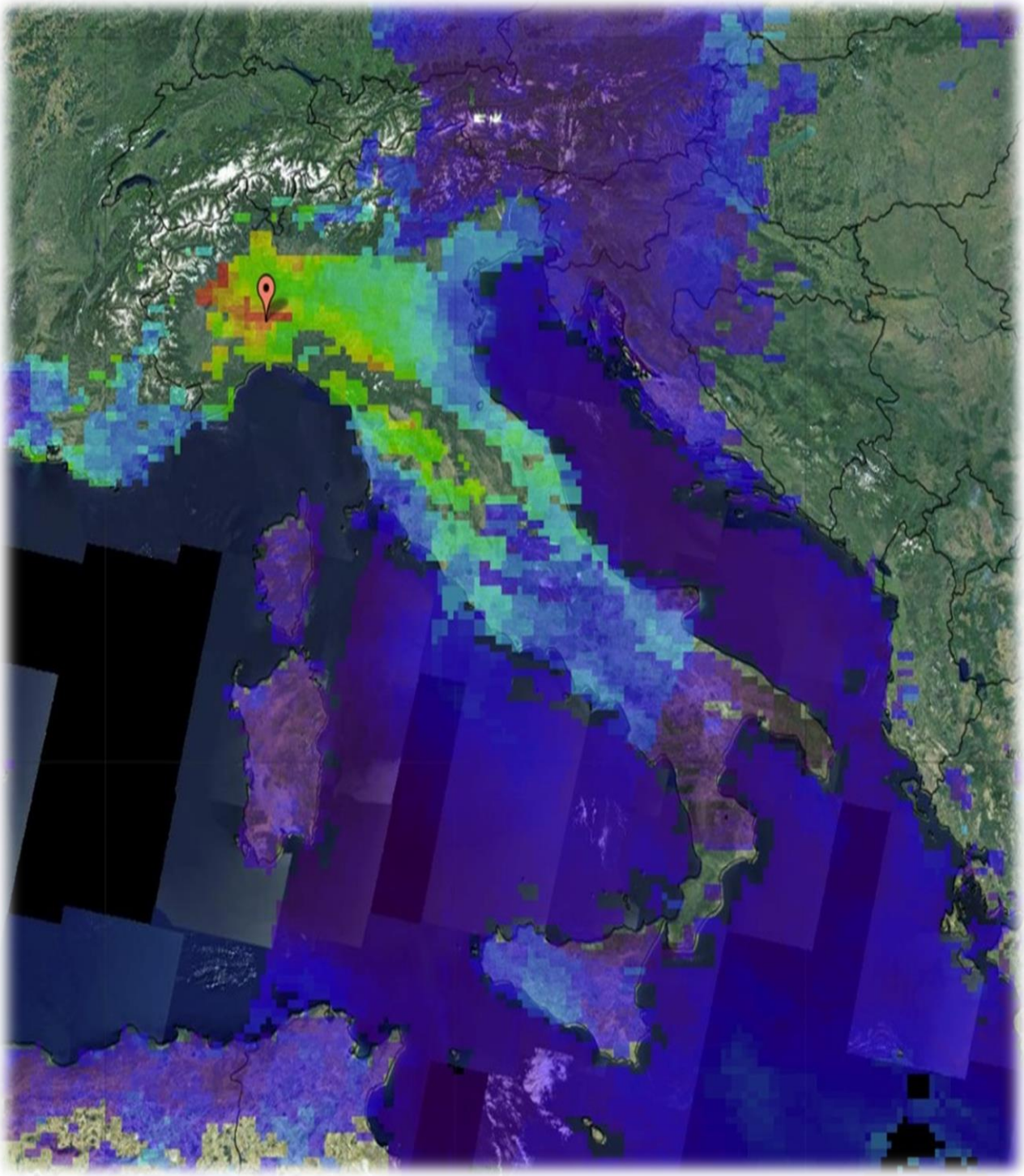


EXERCISE - 05

EARTH OBSERVATION DATA ANALYSIS

Vikranth Ale	1873995

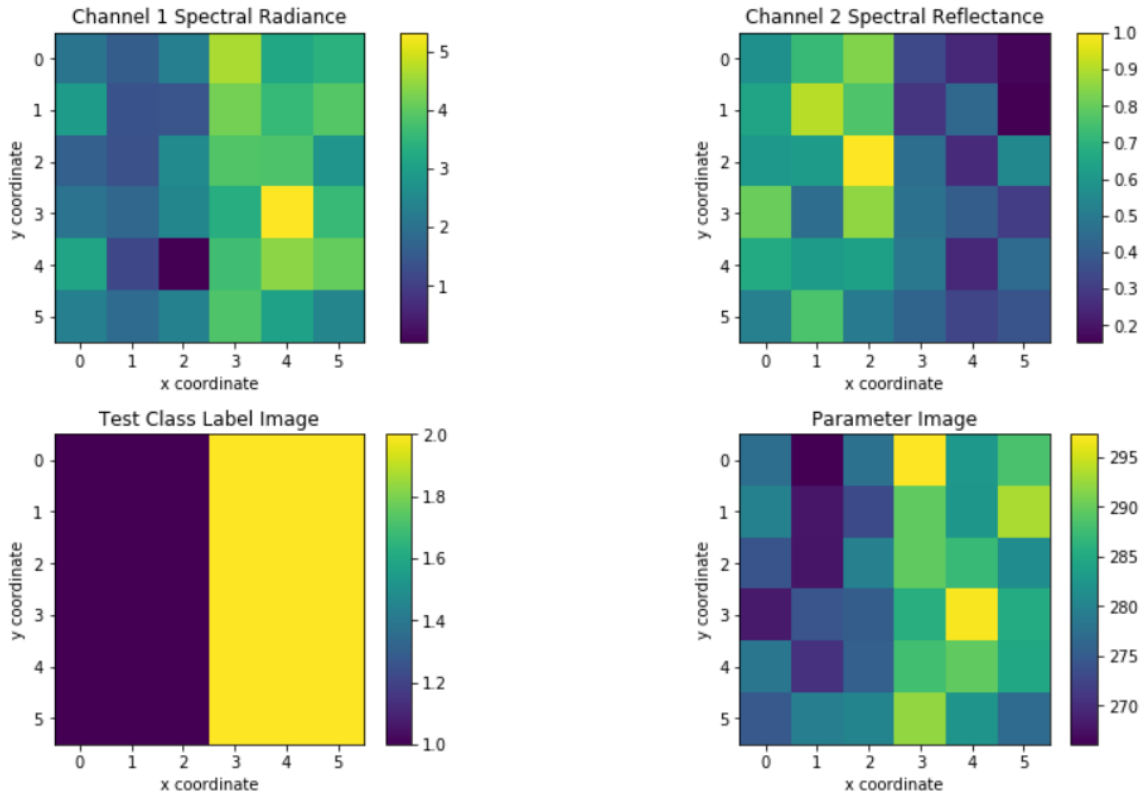
EARTH OBSERVATION: IMAGE DATA PROCESSING & RETRIEVAL

INTRODUCTION

For this exercise 5, I have worked on data received from a satellite carrying thermal infrared(TIR) and visible(VIS) spectroradiometer that is place on Low-Earth-Orbit(LEO) and having two channels. Channel 1 is measures spectral radiance and channel 2 measures spectral reflectance from a scene of 36 x 36 km^2 at 1-km spatial resolution.

Apart from channels, the dataset also contains images of geocoded in-situ 2 class labels with mid to high altitude particle clouds and sea water surface along with surface temperature expressed in kelvins.

