GEOG357: Remote sensing Lab 2
Unsupervised classification

General Goal:

This lab involves identifying land cover types by using algorithms to group pixels with similar multi-spectral (MS) Digital Numbers (DN) / spectral signatures, separated by MS natural breaks.

Open your copy of **pg20july2023.pix** from lab1, and display SWIR-NIR-Red bands in RGB
.. It will become a tad annoying how often you have to remind PCI you are using this file :)

1. Classification Introduction

Classification involves identifying a set of **unique** spectral signatures (DNs) for a feature type in the image scene. We want our classifier to take advantage of maximum information content available in the image, so we choose relatively uncorrelated bands to run the classification e.g. 654 and NOT 321

BAND CORRELATION
You should notice the following for Landsat (and other sensors) data:

1. VIS (visual) bands show cultural features and in detail
2. NIR shows the land/water boundary sharply. Water appears black
3. SWIR shows dryness or inverse of moisture (high DN = low moisture)

Ensure you are zoomed out to show the full extent – symbol to left of zoom options

To view correlation between any selected band pair, Select **layer -> scatterplot**
Right-click on filename -> scatterplot does the same thing

Look at the correlation (r) between the following bands below. The more they are correlated, the closer the graph shows a straight line (then one band is nearly a linear function of the other).

try these and check you get these values

Green v Red r = 0.98

Red v SWIR1 r = 0.66

SWIR1 v SWIR2 (6 and 7) r= 0.94

you should see a high 'r' between visible bands, and less with IR (except 6 v 7)
r values correspond to the image data displayed – values **will vary if you are zoomed in**

Q1a: what is the r value between the Red and NIR bands ?
1b: what is the r value between NIR and SWIR1

2. Unsupervised Classification algorithms

You will now see the difference between a band (recorded by the sensor) and a channel – which can store a band, but also any other data generated by the user. We will add empty channels to store classification results, 16-bit to match the band format.

You are ready to classify:

- Analysis -> Image Classification -> Unsupervised
- Select the file to use = your 2025 .pix file
- Select New session
- Click on add layer and add five (6) more 16-bit unsigned (channels)   ->add   ->close

- We then need to specify the display, input bands and output CHANNEL
- Select TM bands 6,5,4 (R, G, B) as input channels (and tick in input column)
- The colour ticks in the first 3 columns are simply what is displayed
- Select the first empty channel (8) as output *
- NOTE *: the designated output channel will be overwritten, so if you specify a band number (1-7) you will LOSE the band data - **always** double check output channel number - it should be an empty or not-needed channel. - but it is recoverable if you mess up.
- click OK

a. K-Means

In the new Classify window, select these options:

- Algorithm: **K-Means**
- Max class: 6
- Max Iteration: 6
- Min Threshold: leave as is
- Max Sample Size: leave as is (with a bigger scene, we might specify a subset)
- Show Report button: 'on' (depressed)
- OK

.. this shows the report for the 6 clusters, # of pixels in each, and average DNs for bands 4,5,6
.. image displays in 'PC' (pseudocolour), the new DNs are 1-6 (one number for each cluster). You can see which colour corresponds to which cluster number by querying some locations - you may need to highlight the new channel on left if the DN display is still on the RGB.

This report could be saved if needed for future work (but is not needed here)

Can you visually identify the clusters as classes approximately? (do a quick check)

tick the PC (classification layer) off and on to view the classification then the composite.

View the report also - it gives the number of pixels in each class, and the mean DNs

It should be poor as there are too few classes - and iterations

Right-click the **Classification MetaLayer** and select **run classification**

Once more, Right-click the **Classification MetaLayer** and select **run classification**

This time, try the default **16 classes/clusters and iterations** then run classification again. This may seem better – click classification Metalayer off and on to compare with the image. Still some empty clusters, but we'll keep it (in channel 8).
Save the classification report file: use the save button and call it **kmeans.txt**
Note the other options when you right-click the Metalayer e.g. accuracy assessment – see Lab 3.

