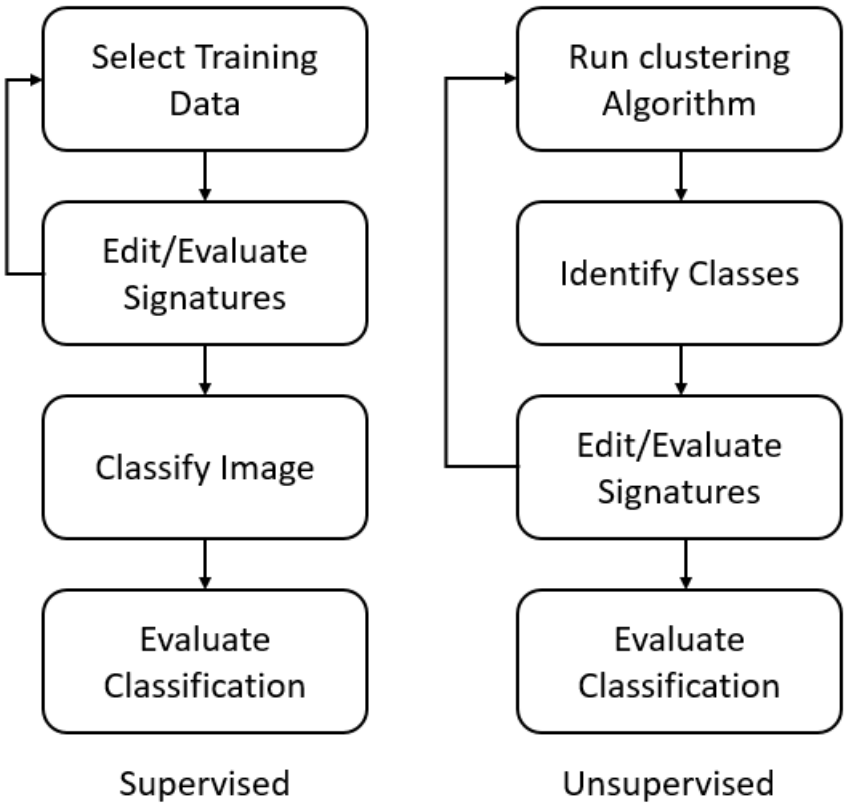
Exercise 27: Supervised Classification

In this exercise, you will conduct a supervised classification using machine learning methods implemented in ArcGIS Pro. Specifically, you will compare the results of support vector machines (SVM) and random forests (RF) classifications using a Sentinel-2 images of Vancouver, British Columbia.

When classifying an image, two broad methods are available: unsupervised classification and supervised classification. Unsupervised classification requires that the image be clustered into spectral classes using a clustering algorithm, such as k-mean or ISODATA. The analyst must then label these spectral classes as informational classes. In contrast, supervised classification requires that the analyst collect training data, or examples of the informational classes of interest. These are generally collected at point or pixel locations or as polygons. An algorithm is then used to classify the image using the training samples as examples. Here, we will focus on supervised classification.



Once a classification is produced, this result is generally validated using a confusion matrix. The confusion matrix compares the classification of a location on a map (such as a pixel) with the actual classification of that

location. The goal is to compare the results of a classification output to the correct classification. This is done by comparing your output to validation data. These are data that have been produced using field sampling, manual image interpretation, or some method deemed to be significantly more accurate than the classification. So, you are comparing your classification to a dataset of higher quality. Note that we try to avoid the term "ground truth." Instead, we use the term "validation" or "reference" data. This is because all data have some error, so no dataset represents the "truth." Validation data have been provided for you to use in this exercise, so you will not need to produce your own.

A confusion matrix is designed such that the columns represent the validation data and the rows represent the classification. An example confusion matrix is shown below.

		Reference Data							
		Forested	*Pasture/ Grass*	*Barren*	*Cropland*	*Developed*	*Water*	**Row Total**	**User's Accuracy**
Classified Data	*Forested*	A							
	Pasture/ Grass					B			
	Barren								
	Cropland								
	Developed								
	Water			C					
	Column Total								
	Producer's Accuracy								

In this matrix, the cell labelled as A would represent samples that were forest and were correctly classified as forest. So, these are correct classifications. Cell B would represent samples that were developed but were misclassified as pasture/grass. Cell C would represent samples that were barren but were misclassified as water. The diagonal shown in gray represents correct classifications. All cells off the diagonal represent misclassifications or error.

From the error matrix, a variety of measures can be calculated.

From the error matrix, you can calculate an overall accuracy. The overall accuracy is defined as follows and it represents the percentage of samples that were correctly classified:

$$\frac{\text{Number of Features Correctly Classified}}{\text{Total Number of Features}} \text{ X } 100$$

You can also calculate class accuracies as user's accuracy (1 - commission error) and producer's accuracy (1 - omission error). User's accuracy (1 - commission error) specifically quantifies errors of commission, samples that were included in the wrong class (for example, a sample that was forest was included in the water class). In contrast, producer's accuracy (1- omission error) quantifies errors of omission, samples that are not in the right class (for example, a forest sample was not included in the forest class).

