IMPACT OF FEATURE SELECTION ON SEGMENTATION OF HYPERSPECTRAL IMAGES OF ATMOSPHERIC CLOUDS

Ben T. Nguyen

Auburn University at Montgomery

2023

IMPACT OF FEATURE SELECTION ON SEGMENTATION OF HYPERSPECTRAL IMAGES OF ATMOSPHERIC CLOUDS

by

Ben T. Nguyen

A thesis submitted to the Graduate Faculty of
Auburn University at Montgomery
in partial fulfillment of the
requirements for the Degree of
Master of Science

in

Computer Science

Montgomery, Alabama
26 July 2023

Approved by

Olcay Kursun Digitally signed by Olcay Kursun Date: 2023.07.22 14:23:57 -05'00'
Dr. Olcay Kursun
Thesis Director

Randy D. Russell Digitally signed by Randy D. Russell Date: 2023.07.24 09:49:05 -05'00'
Randy Russell
Thesis Co-Director

Lei Wu Digitally signed by Lei Wu Date: 2023.07.22 20:47:37 -05'00'
Dr. Lei Wu
First Reader

Digitally signed by Hua Yan Date: 2023.07.24 11:04:54 -05'00'
Hua Yan
Second Reader

Digitally signed by Matthew Ragland Date: 2023.07.25 13:54:02 -05'00'
Dr. Mathew Ragland
Associate Provost

ACKNOWLEDGEMENTS

This work was supported, in part, by NSF under Grant No. 2003740.

I want to give my thanks to everyone that has helped me with this thesis. I have learned so much about machine learning and hyperspectral images as part of this NSF funded research project. I would like to express my sincere appreciation to my advisors Dr. Kursun and Dr. Russell. They both have supported me all the time throughout this thesis. They will not hesitate to check up on me to see the progress. They continued to support me when I was under restricted schedule due to the Air Force's training. They willingly set up Zoom meeting in their free time over the weekend to help me. My heartfelt gratitude will also be extended to Ryan Vales, a senior undergraduate student at Auburn University at Montgomery and a student researcher on the NSF grant. His invaluable assistance during the experiment, particularly with the analysis portion, helped my work tremendously.

In addition, I would like to convey my thanks to all the professors of Computer Science at Auburn University at Montgomery. Without them, I would not have become the person I am today. I would like to especially thank Dr. Semih Dinc and Dr. Gao. They taught me so many things throughout my undergraduate at AUM.

Finally, I would like to acknowledge all my friends and family for their unwavering support to me during my school years. I was able to achieve so many things because of them. Their support and encouragement pushed me through every obstacle for the past years, I am deeply in their gratitude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Segmentation of hyperspectral images of sky/clouds is an important step in studying the scattering of sunlight by clouds. To achieve this challenging task of automatically analyzing the image collected by AUM Hyperspectral Imaging Team as part of an NSF grant, this thesis studies a combination of feature selection and clustering algorithms. Hyperspectral cameras are spectrometers which gather high resolution spectral information at each pixel of an image. In contrast to RGB images, hyperspectral images contain many narrow wavelength bands. The Resonon PIKA XC2 hyperspectral camera used in this study for imaging sky and clouds has a spectral resolution of only 1.3 nm and produces spectra with 462 bands. This large number of bands, as opposed to the regular RGB color images, serves as features to the clustering analysis. It is a challenging task to determine the ground truth and we can use machine learning clustering algorithms with different settings to create alternative versions. More importantly, in this thesis, I was interested in which features would be the most interesting to keep for the ability to maximally reproduce the clustering results with fewer features. With feature selection, we can see which bands could be potentially excluded from future study due to their lack of similarity to the ground truth. The dataset used contained two hyperspectral images with clouds and clear-sky pixels in them. After feature selection applied to the first dataset image, we have bands 245, 175, and 269 selected as they show the highest similarity of 96% to the 3-cluster ground truth. For 4 clusters, bands 190, 267, and 163 have a similarity of 95.72%. Lastly for 5 clusters, bands 268, 149, and 193 have a similarity of 95.11%. When the same concept is applied to the other hyperspectral image, the results vary, but we conclude that wavelength region 741 – 758 nm (bands 267 – 280) might be a good choice for clustering since they are frequently selected in the feature selection process.

CHAPTER 1: INTRODUCTION TO HYPERSPECTRAL IMAGES

The segmentation of hyperspectral sky/cloud images is a crucial process in examining sunlight scattering by clouds [4-7,18,24,28]. Hyperspectral imaging, renowned for capturing rich information, has gained popularity in climate change research. Notably, NASA's EMIT mission utilizes an advanced imaging spectrometer, showcasing the application of hyperspectral imaging in climate change analysis [22]. Facilitating the automatic analysis of images collected as part of an NSF grant [18] by the AUM Hyperspectral Imaging Team, this thesis explores the combination of clustering algorithms with feature selection [14,19,27] and cluster validity indices [1,15,25] to address this challenging task.

Ordinary RGB cameras use a grid of colored filters superposed on top of the individual photodiodes comprising the camera's light sensing array. Such an arrangement is referred to as a Bayer filter. The usual arrangement of filters within the grid is shown in Figure 1.1. The red, green, and blue channel counts for a given image pixel are determined by the response of a collection of photodiodes in close proximity to one another. Typical response curves for the colored filters are shown in Figure 1.2 [13]. Note the large overlap of the wavelength bands transmitted by the filters.

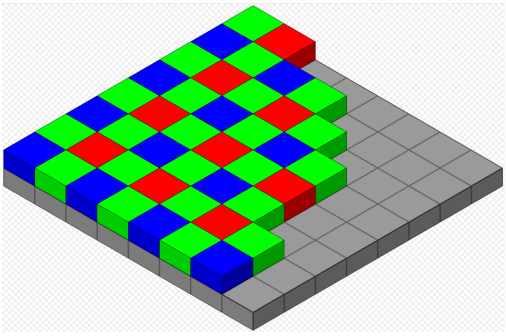

Figure 1.1 Bayer filter grid of colored filters on sensor pixels.

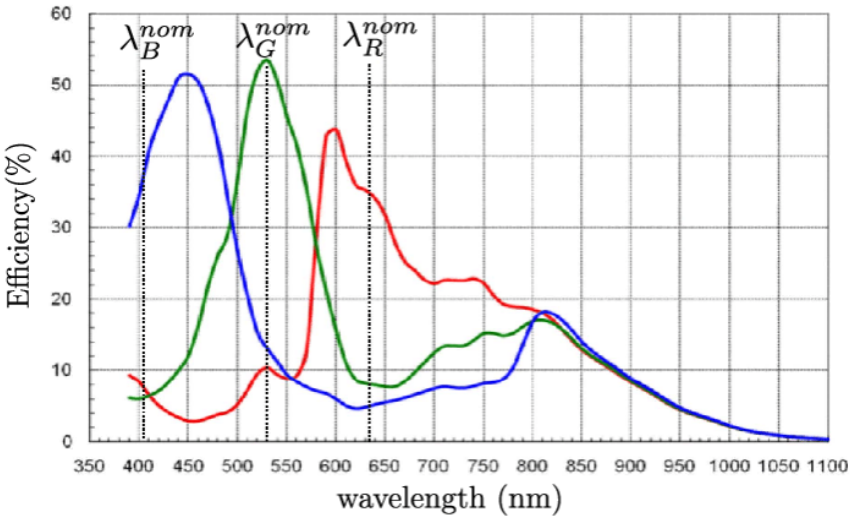

Figure 1.2 Spectral response of a typical Bayer filter

Hyperspectral cameras are spectrometers which gather high resolution spectral information at each pixel of an image. The Resonon PIKA XC2 hyperspectral camera [26] used in this study has a spectral resolution of only 1.3 nm. High spectral resolution is made possible by the use of a PGP (prism-grating-prism) to decompose light passing through the slit of the spectrometer into its constituent wavelengths (see Figure 1.3 [8]). In contrast to RGB images, hyperspectral images contain a large number of wavelength bands and the bands do not overlap. The PIKA XC2 produces spectra with 462 bands.

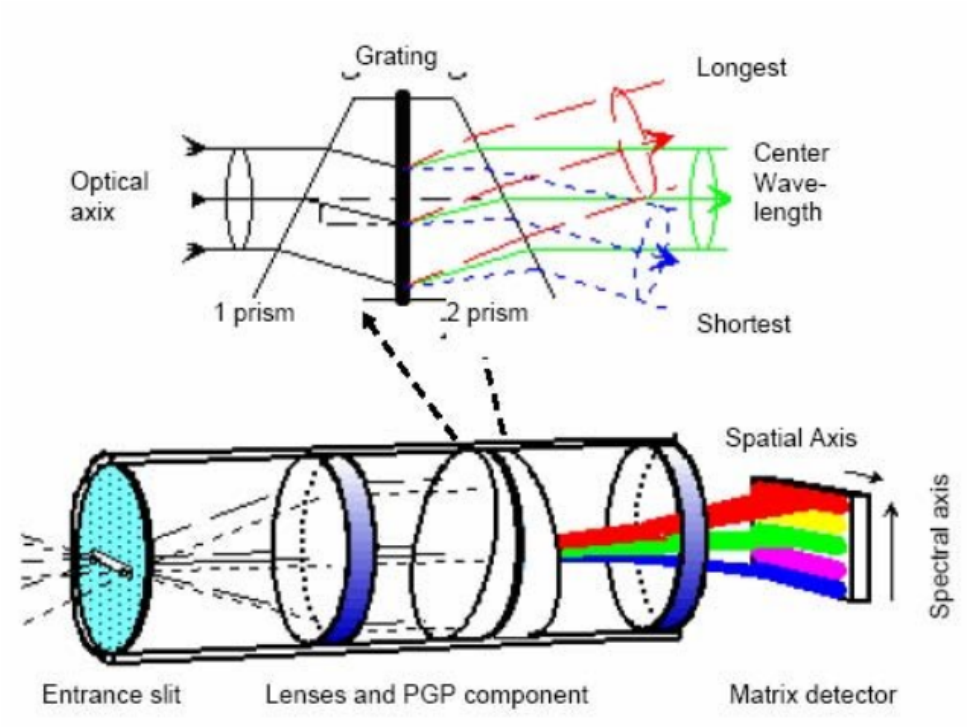

Figure 1.3 Prism-grating-prism spectrometer

Hyperspectral images are often referred to as “cubes” because they can be thought of as many two-dimensional gray scale images stacked on top of one another (see Figure 1.4 [12]).

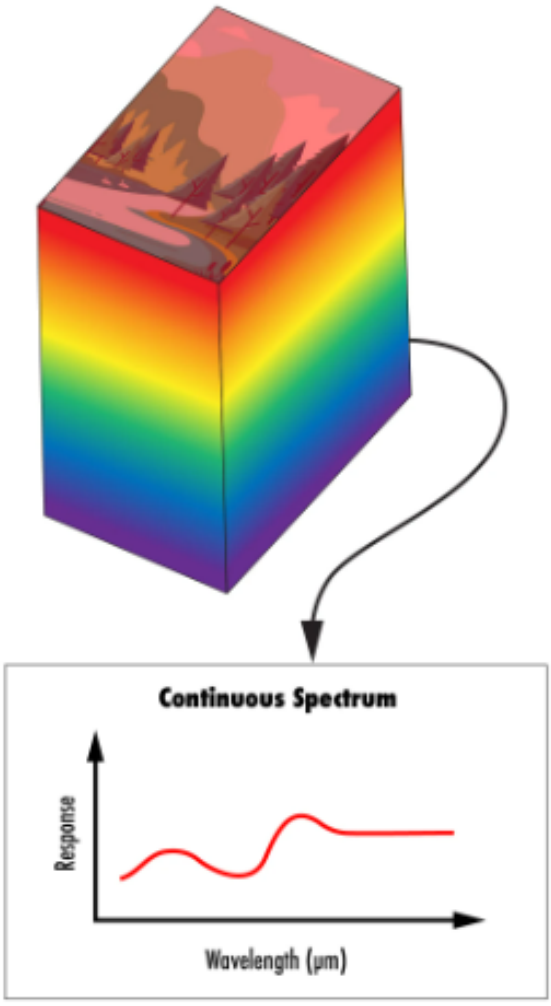

Figure 1.4 Hyperspectral cube

CHAPTER 2: DATA COLLECTION AND PRE-PROCESSING

The Resonon Pika XC2 hyperspectral camera [26] is a scanning spectrometer that uses a "push-broom" method to produce images. The camera must be panned across the target and the narrow individual images taken through the spectrometer slit stitched together to obtain a full image. A 17 mm focal length camera lens with a fast focal-ratio of 1.4 was used in this study. Using that lens, the field of view in the long direction of the camera's slit is 30.8 degrees. The field of view along the short direction of the slit, called the integration field of view, is only 0.71 mrad (see Figure 1.4, which was obtained from Resonon_Product_Catalog_June_2022 [26]). An image covering a 90-degree range in azimuth requires 4402 individual exposures. Since each exposure contains 1600 spatial pixels, a complete 90-degree image has dimensions (1600 x 4402) spatial pixels x 462 wavelength bands.

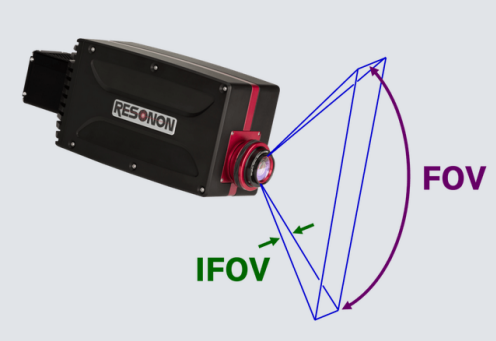

Figure 2.1 Field of view (FOV) and integration field of view (IFOV)

The camera is attached to a stage that pans the camera across the target by turning the camera slowly around a vertical axis as exposures are accumulated (see Figure 2.2). An Oben VH-R2 tilt head is used to keep the elevation of the camera constant during the scan. This allows the solid angle subtended by each pixel in a sky image to be easily determined.

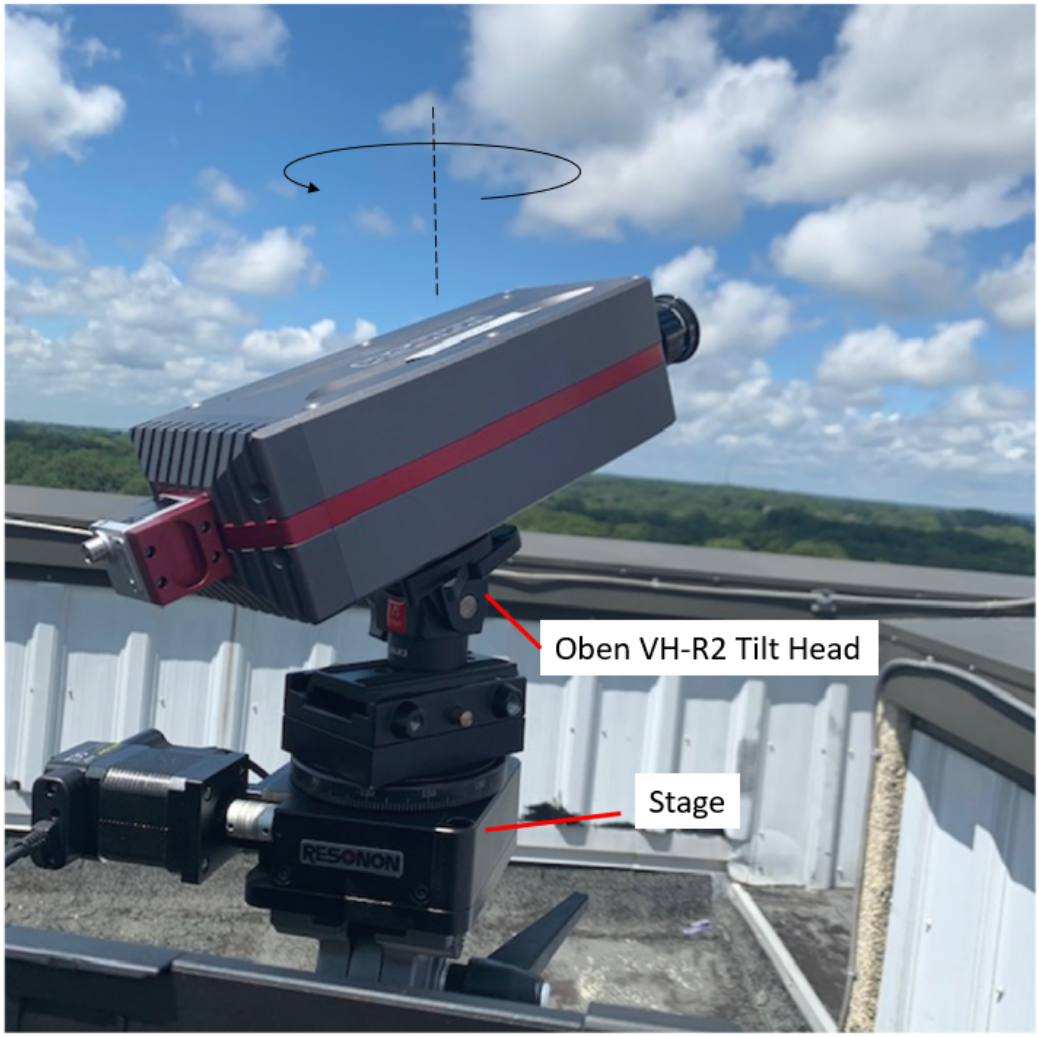

Figure 2.2 Camera stage and tilt head

It is possible for the stage motor to step through small increments of the angle, so the individual exposures fit seamlessly side-by-side, creating a complete image. However, this process is very time consuming. Several minutes are required to collect images covering a reasonable field of view. Instead, the camera is panned continuously. In order to achieve an image that has the correct aspect ratio and the minimum amount of blurring due to the camera motion it is necessary to synchronize the rate at which the camera pans with the framerate (the rate at which the camera accumulates exposures). For our hyperspectral camera, the correct relationship between frame rate(FR) and scan rate(SR) is

$$SR = 0.02035\ FR$$

The scan rate (SR) is in units of degrees per second and the framerate (FR) is in units of frames per second. For sky images, a frame rate of 45 frames per second and a scan rate of 0.92 degrees per second are a good compromise between special resolution and the time required to acquire images. The acquisition time for a full 90-degree image is about 1.6 minutes. This time scale is comparable to the time over which rapidly evolving clouds undergo significant changes.

