



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

***Lithological mapping of Koutala island (Lavrio, Attiki) using
machine learning methods on multispectral data***

Κωνσταντίνος Τσαμκόσογλου

Επιβλέπων: Κουτρούμπας Κωνσταντίνος, Διευθυντής Ερευνών, ΙΑΑΔΕΤ, ΕΑΑ

ΑΘΗΝΑ

Φεβρουάριος 2024



NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

**SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS**

**INTERDEPARTMENTAL PROGRAM OF POSTGRADUATE STUDIES IN
DATASCIENCE AND INFORMATION TECHNOLOGIES**

MSc THESIS

***Lithological mapping of Koutala island (Lavrio, Attiki) using
machine learning methods on multispectral data***

Konstantinos Tsamkosoglou

Supervisor: Koutroumbas Konstantinos, Research Director, IAASARS, NOA

ATHENS

February 2024

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Lithological mapping of Koutala island (Lavrio, Attiki) using machine learning methods
on multispectral data

Κωνσταντίνος Τσαμκόσογλου

A.M: DS2.20.0008

ΕΠΙΒΛΕΠΩΝ: Κουτρούμπας Κωνσταντίνος, Διευθυντής Ερευνών, ΙΑΑΔΕΤ, ΕΑΑ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Κουτρούμπας Κωνσταντίνος, Διευθυντής Ερευνών, ΙΑΑΔΕΤ, ΕΑΑ

Πικράκης Άγγελος, Επίκουρος Καθηγητής, Τμ. Πληρ/κής, ΠΑΠΕΙ

Συκιώτη Όλγα, Κύρια Ερευνήτρια, ΙΑΑΔΕΤ, ΕΑΑ

Φεβρουάριος 2024

MSc THESIS

Lithological mapping of Koutala island (Lavrio, Attiki) using machine learning methods
on multispectral data

Konstantinos Tsamkosoglou

S.N: DS2200008

SUPERVISOR: **Koutroumbas Konstantinos**, Research Director, IAASARS, NOA

EXAMINATION COMMITTEE:

Koutroumbas Konstantinos, Research Director, IAASARS, NOA

Pikrakis Aggelos, Assistant Professor, Dept. of Informatics, UNIPI

Sykioti Olga, Senior Researcher, IAASARS, NOA

February 2024

Abstract

In recent decades, there has been a rapid advancement in the utilization of Earth Observation (EO) data in geology, driven by a growing interest in its application to identify potential sites associated with hydrothermal alteration and ore deposits. This development has garnered increasing attention due to its potential for substantial time and cost savings. In the present study, the target of interest is a small island called Koutala near the city of Lavrion (Attiki, Greece) and the aim is (a) to identify granitoid intrusions and schist formations on its surface and (b) to detect the associated alteration minerals. To this end, two high-resolution satellite datasets depicting the area of interest, taken from the Sentinel-2 and WorldView-3 missions, are utilized (the data sets differ in their spatial and spectral characteristics). Two different machine learning methods, namely clustering and spectral unmixing, were applied to extract geological information from the island.

Clustering was applied to both datasets to delineate regions with similar spectral signatures, aiming to identify granitoid and schist formations, as is referred on previous research insights [1]. In this framework, a novel clustering algorithm named SHC was introduced. SHC has been tailored especially for multispectral data. It takes advantage of the derivative of each pixel's spectral signature, and outperforms traditional off-the-shelf clustering algorithms, like K-means and hierarchical methods. The SHC algorithm demonstrated improved accuracy in identifying granitoid intrusion areas, especially in the challenging lower spatial resolution context of the Sentinel-2 dataset and in general yield to more homogeneous clusters (in terms of spectral characteristics).

Additionally, various linear spectral unmixing methods were explored in the Sentinel-2 dataset, taking into account its larger number of spectral bands and spectral positions compared to WorldView-3 data, to detect the associated alteration minerals on the surface of the island. Despite the dataset's relatively low spatial resolution for this type of study, alteration minerals with high probability of presence (having as reference previous search insights [1]) were accurately identified by most algorithms.

SUBJECT AREA: Machine learning on satellite data in geology

KEYWORDS: Clustering, Spectral-unmixing, Sentinel-2, WorlView-3 VNIR

Περίληψη

Τα τελευταίες δεκαετίες, υπήρξε μια γρήγορη πρόοδος στην επεξεργασία δορυφορικών δεδομένων Παρατήρησης της Γης στη γεωλογία, οδηγούμενη από το αυξανόμενο ενδιαφέρον για την εφαρμογή τους στην αναγνώριση πιθανών θέσεων που σχετίζονται με την υδροθερμική εξαλλοίωση και την παρουσία ορυκτών υδροθερμικής εξαλλοίωσης. Αυτή η μέθοδος έχει κερδίσει αυξανόμενη προσοχή λόγω της δυνατότητας που προσφέρει εξοικονόμησης χρόνου και πόρων. Στην παρούσα μελέτη, το αντικείμενο ενδιαφέροντος είναι το μικρό νησί που ονομάζεται Κουτάλα κοντά στην πόλη του Λαυρίου και ο στόχος είναι (α) η αναγνώριση γρανιτικών διεισδύσεων και σχιστόλιθου στο νησί και (β) η ανίχνευση των σχετικών ορυκτών εξαλλοίωσης. Για τον σκοπό αυτό, χρησιμοποιούνται δύο σύνολα υψηλής χωρικής ανάλυσης δεδομένων από τις δορυφορικές αποστολές Sentinel-2 και WorldView-3 που απεικονίζουν την περιοχή ενδιαφέροντος (τα δεδομένα διαφέρουν στα χωρικά και φασματικά χαρακτηριστικά τους). Δύο διαφορετικές μέθοδοι μηχανικής μάθησης εφαρμόστηκαν για την εξαγωγή γεωλογικών πληροφοριών από το νησί: η ομαδοποίηση (clustering) και ο φασματικός διαχωρισμός (spectral unmixing).

Το clustering εφαρμόστηκε και στους δύο τύπους δεδομένων με στόχο να ανιχνευτούν περιοχές με παρόμοιες φασματικές υπογραφές που αντιστοιχούν σε γρανιτικές διεισδύσεις και σχιστόλιθους, όπως έχει αναφερθεί σε προηγούμενες έρευνες [1]. Σε αυτό το πλαίσιο, ένας νέο αλγόριθμος clustering με την ονομασία SHC υλοποιήθηκε. Ο SHC έχει σχεδιαστεί ειδικά για πολυφασματικά δεδομένα. Εκμεταλλεύεται την παράγωγο της φασματικής υπογραφής κάθε εικονοστοιχείου (pixel) και υπερέχει των παραδοσιακών αλγορίθμων clustering, όπως ο K-means και οι ιεραρχικές μέθοδοι. Ο αλγόριθμος SHC επέδειξε βελτιωμένη ακρίβεια στην αναγνώριση περιοχών με γρανίτη, λαμβάνοντας υπόψη την σχετικά χαμηλή χωρική ανάλυση των δεδομένων Sentinel-2 για τέτοιου τύπου μελέτες και γενικά είχε ως αποτέλεσμα πιο ομοιόμορφα cluster (όσον αφορά τα φασματικά χαρακτηριστικά τους).

Επιπλέον, εξερευνήθηκαν διάφορες μέθοδοι spectral unmixing στα δεδομένα Sentinel-2, λαμβάνοντας υπόψη τον μεγαλύτερο αριθμό φασματικών καναλιών σε διαφορετικές

θέσεις σε σύγκριση με τα δεδομένα WorldView 3 VNIR, για την ανίχνευση των ορυκτών εξαλλοίωσης στο νησί. Παρά τη χαμηλή χωρική ανάλυση των δεδομένων Sentinel-2 για τέτοιου τύπου μελέτες, τα ορυκτά εξαλλοίωσης με υψηλή πιθανότητα παρουσίας στην επιφάνεια του νησιού αναγνωρίστηκαν με ακρίβεια από τους περισσότερους αλγορίθμους, με βάση προηγούμενες έρευνες [1].

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανική μάθηση σε δορυφορικά δεδομένα

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ομαδοποίηση, Φασματικός διαχωρισμός, Sentinel-2, WorldView-3 VNIR

ACKNOWLEDGMENTS

I am sincerely thankful for the guidance, support, and valuable insights provided by Dr Konstantinos Koutroumbas and Dr Olga Sykioti throughout my MSc thesis. Their motivation, their expertise, and the continuous support played crucial role in the completion of my MSc thesis.

TABLE OF CONTENTS

1. Introduction.....	14
2. Materials.....	16
2.1 Study area.....	16
2.2 Satellite Data.....	17
2.2.1 Sentinel-2.....	17
2.2.2 WorldView-3.....	19
3 Methods	21
3.1 The concept of the spectral signature.....	21
3.2 Continuum removal on reflectance spectra	21
3.3 Spectral signature derivative calculation.....	22
3.4 Fréchet distance between curves	23
3.5 Clustering algorithms	24
3.5.1 Partitional Algorithm - K-means	24
3.5.2 Hierarchical Algorithms	25
3.6 Unmixing spectral signature algorithms	29
3.6.1 Linear unmixing.....	29
3.6.2 Nonlinear unmixing	30
4. Methodology.....	31
4.1 Clustering	31
4.1.1 Sequential clustering.....	32
4.1.2 Hierarchical clustering.....	35
4.2 Spectral Unmixing	36
4.2.1 Endmembers definition	37
4.2.2 Linear Unmixing methods	38
5. Results.....	40
5.1 Identification of granitoid intrusions and schist formations (Clustering approach).....	40
5.1.1 Sentinel-2 dataset	40
5.1.2 WorldView-3 VNIR dataset	41
5.2 Detection for alteration minerals (Spectral unmixing approach).....	43
5.2.1 Sentinel-2 dataset	43
6. Discussion	46
6.1 Identification of granitoid and schist formations (Clustering approach)	46
6.1.1 Sentinel-2 dataset	47
6.1.2 WorldView-3 VNIR dataset	66
6.2 Detection for alteration minerals (Spectral unmixing approach).....	70
6.2.1 Reflectance spectra	70
6.2.2 Continuum-removed spectra.....	71

7. Conclusion	73
8. Data and code availability	75
References.....	75

LIST OF TABLES

Table 1: Spectral characteristics of Sentinel-2 and WorldView-3 VNIR data.....	17
Table 2: Reflectance and corresponding continuum removed ($1 - S_{cr}$) spectral signatures of the mineral endmembers used in this study.	38
Table 3: Overall methods used for the lineal spectral unmixing, along with their abbreviations.	39
Table 4: A sample with the best clustering results for the Sentinel-2 dataset.....	41
Table 5: A sample with the best clustering results for the WorldView-3 VNIR case.	42
Table 6: Spectral unmixing results for the Sentinel-2 reflectance spectra.....	44
Table 7: Spectral unmixing results for the Sentinel-2 continuum removed spectra case ($1 - S_{cr}$).....	45
Table 8: Sentinel-2 image granitoid pixels locations.....	48
Table 9: Spectral signatures of granitoid pixels for both reflectance and continuum-removed spectral values.....	48
Table 10: K-means - reflectance spectra - granitoid pixels vs non granitoid pixels at the clusters.	49
Table 11: K-means - reflectance spectra - granitoid pixels positions at the clusters.	49
Table 12: K-means - reflectance spectra – signatures of clusters containing granitoid pixels. Each distinct spectral signature within a cluster is represented by a unique color.	50
Table 13: Hier-Fréchet – reflectance spectra - granitoid pixels vs non granitoid pixels at the clusters.	51
Table 14: Hier-Fréchet – reflectance spectra - granitoid pixels positions at the clusters.	51
Table 15: Hier-Fréchet – reflectance spectra – signatures of clusters containing granitoid pixels. Each distinct spectral signature from the pixels is represented by a unique color.....	53
Table 16: SHC algorithm - reflectance spectra- granitoid pixels vs non granitoid pixels at the clusters.	53
Table 17: SHC algorithm - reflectance spectra- granitoid pixels positions at the clusters.	54
Table 18: SHC algorithm - reflectance spectra- signatures of clusters containing granitoid pixels. Each distinct spectral signature within a cluster is represented by a unique color.	56
Table 19: K-means - continuum removed spectra - granitoid pixels vs non granitoid pixels at the clusters.	57
Table 20: K-means - continuum removed spectra - granitoid pixels positions at the clusters.....	57
Table 21: K-means – continuum removed spectra – signatures of clusters containing granitoid pixels. Each distinct spectral signature within a cluster is represented by a unique color.	58
Table 22: Hier-Fréchet – continuum removed spectra - granitoid pixels vs non granitoid pixels at the clusters.	59
Table 23: Hier-Fréchet – continuum removed spectra - granitoid pixels positions at the clusters.	59
Table 24: Hier-Fréchet – continuum removed spectra – signatures of clusters containing granitoid pixels.	61
Table 25: SHC algorithm - continuum removed spectra - granitoid pixels vs non granitoid pixels at the clusters.....	62
Table 26: SHC algorithm - continuum removed spectra - granitoid pixels positions at the clusters.	62
Table 27: SHC algorithm – continuum removed spectra- signatures of clusters containing granitoid pixels.	64
Table 28: Signatures of clusters of all the algorithms in the WorldView-3 VNIR dataset.....	69
Table 29: Mineralogy of the lithologies present in the study area according to XRD analysis results on the four samples collected in the field..	70

LIST OF FIGURES

Figure 1: Location of the study area	16
Figure 2: True color composite of the Sentinel-2 image of the study area.	18
Figure 3: Masked, subset pseudo-color composition of the Sentinel-2 subset image of the Koutala islet.....	19
Figure 4: True color composite of the WoldView-3 VNIR image used in this study.....	20
Figure 5: Masked, subset true color composition of the WorldView-3 VNIR image.....	20
Figure 6: A pixel spectral signature example.....	21
Figure 7: Reflectance spectrum with the continuum and the continuum-removed spectrum....	22
Figure 8: Derivative of a Sentinel-2 spectral pixel.....	23
Figure 9: Example of discrete Fréchet distance.....	24
Figure 10: Example of a dendrogram.....	27
Figure 11: Linear unmixing method visualization.....	30
Figure 12: Flow chart of the SHC algorithm.....	32
Figure 13: Google Earth high resolution image of the island.	46
Figure 14: RGB-image (produced by the Sentinel-2 image) georeferenced to match the Google earth image.	47
Figure 15: WorldView-3 VNIR masked RGB image where the red squares represent the granitoid areas.....	66
Figure 16: Signatures of granitoid and non- granitoid pixels of the WorldView-3 VNIR image.....	67

1. Introduction

According to the United States Geological Survey “Remote sensing is the process of detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance (typically from satellite or aircraft). Special cameras collect remotely sensed images, which help researchers “sense” things about the Earth” [2]. According to [3], the benefits of the use of remote sensing are (among others) the ability “to collect information over large spatial areas; to characterize natural features or physical objects on the ground; to observe surface areas and objects on a systematic basis and monitor their changes over time; and the ability to integrate this data with other information to aid decision-making”.

Machine learning is the process of extracting information from the data in an automated way. It is a branch of the Artificial Intelligence field and, nowadays, it is used in almost any sector of the human activity. Machine learning algorithms, offer valuable capabilities for analyzing vast areas, including object classification, detection of temporal changes, data fusion, cloud removal, and spectral analysis using satellite or aerial imagery [4].

Machine learning has dynamically entered to the remote sensing area, in order to aid to the more effective and reliable processing of the huge amount of data gathered in various remote sensing contexts, most of them depicting the earth's surface. The essential aim of machine learning in the remote sensing framework is the recognition of patterns, by identifying/highlighting both more obvious and less obvious feature correlations in the data. This aids end-users in comprehending collected data and finding advanced solutions in solving problems related to natural environment (e.g., agricultural areas classification, lithological classification/identification).

Referring to satellite data, there are several types of them, such as the optical hyperspectral/multispectral¹ imaging systems (e.g. Sentinel-2, WorldView-3, ASTER, Landsat series, Hyperion, EnMAP). These datasets have different spectral and spatial resolutions, offering the potential to extract information about the composition and

¹ The main difference between multispectral and hyperspectral is the number of bands and the spectra of electromagnetic radiation that each band contains.

characteristics of various materials. [5]

The majority of the machine learning techniques that are used in remote sensing data, are applied on the image pixels.

Two famous machine learning techniques (that are also used in remote sensing) are clustering and spectral unmixing.

Clustering is the process of grouping more similar objects into the same group and less similar objects into different groups, according to a predetermined proximity measure [6]. The goal of applying clustering in remote sensing data is to identify homogenous areas in the image.

On the other hand, spectral unmixing relies on the assumption that the spectral signature of a specific pixel in a remote sensing image is a combination/mix of the (spectral signatures of the) materials that lie in the area of interest. The aim is to identify for each pixel in an image, the degree to which (the spectral signature of) each material contributes to the formation of (the spectral signature of) the pixel. [7]

The approach allows for a quantitative analysis of the materials met in the image.