The user's accuracy for a specific class can be calculated as follows:

$$\frac{\text{Number of Features of Specific Class Correctly Classified}}{\text{Row Total}}$$

The producer's accuracy for a specific class can be calculated as follows:

$$\frac{\text{Number of Features of Specific Class Correctly Classified}}{\text{Column Total}}$$

Lastly, chance adjusted agreement can be calculated using the Kappa statistic. This is a correction because sometimes you get samples correct just by random chance. So, Kappa offers a chance adjusted agreement. It is generally expressed as a ratio (0-1) as opposed to a percentage (0-100%).

Here is the formula:

$$\frac{[\text{Total Number of Features} * \text{Number of Correct Features}] - [\text{Sum of All Row Totals} * \text{Column Totals}]}{[\text{Total Number of Features squared}] * [\text{Sum of All Row Totals} * \text{Column Totals}]}$$

In this exercise, you will use Sentinel-2 data, which is currently provided at no cost by the European Space Agency (ESA). These data can also be obtained from the United States Geologic Survey EarthExplorer. The table below provides a description of these data.

Sentinel-2 Bands	**Center Wavelength (μm)**	**Spatial Resolution (m)**
Band 1: Coastal Aerosol	0.443	60
Band 2: Blue	0.490	10
Band 3: Green	0.560	10
Band 4: Red	0.665	10
Band 5: Red Edge	0.705	20
Band 6: Red Edge	0.740	20
Band 7: Red Edge	0.783	20
Band 8: Near Infrared (NIR)	0.842	10
Band 8A: Narrow Near Infrared (NIR)	0.865	20
Band 9: Water Vapor	0.945	60
Band 10: Shortwave Infrared Cirrus	1.375	60
Band 11: Shortwave Infrared (SWIR)	1.610	20
Band 12: Shortwave Infrared (SWIR)	2.190	29

In this exercise, we will only make use of bands 2, 3, 4, 8, 11, and 12. These bands were stacked into a six channel composite at a 10 m spatial resolution. Here are the designations:

Layer_1: Blue (Band 2)

Layer_2: Green (Band 3)

Layer_3: Red (Band 4)

Layer_4: NIR (Band 8)

Layer_5: SWIR (Band 11)

Layer_6: SWIR (Band 12)

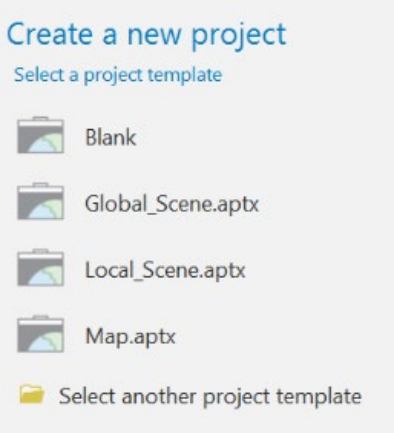
Topics covered in this exercise include:

1. Interpreting imagery to collect training data
2. Executing machine learning supervised classification
3. Interpret a confusion matrix
4. Compare the results of different supervised classification methods using the confusion matrix

Step 1. Create and Prepare a New Project

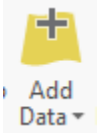
First, you will need to create a new project in which to work.

- Open ArcGIS Pro. This can be done by navigating to All Apps followed by the ArcGIS Folder. Within the ArcGIS Folder, select ArcGIS Pro. Note that you can also use a Task Bar or Desktop shortcut if they are available on your machine.
- Once ArcGIS Pro launches, select **Map.aptx** under Create a new project on the right side of the page.
- In the Create a New Project Dialog Box, name your new project **Exercise_27** and save it a location of your choosing.



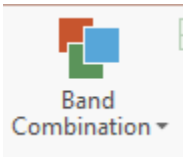
You have now created a new project. Since you used the **Map.aptx** project template, a map has already been added, but it does not contain any data layers other than a basemap. So, you will need to add the required data.

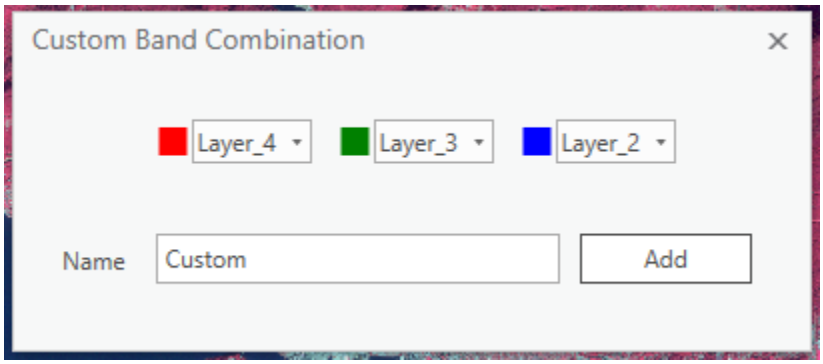
- Download the **Exercise_27** data from https://www.wvview.org/. All lab materials are available on the course webpage and linked to the exercise. You will need to extract the compressed files and save it to the location of your choosing.
- Using the Add Data button, add the following files from the downloaded **Data** folder: **validation_points.shp** and **sentinel_vancouver.img**.
- You can make the **validation_points** layer not visible, as you will not need it until the end of the exercise.