Because of large differences in the radiances coming from different sky regions, particular care must be taken in the choice of integration time (also called exposure time) to avoid the saturation of sensor pixels in the regions of bright clouds. A typical integration time is about 12 milliseconds, but satisfactory integration times can range from about 8 to 15 milliseconds depending on sky conditions.

As the image is being acquired, the Spectronon software which runs the camera builds a "waterfall" image by sequentially displaying the acquired frames as a horizontal line until eventually a depiction of the entire image is displayed. The displayed image is referred to as a render, because the hyperspectral image is rendered as an RGB image in which the channel counts at three select bands of the hyperspectral image are used to determine the values of R, G, and B. The default channels used to determine R, G, and B are at wavelengths of 643.1 nm, 548.8 nm, and 461.6 nm, however, the Spectronon software allows any three channels in the hyperspectral image to be used in displaying the render. By choosing certain channels, it is sometimes possible to bring our features in the hyperspectral image not readily apparent in the standard render.

The CMOS detector of the Pika XC2 has a different sensitivity to light energy at different wavelengths. Also, the glass in the camera's optics absorbs light energy differently at different wavelengths. As a result, the "raw" images acquired by the camera must be corrected for this instrument error to obtain images for which the channel counts are proportional to monochromatic radiance (the rate at which light energy at a particular wavelength is received per unit area per solid angle subtended by the region detected by the pixel). Resonon's Spectronon software can be used to make the appropriate correction. A calibration file provided by the manufacturer of the camera is used in the image calibration process. Figure 2.3 shows a comparison between the spectra of raw and radiance cloud images.

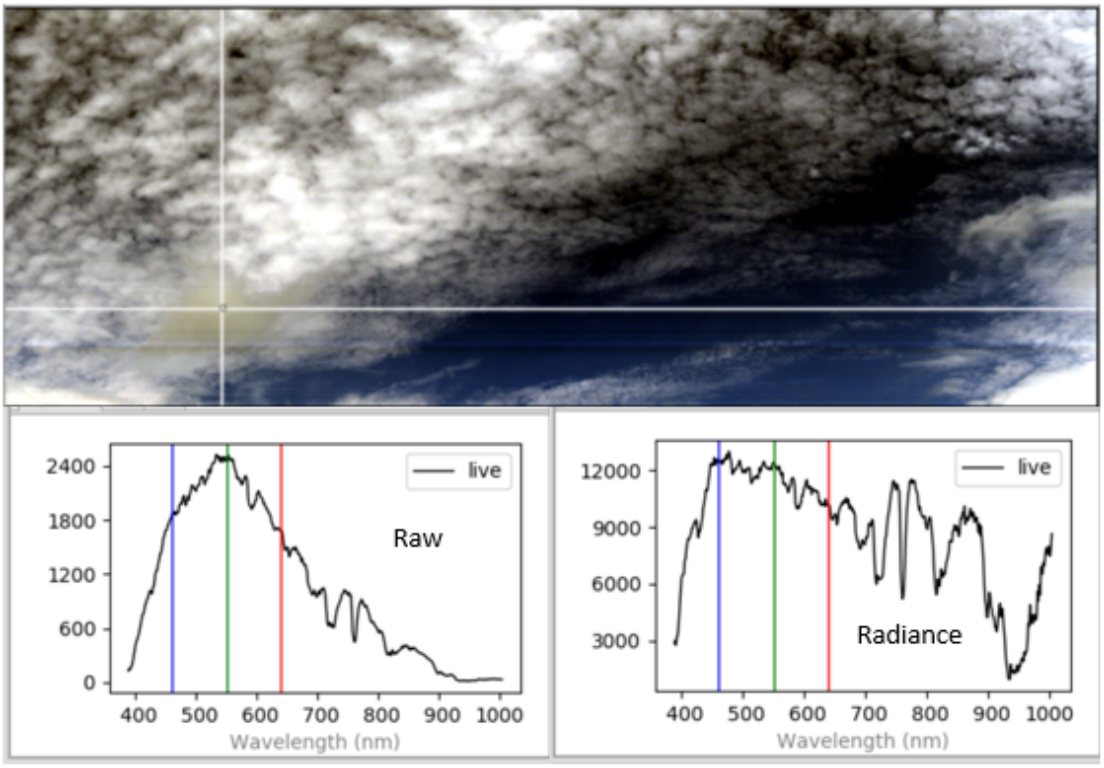

Figure 2.3 Comparison between a raw spectrum and a radiance spectrum for a cloudy pixel.

CHAPTER 3: PREVIOUS WORK

This thesis will explore several methods that have potential application in automated systems for determining the fraction of sky cover by cloud. Such systems are useful for solar energy engineers in making short term forecasts of available solar energy as well as in determining the solar energy potential of locations being explored as possible sites for solar energy collectors [3, 30]. The amount of sunlight available to plants for photosynthesis is strongly dependent on the amount of cloudiness, so a record of cloudiness is useful in agriculture and forest management [11, 20]. Also, cloud fraction is often a required input into short-term weather forecast models, and a detailed and precise record of cloud cover is valuable to atmospheric scientists studying the effects of cloud cover on climate [2].

3.1 NSF-Funded Hyperspectral Image Segmentation Project

This thesis is part of a National Science Foundation (NSF) - funded project being undertaken by the AUM Department of Computer Sciences and Computer Information Systems in conjunction with the University of North Georgia [18]. The project centers around the development of a real-time, three-layer framework for hyperspectral image segmentation using machine learning models and optimized for high-performance computation.

Hyperspectral images have both high spatial and spectral resolution. Analysis of such images poses a significant computational challenge due to the enormous amount of data they contain. However, their rich information content makes them more useful to scientists compared to ordinary RBG images. By leveraging the power of high-performance computing tools and machine learning techniques, this project seeks to provide an efficient solution to the complex problem of hyperspectral image segmentation. The framework is constructed in a multi-layered design, where each subsequent layer enhances the accuracy of the results of the previous layers.

One significant application of this project is the development of methods for the characterization of cloud climate using hyperspectral and multi-spectral imaging. Cloud fraction, the portion of the sky covered by cloud, is an important facet of cloud climate. To determine cloud

fraction, sky images must be accurately segmented into clouds and clear sky. Once a hyperspectral image has been segmented, the image's spectral content can be used to determine the amount of solar radiation originating from scattering by cloud particles.

A second application involving the segmentation of sky images is the forecasting of solar radiation. Short-term predictions of the shading of solar arrays by cloud requires the segmentation of multiple images in succession to predict cloud motion.

The implementation of this project is broken into four distinct phases, with this thesis primarily contributing to the first two phases:

Phase-1. Data Collection and Analysis

Phase-2. Building Classification Models and Prototyping

Phase-3. HPC Integration and Software Design

Phase-4. Software Implementation, Testing, and Deployment

3.2 Hyperspectral Image Classification Using Cluster Ensemble-Based Categorical Features

As part of his image segmentation project, Giovanni Bellio's master's thesis [6] proposes a unique approach to hyperspectral image pixel classification. The approach relies on a cluster-ensemble-based categorical feature extractor and a categorical boosting classifier that utilizes these features. Instead of traditional feature selection, which can be computationally costly due to the high dimensionality of hyperspectral data, a sliding window technique was applied to generate a diverse set of clustering runs. By adopting clustering algorithms like K-means as preprocessing tools, Giovanni was able to simplify and quantize the hyperspectral image datasets, which generally lack categorical features. This preprocessing, performed through multiple clustering runs, converted each pixel's cluster membership into categorical 'super-features'. That is, the cluster indices produced from these runs were then employed as categorical features for the categorical-boosting classifier. This approach allows for a richer data representation by employing

multiple clusterings to augment the dataset, thus enhancing the classification accuracy of the boosting ensembles.

Giovanni's method is advantageous for the segmentation of high-resolution hyperspectral sky images which contain hundreds of features comprised of a sequence of narrow wavelength bands. Furthermore, his approach has the potential for easy integration into high-performance computing frameworks due to its inherent parallelism. This characteristic, coupled with its enhanced classification performance, makes it an appealing and effective strategy for hyperspectral image classification. Giovanni's work provided several contributions to the overall NSF project, specifically assisting in the algorithm development for Phase 2 - Building Classification Models and Prototyping. It also resulted in a publication [7].

3.3 Impact of Feature Selection and Spectral Normalization on Hyperspectral Image Segmentation through Gray Level Image Thresholding

In her undergraduate honors thesis, Derienne Black explored the effects of feature selection and spectral normalization on the segmentation of hyperspectral images [5]. Her study was conducted as part of the hyperspectral image segmentation project sponsored by NSF and focused on the application of thresholding techniques to gray level images. Binarization is not as powerful as the clustering algorithms used in the research this thesis builds upon, but Black's research focused on evaluating the effects of normalization on binarization. Her primary objective was to ascertain if normalizing the data by average radiance would enhance segmentation.

Black used the project's hyperspectral imaging system to capture hyperspectral sky images. Data at each image pixel comprises a calibrated spectrum across 462 narrow wavelength bands ranging from 400 to 1000 nm. The normalization process was carried out using the channel corresponding to a wavelength of 586 nm. The monochromatic radiance at that channel has been shown to be proportional to the spectrally averaged radiance over a wide variety of sky conditions. The normalized monochromatic radiance indicates how the energy content of the spectrum varies with wavelength. At a wavelength of 454 nm, the difference in normalized monochromatic radiance between cloud and clear sky pixels was found to be larger than at other wavelengths, so

normalized radiance at a wavelength of 454 nm was selected to produce gray level images for segmentation.

Three distinct thresholding techniques were employed in the study: Otsu's method [23], Kapur's entropy thresholding [16], and Kittler-Illingworth minimum error thresholding [17]. Segmentation results were compared using Normalized Mutual Information (NMI) and the Adjusted Rand Index (ARI).

Otsu's method, an automatic thresholding technique, calculates an optimal value to maximize the separability of foreground and background regions in an image. This method assigns pixels with intensity values above the threshold to the foreground and those below to the background, resulting in a binary image. Kapur's entropy thresholding determines an optimal threshold by maximizing the binary image's entropy or information content and is primarily used for image segmentation. It divides the image into two classes based on a threshold value, then calculates the entropy of each class. The threshold value maximizing the total entropy is selected as optimal. Kittler-Illingworth thresholding minimizes the classification error between foreground and background pixels. This technique computes probabilities of the foreground and background for each threshold value and calculates the error. The optimal threshold corresponds to the lowest overall classification error.

Ms. Derienne Black concluded her study by examining the suitability of NMI and ARI for assessing segmentation quality. She found NMI more appropriate for datasets with clusters of varying sizes, while ARI performed better for datasets with similar cluster sizes. Her selection of the Adjusted Rand Index (ARI) as the preferred method for calculating segmentation has been adopted in the current work as well. Her work contributed to the NSF project's Phase 1 - Data Collection and Analysis, particularly by evaluating the influence of data normalization and thresholding techniques on the segmentation of hyperspectral images.

CHAPTER 4: MACHINE LEARNING METHODS AND DATA ANALYSIS

The methodology adopted in this thesis employs the K-Means clustering algorithm [14, 19], the Adjusted Rand Index (ARI) [1,15,25], and sequential feature selection [9,10,14,21,27,29] to analyze two exemplary cloudy image patches from the hyperspectral image dataset that we collected. The aim is to identify the feature (wavelength/band), among the total of 462 bands in the hyperspectral images, that exhibits the highest similarity to the ground truth (the most accurate known label for each pixel in the images). That is, assuming that the ground truth segmentation is obtained by application of K-Means to the whole set of features, which few features could approximate the same partitioning of the image pixels. For example, one can test how well the clustering obtained when using the red and blue wavelengths/bands together matches with that of the ground truth.

K-Means is an unsupervised machine learning algorithm that assigns data points to K clusters, where each data point belongs to the cluster with the nearest mean value. The Adjusted Rand Index (ARI) is a measure that quantifies the similarity between two data clusterings. Being adjusted means that the ARI score is corrected for chance (for example if K is set to the number of data points, then the uncorrected score would give 100% agreement). ARI score typically lies between 0 and 1 and can go slightly below 0 when two clusterings are in disagreement. A high negative value (near 1) is not easy to obtain and may indicate a bug or data leak.

K-Means is one of the most popular clustering algorithms [14,19]. It is a simple to implement algorithm that can take a few lines of code, but it can help analyze the data by revealing its structure/groups/clusters. It is an unsupervised algorithm, which means it does not use class-labels. We used Scikit-learn Python libraries for applying K-Means to our problem. Scikit-learn describes K-Means and its input-output arguments as follows [19].

```
from sklearn.cluster import KMeans
import numpy as np
X = np.array([[1, 2], [1, 4], [1, 0],
              [10, 2], [10, 4], [10, 0]])

kmeans = KMeans(n_clusters=2, random_state=0, n_init="auto").fit(X)

print(f"{X = }")

print(kmeans.labels_)

test_set = [[0, 0], [12, 3]]
print(kmeans.predict(test_set))

print('centers = ', kmeans.cluster_centers_)
```

In this example, our input dataset X contains 6 examples and 2 dimensions each. These points in 2D space will be replaced by points (corresponding to pixels) in 462-dimensional space. Kmeans will return cluster memberships as: [1 1 1 0 0 0].

K-Means also returns the centroid coordinates, which could be useful in our research to know prototypical clear-sky and cloud hyperspectral signature (Refer to Figure 2.3). In this particular example, we have:

```
centers = [[10.  2.]
           [ 1.  2.]]
```

For its application to our HSI (Hyperspectral image) data analysis, we used K-means clustering with K=3 for segmentation (considering clear-sky, thin-clouds, and clouds as the potential sources of clusters). We used sequential feature selection to pick the best bands (we picked the best three in our analysis), and rand-index for cluster validity index (to measure the match between the partitioning based on all the bands and the partitioning based on single bands).

To apply these methodologies, an iterative analysis of two images is required. We use two sample images, R1 and R2, each of the size 401x402x462 (see Figure 3.1 and Figure 3.2). Initially, K-Means is applied to these images to establish the ground truth. The K parameter is set to three,

but we also tried two, four and five clusters to check our assumption of having three main classes in the images is right. The random_state parameter for K-Means is kept at zero for this experiment for reproducibility, but it can be altered as per requirement.