The present study focuses on a geological application on a small island, called Koutala (Lavrio, Attiki, Greece), utilizing multispectral Sentinel-2 and WorldView-3 VNIR (Visible – Near Infrared) remote sensing data. The aim of this study is to investigate the capability of such data (a) to identify/discriminate granitoid intrusions and schist formations on the island and (b) to map related hydrothermal alteration minerals distributed on the surface of the island, using clustering and spectral unmixing methods respectively.

2. Materials

In this chapter, the study area for our application is first introduced. Then, some general information about the nature of the Sentinel-2 and WorldView-3 VNIR data utilized in this study is provided. In parallel, the Sentinel-2 and WorldView-3 VNIR images depicting the Koutala islet are also given.

2.1 Study area

The islet of “Koutala” is located about 5 km NNE of the city of Lavrio (Fig. 1). The islet has a form of a rocky promontory, forming a characteristic tombolo feature with the mainland (in coastal geomorphologic terms). Its size is about 240 m in E-W by 40- 60 m in the N-S direction. [1]

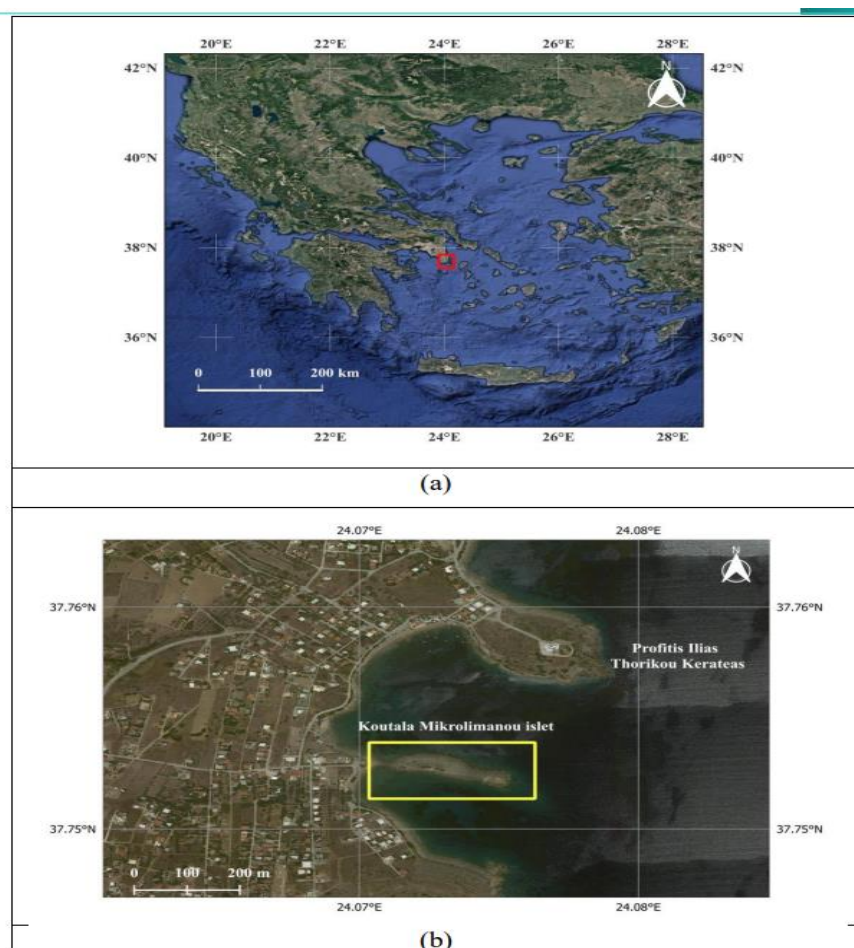


Figure 1: Location of the study area: (a) Lavrion area in Attica (Greece) (red rectangle); (b) Koutala islet in Lavrion area (yellow rectangle). Background image from Google Earth. (Source: [1]).

2.2 Satellite Data

A multispectral image comprises several image layers captured from the same scene, with each layer acquired within a specific wavelength band. [8]

In this study Sentinel-2 and WorldView-3 VNIR data were used. Table 1 presents the spectral characteristics of the two sensors.

Sentinel-2				Worldview-3 VNIR			
Band (Sb)	Centre (nm)	Width (nm)	Res. (m)	Band (Wb)	Centre (nm)	Width (nm)	Res. (m)
1	443	20	10	1	425	50	1.33
2	490	65	10	2	480	60	1.33
3	560	35	10	3	545	70	1.33
				4	605	40	1.33
4	665	30	10	5	660	60	1.33
5	705	15	10				
6	740	15	10	6	725	40	1.33
7	783	20	10				
8	842	115	10	7	832	125	1.33
8A	865	20	10				
9	940	20	10	8	950	180	1.33
11	1610	90	10				
12	2190	180	10				

Table 1: Spectral characteristics of Sentinel-2 and WorldView-3 VNIR data. For each spectral band, its center, width and resolution are provided.

2.2.1 Sentinel-2

Sentinel-2 mission provides high-resolution, multi-spectral imaging with a wide swath. It is designed to support Copernicus Land Monitoring initiatives, encompassing the assessment of vegetation, soil, and water coverage, in addition to the observation of inland waterways and coastal regions. The Sentinel-2 MultiSpectral Instrument (MSI) captures information in 13 spectral bands, from which four bands have a 10-meter spatial resolution, six bands have a 20-meter spatial resolution, and three bands have a 60-meter

spatial resolution. [9]

The image that we used in our study is a Sentinel-2 Level 2A (atmospherically corrected) image with 12 bands (dimensions) resampled to 10m acquired on 19 July 2022 (Fig. 2)

The image was subset to the area of interest with totally 832 pixels, while the pixels corresponding to the sea were masked. (Fig. 3). The number of unmasked pixels is 144 in total, and further processing is exclusively focused on them.



Figure 2: True color composite of the Sentinel-2 image of the study area.



Figure 3: Masked, subset pseudo-color composition of the Sentinel-2 subset image of the Koutala islet. Bands 2,3,4 were used to construct the pseudo color composition respectively.

2.2.2 WorldView-3

WorldView-3, owned by DigitalGlobe, is a commercial Earth observation satellite. It offers various imaging capabilities, including panchromatic imagery with a resolution of 0.31 meters (VNIR), eight-band multispectral imagery at 1.24 meters resolution (VNIR), shortwave infrared imagery at a resolution of 3.7 meters (SWIR), and provides CAVIS data (Clouds, Aerosols, Vapors, Ice, and Snow) at a resolution of 30 meters. [10]

In our study, the image used has 8 spectral bands in the Visible-Near infrared region of the E/M spectrum (VNIR) and 1.33m spatial resolution, and was acquired on 15 January 2022 (Fig.4).

This image underwent atmospheric correction and then subset to our specific area of interest with a total of 20496 pixels. As in the case of the Sentinel-2 image, all the pixels representing the sea were appropriately masked and excluded from subsequent analysis. The remaining pixels were 6510 in total after the masking. (Fig.5)



Figure 4: True color composite of the WorldView-3 VNIR image used in this study.



Figure 5: Masked, subset true color composition of the WorldView-3 VNIR image. Bands 4,3,2 were used to construct the color composition, respectively.

3 Methods

In this section the methods that were used in this study are described.

3.1 The concept of the spectral signature

An important concept in this type of applications is that of the spectral signature. A spectral signature (sometimes called pixel spectrum) refers to the fluctuation in reflectance exhibited by a material in different (consecutive) wavelengths. It essentially represents the reflectance variation as a function of wavelength [11] (Fig.6). Usually, it is depicted as a continuous line connecting consecutive band reflectance values.

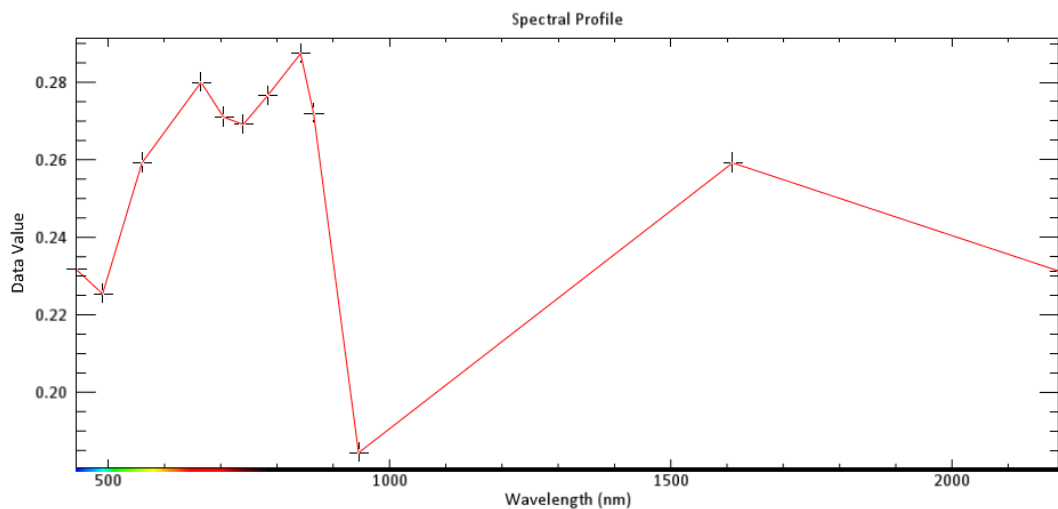


Figure 6: A pixel spectral signature example: black crosses correspond to the wavelength (x axis) and the corresponding reflectance value (y axis) of a Sentinel-2A image.

3.2 Continuum removal on reflectance spectra

The continuum removal method is a technique that standardizes reflectance spectra, enabling the comparison of individual absorption features from a consistent baseline. In this process, the initial and final spectral data values are set to 1.0, ensuring that the first and last bands in the resultant continuum-removed spectrum have this standardized value.

More specifically, for each image pixel, its continuum is removed by dividing the original spectrum with the continuum curve:

$$S_{cr} = \frac{S}{C}$$

where: S_{cr} = Continuum-removed spectrum, S = Original spectrum, C = Continuum curve (Fig. 7) [12]

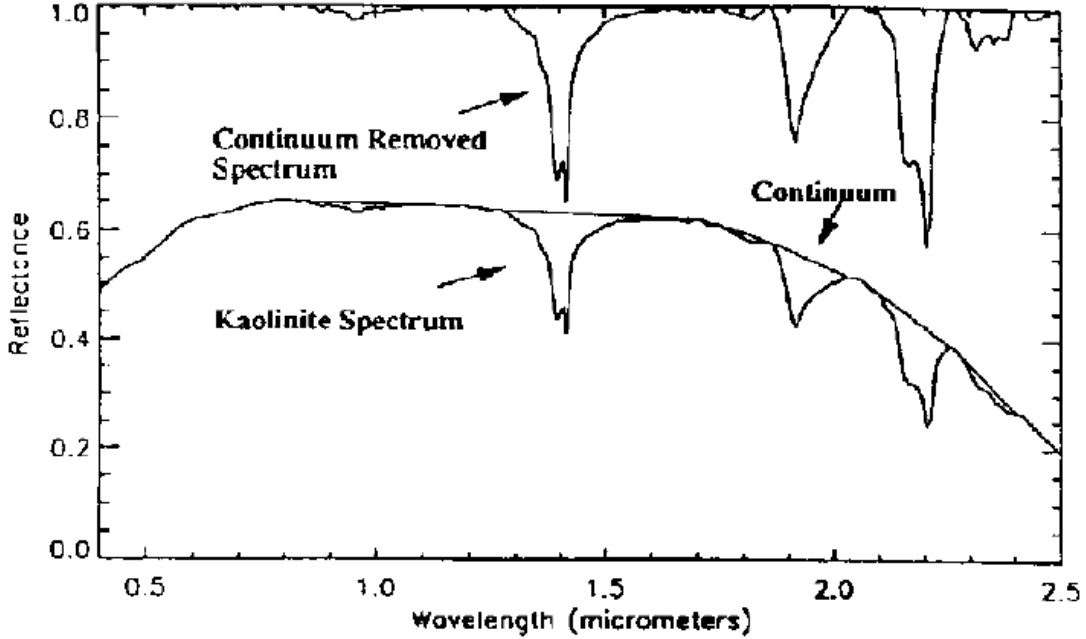


Figure 7: Reflectance spectrum with the continuum and the continuum-removed spectrum (Source: [12]).

In our case $1 - S_{cr}$, values are used so the first band and the last band have value zero.

3.3 Spectral signature derivative calculation

The derivative of a spectral signature is a vector that represents the rate of change of the reflectance value from one band to its next one. This can help us to recognize the rate of change of the reflectance values within a spectral signature. Among the various approaches that can be used to arithmetically approximate the derivative, in this work the derivative of an n -dimensional spectral signature vector $\vec{X} = (x_1, x_2, \dots, x_n)$ is approximated by the $(n - 1)$ dimensional vector (Fig.8) :

$$\vec{Y} = (x_2 - x_1, \dots, x_n - x_{n-1})$$

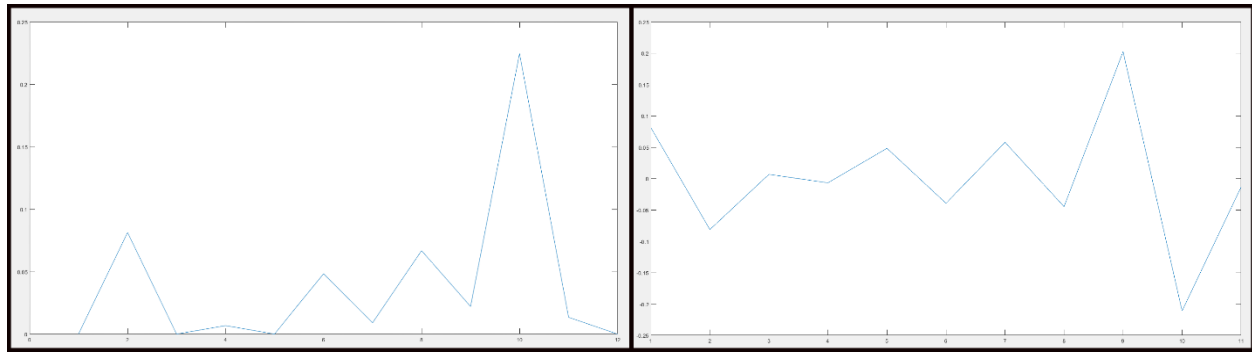


Figure 8 : Derivative of a Sentinel-2 spectral pixel. On left side of the figure a continuum removed spectral pixel ($1 - S_{cr}$) is shown and on the right side the respected derivative of this spectral pixel. On x-axis the number of band and on y-axis the respected reflectance value are presented.

3.4 Fréchet distance between curves

In the field of mathematics, the Fréchet distance is a metric for assessing the likeness between curves, considering both the arrangement and sequence of points along these curves. This distance metric is named in honor of Maurice Fréchet. An intuitive definition of the Fréchet distance is the following: An individual walks along a finite curved route, accompanied by their leashed dog, which follows a distinct finite curved path. Both the person and the dog can adjust their speeds to maintain some slack in the leash, but neither can reverse direction. The Fréchet distance between these two curves quantifies the length of the shortest leash necessary for both the person and the dog to complete their respective paths from beginning to end. It is important to note that this definition remains symmetric regardless of whether the dog is leading or following its owner.

The discrete Fréchet distance, sometimes referred to as the coupling distance, serves as an approximation of the Fréchet metric but is specifically tailored for polygonal curves. In the context of the discrete Fréchet distance, only the positions of the leash matter when its endpoints are positioned at the vertices of the two polygonal curves, never within the interior of an edge. This unique characteristic enables the computation of the discrete Fréchet distance using a straightforward dynamic programming algorithm, making it possible to calculate it in polynomial time. [13]

To visualize this concept figure 9 displays two polygonal curves, namely $[a_1, a_2, a_3]$ and $[b_1, b_2]$. We can identify two possible couplings between these curves:

$[b1\ a1, b2\ a2, b2\ a3]$ and $[b1\ a1, b1\ a2, b2\ a3]$. It's important to note that these couplings must adhere to the requirement that the endpoints of both polygonal curves coincide, respecting the order of the points and preventing backward movement.

The discrete Fréchet distance is determined by selecting the smallest of the maximum pairwise distances within these couplings. In the provided example, the maximum distance found in both couplings occurs at $b2\ a3$, which is equal to two units. Consequently, the minimum of these two maximum distances is also two units. [14]

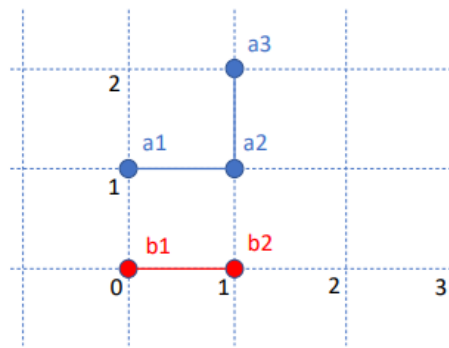


Figure 9: Example of discrete Fréchet distance.

