datasets when selecting radar data for crop mapping, especially crops on steep slopes?
Answer 6: Because SARs are side-looking instruments, these data are affected by topographic and radiometric distortions (layover, foreshortening, and shadowing) in regions of significant topography. It is possible to apply a terrain correction during SAR pre-processing (please refer to Part III). However, it is convenient to access S1 data from ASF that has already had RTC applied. Keep in mind that radiometric distortions in mountainous regions are very difficult to correct, so you will never get a perfect result. Nevertheless, we would recommend that you evaluate the ASF RTC products to determine if these products meet your needs.

Question 7: Can we measure the salinity of soil using biophysical variables for soil brightness? If not, which index may be the proper one to use for soil salinity?
Answer 7: The Soil Adjusted Vegetation Index (SAVI) has been used in some research studies to estimate soil salinity. Please search the literature for studies which have tested SAVI for this purpose.

Question 8: Is it possible to know the final decision rule or rules of the trained random forest classifier?
Answer 8: There are functions in Python that can allow the visualization of the entire tree and all the rules used at each node. However, I would not recommend trying to use them when classifying EO images (many bands) as the trees will be extremely large and the values are continuous. They can be tested with smaller datasets like the example with the fruits in the presentation.
See:
https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c

Question 9: Are OOB score and OOB error different? Could you please provide a clarification?

Answer 9: OOB score is the portion of correct results, while the OOB error is the portion of false results (e.g., OOB error = 1 - OOB score).

Question 10: What if two features have the same distance from the z plane in the SVM classification?
Answer 10: The algorithm never considers only one datapoint but multiple. Therefore, if the situation occurs that two features have the same distance, then other data points are always used to maximize the margin.

Question 11: Are both SVM and K-means based on distance?
Answer 11: Yes. They are both distance-based, although they use very different approaches.

Question 12: Are Sentinel-2 SWIR bands resampled to 10 m or downscaled?
Answer 12: The resolution of the Sentinel-2 SWIR bands is 20 m. We will resample them to 10 m resolution.

Question 13: If the training data is based on the classified image, how will the ML capture the temporal change in crop class?
Answer 13: The training dataset that we have created in Python also contains, for each of our training points, the values of all the bands present in our input dataset.

Question 14: How can we identify outliers in the K-means approach?
Answer 14: In the K-means-based outlier detection technique, the data are partitioned into k groups by assigning them to the closest cluster centers. Once assigned, we can compute the distance or dissimilarity between each object (point) and its cluster center, and pick those with largest distances as outliers.

Question 15: How can I get training data for my country or locality?
Answer 15: You need to contact any local authorities or institutes that may happen to have such data, and can make them publicly available. Each region has its own rules and databases. Collect Earth Online may be an option as well:
https://www.sciencedirect.com/science/article/pii/S1364815218312568
https://collect.earth/home

Question 16: Does SNAP fully support Mac M1 Chip?

Answer 16: Unfortunately I do not know. Please consult the SNAP website or the SNAP forum - https://forum.step.esa.int. Someone may have asked a similar question or you can pose the question to the group.

Question 17: For regression with RF, is it sufficient to perform tuning in n_estimators, m_try, and extract important features using the recursive feature elimination (RFE) algorithm? Assuming that I split the data to 80% for training and 20% for test/validation, is it enough to produce significant accuracy or do I have to optimize tree depth and include the bootstrap aggregation in any case?
Answer 17: That is difficult to say, as the accuracy of the algorithm always depends on many different parameters. In the end, testing will provide the best answer. Bootstrapping can help reduce overfitting.

Question 18: Can we subset by importing a shapefile or json file for a region of interest (ROI)?
Answer 18: In SNAP unfortunately no. You can use online tools or QGIS to convert your json or shapefile to WKT format that can be used in the SNAP Graph interface.
Draw and Convert WKT: http://dev.openlayers.org/examples/vector-formats.html

Question 19: Batch processing seems to take a long time collectively, per se for 5 Sentinel-2 images. I noticed the processing of one-by-one takes less time in total. But, of course, there is less effort needed for batch processing. SNAP tends to also lag a lot when handling large files as noticed from last week's practical week 2 data. Does SNAP require a super high-performance computer to run with ease?
Answer 19: Using the batch processing at the GUI of SNAP is relatively slower than running it from the GPT (command line interface), as it does not release memory between products. Therefore, sometimes you may indeed face slower performance or run out of memory during batch processing. The minimum required RAM to use SNAP is 8GB, but 16GB provides quite good performance. Of course, the more memory available, the faster the processing. There is also the option to go into SNAP through Tools -> Options -> Performance tab, and increase the cache to a size that your device can support. Command line batch processing script using python or shell script can be faster. We have examples of shell scripts on our RUS site.
For example, the Vegetation Monitoring tutorial includes the shell script method: https://rus-copernicus.eu/portal/wp-content/uploads/library/education/training/LAND11_VegetationMonitoring4Agri_Italy_Tutorial.pdf.