VISUALIZATION OF INPUT DATA



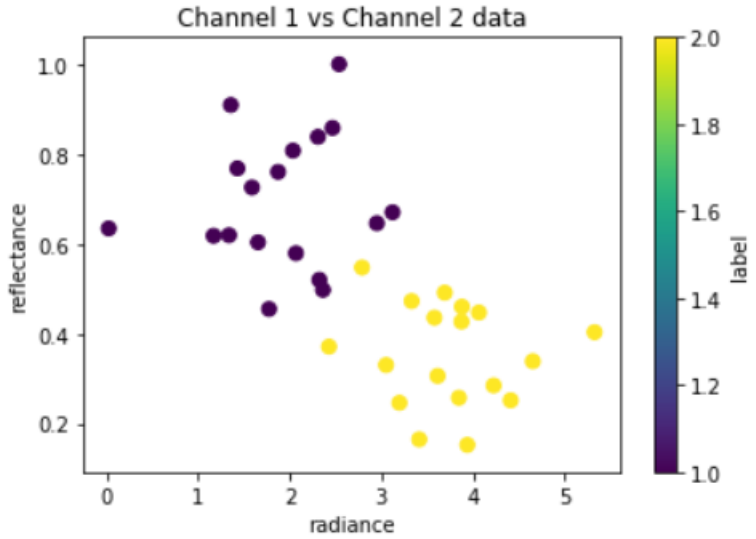
DATASET CONSTRUCTED FROM THE ABOVE IMAGE DATA FOR OUR ANALYSIS:

	radiance	reflectance	label	surface_temp
0	2.068526	0.579204	cloud cover	277.39425
1	1.588779	0.725697	cloud cover	266.11364
2	2.308740	0.838262	cloud cover	277.84215
3	4.651697	0.338762	sea water	297.25935
4	3.197330	0.246227	sea water	282.75330
5	3.413580	0.164858	sea water	288.30717
6	2.951153	0.645951	cloud cover	279.79463
7	1.358214	0.909161	cloud cover	267.95825
8	1.429915	0.768129	cloud cover	273.07966
9	4.225014	0.284818	sea water	289.70766

In the above table, the first column represents the channel 1 spectral radiance data, second column represents channel 2 spectral reflectance data, third column represents if the image represents cloud cover or sea water surface and the last column represents the surface temperature in Kelvins.

QUESTIONS:

1. **Plot the acquired spectral reflectance vs and spectral radiance and possibly superimpose the class label to each observation, explaining why you expect these classes can represent particle cloud and sea surface radiative signature**



We know that for clouds, reflectance is more and radiance is less For seawater, reflectance is less and radiance is more

The same can be observed after analyzing both channels and their surface temperature, Sea water has generally more surface temperature due to absorption of solar radiation and warms up. So, it is justified that they are classified correctly due to their radiative signature.

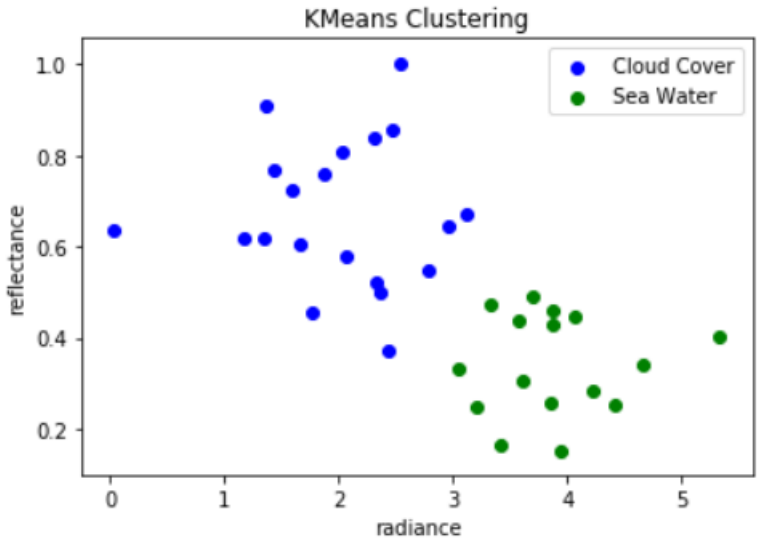
2. **Perform a Lloyd's k-means unsupervised classification and, by using the in-situ class labels, compute the confusion (or contingency) matrix (CM).**