The goal of this analysis is to identify which bands give the closest ARI score to the ground truth in our HSI images. Since HSI provides a wealth of spectral information per pixel, it leads our machine learning clustering methods to run on hundreds of wavebands and presents a challenge known as the "curse of dimensionality" [10,14], which requires employing feature selection algorithms to mitigate computational burdens and improve prediction accuracy. Band selection is a crucial process to optimize the use of the rich spectral information and sequential feature selection is an efficient algorithm that can be used to reduce redundancy among spectral bands while attempting to maintain the original information of the image [9,21,29].

Following the establishment of the ground truth, two arrays, 'labels' and 'scores', are initialized to record the labels of the data points and the overall partitioning's similarity scores to the ground truth as each band is iterated. We need the labels to calculate the ARI scores, which are then saved in 'scores'. In sequential feature selection, after the first band is selected, it stays in. Thus, for the second-best band, the search is for the feature that best complements the first/already selected feature to improve the ARI similarity score with the ground truth. Therefore, in the first pass all 462 bands are tested (K-Means and ARI) to identify the best band, in the second pass 461 remaining bands are tested to identify the second band that complements the first one, and in the third pass, 460 remaining bands are tested to identify the third band that complements the first two. This is an $O(n^2)$ process if all bands were to be ranked but generally just a few features are needed.

More specifically, the algorithm consists of two for-loops. The outer loop runs just a few times until sufficiently many features are selected. The goal of the inner loop is to iterate each band to compare which band helps obtain the most similarity to the ground truth (the band is tested along with the already selected features to see how much added benefit it gives):

```
#Input: cube[:num_rows, :num_cols, :462]
#Output: selected_features[:3]

gt = k_means.fit(cube).labels_
selected_features = []
num_bands = 462
num_select = 3

for best in range(num_select):
    scores = zeros(num_bands) #initialize all scores to zero
    for i in range(num_bands):
        if i in selected_features:
            continue  #skip already selected features

        testing = selected_features + [i]
        data = cube[:,:,testing].reshape(num_rows*num_cols,1)
        testing_partition = k_means.fit(one_band).labels_
        scores[i] = adjusted_rand_score(gt, testing_partition)
     best_band = argmax(scores) #find the best band
     selected_features.append(best_band)
```

It is a Python implementation detail, but it is important to note that we used ‘testing’ as a new set of features/bands to test its suitability for K-Means to approximate the ground_truth. Note that we used ‘+’ operator, otherwise ‘.copy()’ should be used, to avoid mutating the selected_features list.

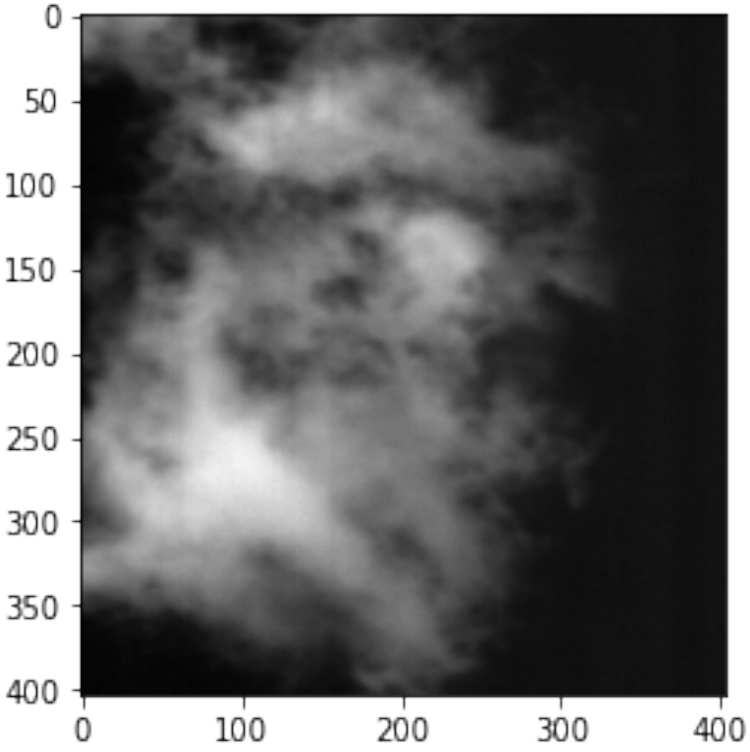

Figure 4.1 The hyperspectral R1-image

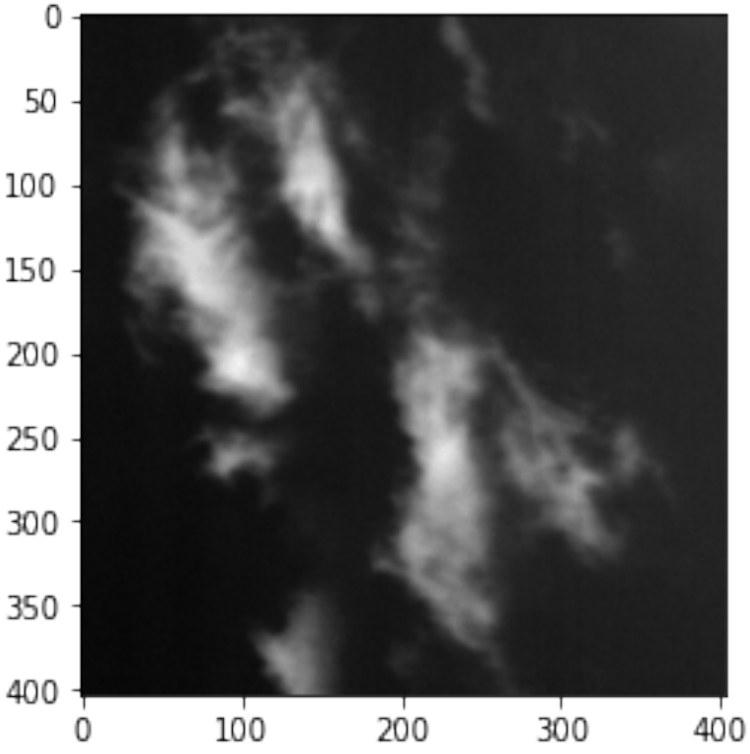

Figure 4.2 The hyperspectral R2-image

CHAPTER 5: EXPERIMENTAL RESULTS

In this chapter, we will discuss the effects of sequential feature selection on the quality of the clustering. As explained in Chapter 3, ARI is used to measure the quality of a feature subset, measuring the match between clustering obtained when using this subset and the ground truth obtained by clustering the image with all the 462 bands. Table 5.1 and table 5.2 show the results of the algorithm. In addition, Figures 5.1 through 5.21 consist of all of R1's experiments that correspond to Table 5.1, and Figures 5.22 to 5.39 correspond to Table 5.2.

Let us first discuss what Table 5.1 and Table 5.2 show. These tables contain the main results of the experiments on the two exemplary HSI images, Figures 3.1 and 3.2, respectively. The row shows which band is the highest result for that certain pass in addition to its wavelength and its score. The column shows the number of clusters for that particular picture. For the first pass of three clusters, the result shows band-245 (corresponding to 712.79 nm wavelength) has a 93.00% similarity to the 3-cluster ground truth (the clusterings of the ground-truth and band-245 are shown in Figure 5.1 and Figure 5.2, respectively).

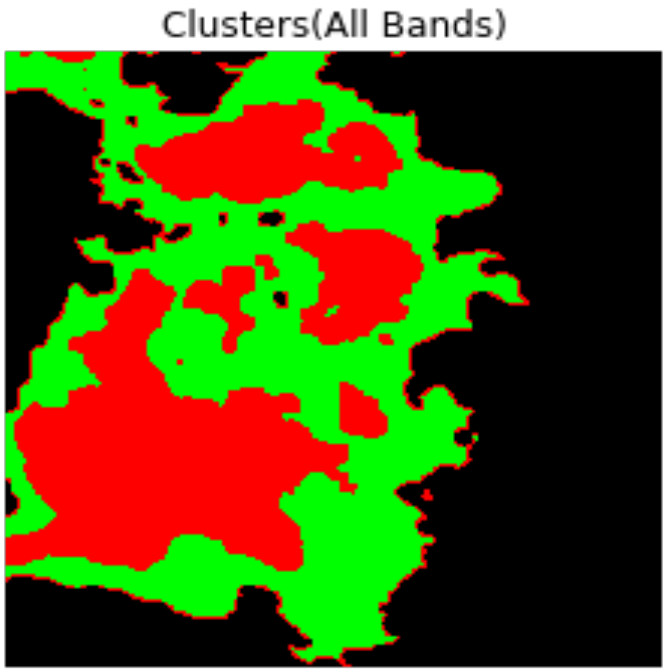

Figure 5.1 The 3-cluster ground truth for the R1-image

The partitioning shown in Figure 5.1 is well approximated by band-245 as can be seen in Figure 5.2. In order to visually compare Figure 5.1 and Figure 4.2 and to see that they are similar to each other, we need to make the cluster colorings correspond to each other. There are three clusters, hence three different colors to distinguish them. Different colors represent different clusters, but these color assignments are arbitrary.

When K-Means is applied to an image, it goes through every pixel and labels them as either this pixel is with cluster 0, cluster 1, or cluster 2 based on where the cluster centroids converged in the training (fit). However, as K-Means clustering is an unsupervised process, it is hard to guarantee the cluster colors correspond to each other. As the algorithm starts with random centers, the labelling is arbitrary; that is, next time we run K-Means on the same exact data, it can number these clusters differently (and will end up with a slightly different partitioning anyway in most cases – that is why scikit-learn runs K-Means a number of times and reports the best run).

In addition to the difficulty of visually quantifying the agreement between two clusterings, we also face the difficulty of having 462 of such clusterings to compare with the ground truth. Therefore, we used ARI scores for individual bands and picked the one that has the maximum score. Figure 4.3 shows that ARI plot.

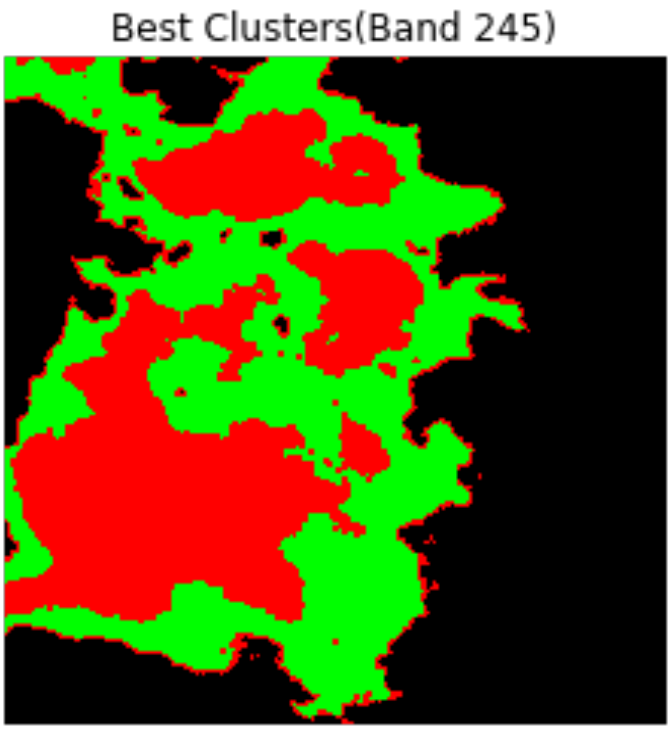

Figure 5.2 The 3-cluster partitioning with band-245 on the R1-image.

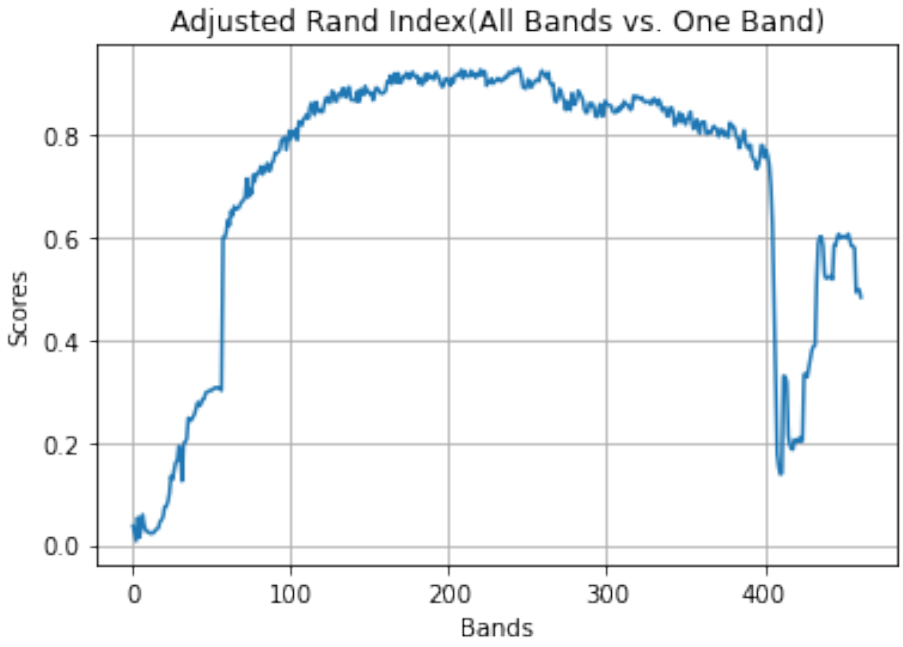

Figure 5.3 ARI-scores of the first pass for the 3-cluster feature selection on the R1-image.

This feature selection process needs to continue for selecting the second-best feature and then the third-best feature such that they have the maximal performance when used together. As the number of figures to show how these decisions affect the clusterings is large, we created Table 5.3 and Table 5.4 to show which figure explains what.

K for Ground Truth	First Pass	Second Pass	Third Pass
3 Clusters	band-245 (712.79 nm) 93.00% Similarity	band-175 (619.49 nm) 93.39% Similarity	band-269 (744.91 nm) 96.00% Similarity
4 Clusters	band-190 (639.43 nm) 91.42% Similarity	band-267 (744.91 nm) 94.92% Similarity	band-163 (603.55 nm) 95.72% Similarity
5 Clusters	band-268 (712.79 nm) 90.55% Similarity	band-149 (584.99 nm) 94.57% Similarity	band-193 (643.42 nm) 95.11% Similarity

Table 5.1 Summary of the feature selection results on the R1-image.

Table 5.1 shows that when we used the 3-cluster ground-truth obtained on all the bands is best matched by band-245, and with the selection of the second band (band-175) this match percentage (ARI-score) goes up to 93.39%, and finally when band-269 is selected then the overall ARI (using bands 245, 175, and 269 together) goes up to 96%. A similar table is provided for the HSI image R2 in Table 5.2. Also, to find the clustering results on the R1-image and ARI-plots, the reader refers to Table 5.3 as the map of the figures. Table 5.4 serves the same purpose for following the results on the R2-image.

K for Ground Truth	First Pass	Second Pass	Third Pass
3 Clusters	band-118 543.96 nm 86.78% Similarity	band-133 563.8 nm 89.88% Similarity	band-280 759.6 nm 91.35% Similarity
4 Clusters	band-134 562.12 nm 88.06% Similarity	band-175 619.49 nm 90.65% Similarity	band-280 759.6 nm 91.63% Similarity
5 Clusters	band-175 619.49 nm 87.05% Similarity	band-178 623.47 nm 89.71% Similarity	band-33 432.47 nm 91.42% Similarity

Table 5.2 Summary of the feature selection results on the R2-image.

K for Ground Truth	Ground Truth	Pass-1 Clusters	Pass-1 Scores	Pass-2 Clusters	Pass-2 Scores	Pass-3 Clusters	Pass-3 Scores
3 Clusters	5.1	5.2	5.3	5.4	5.5	5.6	5.7
4 Clusters	5.8	5.9	5.10	5.11	5.12	5.13	5.14
5 Clusters	5.15	5.16	5.17	5.18	5.19	5.20	5.21

Table 5.3 Figure-navigation table for the results on the R1-image.