3.5 Clustering algorithms

As it has been also stated in the introduction, the aim of a clustering algorithm is to assign more similar data vectors to the same group and less similar data vectors to different groups (in terms of a predetermined proximity measure). Clustering algorithms can be roughly categorized as either hierarchical or partitional. In hierarchical algorithms, clusters are built step by step, building upon previously formed clusters. On the other hand, partitional algorithms produce a single clustering for the data set of interest and (most of them) are less computationally demanding, compared to the hierarchical algorithms.

3.5.1 Partitional Algorithm - K-means

A celebrated paradigm of partitional algorithms that is used very often in practice (although its age exceeds the six decades) is the K-means algorithm. This algorithm represents each cluster with a representative vector (also called, representative, or

center, or centroid of the cluster) and its aim is to place each such representative to a region that is dense in data. Then, it assigns each data point to the cluster whose center, is closest to it, in terms of the squared Euclidean distance measure. It turns out that the centroid represents the average position of all the data points within the cluster. This means that for each dimension, the centroid's coordinates are calculated as the arithmetic mean of all the corresponding coordinates of the points in the cluster. The algorithm is described below:

Let $X = (x_1, x_2, \dots, x_n)$ be the set of data points and $V = (v_1, v_2, \dots, v_c)$ be the set of centers.

- Initialize randomly the c cluster centers.
- **(A)** Compute the distances of each data point from all the cluster centers.
- Assign each data point x_i to the cluster C_j whose center is closest to x_i .
- Reestimate the cluster center v_j of each cluster C_j using the formula:

$$v_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i$$

where, n_j is the number of data points in cluster C_j .

- Repeat from step (A), until no data point is reassigned to a different cluster. [6]

3.5.2 Hierarchical Algorithms

The hierarchical algorithms, produce sequentially a hierarchy of clusterings. They are further divided into agglomerative and divisive clustering algorithms.

Agglomerative clustering algorithms: In the case, the initial clustering consists of N clusters (each one containing a single data point) and the algorithm proceeds in the definition of the next clusterings, by merging at each level of the hierarchy the two most similar clusters, until the final clustering consisting of a single cluster (the whole dataset) is reached.

In more detail, the agglomerative hierarchical algorithms have the following steps:

1. **Initialization:** The initial clustering (0-th clustering level) consist of N clusters, each one containing a single data vector.

(A) At the t -th clustering level:

2. **Pairwise Distance Calculation:** Compute the distance between all pairs of clusters. This often involves the use of distance metrics like Euclidean distance or Manhattan distance, or another distance metric.
3. **Merging the Closest Clusters:** Identify the two clusters that are closest to each other and merge them into a single cluster.
4. **Updating the Distance Matrix:** After the merging, update the distance matrix² to include the distances between the newly formed cluster and each one of the remaining clusters. The method for updating distances depends on the chosen linkage criterion, such as single linkage, complete linkage, or average linkage that are described in detail below.
5. **Iteration:** Go to (A), until a predetermined stopping condition is met. This condition can involve achieving a specified number of clusters, reaching a distance threshold beyond which clusters are not merged, or another criterion tailored to the problem under study.
6. **Output:** The result of agglomerative clustering is typically represented as a dendrogram, a tree-like structure illustrating the sequence of cluster mergers (Fig.10). To obtain the desired number of clusters, one should cut the dendrogram at an appropriate level.

² The matrix that contains the distances of all pairs of clusters.

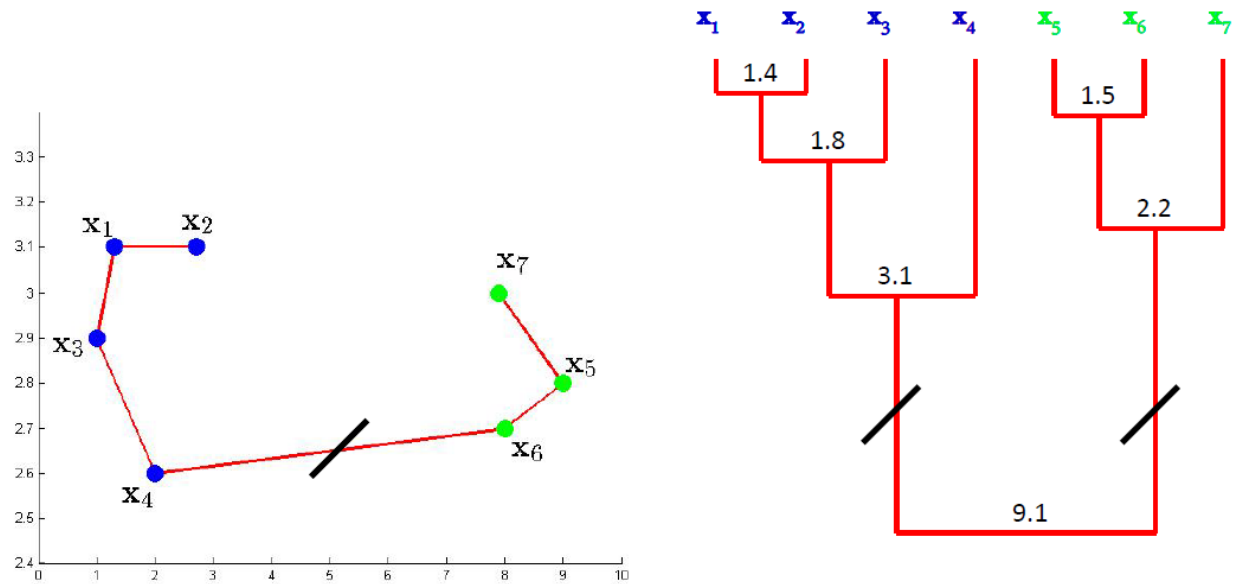


Figure 10: Example of a dendrogram. On the left side, the dataset is displayed. On the right side, the dendrogram is depicted along with the corresponding distances. The dendrogram is cut when the distance reaches 9.1, resulting in the formation of two clusters.

Divisive hierarchical clustering: Algorithms of this kind proceed in the opposite way compared to the agglomerative ones. They start with the single cluster clustering and proceed by dividing at each level the cluster with the smallest internal coherence.

The divisive hierarchal algorithms have the following steps:

1. **Initialization:** The initial clustering (0-th clustering level) consist of a single cluster, which is actually the whole data set.

(A) At the t -th clustering level:

2. **Pairwise Distance Calculation:** For each cluster, determine its partition to two sub-clusters, so that these sub-clusters to have the maximum possible dissimilarity (or minimum possible similarity).
3. **Cluster Splitting:** Among all the clusters at the current clustering level, select the one whose associated two sub-clusters exhibit the maximum possible dissimilarity and replace it with its two subclusters. Thus, the resulting clustering has now one cluster more than the previous clustering.

4. **Iteration:** Go to (A), until a predetermined stopping condition is met. This condition can involve achieving a specified number of clusters, reaching a distance threshold beyond which clusters are not merged, or another criterion tailored to your problem.

In hierarchical algorithms, a pivotal concept lies in determining how to calculate the distance between clusters as the algorithm progresses. Various methods have been devised to address this issue, with the most prevalent types being:

- **Maximum or complete linkage clustering:** The algorithm calculates all dissimilarities between each element in cluster 1 and every element in cluster 2, selecting the highest value (i.e., maximum) from these dissimilarities to represent the distance between the two clusters. This approach often leads to the formation of more compact clusters.
- **Minimum or single linkage clustering:** The algorithm calculates all pairwise dissimilarities between the elements in cluster 1 and those in cluster 2, choosing the smallest dissimilarity as the linkage criterion. This method often results in the formation of elongated, less compact clusters.
- **Mean or average linkage clustering:** The algorithm computes dissimilarities between all pairs of elements in cluster 1 and cluster 2, using the average of these dissimilarities as the measure of distance between the two clusters.
- **Centroid linkage clustering:** It calculates the dissimilarity between the centroid of cluster 1 and the centroid of cluster 2.
- **Ward's minimum variance method:** It aims to minimize the increase in variance within the newly formed cluster when two clusters are merged. This method is known for producing relatively balanced and compact clusters. [6]

3.6 Unmixing spectral signature algorithms

In both multispectral and hyperspectral imagery, the spectral signature of a single pixel usually corresponds to a mixture of reflectance spectra from multiple materials (endmembers), with the mixture coefficients (each one associated with a material) indicating the relative contribution of each constituent material to the formation of the pixel spectral signature. These coefficients offer insight into the abundances of the composing materials within the pixel.

3.6.1 Linear unmixing

The linear unmixing is based on the assumption that each mixed pixel is expressed as a linear combination of n endmembers weighted by their corresponding abundances. A spectral image of k pixels and b bands can be represented as a $b \times k$ matrix, whose columns are the spectral signatures of the pixels (the rows corresponding to the spectral bands), that is:

$$Y = (y_1, y_2, \dots, y_k) \in R^{b \times k}$$

Then (according to the linear mixing hypothesis)

$$Y = \theta \cdot W + E$$

where θ is a $b \times n$ matrix whose columns are the spectral signatures of the n materials, $W = (w_1, w_2, \dots, w_n)$ is a $n \times k$ matrix, whose j -th row is the abundance vector associated with the j -th pixel and E is a $b \times k$ matrix which represents the noise.

Linear unmixing typically involves three primary stages: first, estimating the number of endmembers; second, extracting the spectral signatures of these endmembers; and finally, estimating the abundances of these endmembers within each pixel (Fig.11) [15].

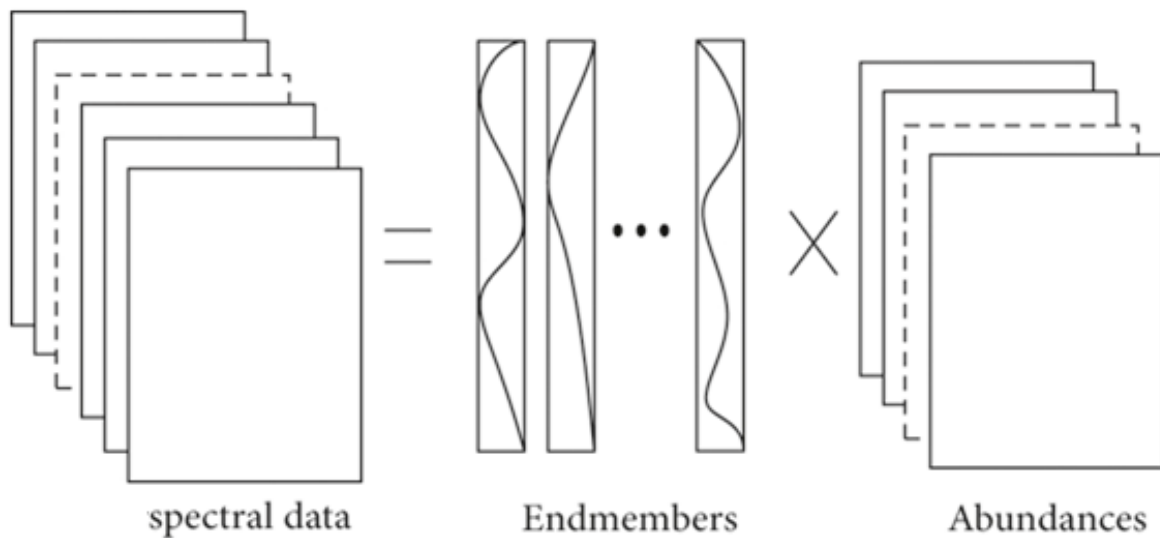


Figure 11: Linear unmixing method visualization. This method can be applied to both multispectral and hyperspectral data.

3.6.2 Nonlinear unmixing

The linear mixture model has demonstrated very good performance in situations where the Earth's surface exhibits extensive, well-defined regions with distinct endmembers. However, its effectiveness diminishes in scenarios characterized by intricate geometric structures and/or intimate mixtures. In such cases, incident light rays can interact with multiple pure materials within a pixel before reaching the sensor, resulting in reflectance spectra that are highly non-linear mixtures of the individual endmember reflectances. In such cases, the linear mixing hypothesis is not valid, and one should resort to nonlinear models. [16]

4. Methodology

In our geological application, we considered two distinct problems.

The first one has to do with the identification of homogeneous regions on the island (granitoid intrusions and schist formations). This problem has been tackled via a novel clustering approach to group the pixels associated to the island into clusters based on common pixel spectral signature characteristics.

The second problem has to do with the. In this case the linear spectral unmixing approach has been utilized.

4.1 Clustering

In our study, a novel clustering methodology has been developed, tailored to the specificity of the problem under consideration. In particular, the Sentinel-2 dataset consists of mixed pixel spectral signatures in a small area combined with significant spatial heterogeneity. Moreover, traditional clustering algorithms treat the pixel spectral signature as a whole and do not focus exclusively on specific spectral characteristics within the signature (e.g. absorptions) that are indicative of the presence of a specific material. In the sequel, a new methodology is presented, where the pixel spectral signatures are transformed, in order to better highlight the differences between different materials.

The algorithm comprises a two-step clustering procedure. First, a sequential algorithm is employed to cluster the pixels into c' groups based on their spectral derivatives. The resulting clustering by this algorithm is next fed to the second algorithm, which is of hierarchical nature. As is well known, the latter algorithm (as all hierarchical clustering algorithms) requires the calculation of the distances between any pair of clusters, C_q, C_r resulted from the sequential algorithm. To this end, the l_1 distances among the spectral derivatives of all possible pairs of pixels $(pixel_i, pixel_s), d_1(pixel_i, pixel_s)$, with $pixel_i \in C_q$ and $pixel_s \in C_r$ are calculated and the maximum of them defines the dissimilarity between C_q, C_r . The produced $c' \times c'$ (symmetric) distance matrix (whose (q, r) element is the distance between the clusters C_q and C_r) is fed to a hierarchical algorithm, along with the

desired number of final clusters, c . In our implementation the Ward algorithm has been used. Due to the two-step clustering approach in the algorithm, we called the algorithm Sequential Hierarchical Clustering (SHC) (see its flowchart in Fig.12)

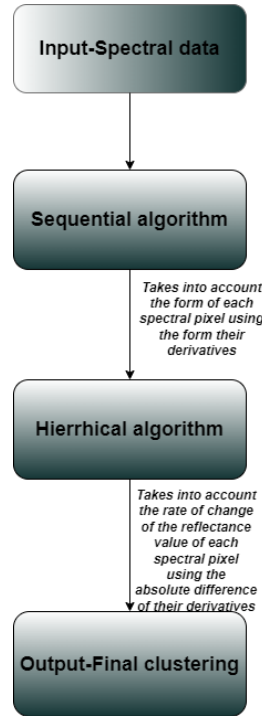


Figure 12: Flow chart of the SHC algorithm.

In the sequel, the proposed clustering methodology is described in detail.

4.1.1 Sequential clustering

The sequential clustering step of the SHC methodology takes a matrix A of size $b \times n$ as input, where n is the number of pixel spectral signatures of the pixels, and b is the number of bands characterizing each pixel spectral signature $\vec{x}_k = [x_1, \dots, x_b]^T$.

$$A = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{b,1} & x_{b,2} & \dots & x_{b,n} \end{bmatrix}$$

In this study n is the number of the unmasked pixels.

The first step in the SHC algorithm is to create the derivative matrix D of size $(b - 1) \times (n)$ from the A matrix. To create the derivative matrix, we approximate the derivative of each spectral pixel vector \vec{x}_k as the vector containing the differences between the values of

consecutive spectral bands. This, gives raise to the following derivative matrix:

$$D = \begin{bmatrix} x_{2,1} - x_{1,1} & \cdots & x_{2,n} - x_{1,n} \\ \vdots & & \vdots \\ x_{b,1} - x_{b-1,1} & \cdots & x_{b,n} - x_{b-1,n} \end{bmatrix}$$

After the calculation of the derivative matrix, the clustering process starts. The first vector from the derivative array is assigned to the first cluster. Each next point is assigned to one of the currently formed clusters, say C_j , if it has the same form of derivative with one of points belonging to C_j , otherwise a new formed cluster is created. The procedure continues sequentially for all the remaining points.