I have constructed the dataset to perform k-means unsupervised classification using the channel data along with surface temperature and classified them in to clusters to identify if the algorithm can correctly classify the data as compared to the given data.

The dataset along with their classified clusters, 1 represents cloud cover and 2 represents sea water.

	radiance	reflectance	surface_temp	cluster
0	2.068526	0.579204	277.39425	1
1	1.588779	0.725697	266.11364	1
2	2.308740	0.838262	277.84215	1
3	4.651697	0.338762	297.25935	2
4	3.197330	0.246227	282.75330	2
5	3.413580	0.164858	288.30717	2
6	2.951153	0.645951	279.79463	1
7	1.358214	0.909161	267.95825	1
8	1.429915	0.768129	273.07966	1
9	4.225014	0.284818	289.70766	2

VISUALIZING THE ABOVE CLASSIFIED CLUSTERS.



COMPUTING THE CONFUSION MATRIX

Confusion matrix for the above data can be calculated as below.

```
array([[16,  2],
       [ 2, 16]], dtype=int64)
```

From the above confusion matrix, we have TP = 16, FP = 16, FN = 2, TN = 2

Our model has True positive = 18 with accuracy score of 94 percent.

Also, for above data the classification report can be calculated as below for each class which represents the precision and recall rate.

	precision	recall	f1-score	support
1	0.89	0.89	0.89	18
2	0.89	0.89	0.89	18
accuracy			0.89	36
macro avg	0.89	0.89	0.89	36
weighted avg	0.89	0.89	0.89	36

3. **From the confusion matrix, compute the probability of detection (POD, sensitivity or true positive), false alarm rate (FAR, false negative) and Cohen's index k.**

   The computations asked in the question can be calculated using the below formulae.

$$SENSITIVITY/TPR = \frac{TP}{TP + FN}$$

$$DETECTION\ RATE\ = \frac{TP}{TP + FP + FN + TN}$$

$$FALSE\ NEGATIVE = \frac{FN}{FN + TP}$$

$$FALSE\ DISCOVERY\ RATE\ = \frac{FP}{FP + TP}$$

   Substituting the values obtained from confusion matrix in the above formulae.

The results are shown in the below table.

S.No	Computation	Value
1	Sensitivity	0.9
2	Detection Rate	0.5
3	False Negative Rate	0.1
4	False Discovery Rate	0.1

COHEN'S KAPPA COEFFICIENT (K)

Manual calculation of Cohen's kappa coefficient score using the below formulae.

$$pYES = \frac{TP + FP}{TP + FP + FN + TN} * \frac{TP + FN}{TP + FP + FN + TN}$$

$$pNO = \frac{FN + TN}{TP + FP + FN + TN} * \frac{FP + TN}{TP + FP + FN + TN}$$

$$po = \frac{TP + TN}{TP + FP + FN + TN}$$

$$pe = pYES + pNO$$

Using the above formulae, we can calculate the Cohen coefficient (K) by

$$k = \frac{po - pe}{1 - pe}$$

Substituting TP = 18, FP=0, FN=2, TN=16 in the above formulae, the value of cohen coefficient is **k = 0.777777777.**

So, based on above Cohen coefficient we can say that our results are almost in perfect agreement.

Our calculation of Cohen coefficient using scikit learn library also yielded same result.