K for Ground Truth	Ground Truth	Pass-1 Clusters	Pass-1 Scores	Pass-2 Clusters	Pass-2 Scores	Pass-3 Clusters	Pass-3 Scores
3 Clusters	5.22	5.23	5.24	5.25	5.26	5.27	5.28
4 Clusters	5.29	5.30	5.31	5.32	5.33	5.34	5.35
5 Clusters	5.36	5.37	5.38	5.39	5.40	5.41	5.42

Table 5.4 Figure-navigation table for the results on the R2-image.

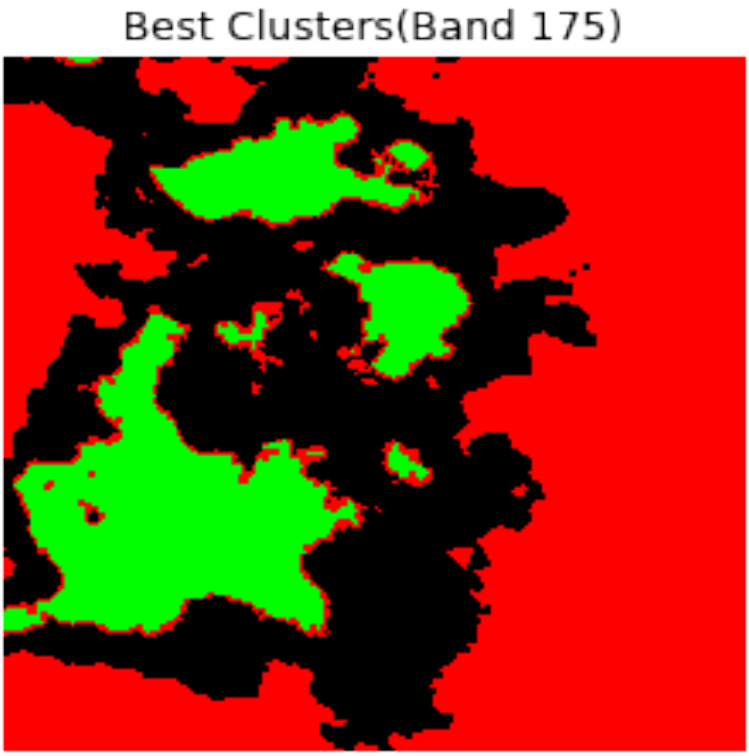

Figure 5.4 The 3-cluster partitioning with band-245 and band-175 on the R1-image.

As can be clearly seen from Figures 5.4 and 5.6, the clusterings match ground truth better with the selection of the second and third features into the “selected_features” subset, respectively.

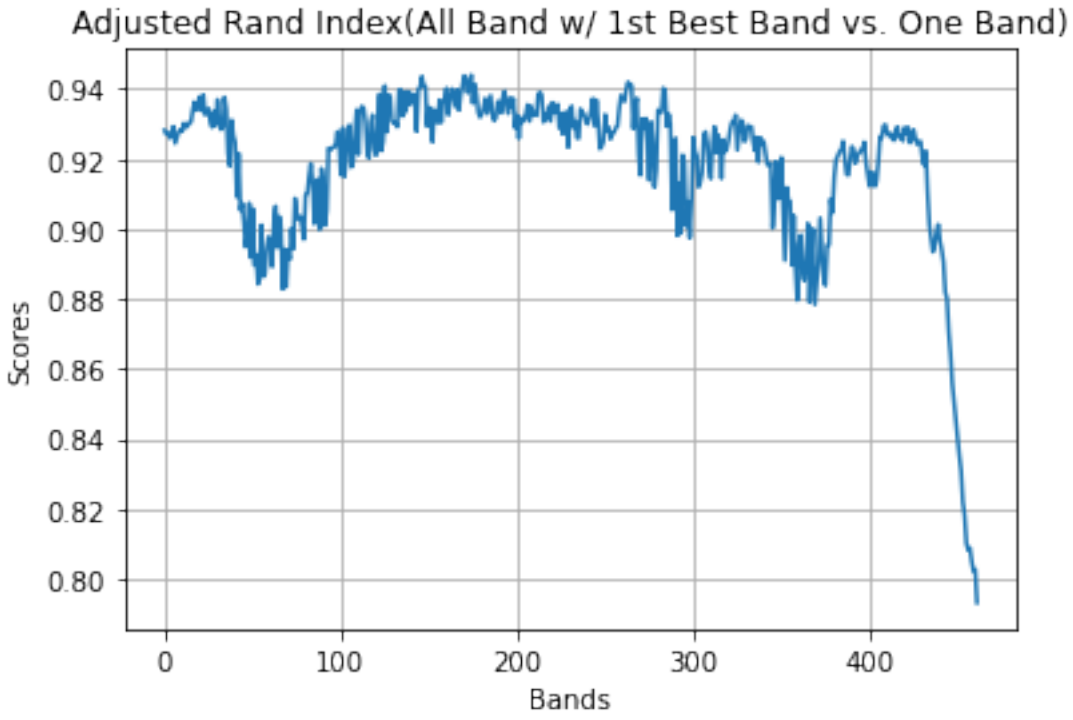

Figure 5.5 ARI-scores of the second pass for the 3-cluster feature selection on the R1-image.

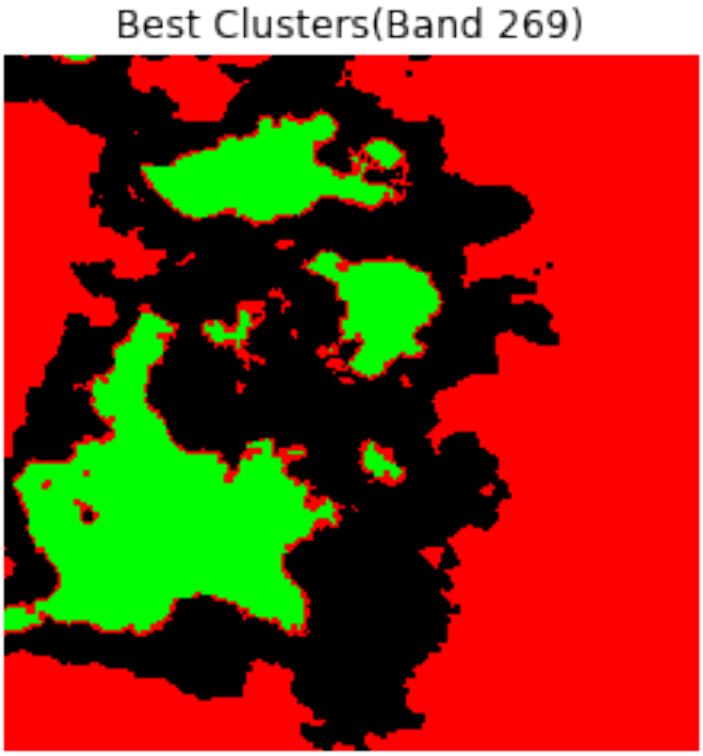

Figure 5.6 The 3-cluster partitioning with band-245, band-175, and band-269 on the R1-image.

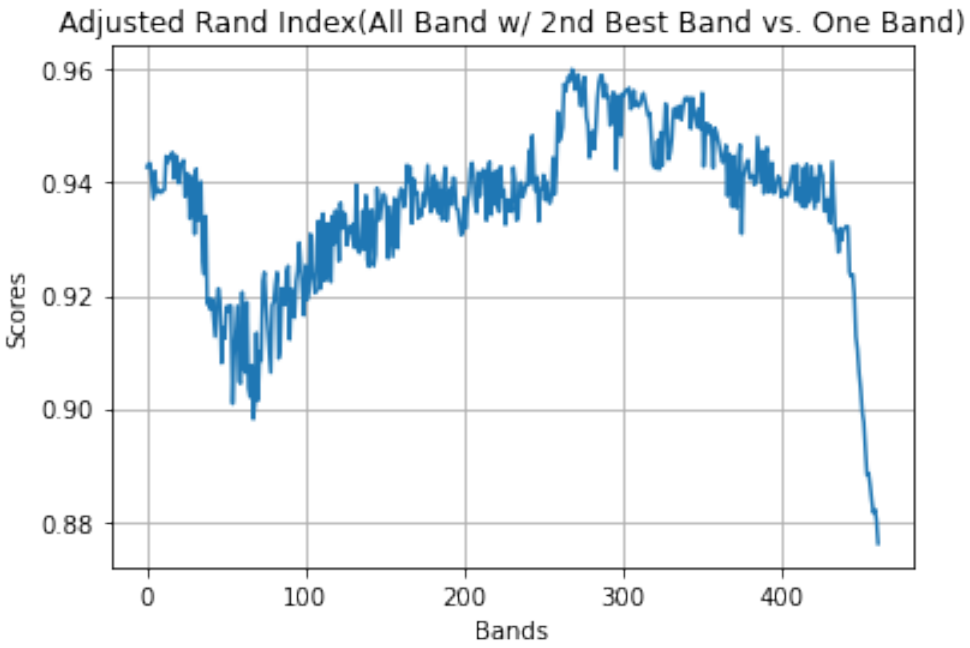

Figure 5.7 ARI-scores of the third pass for the 3-cluster feature selection on the R1-image.

Once the labels are obtained from both sides, the algorithm performs sequential feature selection using adjusted rand index to compare labels. The relationship between each band and the ground truth in the first pass is shown by Figure 5.3.

A significant improvement is shown when we apply the second pass, especially for the first 100 bands. When band-245 combined with band-175 it achieves a 93.39% score, that is 0.39% increase compared to the first pass. Bands after 400 show an increase in scores, unlike the first pass. In the third pass for three cluster partitioning, the scores started higher than the second pass, but they maintained an inverse relationship as the score increases as passes are being iterated and a decrease of scores from the first pass, band-269 perform best with 96.00% score whereas it performed 86.53% in the first pass.

For the four clusters, the algorithm finds band-190 has the highest score of 91.42%. However, Figure 5.3 and Figure 5.10 have similar pattern to each other, Figure 5.10 shows top scores, which are more consistent from band-100 to band-400. Four clusters generally produce better results.

CHAPTER 6: CONCLUSION

In conclusion, this study reaffirms the remarkable capacity of hyperspectral images for atmospheric cloud segmentation when leveraged with effective machine learning methodologies. This thesis demonstrates the use of K-Means clustering algorithm, the Adjusted Rand Index (ARI) cluster validity index for comparing two clusterings, and sequential feature selection techniques for dimensionality reduction, all combined for identifying the optimal wavelengths for atmospheric cloud detection and clustering. In particular, K-means have proven beneficial for segmenting images into similarity-based pixel groups. Nevertheless, the high dimensionality of hyperspectral datasets presents a significant challenge.

By employing feature selection algorithms, we can effectively prune redundant information while preserving the integrity of the clustering results to the maximum extent possible. We can assess this consistency using the rand-index, which ranks the bands in order of their importance. Our results highlight the wavelength region 741 – 758 nm (bands 267 – 280) as a potentially effective choice for clustering, given their frequent selection during the feature selection process. Furthermore, the synergistic relationship observed among the top three bands, specifically bands 118, 133, and 280, provides a higher ARI, reflecting their complementarity.

As future work, we should understand the initial dip observed within the first 100 bands during the first and second passes. By investigating these anomalies, we can further refine our approach and improve our understanding of these complex hyperspectral datasets.

REFERENCES

[1] ARI Adjusted Rand Index. Scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html#sklearn.metrics.adjusted_rand_score (Accessed: July 3, 2023).

[2] Arking, A., 1991. The radiative effects of clouds and their impact on climate. Bul. Amer. Meteor. Soc. 72 (6) 795-814. https:/doi.org/10.1175/1520-0477(1991)072<0795:TREOCA>2.0.CO;2

[3] Bernecker, D., Riess, C., Angelopoulou, E., Hornegger, J., 2014. Continuous short-term irradiance forecasts using sky images. Solar Energy 110, 303-315. http://dx.doi.org/10.1016/j.solener.2014.09.005

[4] Black, D., Russell, R., Kursun, O. (2023). The Effect of Spectral Normalization on the Segmentation of Hyperspectral Sky Images Using Thresholding. AUM Undergraduate Research Symposium, Montgomery, AL, April 2023.

[5] Black, D. (2023). Thesis Director: R. Russell, Thesis Codirector: O. Kursun. Effect of Wavelength Selection on the Segmentation of Hyperspectral Images Using Thresholding of Gray Level Images. Honors Thesis. Auburn University at Montgomery.

[6] Bellio, G. (2022). Thesis Director: O. Kursun, Thesis Codirector: R. Russell. Boosting With Original and Clustered Categorical Features for Machine Learning on Large Datasets. Master's Thesis. Auburn University at Montgomery. Retrieved from https://digitalarchives.aum.edu/theses/Bellio

[7] Bellio, G., Russell, R., Kursun, O. (2023). Boosting with Multiple Clustering Memberships for Hyperspectral Image Classification. In IEEE SoutheastCon, Orlando, FL, USA.

[8] Campbell, M. V., Fischer, R. L., Pangburn, T., & Hardenberg, M. J. (2005). Using high spatial resolution digital imagery. (ERDC TR-05-1). Engineer Research and Development Center.

[9] C. Sheffield, “Selecting band combinations from multi spectral data,” Photogrammetric Engineering and Remote Sensing, vol. 58, no. 6, pp. 681–687, 1985.

[10] Dai, Q., Cheng, J.-H., Sun, D.-W., & Zeng, X.-A. (2015). Advances in feature selection methods for hyperspectral image processing in food industry applications: a review. Critical Reviews in Food Science and Nutrition, 55(10), 1368-1382. https://doi.org/10.1080/10408398.2013.871692

[11] Durand, M., Murchie, E. H., Lindfors, A.V., Urban, O., Aphalo, P.J., 2021. Diffuse solar radiation and canopy photosynthesis in a changing environment. Agric. For. Meteorol. 211, 108684. https://doi.org/10.1016/j.agriformet.2021.108684

[12] Edmund Optics. Hyperspectral and Multispectral Imaging. Retrieved 07/03/2023 from https://www.edmundoptics.com.sg/knowledge-center/applications/imaging/hyperspectral-and-multispectral-imaging/

[13] Flasseur, O., Fournier, C., Verrier, N., & Denis, L. (2017). Self-calibration for lensless color microscopy. Applied Optics, 56(13), F189-F199. https://doi.org/10.1364/AO.56.00F189

[14] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[15] Hubert, L., & Arabie, P. (1985). Comparing Partitions. Journal of Classification, 2(1), 193-218. https://doi.org/10.1007/BF01908075

[16] Kapur, J. N., Sahoo, P. K., & Wong, A. K. C., 2006. A new method for gray-level picture thresholding using the entropy of the histogram. Computer Vision, Graphics, and Image Processing 29 (3), 273-285. https://doi.org/10.1016/0734-189X(85)90125-2

[17] Kittler, J., & Illingworth, J., 2003. Minimum error thresholding. Pattern Recognition 19 (1), 41-47. https://doi.org/10.1016/0031-3203(86)90030-0

[18] Kursun, O. (2020-2024). NSF RUI Collaborative Research: CDS&E. “A Modular Multilayer Framework for Real-Time Hyperspectral Image Segmentation” (Grant No. 2003740).

Auburn University at Montgomery. Budget: $207,664. Role: Principal Investigator (starting from May 2022).

[19] K-Means. Scikit-learn. Available at: https://scikit-learn.org/stable/modules/clustering.html#k-means (Accessed: July 3, 2023).

[20] Lozano, I.L., Sanchez-Hernandez, G., Guerrero-Rascado, J.L., Alados, I., Foyo-Moreno, I., 2022. Analysis of cloud effects on long-term global and diffuse photosynthetically active radiation at a Mediterranean site. Atmos. Res. 268, 106010. https://doi.org/10.1016/j.atmosres.2021.106010

[21] Mou, L., Saha, S., Hua, Y., Bovolo, F., Bruzzone, L., & Zhu, X. X. (2022). Deep Reinforcement Learning for Band Selection in Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14.

[22] NASA. (2022, October 12). NASA Dust Detective Delivers First Maps From Space for Climate Science. Retrieved from https://www.jpl.nasa.gov/news/nasa-dust-detective-delivers-first-maps-from-space-for-climate-science

[23] Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics SMC-9 (1), 62-66.

[24] Pham, T., Russell, R. (2022). A Program of Solar Radiation and Cloud Measurements. AUM Undergraduate Research Symposium, April 2022.