SHC Algorithm step-1 Sequential clustering

```

1:  $m = 1$  number of clusters
2: Assign the first point to the cluster  $C_m = \{\vec{x}_1\}$ 
3: for  $i = 2, \dots, N$  do
4:   for  $j = 1, \dots, m$  do
5:     if  $\vec{x}_i$  has the same spectral form with a  $\vec{x}_k \in C_j$  then
6:        $C_j := C_j \cup \{\vec{x}_i\}$ 
7:       break
8:     Else
9:       if  $j = m$  then
10:         $m = m + 1$ 
11:         $C_m = \{\vec{x}_i\}$ 
12:      end if
13:    end if
14:  end for
15:end for

```

To compare the spectral form between two spectral pixels \vec{x}_i, \vec{x}_j we utilize their spectral derivatives $D_i = [D_{i1}, \dots, D_{i,b-1}]^T$, $D_j = [D_{j1}, \dots, D_{j,b-1}]^T$.

The vectors \vec{x}_i and \vec{x}_j are considered spectrally similar if the respective values of D_{ik} and D_{jk} are of the same sign, for $k = 1, \dots, b - 1$. However, apart from the sign of D_{ik} and D_{jk} , their size should also be taken into account, since differences between near zero values are not considered as indication of dissimilarity. In the light of this observation, we consider that we have similarity in the following cases:

- Both D_{ik} and D_{jk} have the same sign and their sizes, $|D_{ik}|$ and $|D_{jk}|$, are “large”

(greater than a user-defined threshold, $threshold_1$).

- Both D_{ik} and D_{jk} have the same sign, their sizes, $|D_{ik}|$ and $|D_{jk}|$, are “small” (less than $threshold_1$) and their $l1$ distance is “small” (less than a user-defined threshold, $threshold_2$).
- The size of both D_{ik} and D_{jk} is “small” (less than a user-defined threshold, $threshold_3$).

If any other case occurs for any pair D_{ik} and D_{jk} , $k = 1, \dots, b - 1$, we consider that \vec{x}_i and \vec{x}_j are spectrally dissimilar.

Increasing the $threshold_1$ in the SHC algorithm results in more clusters, while increasing the $threshold_2$ and the $threshold_3$ results in fewer clusters.

The above rationale is summarized to the next pseudocode algorithm.

SHC Algorithm step-1 Spectral form similarity

```

1: spectral_similarity = true
2: for  $k = 1, \dots, b - 1$  do
3:   if ( $D_{ik} > threshold\_1$  and  $D_{jk} > threshold\_1$ )
4:     or ( $D_{ik} < -threshold\_1$ 
5:       and  $D_{jk} < -threshold\_1$ ) then
6:     spectral_similarity = spectral_similarity  $\wedge$  true
7:   Else if ( $0 < D_{ik} < threshold\_1$  and  $0 < D_{jk} < threshold\_1$  and
8:      $d_1(D_{ik}, D_{jk}) < threshold\_2$ )
9:     or ( $-threshold\_1 < D_{ik} < 0$  and  $-threshold\_1 < D_{jk} < 0$  and
10:     $d_1(D_{ik}, D_{jk}) < threshold\_2$ ) then
11:    spectral_similarity = spectral_similarity  $\wedge$  true
12:   Else if  $|D_{ik}| < threshold\_3$  and  $|D_{jk}| < threshold\_3$  then
13:     spectral_similarity = spectral_similarity  $\wedge$  true
14:   Else
15:     spectral_similarity = spectral_similarity  $\wedge$  false
16:
17:   end if
18: end for

```

4.1.2 Hierarchical clustering

The hierarchical clustering component of the SHC algorithm takes as input the c' clusters formed by the sequential algorithm, along with the labels of clusters to which each spectral signature belongs, and the desired final number of clusters, c . Each cluster, C_j , is represented as a matrix with dimensions $b \times n_j$, where n_j is the number of pixels within C_j , and b is the number of bands. The cluster labels are represented by a n -dimensional vector, so that its i -th position containing the label of the cluster to which the data vector x_i belongs.

The algorithm starts by calculating the distance matrix between the clusters taken as an input. To calculate the distance between two clusters, the algorithm utilizes the maximum $l1$ distance of the derivative vectors between all pair of pixels belonging to the respective clusters (see Box below).

SHC Algorithm step-2 Distance between clusters C_j and C_k , $j, k = 1, \dots, c'$

```

1: Calculate the derivative  $b \times n_j$  matrix  $D_j$  (the derivative vectors of the pixels in
clusters  $C_j$  are in the columns of  $D_j$ ).
2: Calculate the derivative matrix  $D_k$  with  $b \times n_k$  dimensions (the derivative vectors
of the pixels in clusters  $C_k$  are in the columns of  $D_k$ ).
3: Initialize  $(n_j \cdot n_k)$ dimensional vector  $P$  to zero
4:  $m = 1$ 
5: for  $i = 1, \dots, n_j$  do
6:   for  $s = 1, \dots, n_k$  do
7:     distance = 0
8:     for  $r = 1, \dots, b$  do
9:       distance = distance +  $|D_{ir} - D_{sr}|$ 
10:    end for
11:     $P(m) = \text{distance}$ 
12:     $m = m + 1$ 
13:  end for
14: end for
15:  $d(C_j, C_k) = \max P$ 

```

After constructing the distance matrix of size $c' \times c'$ (c' is the number of clusters resulted from the sequential algorithm), the SHC algorithm proceeds with the execution of a hierarchical algorithm taking into account the desired final number of clusters, c . Following the execution of the hierarchical algorithm, a c' -dimensional vector is returned as output

containing the new labels of clusters in which each cluster from the first clustering refers to (in total c different cluster labels). The final step of the algorithm involves creating the final clustering of the points by combining the vector with the cluster labels from the first clustering (sequential algorithm) and the vector with the cluster labels from the second clustering (hierarchical algorithm) (see Box below).

SHC Algorithm step-2 Combine clusterings to create the final clustering

```

1: Input: an  $n$ -dimensional vector  $S$  (its  $i$ -th element is the cluster label of the cluster
   where the  $i$ -th data vector has been assigned from the sequential clustering).
2: Input: A  $c'$ -dimensional vector  $H$  where  $c'$  is the number of clusters from the
   sequential clustering (its  $j$ -th element is the cluster label of the cluster where the  $j$ -
   th cluster resulted from the sequential algorithm. The value of the cluster label has
   been assigned from the output of hierarchical algorithm and there are in total  $c$ 
   different cluster labels)
3: Initialize the  $n$ -dimensional vector  $P$  to zero
4: for  $i = 1, \dots, c$  do
5:     Determine the positions  $j$  of  $H$  for which  $H(j) = i$  and accumulate their
     respective position indexes into a vector  $L$ 
6:     for  $j = 1, \dots, \text{size}(L)$  do
7:         Determine the positions  $q$  of  $S$  for which  $S(q) = L(j)$  and
         accumulate their respective position indexes into a vector  $W$ 
8:         for  $m = 1, \dots, \text{size}(W)$  do
9:              $P(W(m)) = i$ 
10:        end for
11:    end for
12: end for
13: end for

```

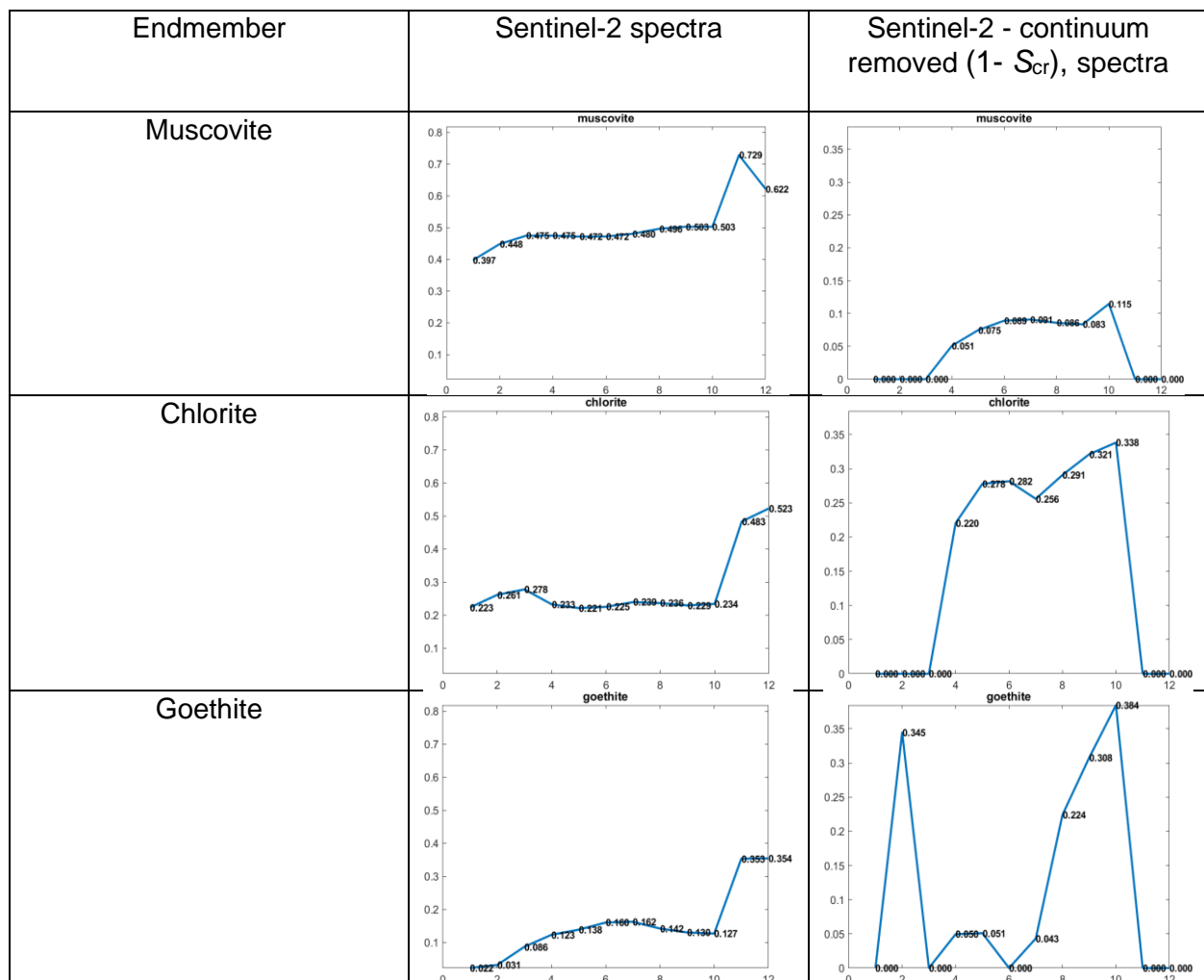
4.2 Spectral Unmixing

The second problem considered in this work, which has to do with the potential detection of alteration minerals, only established linear spectral unmixing methods were employed, to determine the mineral abundances in the pixels on the island.

Spectral unmixing was exclusively performed on the Sentinel-2 dataset due to the presence of a greater number of spectral bands, including two shortwave (SWIR) bands (11, 12). In a future work the spectral unmixing can be investigated into the World-View-3 dataset as well.

4.2.1 Endmembers definition

The endmembers used were selected from the USGS spectral library resampled to the Sentinel-2 spectral bands. The minerals selected as endmembers are (i) muscovite, (ii) chlorite, (iii) goethite, and (iv) pyrochroite. This choice is based on prior research, which provides evidence of the presence of these minerals on the island [1]. The table below illustrates the (a) reflectance and (b) the continuum removed spectral signatures of the four endmembers that have been resampled to the Sentinel-2 spectral bands.



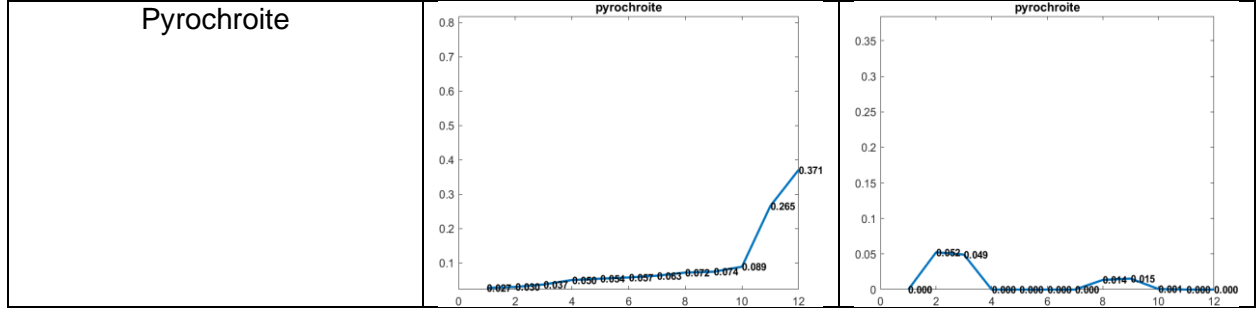


Table 2: Reflectance and corresponding continuum removed (1- S_{cr}) spectral signatures of the mineral endmembers used in this study.

4.2.2 Linear Unmixing methods

Given that each pixel in spectral image $\vec{y} = [x_1, \dots, x_b]^T$ can be described as $\vec{y} = \sum_{i=1}^p w_i * \vec{\theta}_i$ where $\vec{\theta}_i$ represents the i -th endmember, w_i is the abundance of each endmember, b is the number of bands and p is the number of the endmembers our objective is to calculate the abundance values w_i for each pixel.

Using the least squares cost function, we can model the problem using the formula where n is the number of the total pixels:

$$J = \sum_{i=1}^n (\vec{y}_i - \sum_{j=1}^p w_j * \vec{\theta}_j)^2$$

where J should be minimized respect to the w_j, \dots, w_p for every \vec{y}_i

The solution of this problem can be expressed using matrices with the formula:

$$W = (\Theta^T * \Theta)^{-1} \Theta^T * Y$$

where:

- Θ is a $b \times p$ matrix containing all the θ endmembers vectors.
- Y is $b \times n$ matrix containing all the y pixel vectors.
- W is $p \times n$ matrix containing the abundances values for each pixel.

Due to the specific characteristics of the problem, the abundance values are expected to be greater or equal than zero, and their sum should equal 1. In the case of a least squares solution, any abundance values that turn out to be negative are typically modified to be

zero. This adjustment ensures that the abundance values remain non-negative, adhering to the constraints imposed by the problem's nature.

By imposing constraints on the objective function J to ensure that abundances sum to one and are non-negative, the problem can be solved effectively using iterative methods. More specifically, the interior-point optimization algorithm [17] of MATLAB function `fmincon` was used to solve the problem.

Finally, to mitigate the risk of overfitting, the problem can also be addressed using Lasso regularization [18]. This regularization technique helps prevent overfitting by incorporating a penalty term into the optimization formula:

$$J = \sum_{i=1}^n (\vec{y}_i - \sum_{j=1}^p w_j * \vec{\theta}_j)^2 + \lambda \sum_{j=1}^p |w_j|$$

The problem was solved using the lasso MATLAB function and the selection of the λ value was guided by the condition that the solution maintains abundance values greater or equal than zero while ensuring that the norm of the abundance values is maximized among various choices of λ . This approach helps strike a balance between regularization to prevent overfitting and retaining physically meaningful abundance values during the spectral unmixing process.

In the following table the methods used for the unmixing process are summarized.

Method	Constraint	Abbreviation
Least squares	No constraint	U-LS
Least squares	Non negativity constraint	N-LS
Least squares	Sum to 1 constraint	S-LS
Least squares	Non negativity constraint and Sum to 1 constraint	NS-LS
Lasso	l1-norm	LASSO

Table 3: Overall methods used for the lineal spectral unmixing, along with their abbreviations.

5. Results

In this chapter the results from the applied machine learning methods (clustering and spectral unmixing) on the two problems ((a) identification and mapping of granitoid intrusions and schist formations and (b) detection for alteration minerals in Koutala islet will be demonstrated.

5.1 Identification of granitoid intrusions and schist formations (Clustering approach)

In this section, the results generated by the SHC algorithm will be presented, encompassing both the reflectance and continuum-removed spectral cases. For benchmarking purposes, the results obtained from the K-means algorithm and the hierarchical complete link algorithm using the Fréchet distance as distance metric will also be demonstrated. Given the large volume of results, only a subset of the results will be presented and highlighted.

5.1.1 Sentinel-2 dataset

The SHC algorithm was executed with specific threshold values as follows:

- For the reflectance spectra case, *threshold_1* and the *threshold_2* were set to 0.005, and the *threshold_3* was set to 0.002.³
- In the continuum removal spectra case, *threshold_1* was assigned a value of 0.004, and *threshold_2*, *threshold_3* were set to 0.002.

For both the reflectance and continuum removal spectra cases, the K-means algorithm was executed 1000 times, each time with different initial cluster center configurations. The best solution was selected by identifying the run that resulted in the minimum value of its associated cost function. The results are shown in table 4.

³ Recall that these thresholds are in the part of the algorithm.