4. Determine how much variance is explained by the first principal component of the acquired channels

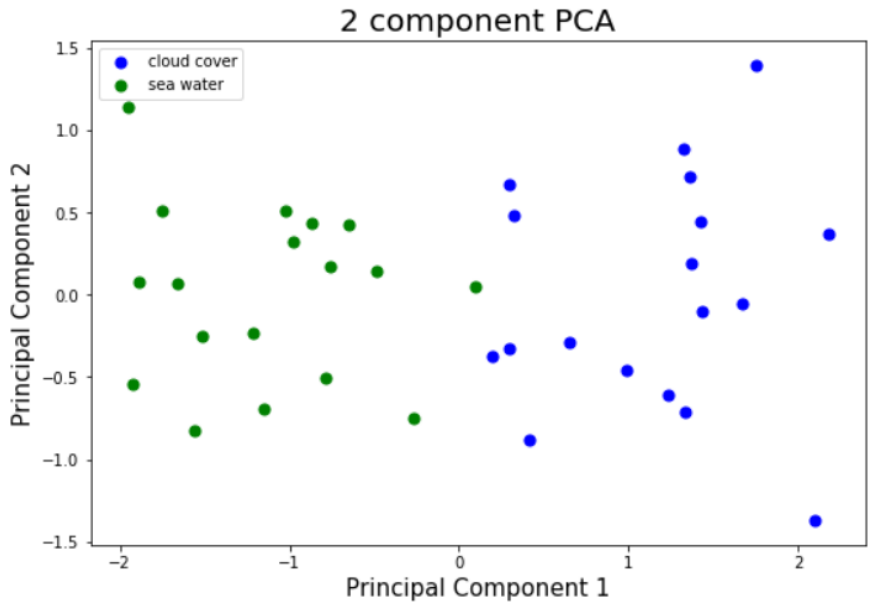
To perform Principal component analysis, I have used the data from both channels and standardized it to get better results.

After the data is ready, I have fed this data into PCA model and obtained the following result.

	PCA 1	PCA 2	label
0	0.647820	-0.288520	cloud cover
1	1.430294	-0.099794	cloud cover
2	1.357908	0.718862	cloud cover
3	-1.747703	0.512950	sea water
4	-1.154464	-0.693770	sea water
5	-1.558009	-0.829678	sea water
6	0.322896	0.478915	cloud cover
7	2.181125	0.365685	cloud cover
8	1.669255	-0.057445	cloud cover
9	-1.662484	0.070097	sea water

In the first row, the value for PCA 1 = 0.64 and PCA 2 = -0.288 which means to classify the class into label, the magnitude and direction on PCA 1 strongly tells that it is responsible to classify the corresponding variable which in this case is cloud cover.

VISUALIZATION OF PRINCIPAL COMPONENT ANALYSIS




```
Explained variance of PCA is given as [0.81709233 0.18290767]
```

From the above result, we can say that the First principal component of acquired channels comprises about 81.7 % variance and the second principal component comprise about 18.2 % variance.

Together both components attribute to 99.9 % of information.

5. **By using the first principal component, re-perform a Lloyd's k-means unsupervised classification and then compute the confusion matrix (CM) and Cohen's index.**

Applying K-mean Unsupervised classification using the first Principal component

Our dataset now looks like this

	radiance	reflectance	surface_temp	PCA1
0	2.068526	0.579204	277.39425	0.647820
1	1.588779	0.725697	266.11364	1.430294
2	2.308740	0.838262	277.84215	1.357908
3	4.651697	0.338762	297.25935	-1.747703
4	3.197330	0.246227	282.75330	-1.154464

After performing k-means our dataset is transformed into like this as per their clusters.

	radiance	reflectance	surface_temp	PCA1	cluster
0	2.068526	0.579204	277.39425	0.647820	1
1	1.588779	0.725697	266.11364	1.430294	1
2	2.308740	0.838262	277.84215	1.357908	1
3	4.651697	0.338762	297.25935	-1.747703	2
4	3.197330	0.246227	282.75330	-1.154464	2
5	3.413580	0.164858	288.30717	-1.558009	2
6	2.951153	0.645951	279.79463	0.322896	1
7	1.358214	0.909161	267.95825	2.181125	1
8	1.429915	0.768129	273.07966	1.669255	1
9	4.225014	0.284818	289.70766	-1.662484	2