[25] Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, 66(336), 846-850. https://doi.org/10.1080/01621459.1971.10482356

[26] Resonon (2022), Product Catalog, Resonon Hyperspectral Imagining Solutions, https://resonon.com/Pika-XC2

[27] Sequential Feature Selector. Scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html (Accessed: July 3, 2023).

[28] Vales, R., Nguyen, B., Russell, R., Kursun, O. (2023). Impact of Feature Selection on Hyperspectral Image Segmentation. AUM Undergraduate Research Symposium, Montgomery, AL, April 2023.

[29] W. Zhang, X. Li, Y. Dou, and L. Zhao, "A geometry-based band selection approach for hyperspectral image analysis," IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 8, pp. 4318–4333, 2018.

[30] West, S.R., Rowe, D.,Sayeff, S., Berry, A., 2014. Short-term irradiance forecasting using skycams: Motivation and development. Solar Energy 110, 188-207. http://dx.doi.org/10.1016/j.solener.2014.08.038.

APPENDIX: CHAPTER 5 FIGURES

This appendix lists the figures that can be navigated by Tables 5.3 and 5.4 to follow the results obtained on hyperspectral image samples called R1 and R2, respectively.

Figure 5.8 below shows Ground truth image for four clusters. Which is well approximated by band-190 for the first pass Shown in Figure 5.9. which is then applied by the band-267 and band-163 for the second and third pass, The respective ARI scores are shown in Figure 5.10, Figure 5.12 and Figure 5.14.

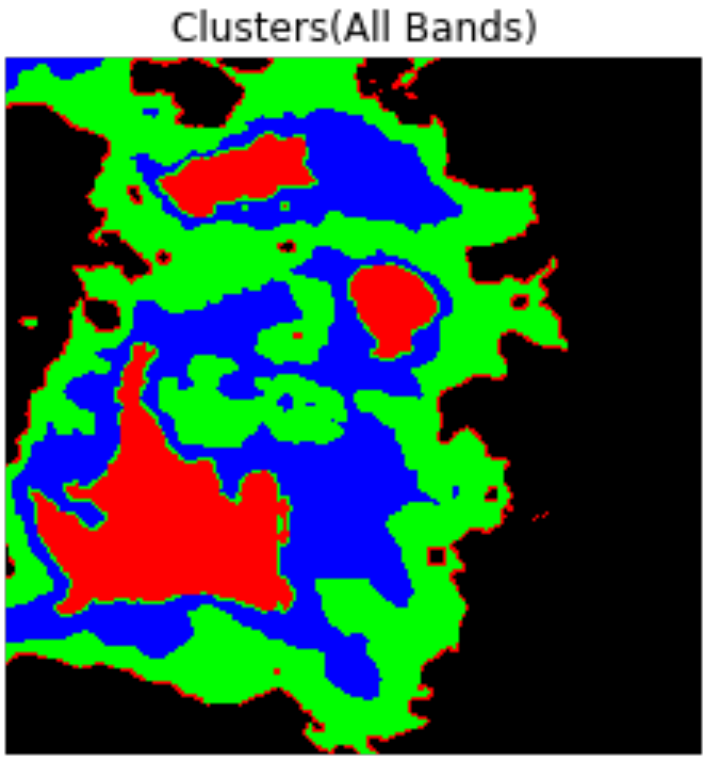

Figure 5.8 The 4-cluster ground truth for the R1-image.

Figure 5.9 The 4-cluster partitioning with band-190 on the R1-image.

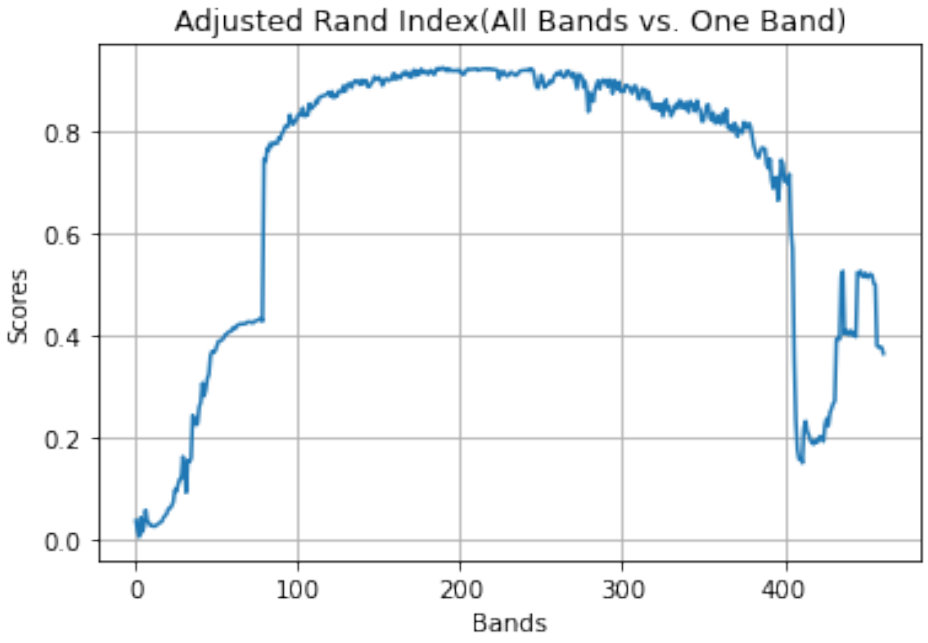

Figure 5.10 ARI-scores of the first pass for the 4-cluster feature selection on the R1-image.

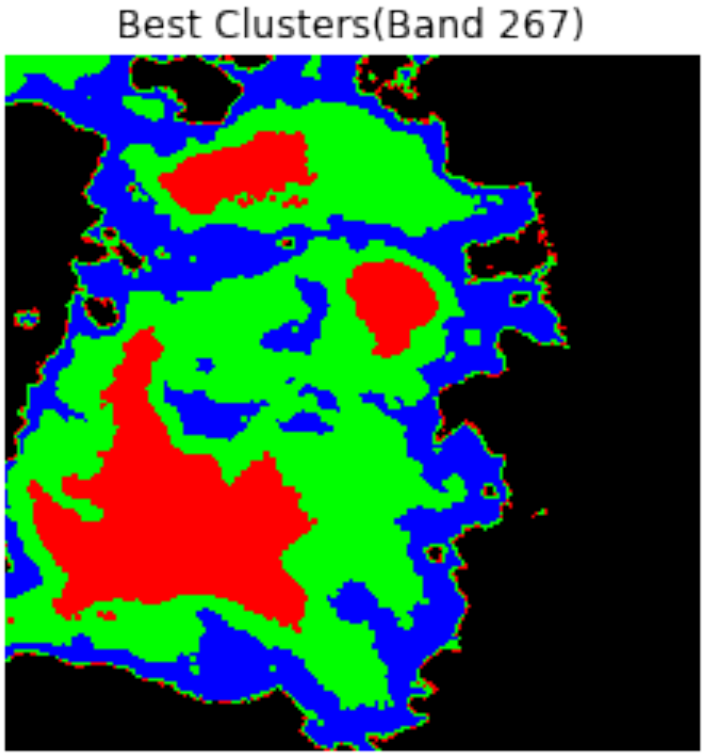

Figure 5.11 The 4-cluster partitioning with band-190 and band-267 on the R1-image.

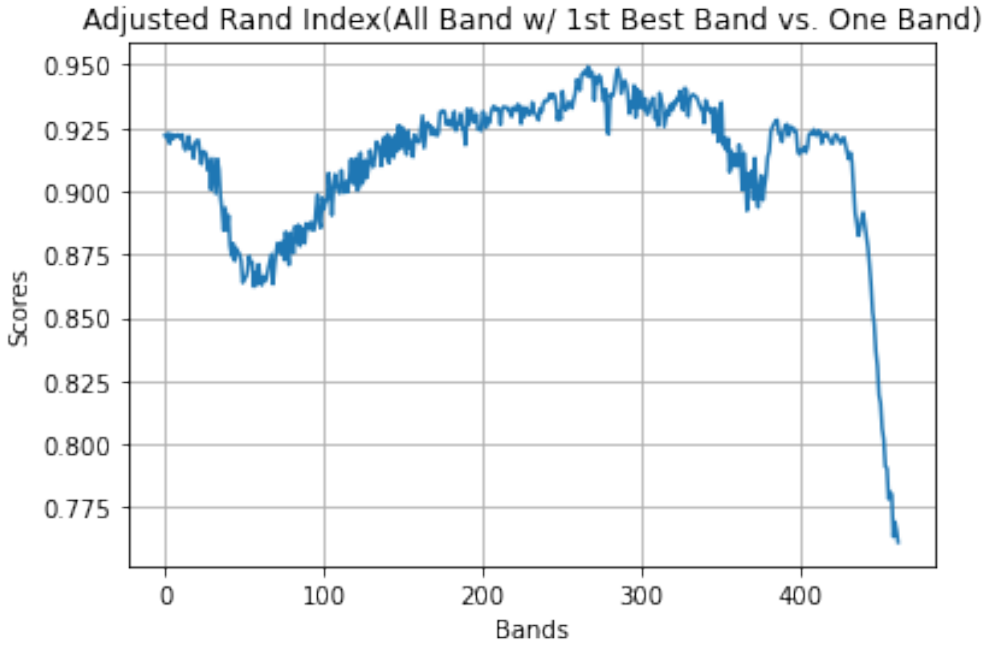

Figure 5.12 ARI-scores of the second pass for the 4-cluster feature selection on the R1-image.

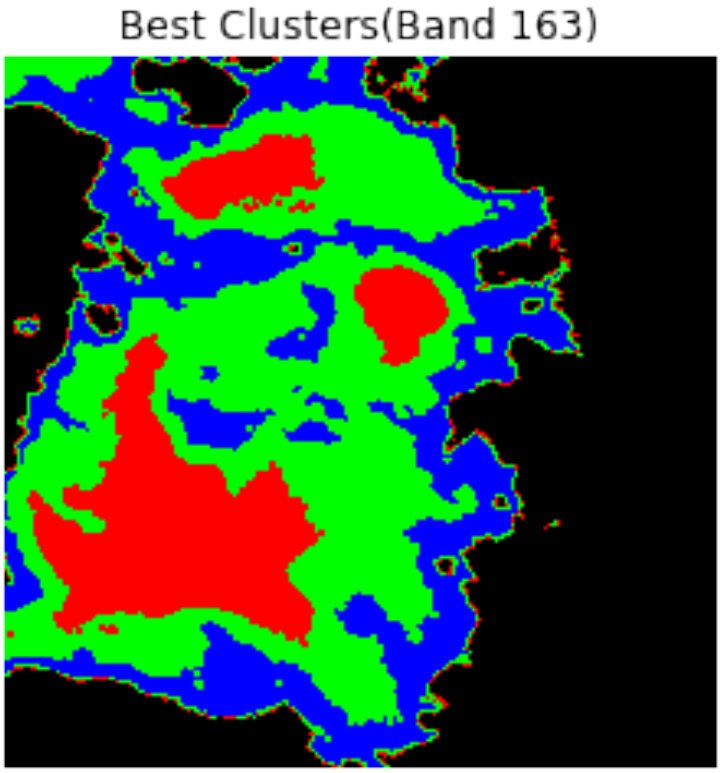

Figure 5.13 The 4-cluster partitioning with band-190 and band-267 and band-163 on the R1-image.

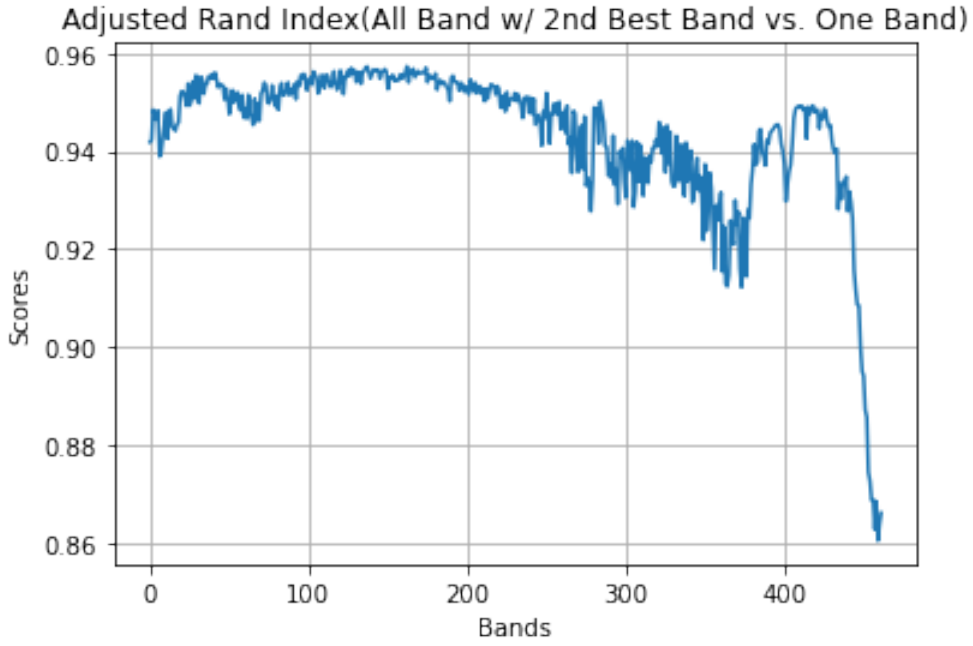

Figure 5.14 ARI-scores of the third pass for the 4-cluster feature selection on the R1-image.

Similarly, the 5-cluster ground truth is well approximated by the band-268, band-149, and band-193 and respective partitioning and the ARI plots are shown in the Figures below. Table 5.3 gives a short idea about the Figures for the clusters and scores on the R1-image.

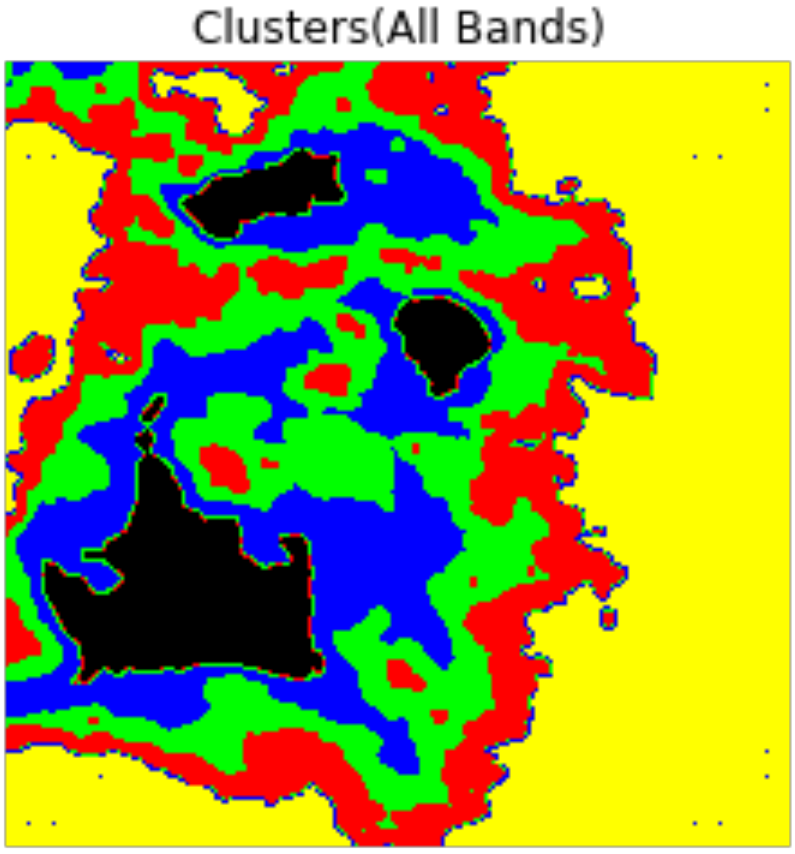

Figure 5.15 The 5-cluster ground truth for the R1-image.