Question 20: I see that RUS-Copernicus has a lot of great training material, but training sessions are only for Europeans. Any chance to make the training information available globally?
Answer 20: RUS Copernicus material (pdf guide and video of the webinar) are publicly available at https://rus-copernicus.eu/portal/the-rus-library/train-with-rus/ and https://www.youtube.com/channel/UCB01WjameYMvL7-XfI8vRIA/videos accordingly. Unfortunately the VMs availability is limited to EU citizens/residents, based on the regulations of the project. The webinars delivered are also open to participants from any region.

Question 21: Doesn't SNAP have its own ML classifier tool?
Answer 21: Yes. Random Forest and a number of other classifiers are also available in SNAP, but less parameters can be optimized and the output dataset does not retain the same class numbers, which can be confusing (for RF). You also have to import training data by class.

Question 22: Is it possible to perform a Spectral Separability analysis in SNAP using distances like Jeffries Matusita?
Answer 22: Unfortunately, SNAP does not have spectral library tools, but perhaps this thread on the SNAP forum may be helpful:
https://forum.step.esa.int/t/classification-based-on-spectral-library/3429/8

Question 23: Thank you so much for a wonderful presentation. Is it possible to provide a .yml file as well, please?
Answer 23: Of course. Please send an e-mail to eotraining@serco.com to request the training kit. It also contains the .yml file and instructions on how to create the environment.

Question 24: Can we get the Jupyter notebook?
Answer 24: Of course. Please send an e-mail to eotraining@serco.com to request the training kit.

Question 25: Can the starting data files that the presenter used be made available to run and validate the Python used? If the combination fails then the Python may not run the same.
Answer 25: Please elaborate. The preprocessed coregistered stack is available from the training webpage. Below Part 4 you will find a link called "Download the Processing

Data". The Training kit can be requested from eotraining@serco.com. These two contain all the data used in the demonstration.

Question 26: For people that are not familiar with coding, is there an alternative to running the data processing in Jupyter lab elsewhere (e.g., SNAP)?
Answer 26: As already mentioned, the algorithms are also available in SNAP, so the majority of the steps can be run there. There are some limitations to the parameters that can be set and also the training data are prepared differently - they are imported separately for each class.

Question 27: What attributes should we provide in the training data shapefile?
Answer 27: The only attributes included in the training data are the point id, gridcode (identifying the class), and the X (UTM_E) and the Y (UTM_N) coordinate in the same coordinate system as the input dataset.

Question 28: Concerning the random selection of training and validation points, how was it done (as I believe you have not clicked thousands of times on a map)?
Answer 28: The training data was prepared in QGIS using the function/tool "Random Points inside the polygons". There you can select the number of samples to be drawn. First you should dissolve the features (polygons) by class - meaning you will have a single multipolygon per class from which you randomly sample.

Question 29: Just wondering if you have developed a module for classes. What is the basis for defining these classes in the notebook?
Answer 29: No, we have not developed a module for classes.

Question 30: For the seams and overlaps when mosaicking, how does one check the quality besides visual assessment?
Answer 30: Visual assessment is your best tool. Advanced mosaicking tools are available in some softwares, which can ensure better results.

Question 31: Will the notebook "PY02_cropMappin" be made available?
Answer 31: Please send an e-mail to eotraining@serco.com to request the training kit and you will receive the notebook as well.

Question 32: Is image normalization important for the Random Forest classification? If yes, is there a way to determine the impact of non-normalized data (training/validation) on the Random Forest classification?
Answer 32: No, normalization is not necessary in Random Forest. It can handle multiscale data.