CONFUSION MATRIX

for the above data, confusion matrix calculated from true and predicted classes is as follows.

```
array([[16,  2],
       [ 2, 16]], dtype=int64)
```

We infer here TP = 16, FP = 2, FN = 2, TN = 16

COHEN COEFFICIENT(K)

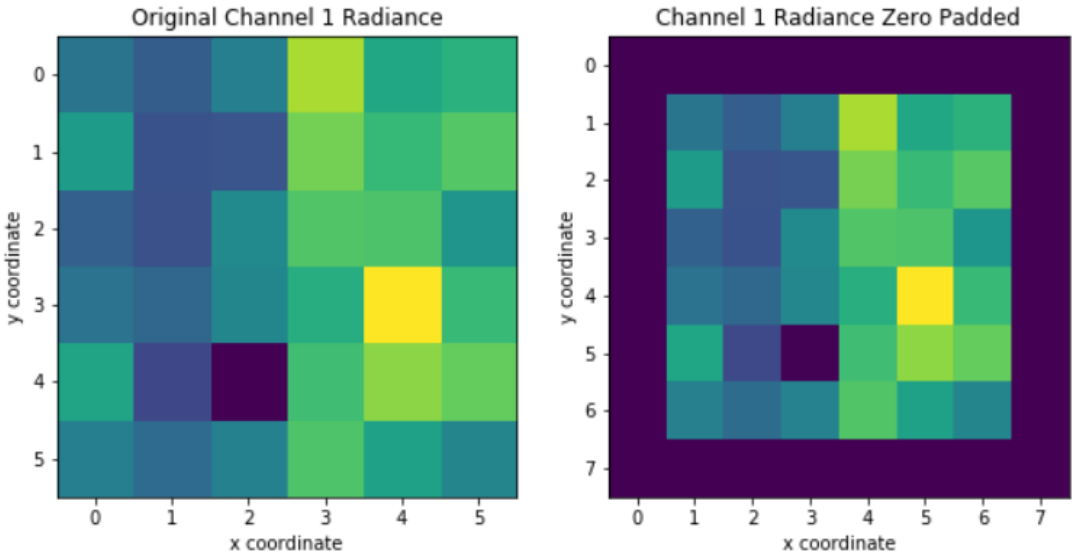
Using Cohen coefficient formulae from previous question we get **k = 0.7777777777**

6. **Apply a (running) average 3x3 numerical filter to both channel-1 and channel-2 images using a zero-padding technique and then display smoothed images comparing them with the original ones.**