You should now change the band combinations for the Sentinel-2 image.

- Make sure the **sentinel_vancouver.img** layer is selected in the Contents Pane.
- Navigate to the Appearance Tab under RASTER LAYER.
- Select Band Combinations.
- In the drop-down list, select Custom.
- Set the red channel to Layer_4 (near infrared), the green channel to Layer_3 (red), and the blue channel to Layer_2 (green).





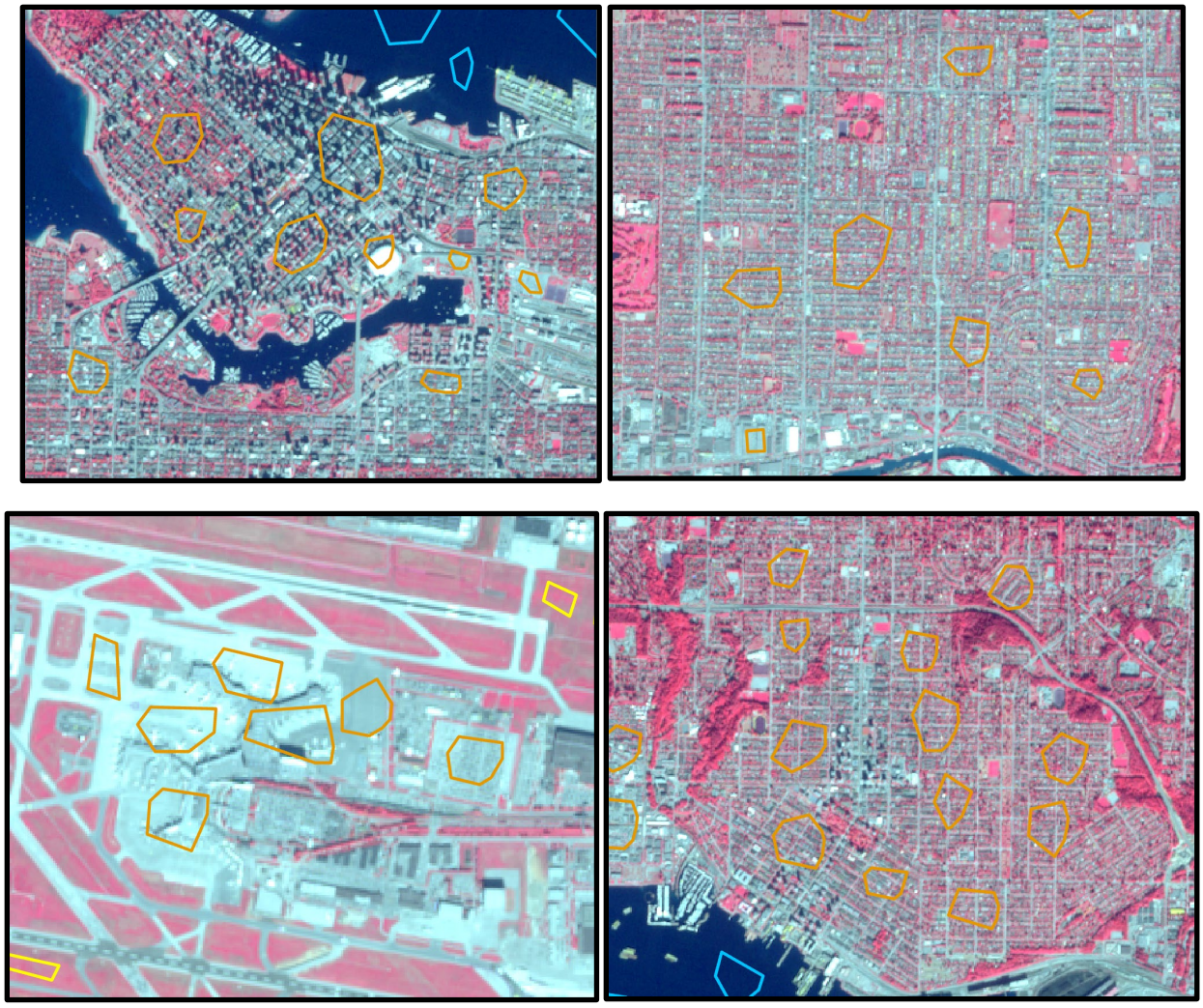
This will produce a standard false color composite. In this image, red areas generally indicated the presence of vegetation, as vegetation is very reflective in the near infrared wavelengths.

Step 2. Prepare to Collect Training Data

Before you can execute a supervised classification, you will need training samples. In this exercise, you will collect training samples as polygons. Here, you will attempt to differentiate five classes.