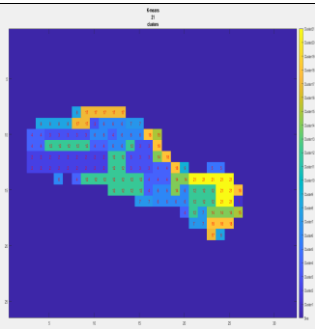
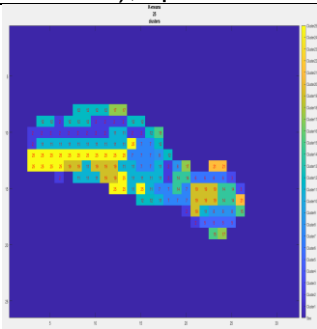
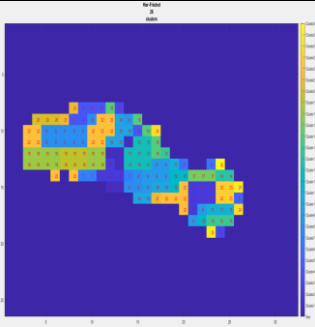
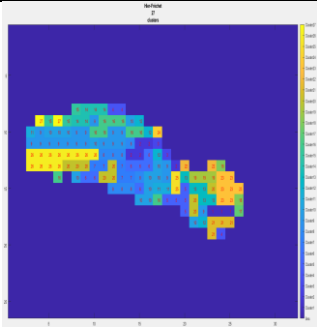
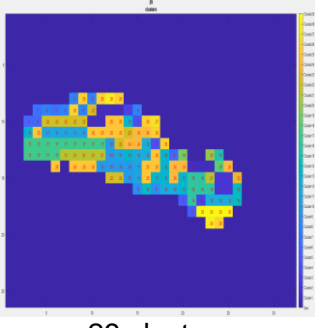
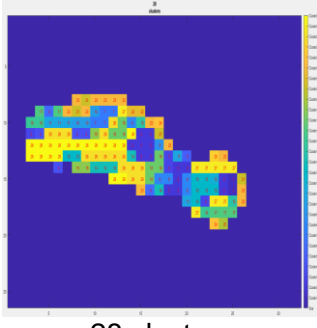
	Sentinel-2 spectra	Sentinel-2 - continuum removed ($1-S_{cr}$), spectra
K-means	 21 clusters	 25 clusters
Hier-Fréchet	 26 clusters	 27 clusters
SHC algorithm	 29 clusters	 28 clusters

Table 4: A sample with the best clustering results for the Sentinel-2 dataset. The numbers at each pixel represent the corresponding cluster label from the output each algorithm.

5.1.2 WorldView-3 VNIR dataset

All the algorithms in the WorldView-3 VNIR dataset were run only for the reflectance spectra since the results from the clustering methods seems to recognize the granitoid clusters (as discussed in the following chapter) without the need of the continuum-removed procedure.

The SCH algorithm was executed using *threshold_1*, *threshold_2* values equal to 0.04 and *threshold_3* value equal to 0.002.

The K-means algorithm was executed 100 times for this dataset, primarily due to the increased time complexity resulting from the large number of pixels involved in this case. The relative results are shown in table 5.

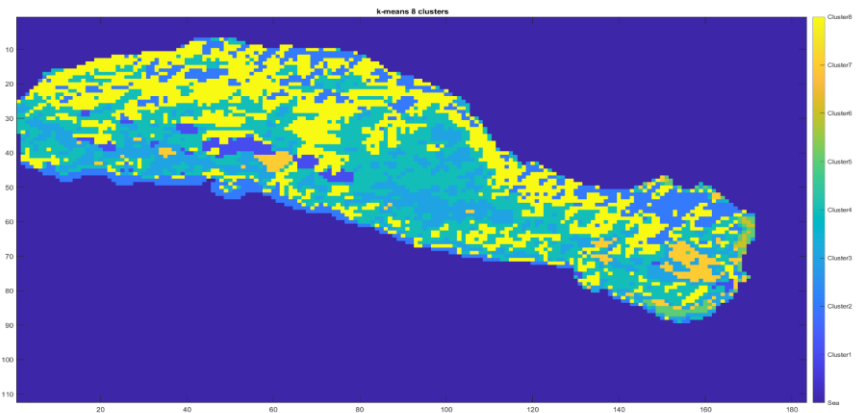
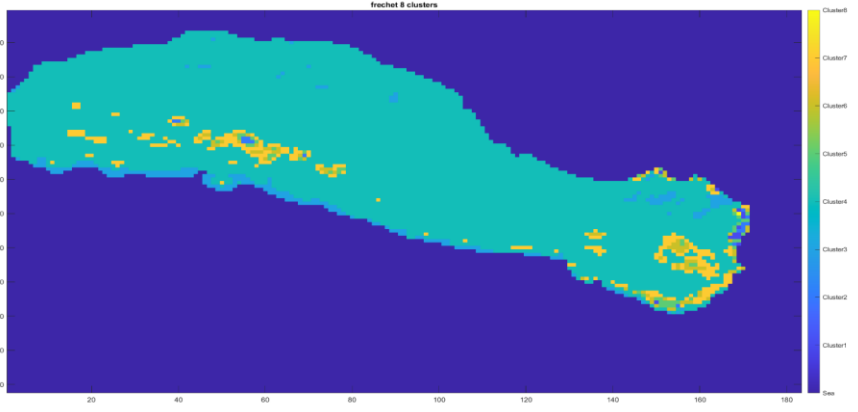
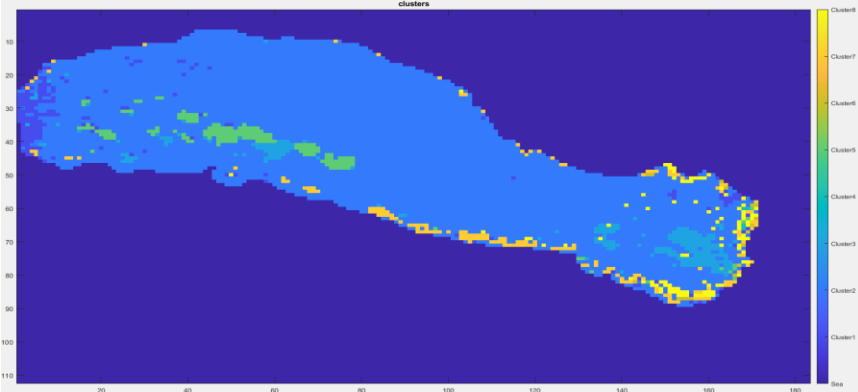
	<div>WorldView-3 VNIR</div>  <div>8 clusters</div>
Hier-Fréchet	 <div>8 clusters</div>
SHC algorithm	 <div>8 clusters</div>

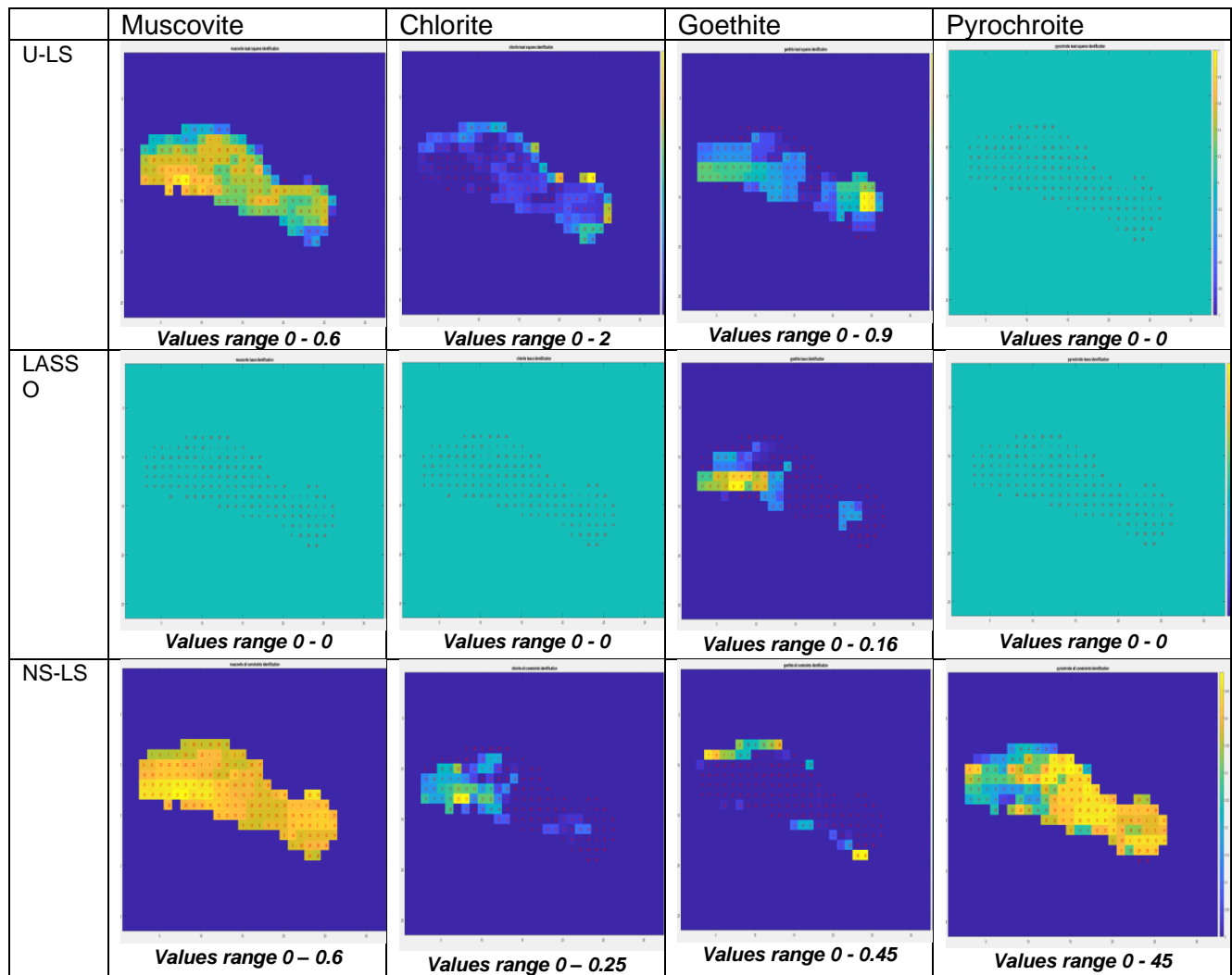
Table 5: A sample with the best clustering results for the WorldView-3 VNIR case.

5.2 Detection for alteration minerals (Spectral unmixing approach)

In this section the results of all the spectral unmixing methods that mentioned in the table 3, are demonstrated. However, the focus will be primarily on the results obtained from these algorithms when applied on the Sentinel-2 dataset. This emphasis to the Sentinel-2 dataset is due to the larger number of spectral bands and their positioning in the spectral spectrum, compared to the WorldView-3 VNIR dataset.

In future research the spectral unmixing in the WorldView-3 VNIR data could also be examined as well.

5.2.1 Sentinel-2 dataset



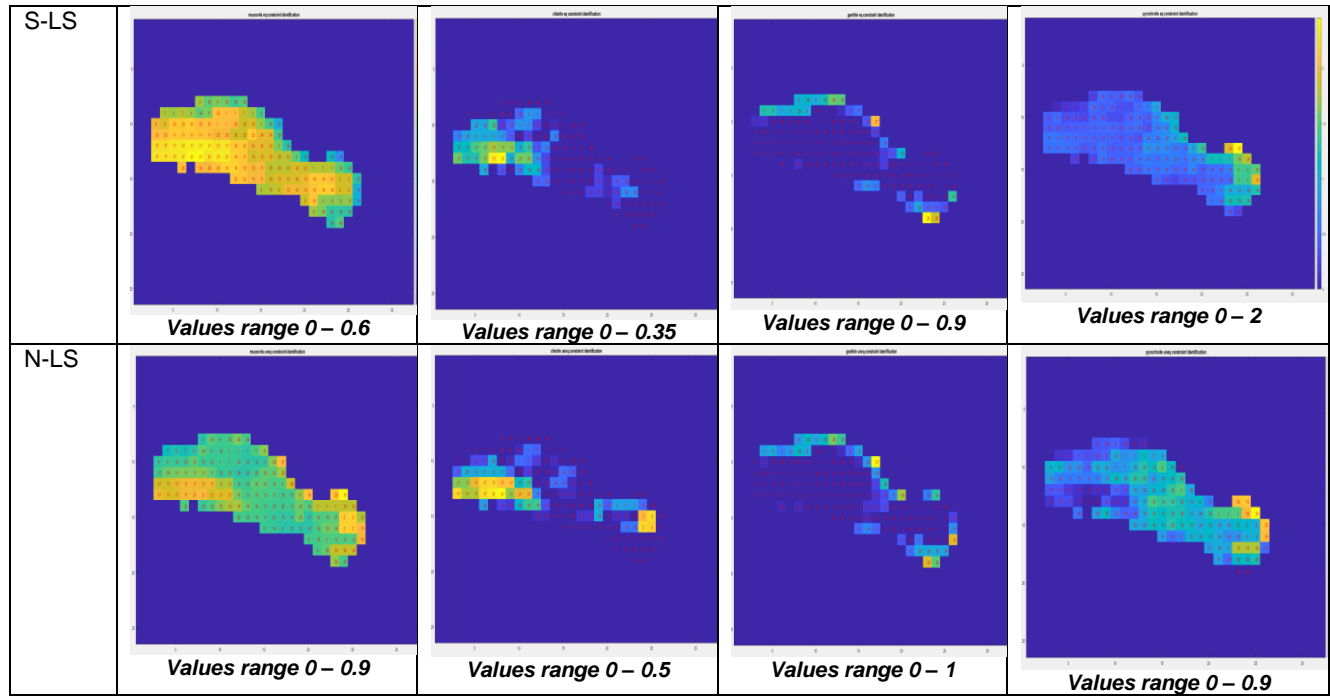
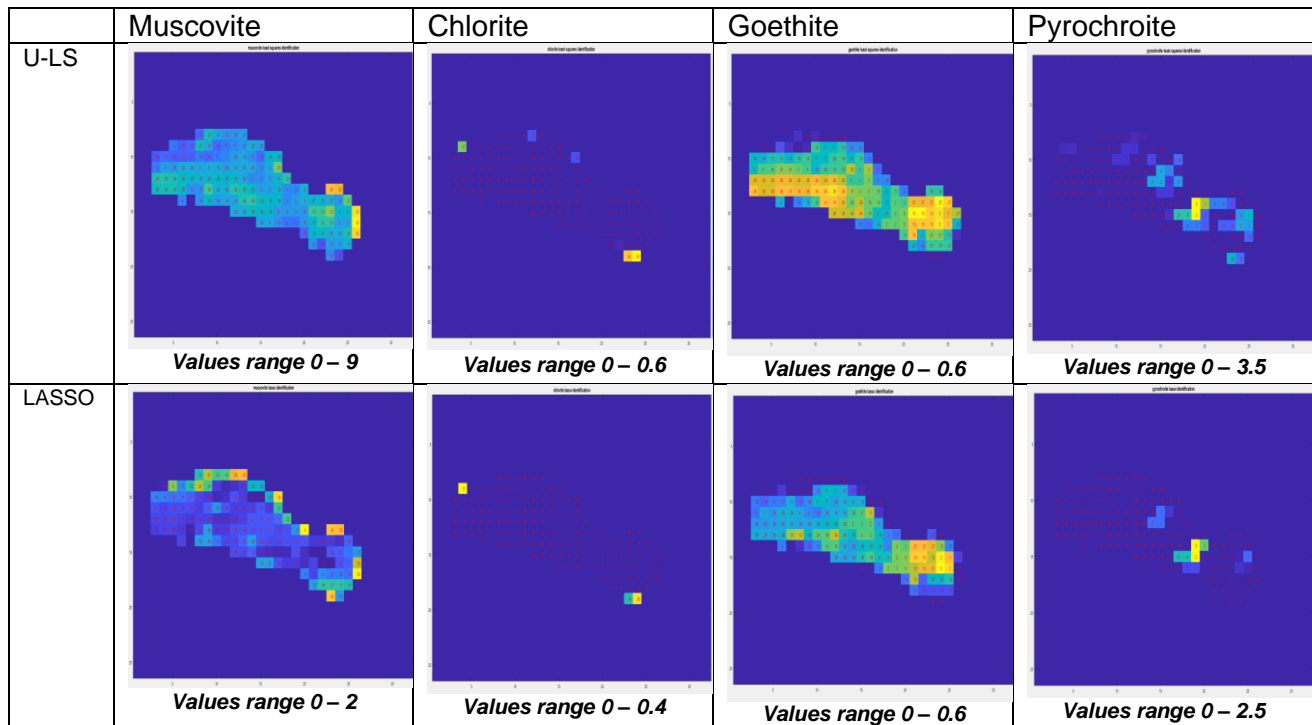


Table 6: Spectral unmixing results for the Sentinel-2 reflectance image. The abundance value for each pixel is represented with a color ranging from light blue (low abundance value) to yellow (high abundance value). Pixels in dark blue correspond to zero abundances. The number at each pixel corresponds to the corresponding cluster label from the output of the SHC algorithm.