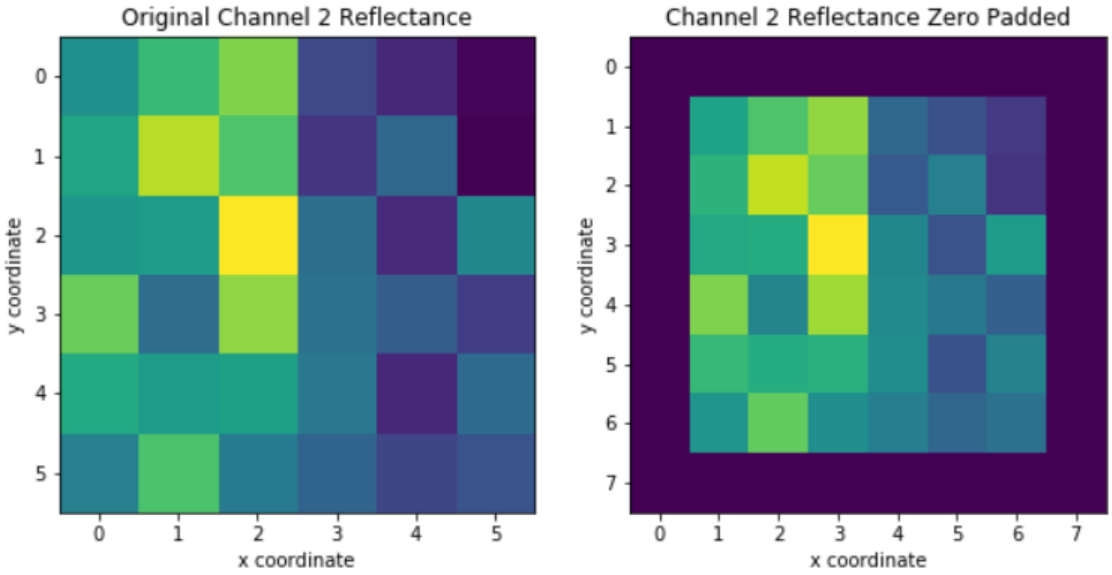
To solve this question, I have zero padded the input channels to perform smoothing operation to the orignal channel images.

Lets first visualize the original channel image and zero padded image side by side.