Now you could match the cluster numbers with land cover types - click each colour on the output display to see which cluster each is, and try to visually match these against the list below. Change some colours for easier reading if you wish - at the very least make the water class blue, and coniferous forest dark green: expand 'output' checkmark and double-click on the legend colours to edit them. [or right-click on classification metacolour – edit PCT]

Note that some shadows get grouped with water – the north facing slopes of the esker ridges north of the Nechako River; next week we will see how this might be corrected.

The classifiers are often challenged to separate water from dark coniferous forest.

Q2. Using the image DNs: which of the three bands shows the biggest difference between these two features and explain why this is so, based on visible, NIR (vegetation) and SWIR (dryness) – you can review the report and/or query some water and forest locations for their DN values.

Cluster -> class identification

Hopefully you can recognise some/most cluster classes from the underlying image and features. Some may be 'hybrid' or mixed, e.g. grassy areas could be regenerating cut areas, city parks, etc.. the bottom 3 clusters above chip piles may intermix. You could apply class names to the clusters, though this only works temporarily on the display; later we'll see a more permanent method.

See which cluster numbers correspond to:

- Water
- Conifers
- Deciduous trees
- Mixed forest
- Industrial – urban
- Residential - urban
- Grass (e.g. soccer fields)
- Agricultural fields
- Cutblocks
- Chip piles (Canfor Mill)

Simplified interpretation-bands654:

- Dark-green: coniferous.
- Light-green: deciduous.
- Pink: arable / agricultural.
    - Yellow: Soccer / golf ?
- Dark-purple: industrial.
- Light-purple: residential.
- Black: sediment-free water.
- Blue: sediment-laden water.

White: brightest pixels

b. Fuzzy k-means

Right-click **Classification MetaLayer** and now select **configure**; change output channel to 9
Be sure to check **fuzzy k-means** for the algorithm (as k-means is the starter default)
Select default 16 / 16 in the clusters / iterations - how does this compare with the K-means?

If you need to re-display the previous k-means classification or any classification in the lab, select **layer->add->pseudocolour** select the k-means channel and finish

It's tough to compare exactly as the colour palettes won't match unless you make them do so -
This one may also mix water and some shadows on the north side of eskers.
Save the report (save as) : fuzzy.txt

Note that it ran slightly slower because the clustering process includes the spread of pixels in each cluster and probability functions, not just which seed point / mean each pixel is closest.

c. ISODATA classifier

Right-click **Classification MetaLayer** and again select configure; change output channel to 10

Now select **Isodata** as the algorithm. Use the 16 default options - Again view the result, and the classification report - almost all clusters should have a fair number of pixels and compare with the previous classification. Save the report: isodata.txt

In order to display both or all 3 classifications, select from the layer dropdown as before:

layer->add->pseudocolour-> next-> select your file and desired class channel -> finish

– the software should have labelled the classification channels suitably. Click the fuzzy / isodata layers off / on to compare with the underlying colour composite … it may be hard to objectively tell, but the isodata slightly retains homogenous areas within the same class – not very scientific, but hey it's only a lab ..

Review subjectively how well each one separates water from forest and coniferous from deciduous and mixed, as well as details in the bowl and residential versus industrial; also did they preserve the airport runways within the same cluster ?

The Isodata does seem to do better in the bowl, but to be sure we'd need to match the two colour palettes and perform an accuracy assessment (but not today – Lab 3 and maybe your project).

Note down your cluster numbers and their most likely corresponding cover type: water, conifers, fields etc.. it's likely some land cover types may include more than one cluster number.

3. Aggregation

Some clusters are too similar to separate and thus desirable to merge/aggregate them: either their spectral or spatial qualities overlap or there is uncertainty e.g. agricultural fields may have 2-3 clusters due to different cover and maybe you even have two clusters for water ? You should try the aggregation tool for at least one pair of clusters: note your cluster numbers and select:

Analysis -> Image classification-> post-classification Analysis -> Aggregation

Select your file (yet again !)