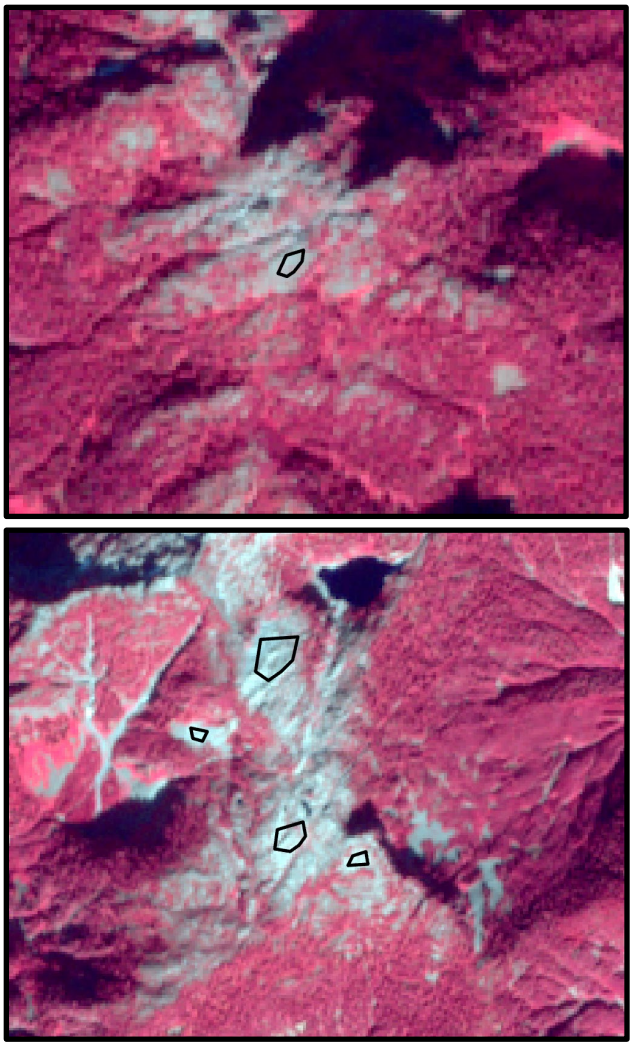
Developed

This class will include all commercial, urban, and residential areas. Common features within developed areas include buildings, roads, yards, and parking lots.



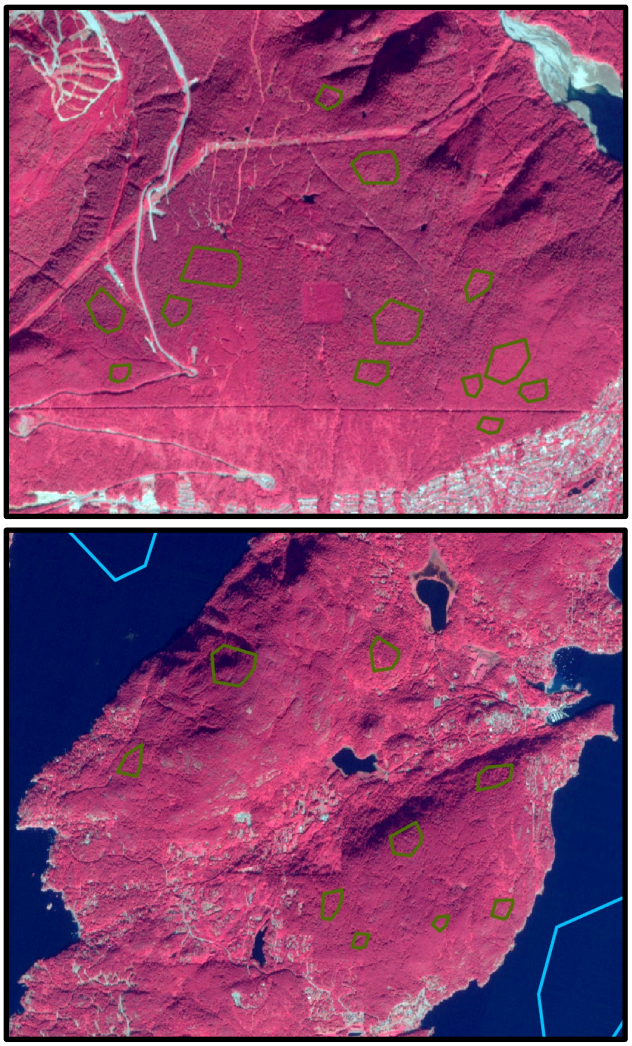
Barren

This class will include non-vegetated areas not associated with development, such as bare rock or soil. This is not a common cover type in this image. However, there are some bare rock surfaces in the mountainous area in the top part of the image.



Forest

This class will include all forested areas in the image. Do not include small stands of trees in residential areas.