CHANNEL 1



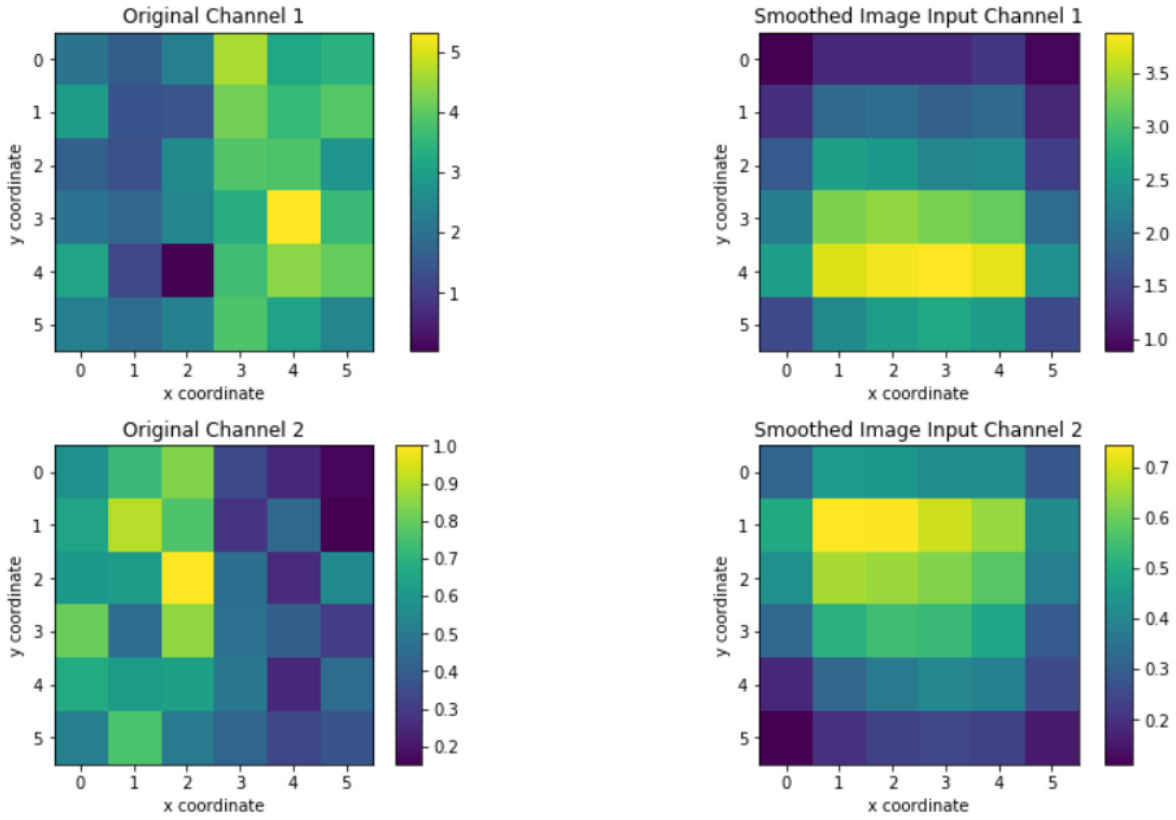
CHANNEL 2



I have applied average filtering technique on the zero padded image with filter size 3 x 3 in which the filter is slided over the image and by convolving the image with a normalized box filter. It simply takes the average of all the pixels under kernel area and replaces the central element with this average.

The obtained central values are used to construct the smoothened images from zero padded channel images.

ORIGINAL IMAGES VS SMOOTHENED IMAGES



We can observe that by performing average filtering the original image has been distorted/blurred a bit with the loss of few details.

7. **By using the smoothed channel images, re-perform a Lloyd's k-means unsupervised classification and then compute the confusion matrix (CM) and Cohen's index (k).**

Buiding the dataset for the smoothed image channels

	smooth_ch1_image1	smooth_ch2_image2
0	0.885186	0.317779
1	1.217820	0.453661
2	1.234863	0.448979
3	1.232959	0.419341
4	1.366716	0.425715

Performing k-mean unsupervised classification on the above data.

	smooth_ch1_image1	smooth_ch2_image2	cluster
0	0.885186	0.317779	1
1	1.217820	0.453661	1
2	1.234863	0.448979	1
3	1.232959	0.419341	1
4	1.366716	0.425715	1
5	0.943307	0.285408	1
6	1.300592	0.496267	1
7	1.915319	0.743260	1
8	1.949930	0.740783	1
9	1.792028	0.696289	1

CONFUSION MATRIX

After performing k-mean and comparing true values with their predicated values, the confusion matrix obtained is as follows.

```
array([[8,  10],
       [7,  11]], dtype=int64)
```

```
From the above matrix, we can infer that

       TP = 11, FP = 10, FN = 7, TN = 8
```

COHEN KAPPA COEFFICIENT(K)

Using the library of scikit learn, the value of cohen kappa coefficient is given as

K = 0.05555555555558

With cohen's coefficient less than zero, it means that the results are in slight agreement with that of results of original channel image data.

8. By using the collocated channel and parameter data, develop a statistical retrieval algorithm (e.g., linear regression) to estimate parameter value from spectral radiance (channel 1) and provide error bias (mean) and standard deviation in kelvins.

To answer this question, I would like to use a linear regression model on the original dataset of the channel image.

Our dataset looks like this.

	radiance	reflectance	label	surface_temp
0	2.068526	0.579204	1	277.39425
1	1.588779	0.725697	1	266.11364
2	2.308740	0.838262	1	277.84215
3	4.651697	0.338762	2	297.25935
4	3.197330	0.246227	2	282.75330

To build a linear regression model I would like to consider only the data from channel 1 & 2 a nd estimate the parameter surface temperature, I have excluded the class labels.

I split the total dataset into 70 % training data and 30 % test data for validation, and trained t he model using training data and validated the model with test data.

This Linear regression model performs decently with an **accuracy score of 89.8 %**

ROOT MEAN-SQUARED ERROR

It is generally calculated using the formula

$$Mean\ Squared\ Error = \frac{1}{N}(yTrue - yPred)^2$$

Where N is the total number of observations.

I have manually written the function for calculating MSE, variance and standard deviation.

For our data, the root mean-squared error is **MSE = 6.58254219629515**
Standard Deviation is given as **Std = 2.0823098461949536**

The overall results are satisfactory but I'm afraid the results can be biased as the dataset is v ery small, if we have more data we can improve the accuracy of the model and minimise the error.