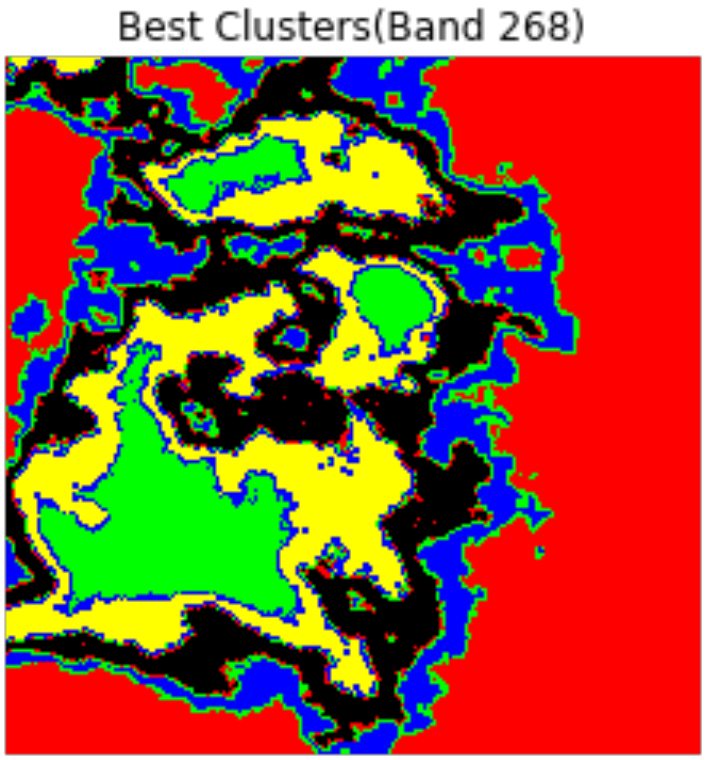

Figure 5.16 The 5-cluster partitioning with band-268 on the R1-image.

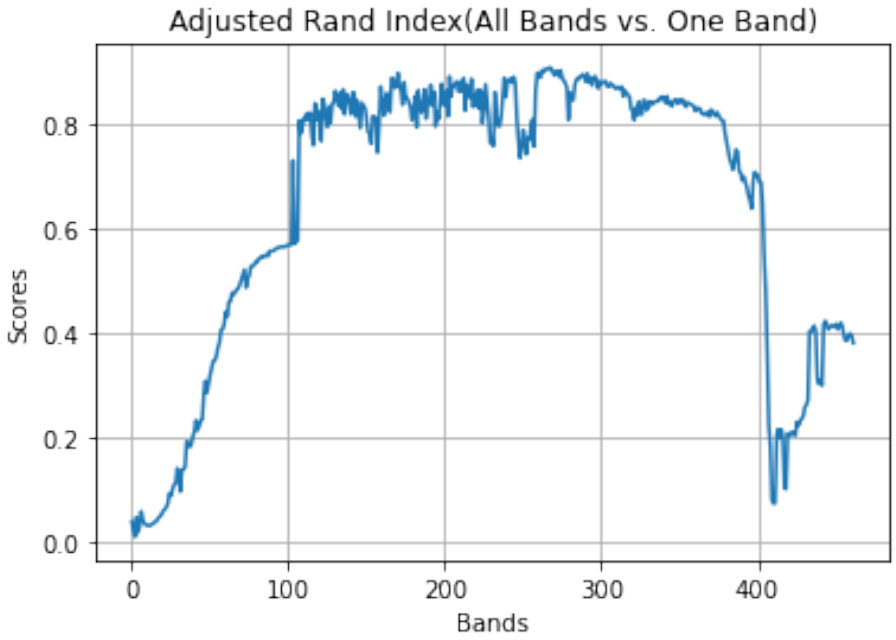

Figure 5.17 ARI-scores of the first pass for the 5-cluster feature selection on the R1-image.

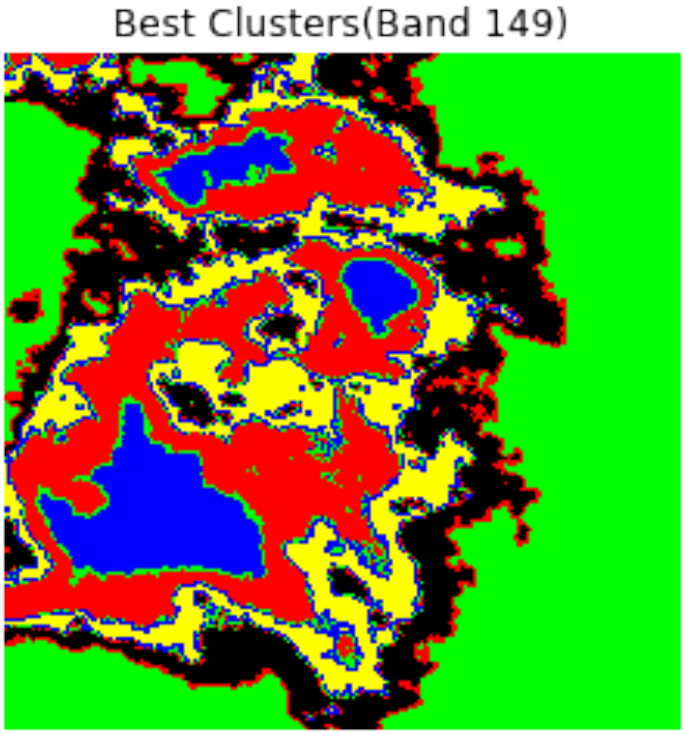

Figure 5.18 The 5-cluster partitioning with band-268 and band-149 on the R1-image.

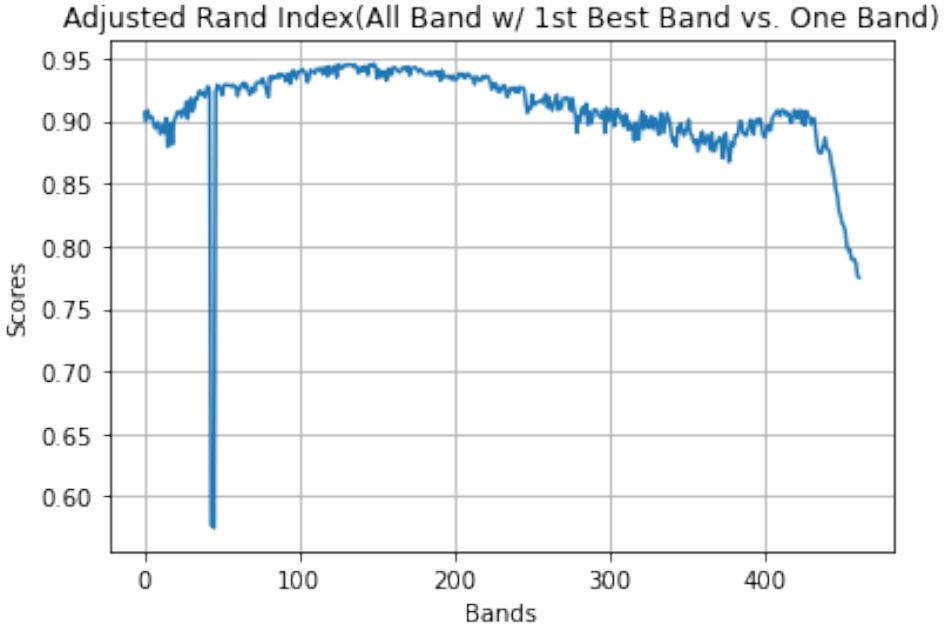

Figure 5.19 ARI-scores of the second pass for the 5-cluster feature selection on the R1-image.

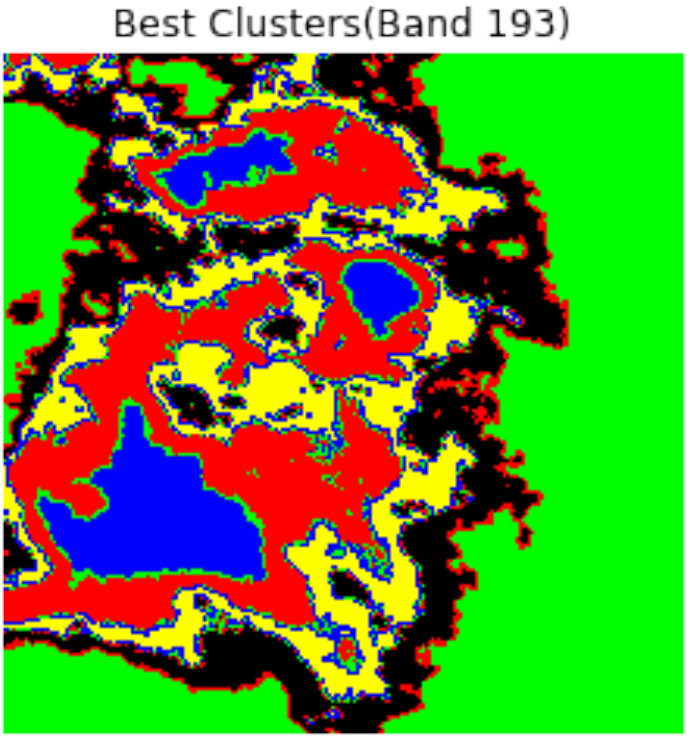

Figure 5.20 The 5-cluster partitioning with band-268, band-149 and band-193 on the R1-image.

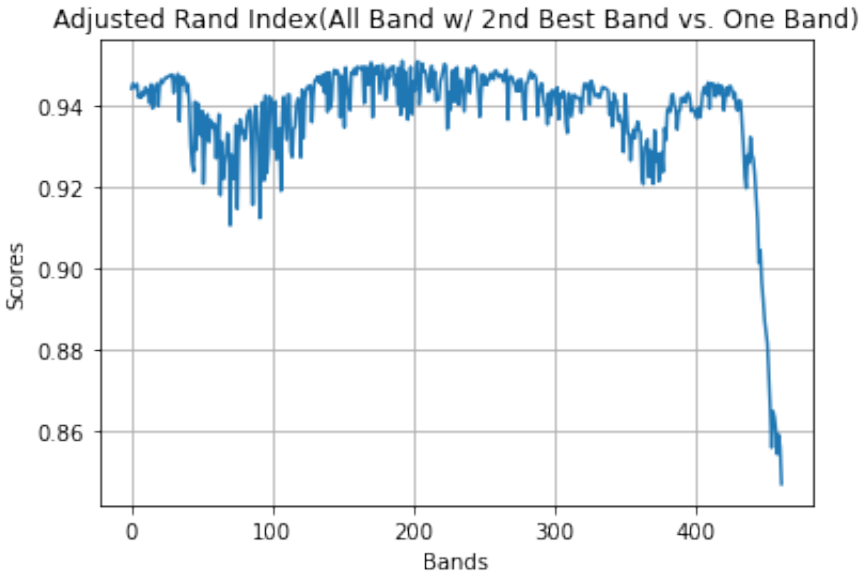

Figure 5.21 ARI-scores of the third pass for the 5-cluster feature selection on the R1-image.

The figures below from 5.22 to 5.28 show 3-cluster partitioning and the identified bands (band-118, band-133 and band-280) for the first, second and third pass and the respective ARI plots. Table 5.4 is the Navigation table for the Figures for the results on the R2-image.

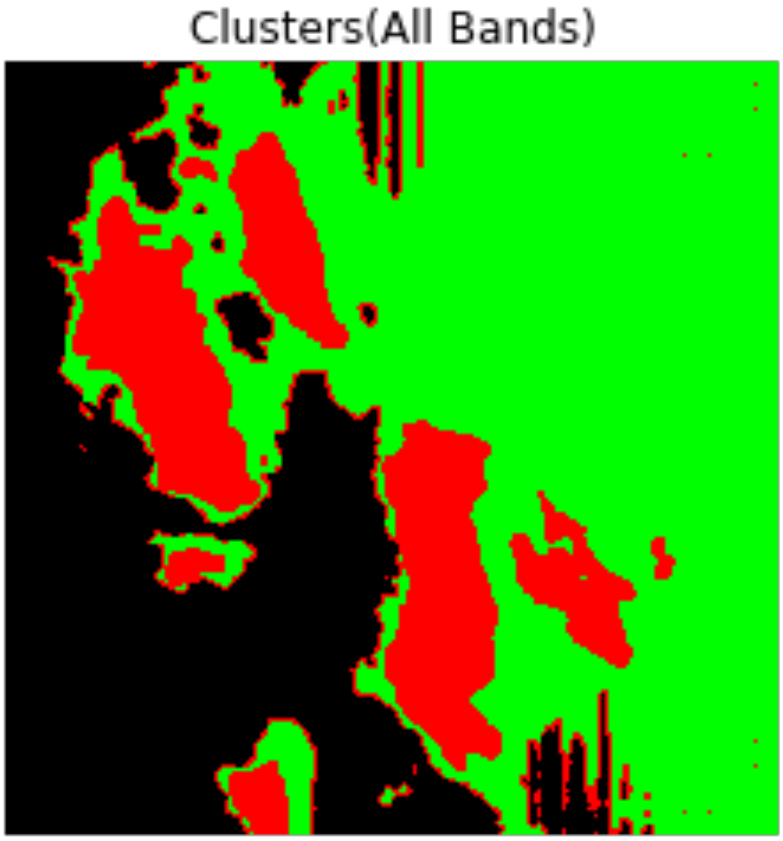

Figure 5.22 The 3-cluster ground truth for the R2-image.

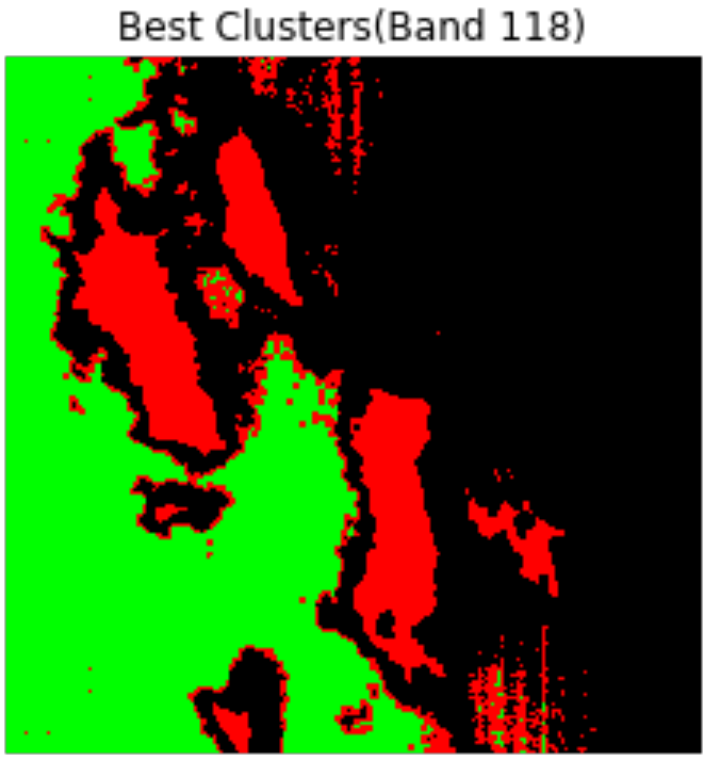

Figure 5.23 The 3-cluster partitioning with band-118 on the R2-image.

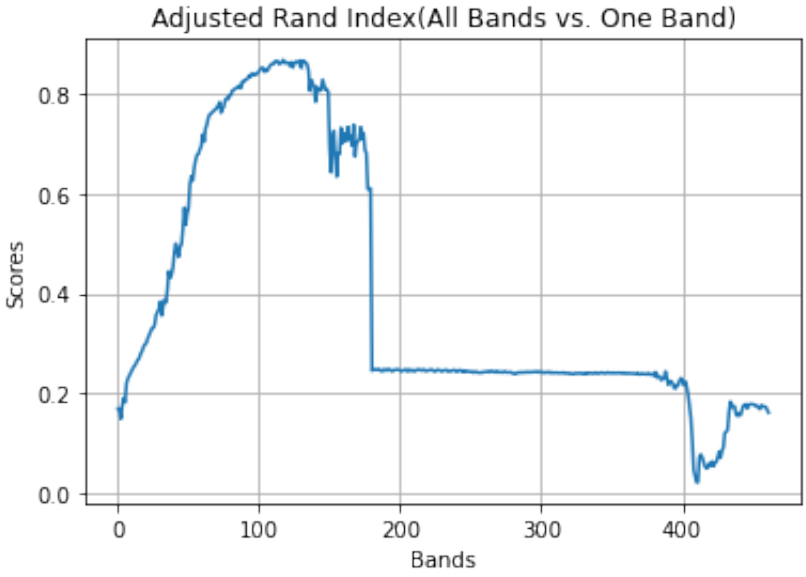

Figure 5.24 ARI-scores of the first pass for the 3-cluster feature selection on the R2-image.