Question 33: How are fields that are double- or triple-cropped dealt with in the classification? If fields are double-cropped with different crops, do you use data from both seasons when classifying one crop?
Answer 33: This will create issues. You could create a multicrop class which needs to be specifically defined for that particular crop combination. But all the training data will change between the seasons. You can't have multiple classes per polygon, so even your training data needs to include this multicrop class.

Question 34: Is there any tool for distinguishing between forest and crop?
Answer 34: If your training data includes forests, then the forest class can be included in the result of your classification.

Question 35: Does snappy (SNAP python interface) support Python 3.4 or later versions? I installed SNAP 8.0 with Python 3.4 and I tried to install snappy with Python 3.7.6 and I couldn't install it.
Answer 35: Snappy supports Python 2.7, 3.3 to 3.6 64-bit (Linux + Darwin) and both 32-bit and 64-bit (Windows), as well as Anaconda distributions. Your issue is likely caused by using version 3.7.6. In this exercise I used version 3.6.13.

Question 36: What should we do for crop lands that have different crop types for different seasons (when defining the training data)?
Answer 36: Unfortunately, you will then need to define the season in a way that it only includes a single crop per field, or you can test defining a specific multicrop class if there is enough training data available and the change occurs at a specific time.

Question 37: How is the dimensionality reduction of data performed for a supervised classification? Is feature engineering part of such analysis? Also, is PCA the preferred way?
Answer 37: Yes, dimensionality reduction can have a positive effect on the classification, even if just reducing the computational time. Regarding the preferred way, I have unfortunately not tested it, so I cannot answer this.

Question 38: What is GRIDCODE? Is it crop type code?
Answer 38: Yes, this is the numerical value assigned to a crop type.

Question 39: What kind of hyper-parameter tuning is performed for the model to avoid overfitting/underfitting?
Answer 39: To avoid overfitting in random forest, the main thing you need to do is optimize a tuning parameter that governs the number of features that are randomly chosen to grow each tree from the bootstrapped data. Typically, you do this via k-fold cross-validation and choose the tuning parameter that minimizes test sample prediction error.

Question 40: Is there a Random Forest routine implemented in SNAP?
Answer 40: Yes there is, but it has limitations. But test it out and maybe it will work for you in your scenario.

Question 41: What is your advice for crop classification (pixel- or object-based classification)?
Answer 41: Both have advantages. I believe object-based classification will provide cleaner results, but can not be implemented in all study areas. Moreover, it is dependent on good segmentation. Pixel classification in general is simpler.

Question 42: For splitting training and validation data, should we choose different polygons/regions? Or is it okay to sample over the same polygons?
Answer 42: In this example I have used different polygons. This is better because it will make your training and validation data truly independent.

Question 43: What is the required machine configuration for running this code? Can it be done in Google Colab?
Answer 43: Unfortunately, I am not familiar with Colab. You need to have the possibility to install the required packages and install SNAP (can be omitted if you supply a list of band names to the code - requires code change). There should be no reason you could not use Colab.

Question 44: Why is it important to have the coordinates UTM E and UTM N in the attributes table?

Answer 44: They are used to extract the values corresponding to the training data points from the input dataset. It uses a list of coordinates.

Question 45: Can we stack images of several years and then run the classification? If so, what will be the effect on accuracy?
Answer 45: Likely the accuracy will be very low, since the crops will change between years. Your training dataset will not be valid for other years. You always need to run new training data each year.

Question 46: Are Random Forest probability maps better as they allow for better control on classification since the default classification takes the probability of 0.5 to bin a particular pixel value? Do you think that the probability approach might be more beneficial for noisy data, especially for Sentinel-1 data? For example, we can set the probability to 0.75 and greater to classify a pixel.
Answer 46: I am not quite sure as I have not tested this.

Question 47: Is there any module to generate shadow and layover masks?
Answer 47: Not one that I am aware of in SNAP.

Question 48: Can one use a spectroradiometer signature for crops in the training phase? If so, how? And which spectral range is used?
Answer 48: I expect you should use the same range as that of the satellite data you are planning to use for the classification. However, I am not an expert on this.

Question 49: Will the addition of texture indices and/or vegetation indices increase the accuracy?
Answer 49: In general, yes. It may also help you to reduce the number of dimensions.

Question 50: How is water masked in the output?
Answer 50: Our data includes a class called ‘masked’ which includes water, wetlands, and urban areas; but you can also include these as separate classes in case your training data contain them.