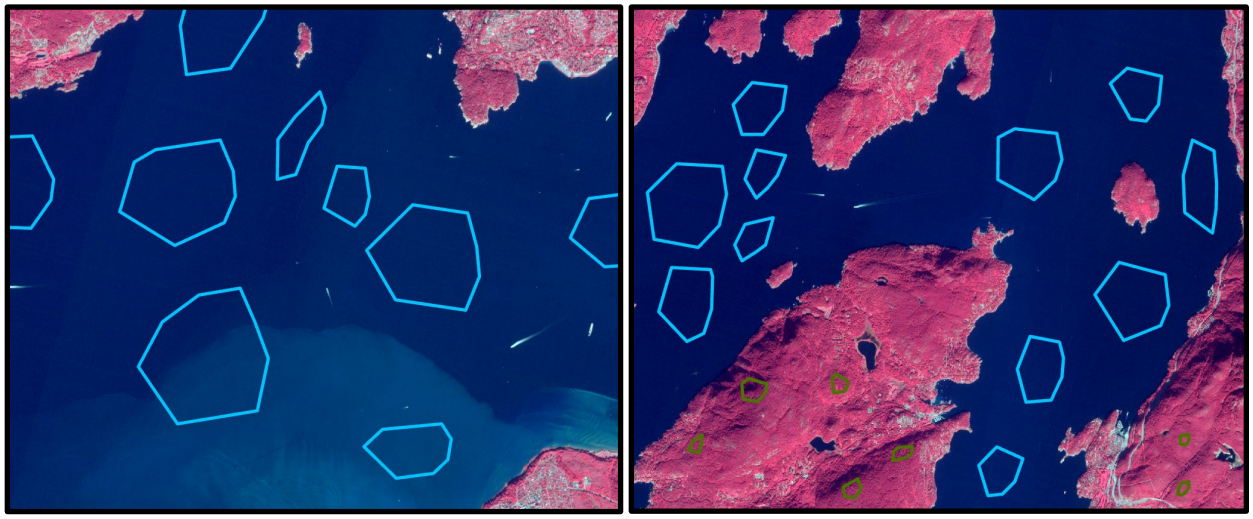


Herbaceous

This class will include vegetated areas that are not trees, such as fields, pastureland, and cropland.

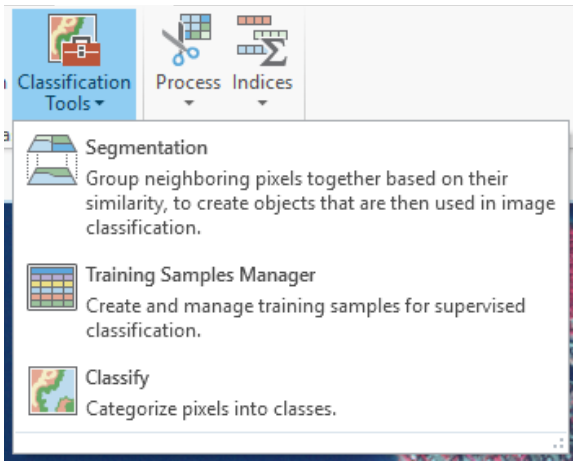
Water

This class will include all water features, such as the ocean, ponds, lakes, and rivers.



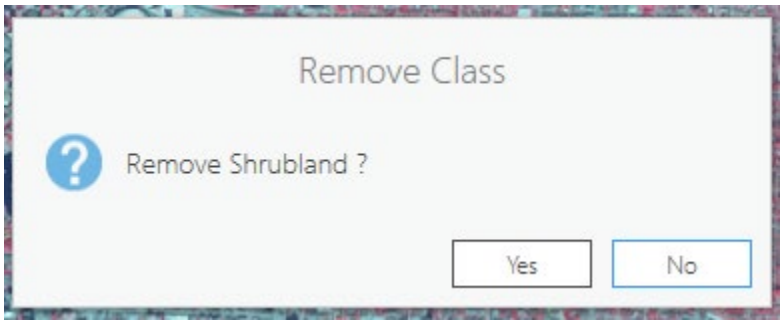
Before you begin collecting training data, you will need to do some preparation.

- Make sure the **sentinel_vancouver.img** is selected in the Contents Pane. Navigate to the Imagery Tab then select Classification Tools.
- Select Training Samples Manager. This will open the Image Classification Pane.



Note that some cover types are already suggested relative to the 2011 National Land Cover Database (NLCD). However, you will not differentiate all of these classes here. So, you will need to remove some of the classes.

- Make sure Shrubland is selected. Click the Remove Class button. In the dialog box, select Yes to remove the class. The class should be removed from the list.

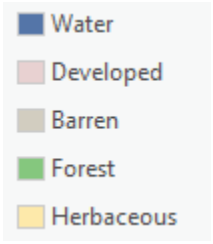


- Repeat this process to remove the Planted/Cultivated and Wetland classes. This should leave only the five classes of interest.

There are subclasses defined under forest. Although this is not necessary, we will also remove these subclasses.