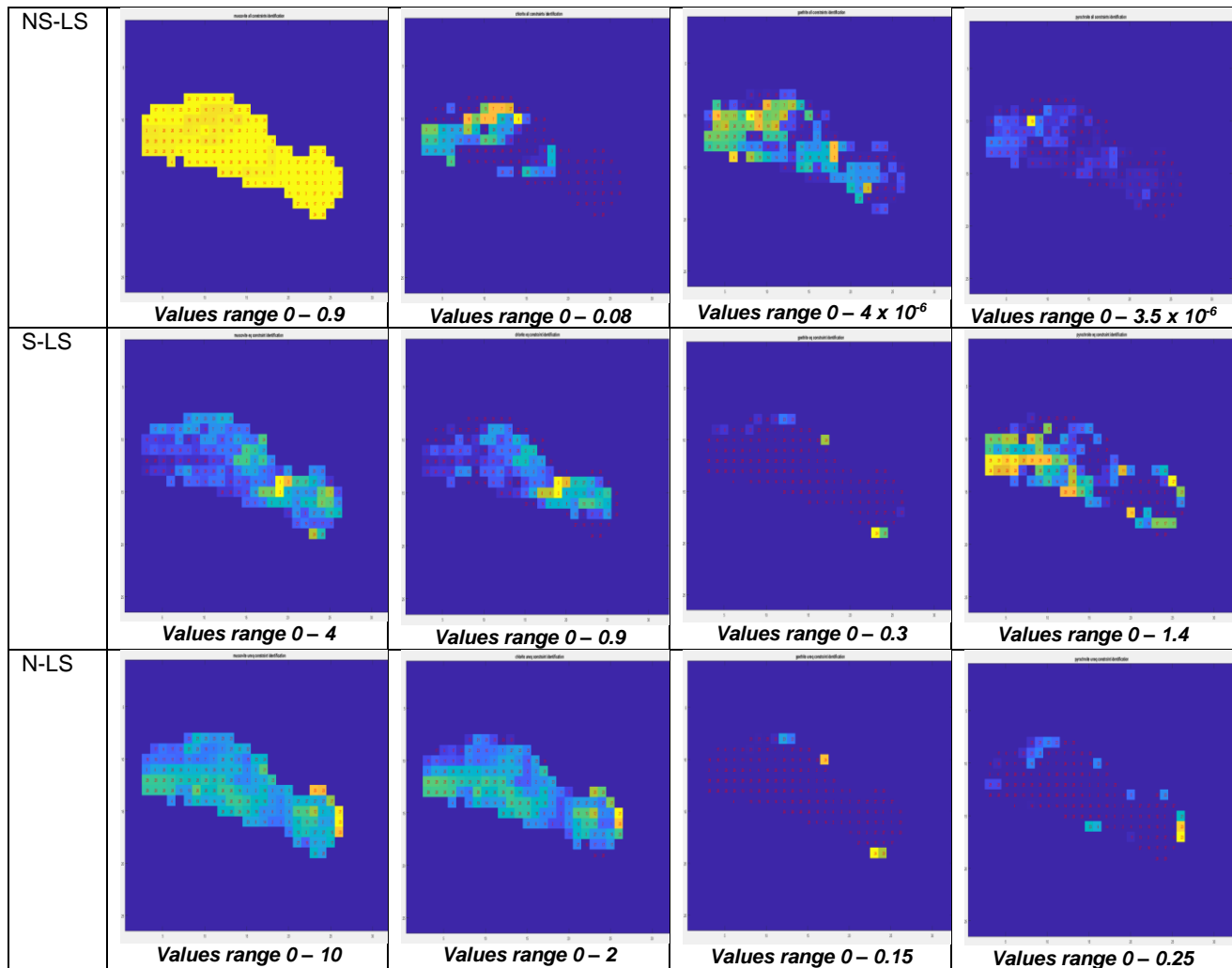


Table 7: Spectral unmixing results for the Sentinel-2 continuum removed image ($1 - S_{cr}$). with a color ranging from light blue (low abundance value) to yellow (high abundance value). Pixels in dark blue correspond to zero abundances. The number at each pixel corresponds to the corresponding cluster label from the output of the SHC algorithm.

6. Discussion

In this section, the results derived from the machine learning methods applied on the Sentinel-2 and WorldView-3 VNIR datasets depicting the Koutala island will be discussed and analyzed.

6.1 Identification of granitoid and schist formations (Clustering approach)

To validate the results of the clustering for this problem, external information from previous research is utilized [1]. This additional information is crucial for assessing the accuracy and relevance of the clustering outcomes. The island contains granitoid intrusions in two separate areas, and it's expected that pixels within each of these locations should ideally belong to the same cluster. (Fig.13) However, due to the significant heterogeneity within the granitoid areas, it is possible for pixels in these regions to be clustered into different groups. This heterogeneity can pose a challenge for the clustering process.



Figure 13: Google Earth high resolution image of the island. With the red rectangles the distinct locations of the granitoids are shown.

6.1.1 Sentinel-2 dataset

To validate the obtained clustering results using the Sentinel-2 image, the related RGB image (Fig.3) is superimposed to match the Google Earth imagery, and the pixels corresponding to the locations of granitoid intrusion are extracted. The red rectangles in the RGB image shown in figure 14, are used to indicate the actual locations of the pixels in the Sentinel-2 image, helping to establish the correspondence between the two images.



Figure 14: RGB-image (produced by the Sentinel-2 image) georeferenced to match the Google earth image. With the red rectangles the granitoid locations are depicted in the Sentinel-2 image. Slight displacements between the background image and the Sentinel-2 image are due to differences in the georeferenced systems and to the very high difference between the spatial resolution between the Sentinel-2 and Google Earth image.

The granitoid pixels in the Sentinel-2 image are summarized in the following table:

Left granitoid area	Right granitoid area
(14,10)	(16,22)
(14,11)	(16,23)
(14,12)	(16,24)
	(16,25)
	(17,22)
	(17,23)

	(17,24)
	(17,25)

Table 8: Sentinel-2 image granitoid pixels locations. Each (row, column) position corresponds to a single pixel in the displayed image. For example, the position (14,10) corresponds to the pixel at row 14, column 10 in the image.

The signatures of the granitoid pixels are demonstrated in the following table 9 both for the reflectance and continuum removal spectral values. The observation of common spectral characteristics within the pixels in the western granitoid area (whose spectral signatures are denoted with a red line in table 9) supports the expectation that they lie in the same cluster. On the other hand, the pixels in the eastern granitoid area (whose spectral signatures are denoted with a blue line in table 9) exhibit spectral variations, indicating the potential need for multiple distinct clusters to accurately represent the diversity within this region.

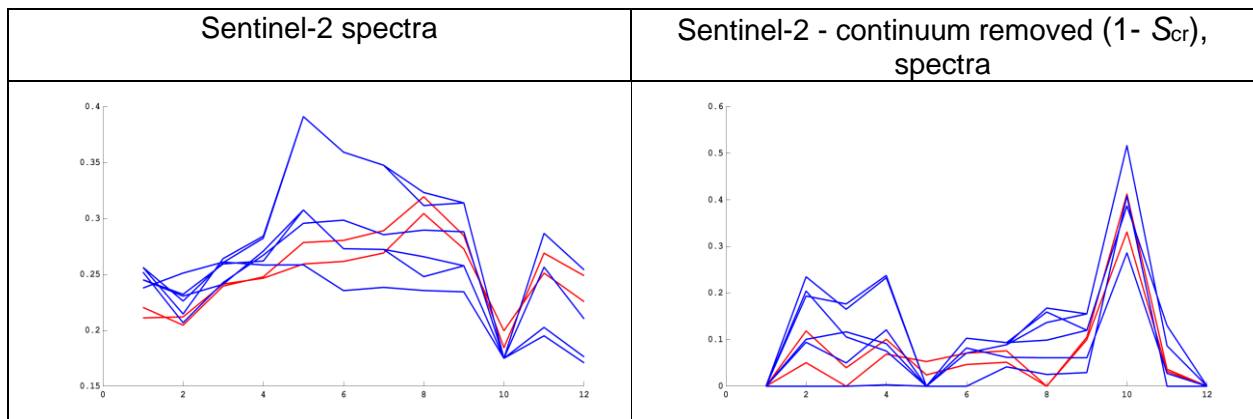


Table 9: Spectral signatures of granitoid pixels for both reflectance and continuum-removed spectral values. The red lines represent the signatures of the granitoid pixels in the western area whereas the blue lines the signatures of the granitoid pixels in the eastern area of the island. Some pixels share the same signature so a line can represent one or more pixels.

6.1.1.1 Reflectance spectra

In this section, the results from the clustering in the reflectance spectra case will be discussed and analyzed. In order to validate the results, a table is provided for each clustering method, that includes the number of granitoid pixels and the number of the non- granitoid pixels at the clusters containing them. Additionally, a table is presented for

each clustering method, indicating the cluster label associated with each granitoid cluster. Finally, a table showing the spectral pixels for these cluster labels is demonstrated. These tables contribute to the comprehensive analysis and validation of the clustering outcomes.

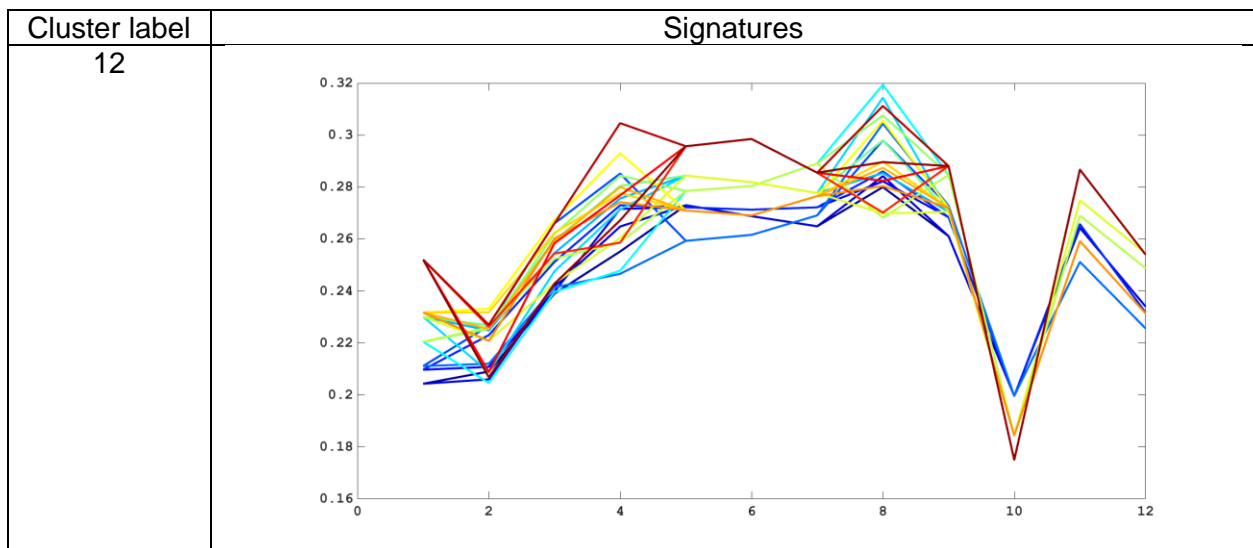
K-means

Number of granitoid pixels in the cluster	Number of non- granitoid pixels in the cluster
5	23
2	7
1	6
3	4

Table 10: K-means - reflectance spectra - granitoid pixels vs non granitoid pixels at the clusters.

Cluster label	Granitoid pixels in the cluster
12	(14,10) (14,11) (14,12) (16,22) (16,23)
21	(16,24) (16,25)
7	(17,22)
14	(17,23) (17,24) (17,25)

Table 11: K-means - reflectance spectra - granitoid pixels positions at the clusters.



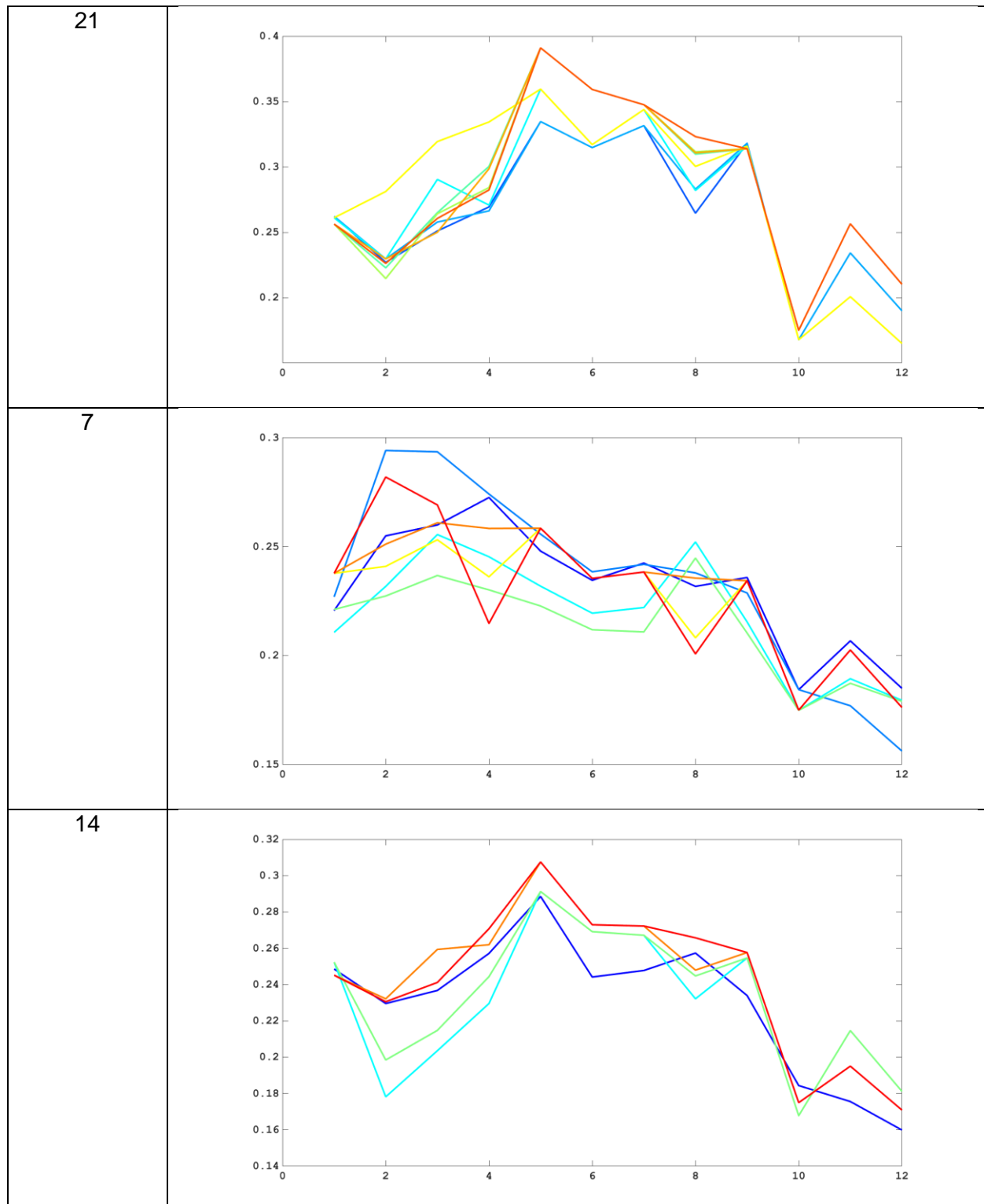


Table 12: K-means - reflectance spectra – signatures of clusters containing granitoid pixels. Each distinct spectral signature within a cluster is represented by a unique color.

As we can see from tables 10, 11, 12 the K-means failed to isolate the granitoid pixels into separate clusters. The cluster 12 for example contains many spectral pixels that they don't have similar spectral form especially in the band 8. The cluster 14 seems to have the most common spectral signatures among the other clusters.