Question 51: How do you tune SVM parameters for image classification?
Answer 51: For SVM parameter tuning see:
https://www.hackerearth.com/blog/developers/simple-tutorial-svm-parameter-tuning-python-r/

It is not an EO image-specific tutorial, however, the tuning/testing method will be the same.

Question 52: In Random Forest, what is the difference between training data, testing data, and validation data?
Answer 52: I am not sure what is meant by testing data. The training data are used for the training of the model and the validation data should be independent from the training data and should be used for testing the performance of the model.

Question 53: I did not understand how to create the training data. Which software do you use to pass them to shapefile? Can the shapefile be both point and polygon? Also, is it okay to create the training data using high resolution satellite imagery, in case I cannot acquire in situ data?
Answer 53: The data for our study area was vectorized. Small polygons were dropped and the number of polygons for the major classes was reduced. The data was split 70/30 to training and validation data. To further limit the number of samples in our training and validation data (due to computational constraints) we dissolved all the polygons belonging to one class and created a sampled point layer.

Question 54: I find trying to learn ML for Earth observation (EO) by myself quite daunting. There is so much information out there, and so many ways to go about it - all with various levels of expertise and complexity, which makes it difficult for me to know how to begin and what tools to start with. Which is why I find this training so great! Would it be possible to provide a list of EO resources/sites that you think are good starting points? Also, sites with a few rules of thumb and cheat sheets for using ML in EO would be great.
Answer 54: There is a new MOOC course available on the Future Learn platform on Artificial Intelligence (AI) for Earth Monitoring which may be a good start for you.
https://www.futurelearn.com/courses/artificial-intelligence-for-earth-monitoring
You can also find a lot of very useful articles and resources on Medium. For example:
https://medium.com/radiant-earth-insights/discoverable-and-reusable-ml-workflows-for-earth-observation-part-1-e198507b5eaa

Question 55: How can one measure accuracy for cluster results?
Answer 55: If you have actual labels, you can compare them with the clusters, assign each cluster a label, and evaluate the performance. Typically purity and NMI (normalized mutual information) are used. The following document (Evaluation of

Clustering) has a detailed explanation. Note that in our training data there is an aggregated class called ‘masked’, which contains water bodies, wetlands, and urban areas. These will be likely represented as different clusters in the clustering results.

Question 56: How do you implement a supervised classification on an image mosaic when covering large regions? Are there any specific steps on mosaicking needed for implementation of Random Forest?
Answer 56: You can perform a RF classification on a mosaic, but a better option is to perform the RF on images separately and mosaic the results. It will likely give you a smoother result.

Question 57: What version of Python did you use? I remember having difficulties getting Snappy to work because it requires an old version of Python (3.5 I think). Was this an issue for you?
Answer 57: I used version 3.6 installed in an Anaconda environment. I have not tested it with older versions. As per the information on the SNAP website, the supported versions are Python 2.7, 3.3 to 3.6 64-bit (Linux + Darwin) and both 32-bit and 64-bit (Windows), as well as Anaconda distributions.

Question 58: How does the classifier work for the intercropping areas?
Answer 58: This will depend on the resolution of the imagery being used. For satellites like Sentinel-2, the pixels in intercropped areas will appear as a blend of both crop types (mixed pixels). In Canada this is not a common farming practice so we do not see it often. When we do, we do not collect an observation point and we let the classifier decide the outcome.

Question 59: Can we run the same classification code in GEE instead of Jupyter notebook? Can we do batch processing in Jupyter notebook?
Answer 59: Unfortunately I do not use GEE, but I see no reason why not if you are able to install all the required packages.

Question 60: How much accuracy can be accepted?
Answer 60: For the accuracy of the AAFC ACI, we consider 85% overall crop accuracy as the minimum that is acceptable (per province).

Question 61: Is seasonal training data important when working on a large time-series dataset for classification?

Answer 61: Yes, training data should be collected to capture any changes in crop types in the Area of Interest. If the time-series spans over seasons where multiple crops will be grown and harvested in the field, then it would be important to provide those observations if possible.

Question 62: If the stack is too heavy to process, can you sub-stack (meaning use several fewer stack images per sub-stack) and get the same results?
Answer 62: When processing classifications over a very large area (such as Canada), we avoid sub-dividing our nominal area (typically a Landsat WRS). There is usually not enough training data to allow for sub-stacking (some field types might be under-represented).