Exercise 7 - Solution

May 21, 2025

LTE, EPFL, Switzerland

Question 1

Briefly discuss the spectral signature of each of the 6 land surface covers

- Forest: Low reflectance in the visible spectrum as these wavelengths are mostly absorbed. Higher reflectance in the near infrared (from bands 6-8).
- Grassland: similar pattern as forest but generally higher reflectance
- Bare soil: signature similar to grassland, higher values in SWIR bands. Bare soil is also difficult to identify on the image and can be more heterogeneous without a very clear spectral signature
- Water: very low reflectance, highest in blue and green visible channels; at 90° angle water does not reflect much radiation
- Cities: general reflectance slightly higher than grassland, highest values in highest bands (SWIR). Cities are very heterogeneous so we do not have a distinct spectral signature
- Clouds: highest reflectance, high reflectance in the visible bands, slowly decreasing towards bands with longer wavelength

For all classes, note that the spectral signature in the graph can vary depending on which point you selected. However, the consistent properties of the spectral signatures of different land covers allows the spectra to be used for classification.

Question 2

How many groups of similar bands do you see (based on online documentation, etc.)? To which regions of the electromagnetic spectrum do these correspond?

Three groups of bands can be seen in the online documentation. The groups correspond

to visible, near infrared and short wave infrared in the EM spectrum.
Bands 1-5: visible spectrum
Bands 6-9: far end of visible red, near infra red
Bands 11 & 12: short wave infrared

Question 3

What can you see? Are there any high correlations between the bands? How does this compare to your answer to question 2?

Bands 1-5: correlation larger than 95%
Bands 6-8: correlation larger than 98%
Bands 11 & 12: 96 % correlation

Apart from bands 9 and 11, all bands show very high correlations to their neighbouring bands. This is expected as the shift in central wavelength is in small increments up to band 9. In the visible representation, the high correlation within each group can be seen as dark squares along the diagonal.

Question 4

Which band contributes the most to component 1? How about the other bands?

Band 1 has the largest contribution to principal component 1 (PC1). However, up until band 9, all bands contribute a similar amount.

Question 5

Which proportion of the variance is represented by the first three components individually? And by the three components together?

The proportion of the variance represented by a given principal component is given by the eigenvalue corresponding to that principal component divided by the sum of eigenvalues.

- PC1 - 90.2 %
- PC2 - 5.8 %
- PC3 - 2.2 %
- PC1-3 - 98.2 %

Question 6

Consider the importance of the PCs and the contribution of the bands to the PCs. If you had to choose only 3 bands, which ones would you choose? Explain your choice.

As the first PC already contains more than 90 % of the variance, we focus on the bands in PC1. From the eigenvectors, bands 1, 2, and 9 have the highest contribution to PC1.

However, as mentioned in question 5, many bands have a similar importance here.

Question 7

What can you say about the reconstructed images? What would happen to the error if different components were used for the reconstruction?

When the first three principal components (0,1,2) are used for the reconstruction, the reconstruction is a good representation of the original image. This is because the first three principal components account for almost all the variance in the image. If principal components of a higher number were used instead of the first three, the error in the reconstruction would be larger.

Question 8

Is 6 an appropriate number of clusters to properly distinguish all the land cover types? Identify the labels that would correspond to each cluster.

We chose to use 6 clusters as there were 6 land classes defined in part 2. However, we see in the clustering that the land classes are not clearly separated into the clusters. Cluster 0 corresponds well with the water class. Clouds are mostly covered by clusters 5, 1 and 4. Highly vegetated areas mostly fall in cluster 4 and less vegetated areas are mostly in cluster 2.

Using more clusters could help with the classification. In this case, multiple clusters would be assigned to each land class.

Question 9

Based on this plot and on your visual inspection (of the training areas, of the true color image), which classes are the least and the most trustworthy and why? Can you explain the differences between the classes that you can see on this plot.

The bare soil class is least trustworthy, as it has the largest standard deviation and a very large overlap with the spectral signatures of most other classes. The urban class also has a very widespread spectral signature. The narrowest signature can be seen for water, which thus has a high likelihood of separating well from the other classes.

The spectral signature for water is very well defined as water is homogeneous. In contrast, bare soil can have different properties based on location and moisture levels, leading to a large range of band intensities at each wavelength.

Question 10

Investigate the mis-classifications. Are there systematic issues with some land cover types? What are possible sources of these errors?

One systematic issue is that the area assigned to bare soil is larger for all 4 methods than for the ground truth. This is likely due to the training area assigned to bare soil, which includes clouds and other land cover types. This falsifies its spectral signature and impacts the training.

Question 11

Compare the results from the different classifiers. Which classifier would you choose?

Based on the mean accuracy indicator, it appears that the random forest classifier performs the best with a value of 0.98. However, KNN also performs well. Visually, the classifications from these two classifiers look very similar.

Question 12

Compare this to your experiences with the unsupervised classification in Australia. What are the advantages and disadvantages?

An advantage of the unsupervised classification is that no training area must be defined in advance. However, as we saw, this can lead to clusters which do not correspond to a particular land type class. The supervised algorithms perform better at identifying our chosen land type classes, but still have some misclassifications.

Question 13

How do you interpret this figure?

For water, the probability of classifying as water is high in the area we expect - where there is a lake. This shows that the water can be correctly classified in a high proportion of cases. The map for water is also quite binary - either water is detected or it is not. This is due to the distinct spectral signature of water which makes it easier to classify. For the other classes, the probability covers a greater range of values between 0 and 1 across the image, showing that misclassifications are more likely than for water.