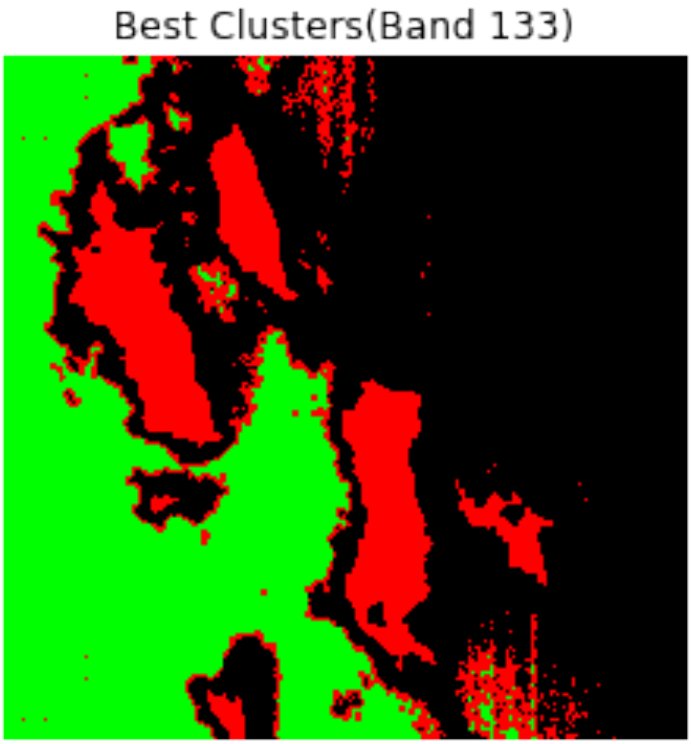

Figure 5.25 The 3-cluster partitioning with band-118 and band-113 on the R2-image.

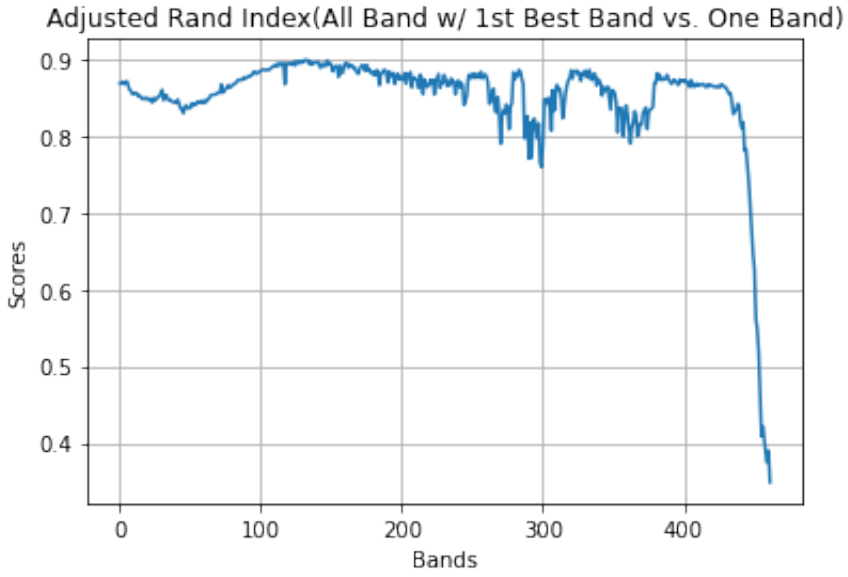

Figure 5.26 ARI-scores of the second pass for the 3-cluster feature selection on the R2-image.

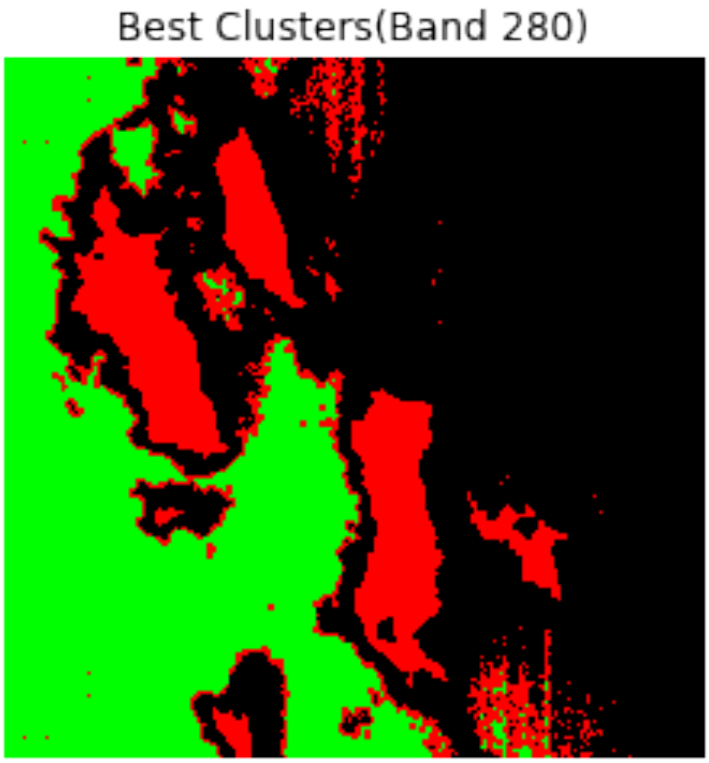

Figure 5.27 The 3-cluster partitioning with band-118, band-113 and band-280 on the R2-image.

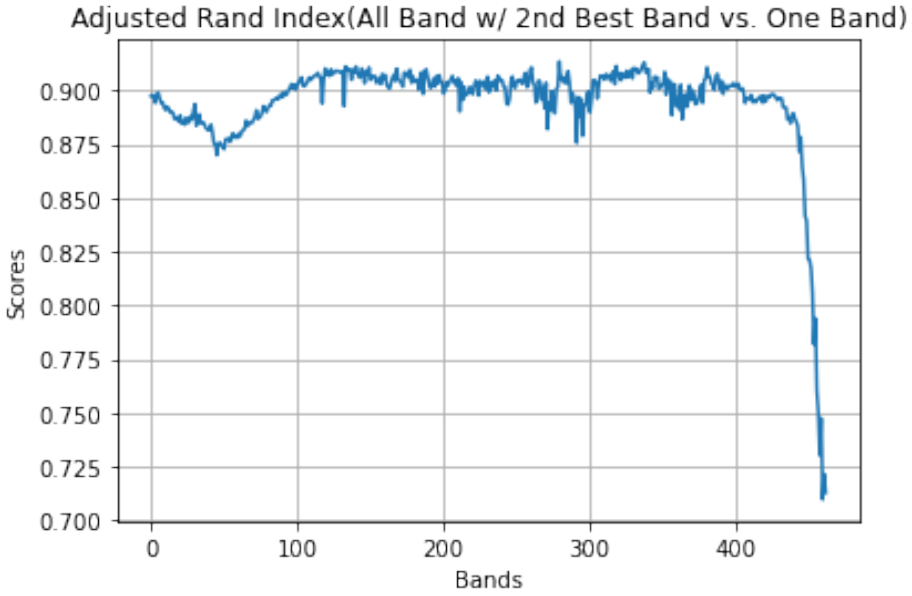

Figure 5.28 ARI-scores of the third pass for the feature selection on the R2-image.

The four-cluster ground truth and the partitioning with band-134 then bands-175 and 280 are shown in Figure 5.30, Figure 5.32 and Figure 5.34 and the ARI scores for the first, second and third pass are shown in Figures 5.31, 5.33, 5.35 respectively. We can see the ARI plot for the second pass in Figure 5.33 and the third pass in Figure 5.35 are quite similar.

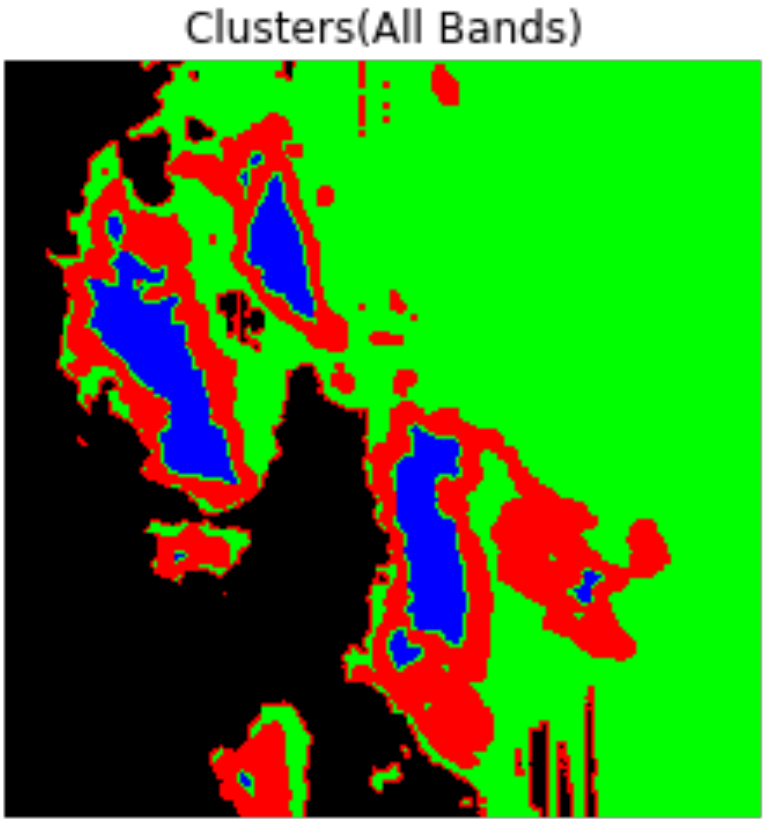

Figure 5.29 The 4-cluster ground truth for the R2-image.

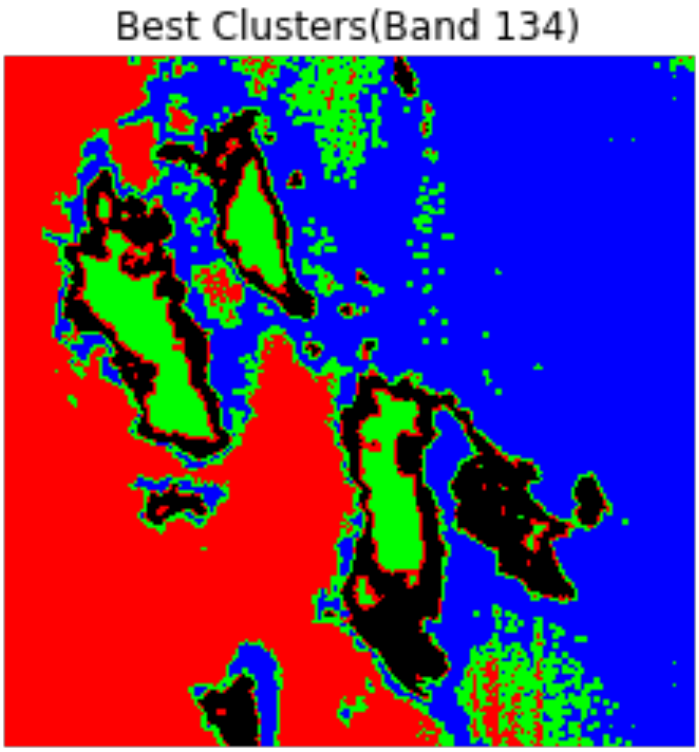

Figure 5.30 The 4-cluster partitioning with band-134 on the R2-image.

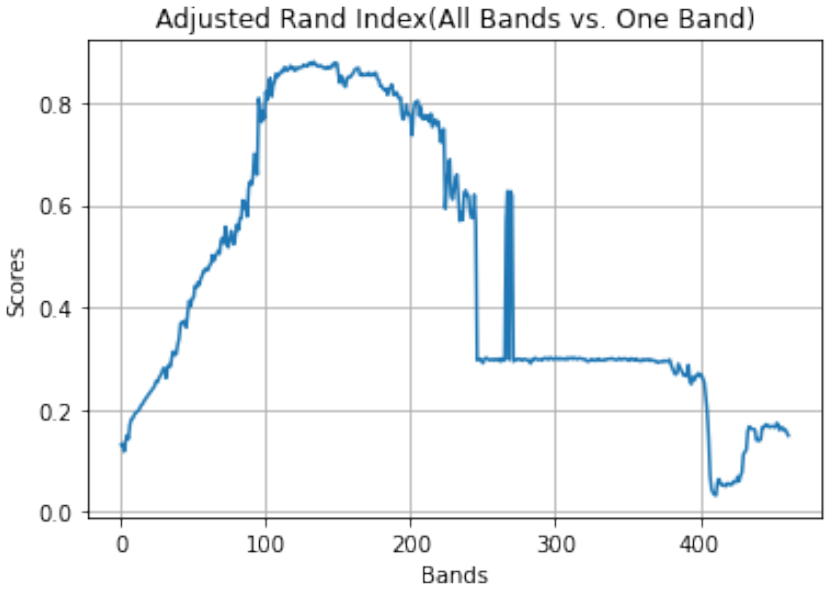

Figure 5.31 ARI-scores of the first pass for the 4-cluster feature selection on the R2-image.

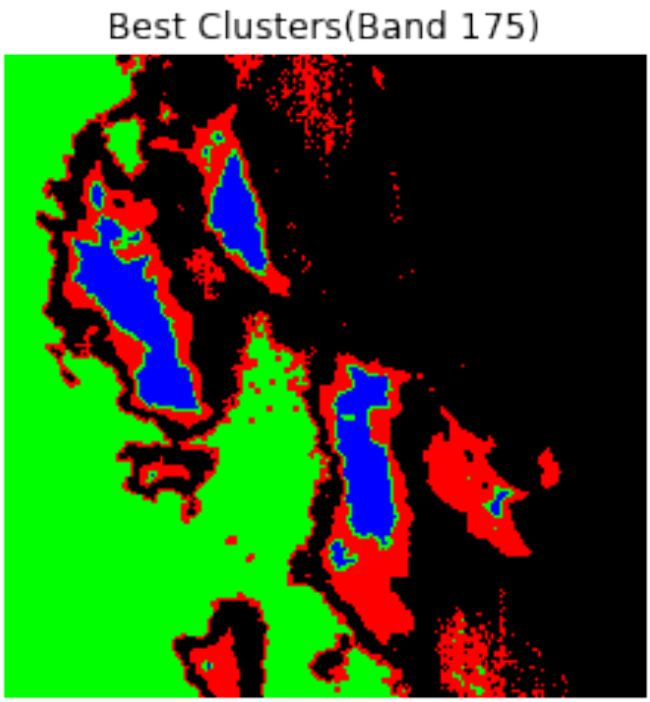

Figure 5.32 The 4-cluster partitioning with band-134 and band-175 on the R2-image.

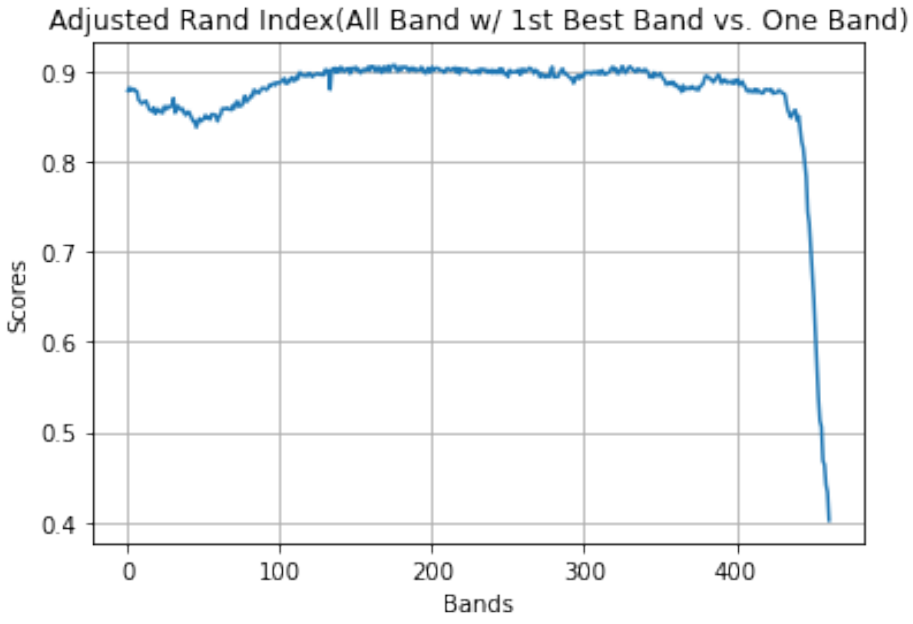

Figure 5.33 ARI-scores of the second pass for the 4-cluster feature selection on the R2-image.