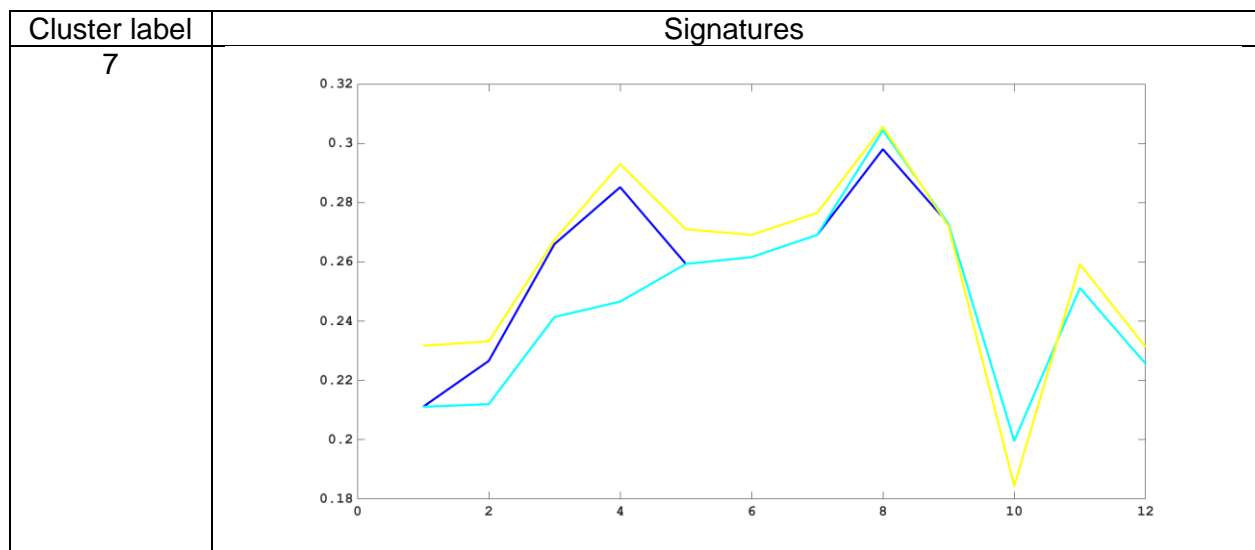
Hier-Fréchet

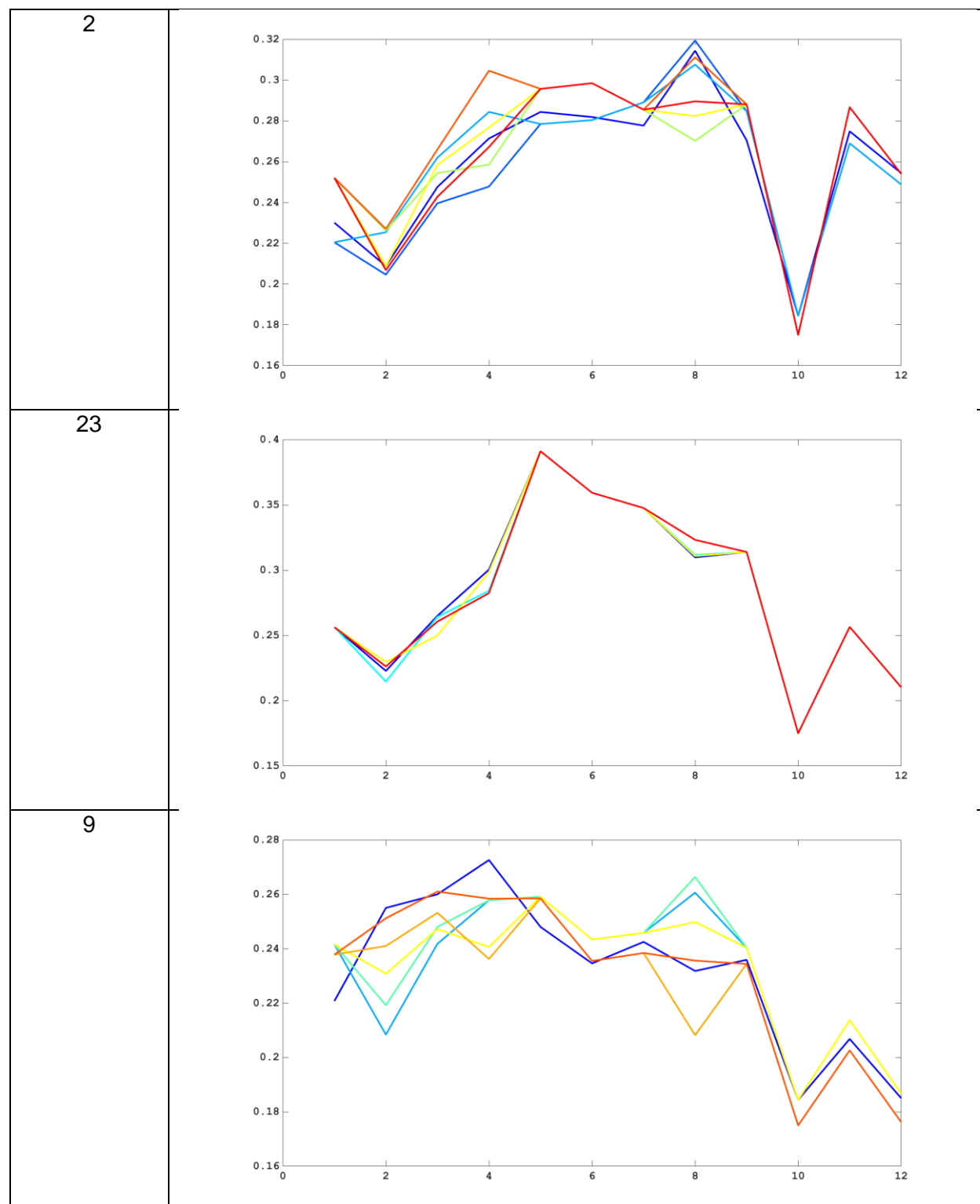
Number of granitoid pixels in the cluster	Number of non- granitoid pixels in the cluster
1	2
4	7
2	2
1	8
3	1

Table 13: Hier-Fréchet – reflectance spectra - granitoid pixels vs non granitoid pixels at the clusters.

Cluster label	Granitoid pixels in the cluster
7	(14,10)
2	(14,11) (14,12) (16,22) (16,23)
23	(16,24) (16,25)
9	(17,22)
11	(17,23) (17,24) (17,25)

Table 14: Hier-Fréchet – reflectance spectra - granitoid pixels positions at the clusters.





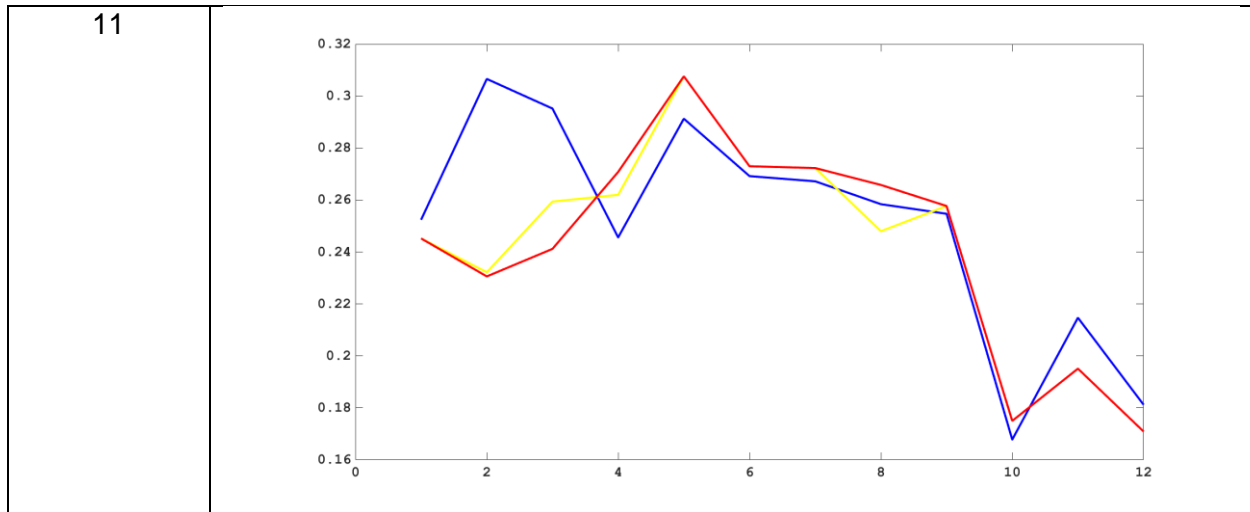


Table 15: Hier-Fréchet – reflectance spectra – signatures of clusters containing granitoid pixels. Each distinct spectral signature from the pixels is represented by a unique color.

The results from the hierarchical complete link algorithm using as distance metric the Fréchet distance seem to be better than the K-means corresponding ones, since the pixels have more similar spectral signatures at the clusters formed. Another observation is that at clusters 7, 23 where the spectral patterns are similar in these clusters, the granitoid pixels are recognized together with non-granitoid pixels. This depicts the complexity of the problem since some non-granitoid pixels share the same spectral signature pattern with the granitoid pixels. In cluster 11, the granitoid pixels are more numerous compared to the non-granitoid pixels. However, it appears that the clustering results are not optimal, as clusters like the one labeled 2 or 9 consist of pixels with varying spectral patterns. This inconsistency suggests that the clustering method may not effectively capture the desired differentiations in the data.

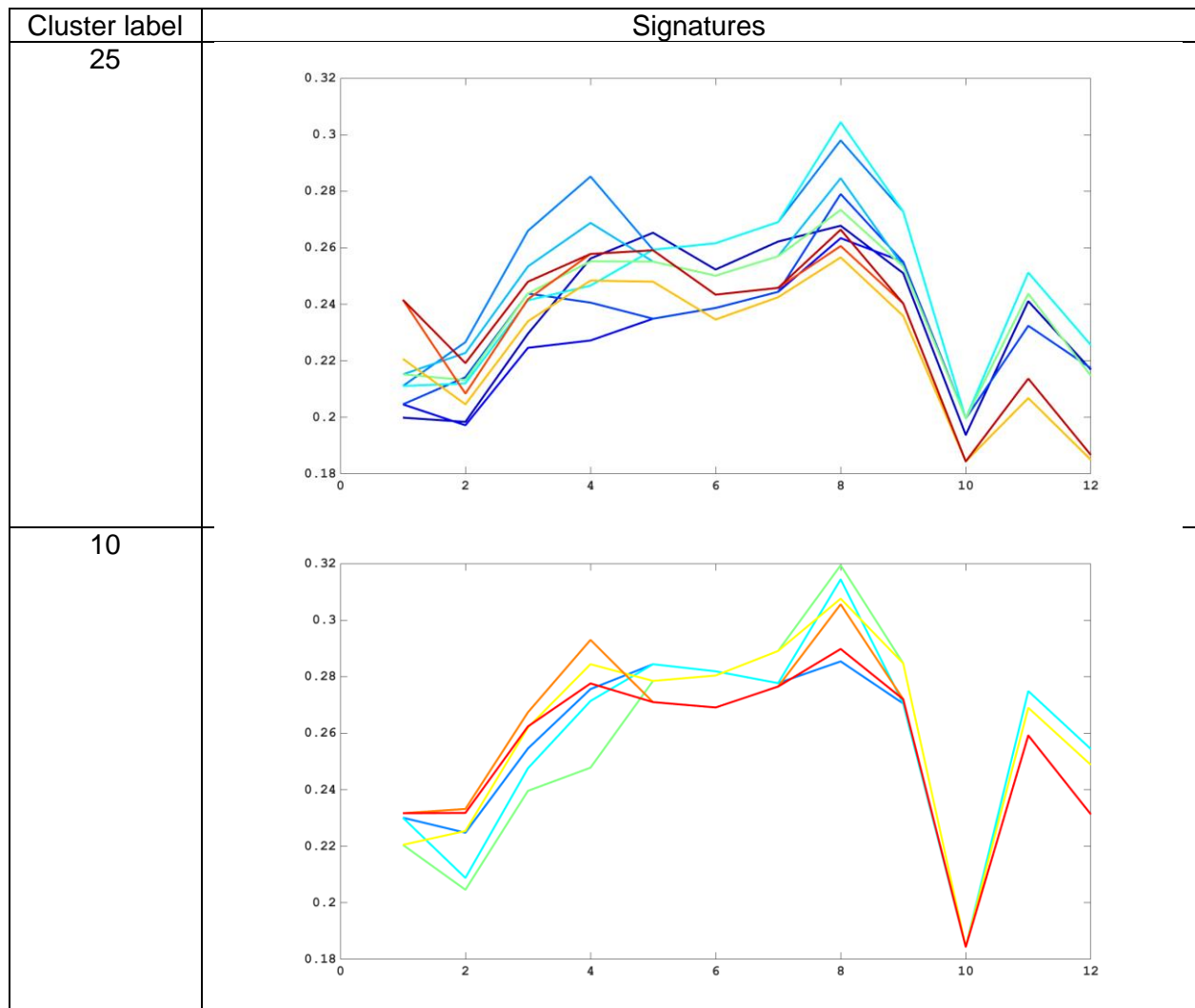
SHC algorithm

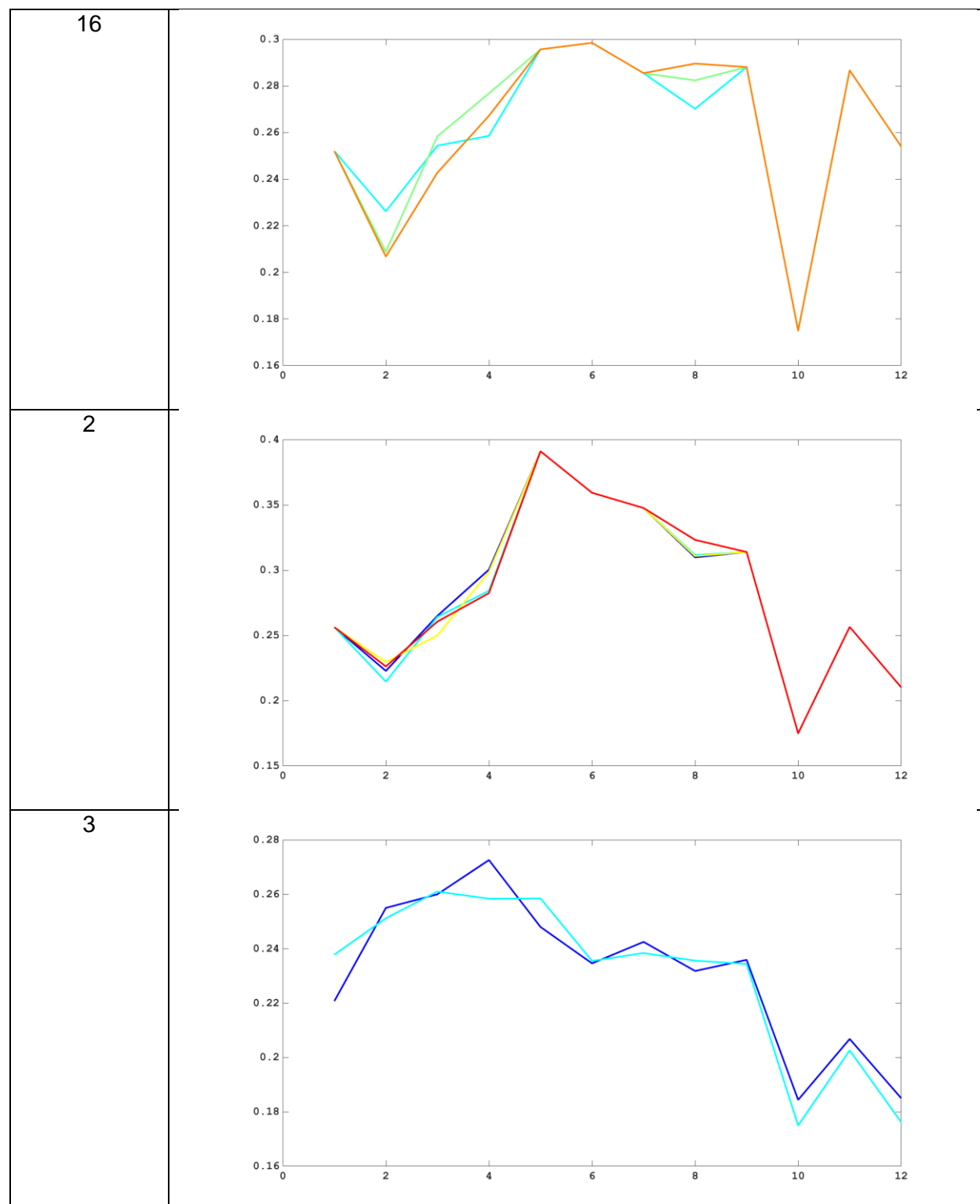
Number of granitoid pixels in the cluster	Number of non- granitoid pixels in the cluster
1	14
2	5
2	4
2	2
1	1
3	5

Table 16: SHC algorithm - reflectance spectra- granitoid pixels vs non granitoid pixels at the clusters.

Cluster label	Granitoid pixels in the cluster
25	(14,10)
10	(14,11) (14,12)
16	(16,22) (16,23)
2	(16,24) (16,25)
3	(17,22)
12	(17,23) (17,24) (17,25)

Table 17: SHC algorithm - reflectance spectra- granitoid pixels positions at the clusters.