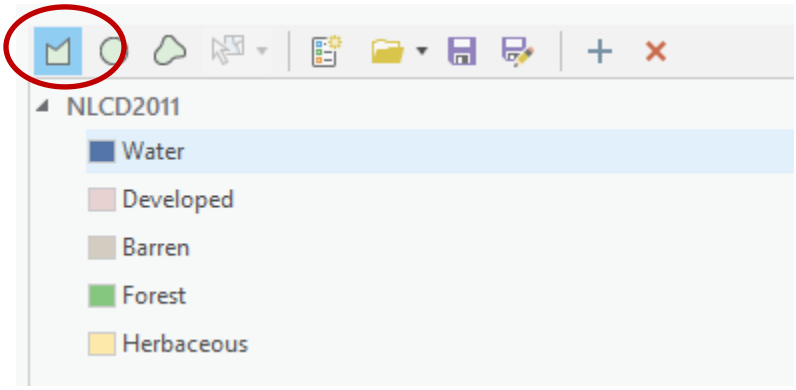
- Click on the drop-down arrow next to the Forest class. Select the three forest types individually then use the Remove Class button to remove them. The drop-down arrow should disappear once you are done since there are now no subclasses.

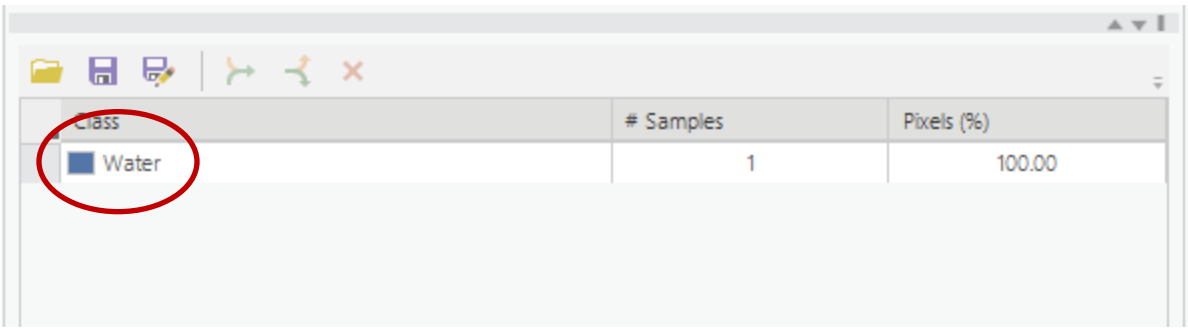
The final legend should look like the following:



You are now ready to start collecting training data to represent the spectral signatures of your classes.

- Collect training data for all of your classes. This can be done by selecting the class of interest from the list, selecting a drawing tool, then drawing a polygon in the map space. We would suggest using the polygon drawing tool, which is the first option in the list. Make sure that all of the pixels in each polygon represent the class of interest. Also, make sure that you capture the full range of spectral signatures for each class. Collect training data across the entire image, not just focused over a limited spatial extent. This will take some time. However, it important to collect quality training data so that the algorithm has the necessary information to classify the image. Use the examples provided above as a guide. To save your training data, use the Save option in the bottom part of the Image Classification Pane. Save them as a feature class in a geodatabase of your choosing.





Note: As you collect training data, you will need to alternate between drawing and navigating in the map space. The navigation tools are available under the Map Tab, and the training data drawing options are available in the Image Classification Pane. So, you will need to navigate through the windows as you draw.

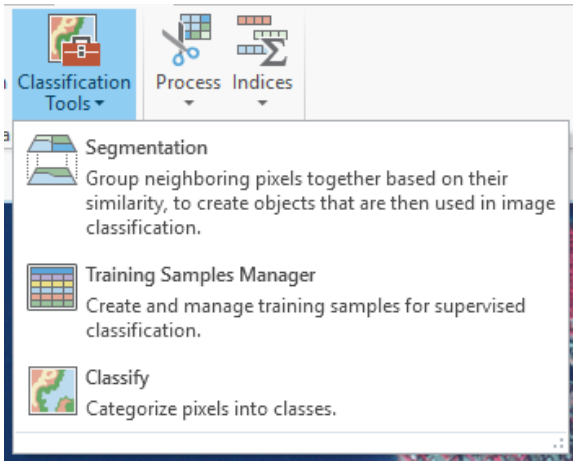
Here is an example of our final training data.

- Once you are done creating training data, remember to save them.

Step 3. Perform Classifications

You are now ready to classify the image. We will start with the SVM algorithm.

- Make sure the **sentinel_vancouver.img** layer is selected in the Contents Pane.
- Navigate to the Imagery Tab then click the Classification Tools button.
- Select Classify. This will open the Image Classification Pane.



- Set the Classifier to Support Vector Machine.
- Set the Training Samples to your training samples that were saved.
- Name the Output Classified Dataset **svm_result.img** and save it to a location of your choosing.
- Name the Output Classified Definition File (.ecd) **svm_definition.ecd** and save it a location of your choosing.
- You do not need to change any additional settings.
- Click Run to execute the classification. This can take some time.

Note: It is possible to include a segmented image in the classification. However, we did not do so here. Also, the performance of the algorithm may be improved by altering the user-defined parameters.

The results from the SVM classification are shown below. Note that this is a categorical raster with five classes differentiated. Your result will likely look different from ours since different training data were used.

You will now execute the RF classification.