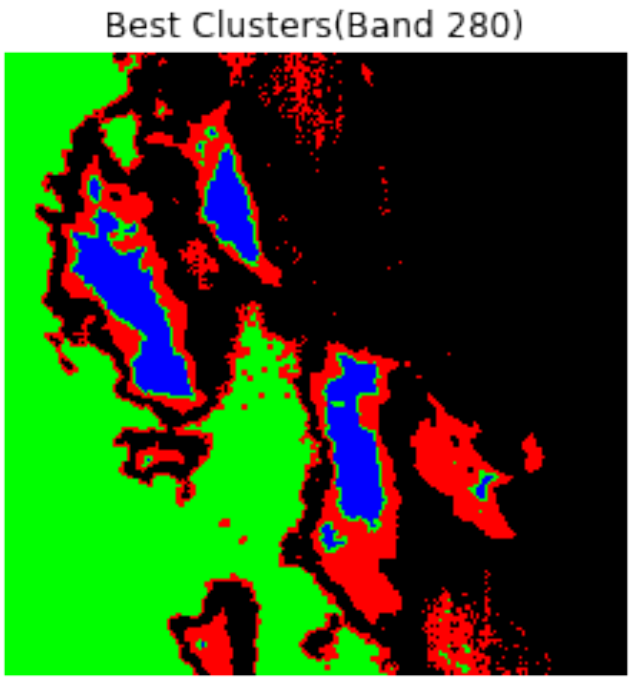

Figure 5.34 The 4-cluster partitioning with band-134, band-175 and band-280 on the R2-image.

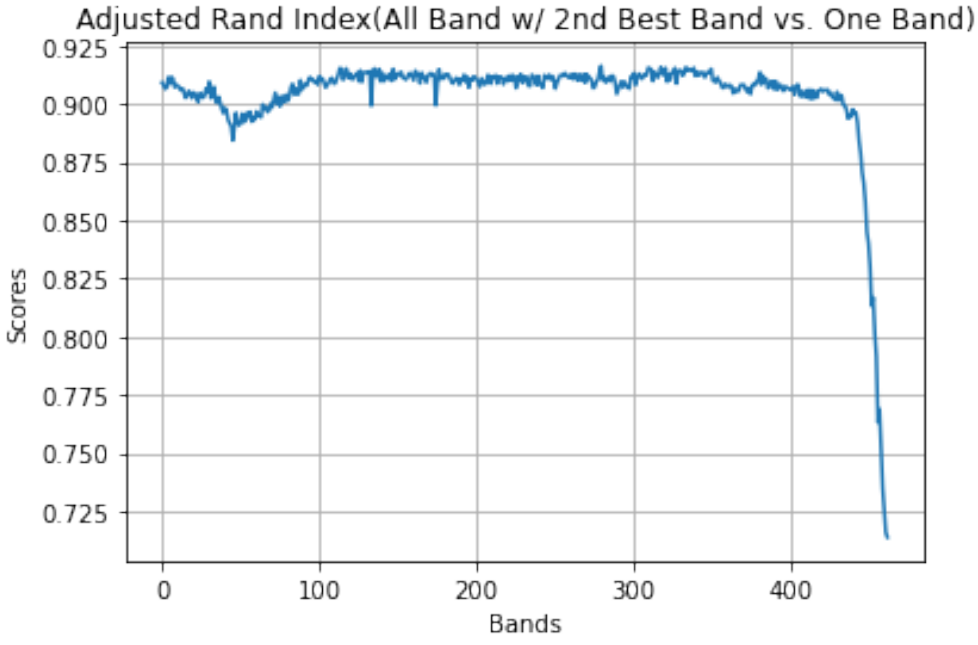

Figure 5.35 ARI-scores of the third pass for the 4-cluster feature selection on the R2-image.

Here, Figure 5.36 is a 5-cluster ground truth for R2-image, In the first pass band-175 is applied which is shown in Figure 5.37 the ARI plot for the first pass in Figure 5.38 shows gradual increase in the first 200 bands. Whereas in the Figure 5.39 and 5.41 when the band-178 and band-33 is applied for the second and third pass, the ARI plot in the Figure 5.40 and Figure 5.42 shows score decreasing at the beginning of the plot until 20 to 50 bands and shown a gradual rise and shown constant almost to end of the plot.

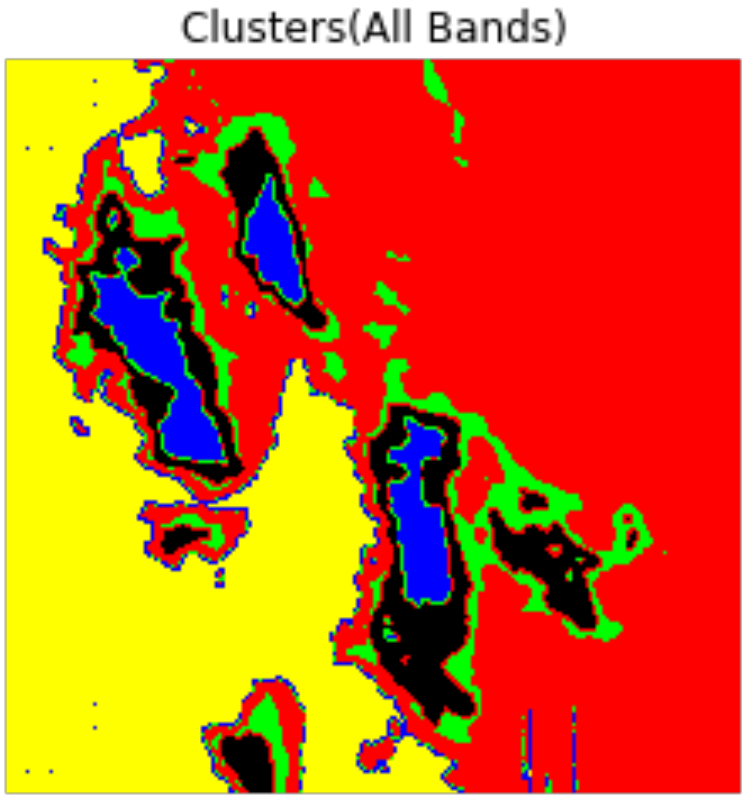

Figure 5.36 The 5-cluster ground truth for the R2-image.

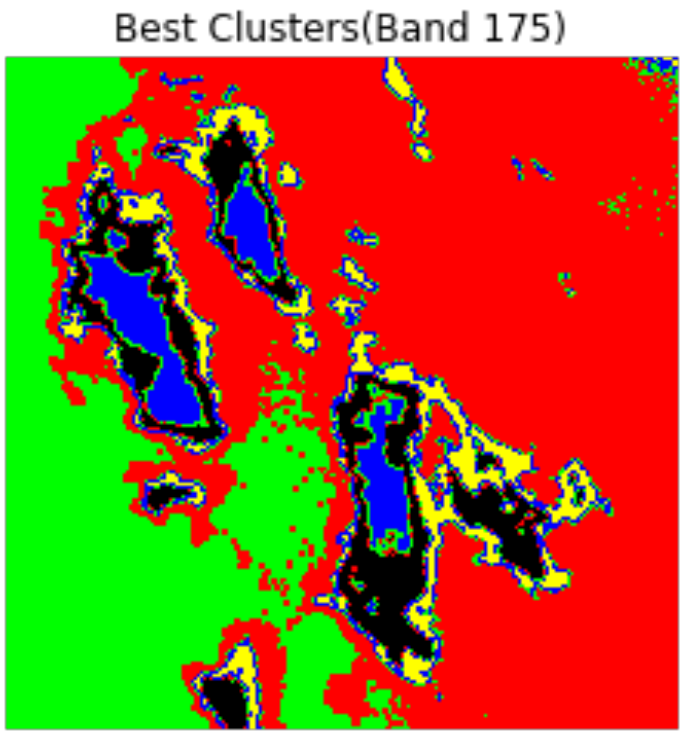

Figure 5.37 The 5-cluster partitioning with band-175 on the R2-image.

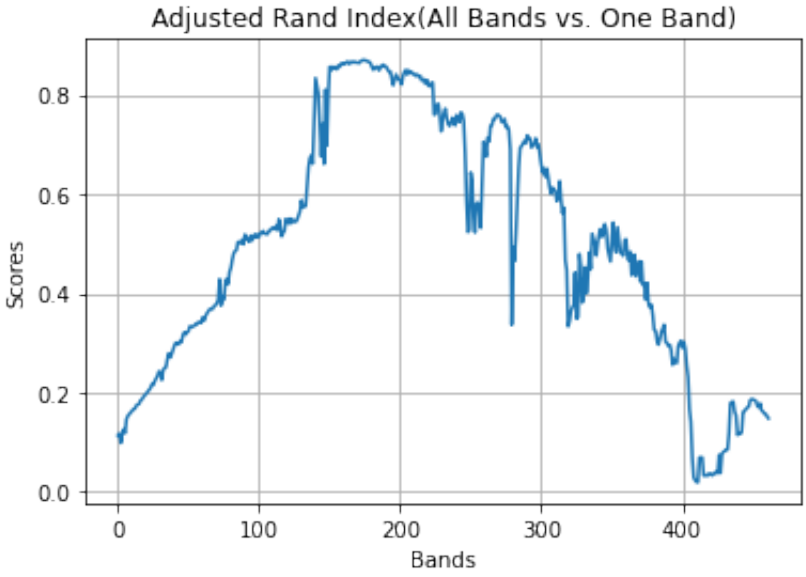

Figure 5.38 ARI-scores of the first pass for the 5-cluster feature selection on the R2-image.

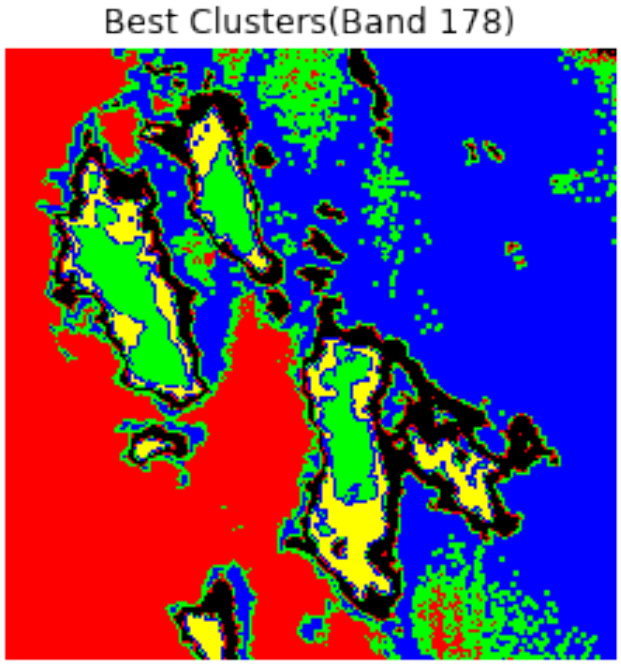

Figure 5.39 The 5-cluster partitioning with band-175 and band-178 on the R2-image.

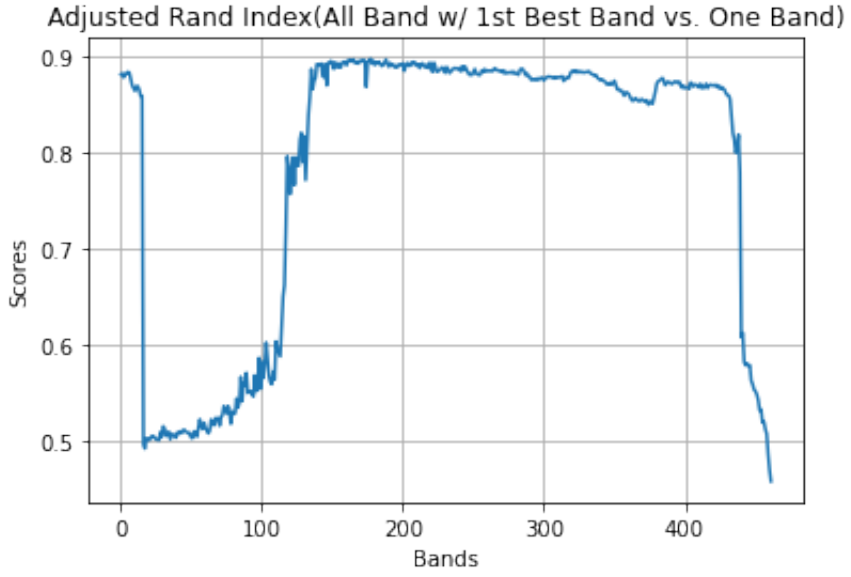

Figure 5.40 ARI-scores of the second pass for the 5-cluster feature selection on the R2-image.

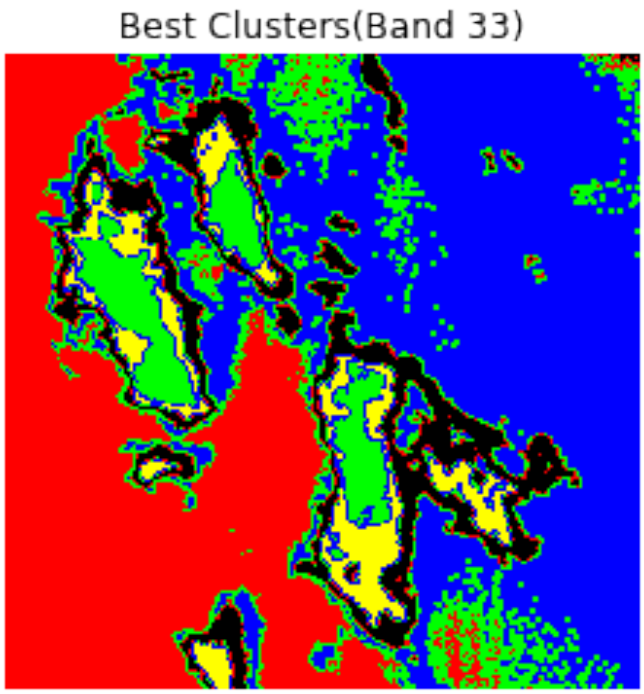

Figure 5.41 The 5-cluster partitioning with band-175, band-178 and band-33 on the R2-image.

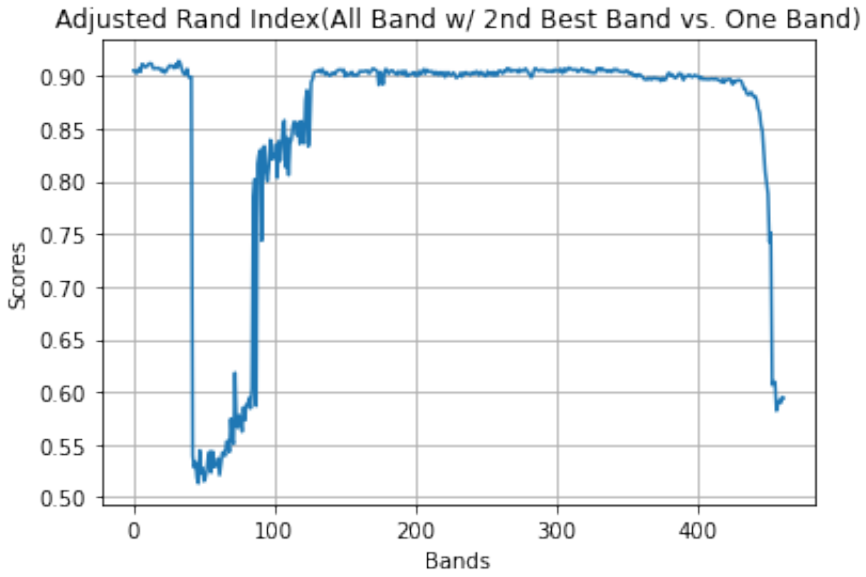

Figure 5.42 ARI-scores of the third pass for the 5-cluster feature selection on the R2-image.