- Input channel: your isodata classification (10)
- Output channel: an empty channel (11)
- Click on New

You can approve the new cluster number to be assigned and its colour, or edit either

- Highlight the input cluster number -> Add
- Then the other cluster number (repeat if more than 2 clusters)
- Apply to output channel
- In order to se that it worked, reload the 'aggregated channel'
- Layer-> Add -> Pseudocolour .. and pick your new channel (11)

.. the two clusters should now be merged.

Now use the 'class labelling' option – this is where you would change the cluster numbers to cover type names – water, fields etc..]

Analysis -> Image classification-> post-classification Analysis -> Class labelling
Go ahead and edit the cluster/class names to match cover types as best you can, Save.

Q3a. Comment on which cluster (excluding water) seems to classify with the minimum confusion and why (flicker between the classification and colour composite) ?
3b. In contrast, which cluster was most challenging to identify and why ?

4. Sieve the classification

The classifications will always have isolated pixels, which are mostly undesirable 'noise'
These can be reduced using the SIEVE algorithm
We'll use the Isodata classification just aggregated but first note - what is the cluster number(s) for water .. this is so you can retain small lakes, as they may be meaningful.
Tools -> Algorithm Librarian
click the 'Find' button and type in **sieve** and then 'Open'
The parameters will be these:
input = ## (= the channel number for your aggregated isodata **classification** (11 ?)
output port should be viewer - PCT (at first)
Let's do a small sieve first– 1 acre which is ~4 pixels (1 acre is about 64 x 64 m)

Input params:
Polygon size threshold= 4
Connectedness - can be 4 or 8 (pixels can be joined vertically/horizontally or also by diagonals)
exclude values list = ##   (where ## is the class number for water – to retain small lakes)
select log tab and run ...

View the result, compared with the unsieved classification - click the sieve on and off and you should see the before and after; now increase polygon size threshold to 11 (= 1 hectare), run.

Try it again with connectedness = 8 – any difference ? maybe a few more polygons preserved – are they useful or just noise ? Go with either one in the next step.
Which is best ? urban areas are tough given the heterogeneity. Go with threshold =11

- run it again, specify your filename under the 'files' tab in sieve, as output at the bottom - make sure you select **pg20july2023.pix**. check the threshold is 11 ->run. This is now saved as a new channel in your file - check which one it is, by viewing your list of channels in the main tab on the left of the image, under 'Files' and expand the rasters list. It should be listed as a sieve. The Sieve can look a bit like a 'staircase' art project, we'll address this in Lab 5.

NOTE (again): NEVER accept the default filename **'Untitled.pix'**. You must have all your data layers in the same file, and you should not EVER have a file named Untitled.pix in your folders.

Examine the histogram for this sieved classification (right-click on this channel in the contents -> histogram and click on the histogram to enlarge … note that the DN values are discrete
(versus continuous for bands), and also nominal data as they represent class types. So most of the stats are not relevant e.g. class/cluster 6 is not twice cluster 3.

Q4a. which class type has the most pixels = covers the largest area ?
4b. which class type has the second most pixels ?

5. Read/skim through this summary on Unsupervised classification:
https://gisgeography.com/unsupervised-classification-remote-sensing

This process is normally referred to as 'semi-automated' rather than automated.
Q5a. Why/where is manual input required ?
Q5b. They say that 'ArcGIS recommends the number of clusters should be at least 10 times larger than the number of bands in your image' (= 30 with 3 bands). I'm not sure I agree, but their image example does conflict with the one we are using and my notes. Why might this be ?

Lab assignment 2 (5%)

Submit answers either as plain text or PDF file via Moodle by Friday 5pm - no need to include the questions, you can send in note form. PLEASE – **no word docs** (I have to download/save)
Next week's lab: Supervised Classification