- Make sure the **sentinel_vancouver.img** layer is selected in the Contents Pane.
- Navigate to the Imagery Tab then click the Classification Tools button.
- Select Classify. This will open the Image Classification Pane.
- Set the Classifier to Random Trees.
- Set the Training Samples to your training samples that were saved.
- Name the Output Classified Dataset **rf_result.img** and save it to a location of your choosing.
- Name the Output Classified Definition File (.ecd) **rf_definition.ecd** and save it to a location of your choosing.

- You do not need to change any additional settings.
- Click Run to execute the classification. This can take some time.

Our RF result is shown below.

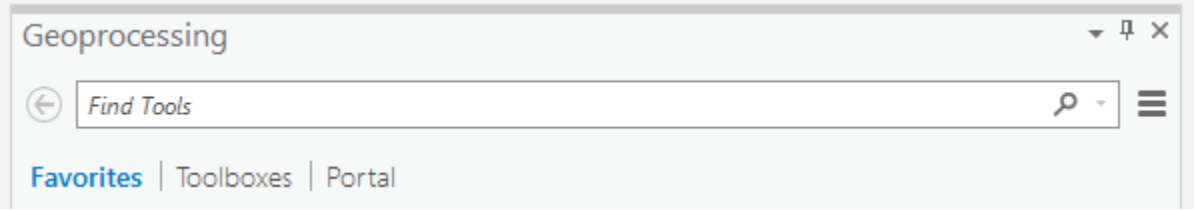
If you compare the two outputs, you will note that they look a bit different. This is because different classifiers were used, which impacts the result. However, it may not be readily obvious which classification yielded the best result. To assess this, we will need to use a confusion matrix.

Step 4. Generate Confusion Matrix

The **validation_points.shp** file provides a set of 1,000 randomly sampled validation data points. All of the points have already been interpreted to label them as a reference class. You will now use these validation samples to assess the accuracy of each classification and compare them. This can be accomplished using the **Extract Multi Values to Points Tool** and the **Compute Confusion Matrix Tool**. You will start with the SVM result.

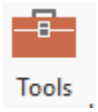
Note: In this lab, we will access tools from ArcToolbox. However, there are many ways to access tools in ArcGIS Pro. For example, some of the more common tools are provided in the Tools list in the Analysis Tab.

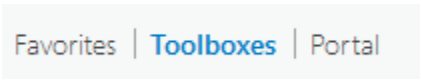
Once you open the Geoprocessing Pane, you can access favorite tools or search for Tools.



We have decided to demonstrate ArcToolbox here so that you get a sense of where the tools are located in the Toolbox directory.

- In the Analysis Tab, select Tools from the Geoprocessing Area. This should open the Geoprocessing Pane.
- In the Geoprocessing Pane, navigate to the Toolboxes.

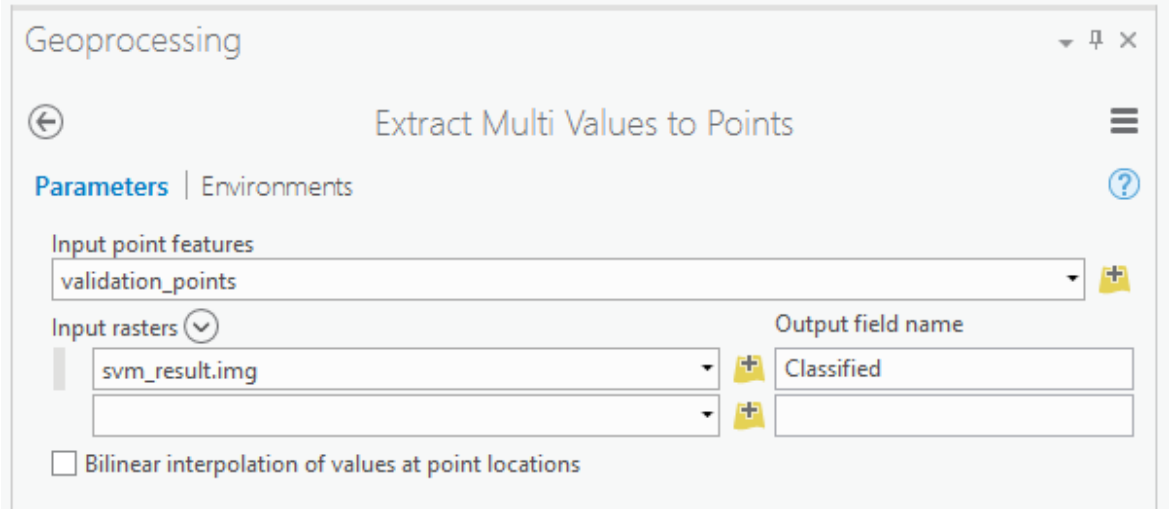




Note: We will not provide these directions for accessing other tools. We will just tell you where to find them within ArcToolbox.

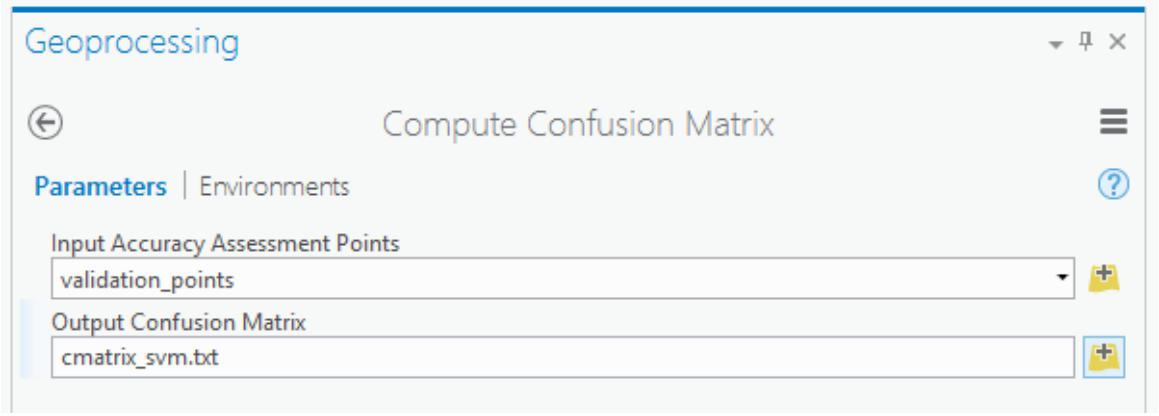
- Navigate to Spatial Analyst Tools followed by the Extraction subtoolbox. Click on the **Extract Multi Values to Points Tool**.
- Set the Input Point Features to the **validation_points** layer.
- Set the Input Raster to **svm_result.img** and name the Output Field "Classified" (without the quotes).

Note: It is important that you name the field "Classified," as this is the field name that the **Compute Confusion Matrix Tool** will look for. Also, the "ground truth" or validation data field must be named "GrndTruth." This field was provided for you.



You can now generate a confusion matrix from the data.

- Navigate to Spatial Analyst Tools followed by the Segmentation and Classification subtoolbox. Click on the **Compute Confusion Matrix Tool**.
- Set the Input Accuracy Assessment Points to the **validation_points** layer.
- Name the confusion matrix **cmatrix_svm.txt** and save it to a location of your choosing. Make sure to include the file extension.
- Click Run to execute the tool. The table should be added to the Contents Pane automatically.



Use the confusion matrix to answer the following questions. Remember that the reference data defines the columns and the classification data defines the rows. Here are how to interpret the codes.

C_0 = Water

C_1 = Developed

C_2 = Barren

C_3 = Forest

C_4 = Herbaceous

Deliverable 1. Make a copy of the SVM confusion matrix to include with your answers. We will need this to grade your answers. (10 Points)

Question 1. What is the reported overall accuracy for the SVM classification? (2 Points)

Question 2. What is the reported Kappa statistic for the SVM classification? (2 Points)

Question 3. What is the producer's accuracy of the forest class for the SVM classification? (2 Points)

Question 4. What is the user's accuracy of the developed class for the SVM classification? (2 Points)

Question 5. For the SVM classification, provide a brief discussion of the sources of error. What classes were most difficult to map? What classes were most commonly confused with each other? (5 Points)

You will now produce a confusion matrix for the RF result.

- First, you will need to delete the "Classified" field from the **validation_points** layer. You can do this by right-clicking on the layer in the Contents Pane then selecting Attribute Table to open the attribute table. You will then need to right-click on the "Classified" field then select Delete to remove it.
- Run the **Extract Multi Values to Points Tool** again. This time use the **rf_result.img** layer. Make sure to name the field "Classified."
- Run the **Compute Confusion Matrix Tool** again. Name the output **cmatrix_rf.txt**. Remember to include the file extension.

Use the resulting confusion matrix to answer the following questions.

Deliverable 2. Make a copy of the RF confusion matrix to include with your answers. We will need this to grade your answers. (10 Points)

Question 6. What is the reported overall accuracy for the RF classification? (2 Points)

Question 7. What is the reported Kappa statistic for the RF classification? (2 Points)

Question 8. What is the producer's accuracy of the forest class for the RF classification? (2 Points)

Question 9. What is the user's accuracy of the developed class for the RF classification? (2 Points)

Question 10. For the RF classification, provide a brief discussion of the sources of error. What classes were most difficult to map? What classes were most commonly confused with each other? (5 Points)

Compare the results to answer the following questions.

Question 11. Based on the accuracy assessment results, which algorithm provided the best result for this classification problem? Explain your reasoning. (5 Points)

Question 12. What are some reasons that we cannot separate these land cover types with 100% accuracy? (5 Points)

Closing Comments

A few comments before we end this exercise. First, land cover classification can be difficult, and it is not always easy to map or separate land cover categories with high accuracy. You will need to consider what accuracy is required for your project and whether or not your result is adequate. Note that there are ways to potentially improve the result, such as incorporating additional data into the analysis, creating more training data, or maybe redefining the classes of interest. It is also possible to generalize or smooth the results, for example you could apply a majority filter or sieving. You could also attempt to include an image segmentation, investigate other algorithms, or try an unsupervised classification method. You could also choose to manually clean up the results using a manual digitizing process.

In short, you may find that these processes require some trial and error and iterations to obtain the desired result.

END OF EXERCISE