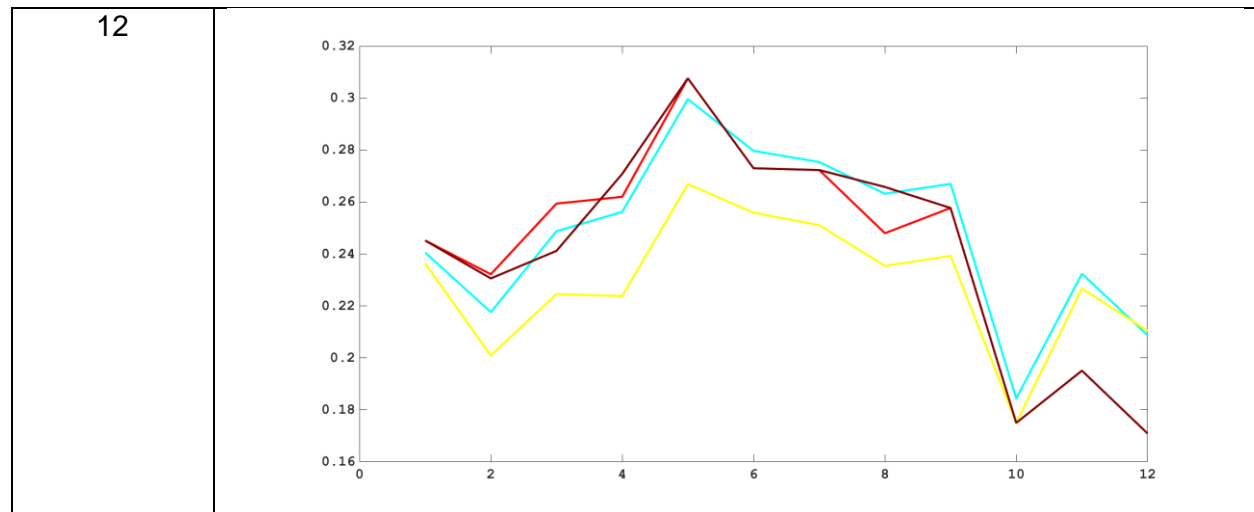


Table 18: SHC algorithm - reflectance spectra- signatures of clusters containing granitoid pixels. Each distinct spectral signature within a cluster is represented by a unique color.

Comparing visually the tables with the signatures of clusters containing granitoid pixels, the first observation from the clustering results of the SHC algorithm is that all formed clusters that contain granitoid pixels consist of pixels with more similar signatures, compared to the clusters produced by the previous methods. This is probably due to the fact that instead of using individual band reflectance values within the algorithm procedure, we use their derivatives at the two steps of the SHC algorithm. Two of the pixels of the western granitoid area (14,11) and (14,12) were grouped into the same cluster with some other non-granitoid pixels near to them (see table 4). The pixel (14,10) was grouped wrongly to another cluster. The clustering of eastern granitoid area resulted in smaller granitoid clusters, with some pixels within the same cluster located near each other in the granitoid area (cluster labels: 16, 2), while others are positioned farther away from the granitoid area (cluster labels: 12, 3) (see table 4). Despite these spatial variations, the spectral signatures within these clusters are quite similar, underscoring the complexity of the area, the possibility of more than one granitoid intrusions and the challenge of accurately clustering pixels in a granitoid area with Sentinel-2 data.

6.1.1.2 Continuum-removed spectra

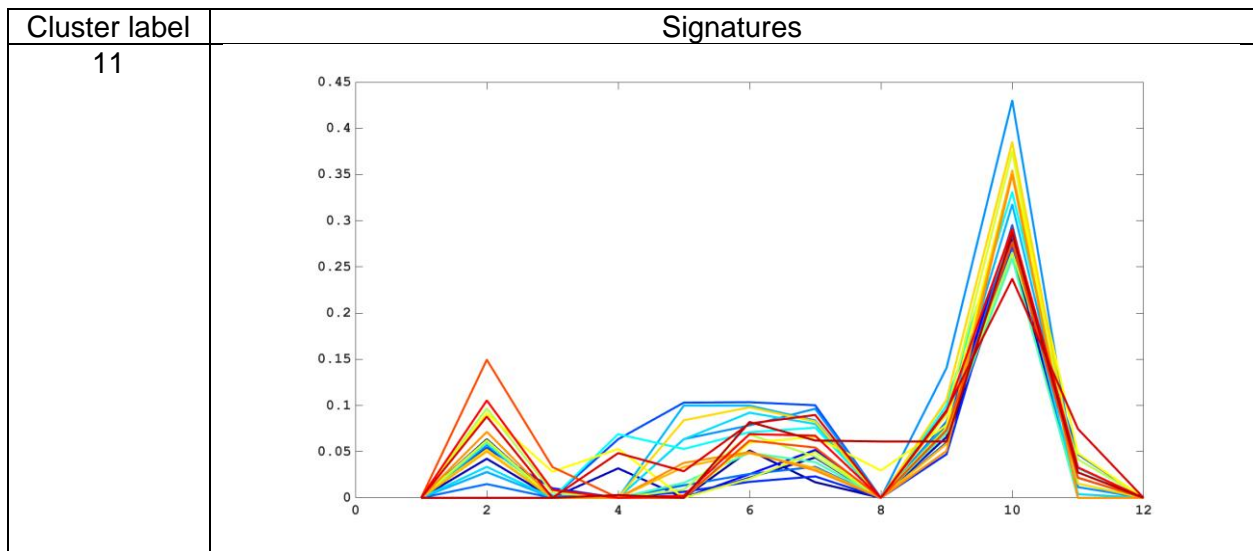
K-means

Number of granitoid pixels in the cluster	Number of non- granitoid pixels in the cluster
2	25
4	10
2	5
3	2

Table 19: K-means - continuum removed spectra - granitoid pixels vs non granitoid pixels at the clusters.

Cluster label	Granitoid pixels in the cluster
11	(14,10) (17,22)
19	(14,11) (14,12) (16,22) (16,23)
14	(14,11) (14,12) (16,22) (16,23)
8	(17,23) (17,24) (17,25)

Table 20: K-means - continuum removed spectra - granitoid pixels positions at the clusters.



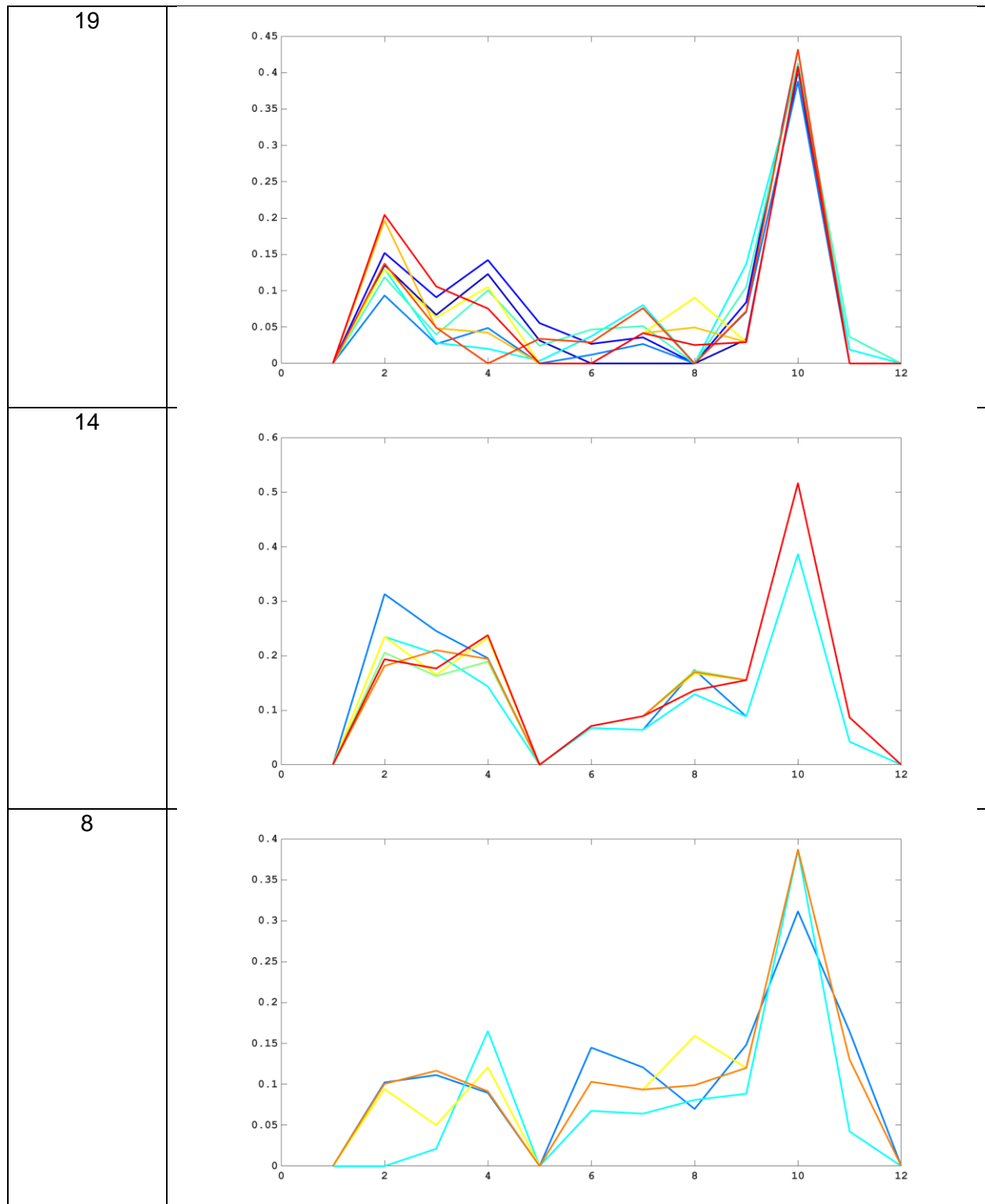


Table 21: K-means – continuum removed spectra – signatures of clusters containing granitoid pixels. Each distinct spectral signature within a cluster is represented by a unique color.

Observing the signatures of clusters containing the granitoid pixels it is evident that the K-means algorithm again failed to distinct signatures with common patterns at the clusters and hence distinguish granitoid clusters.

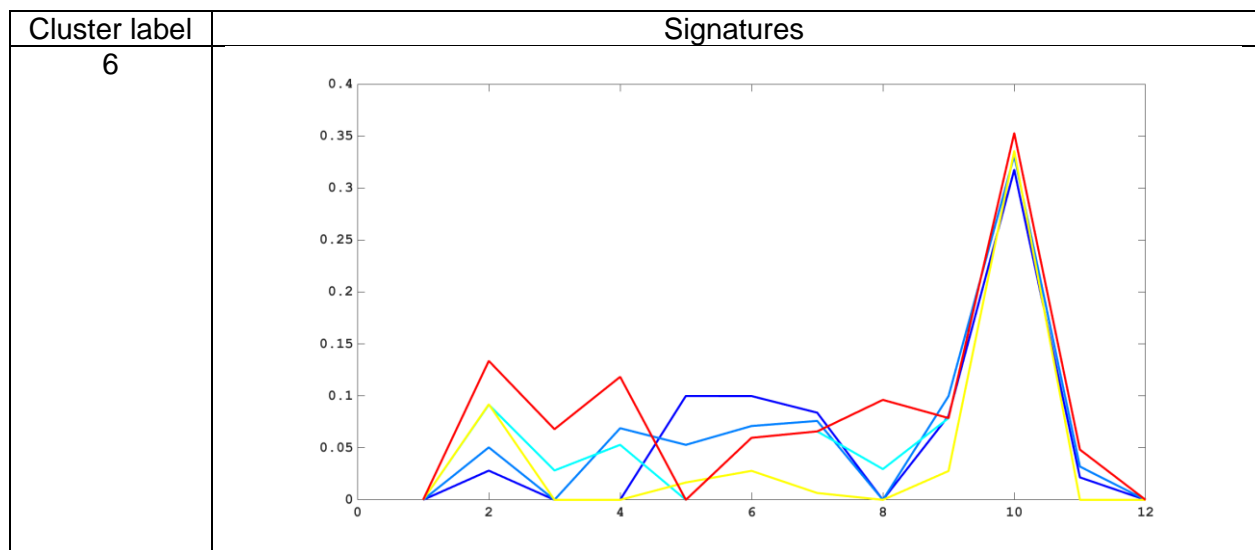
Hier-Fréchet

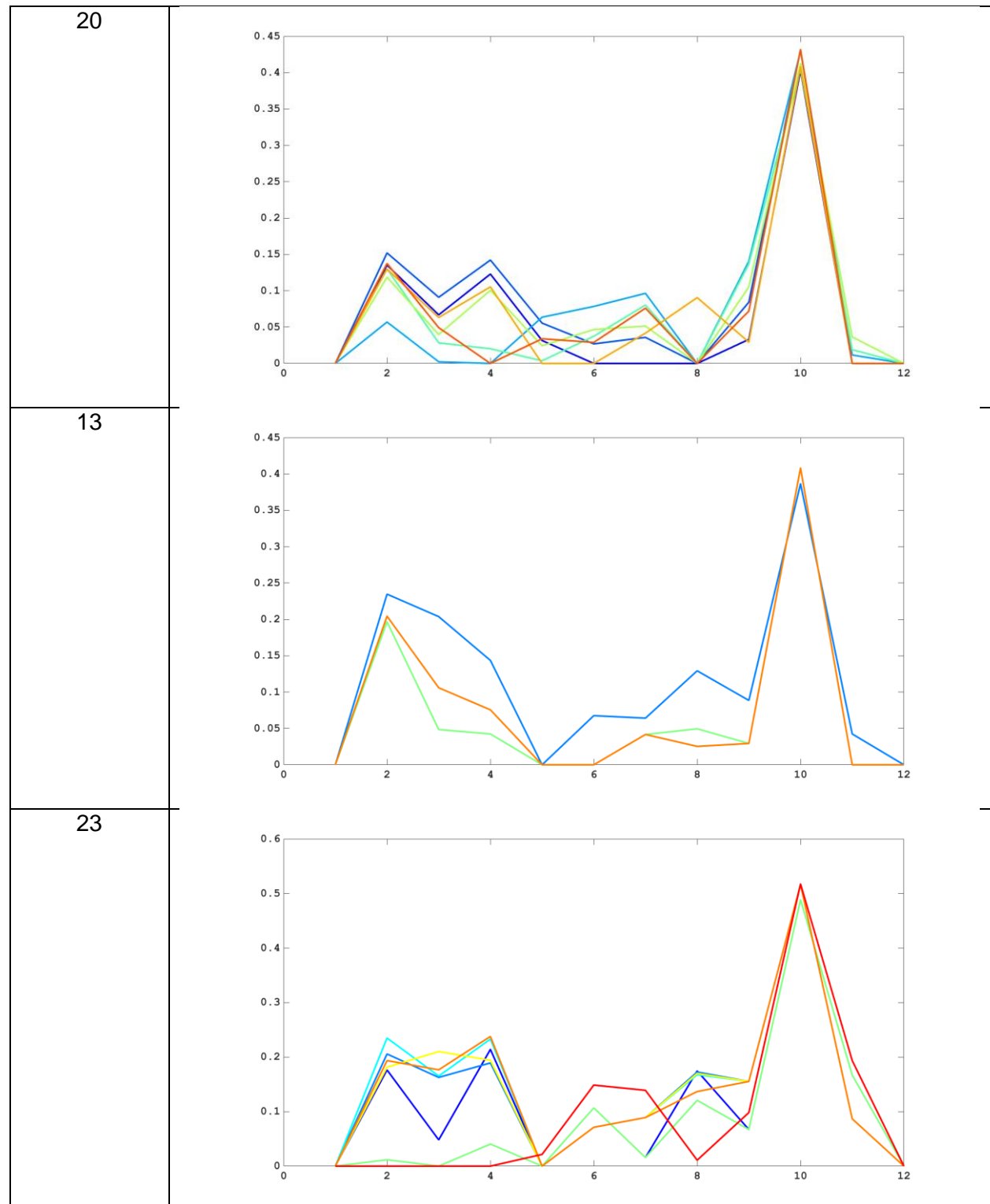
Number of granitoid pixels in the cluster	Number of non- granitoid pixels in the cluster
1	6
2	7
2	3
2	5
1	18
3	0

Table 22: Hier-Fréchet – continuum removed spectra - granitoid pixels vs non granitoid pixels at the clusters.

Cluster label	Granitoid pixels in the cluster
6	(14,10)
20	(14,11) (14,12)
13	(16,22) (16,23)
23	(16,24) (16,25)
9	(17,22)
1	(17,23) (17,24) (17,25)

Table 23: Hier-Fréchet – continuum removed spectra - granitoid pixels positions at the clusters.





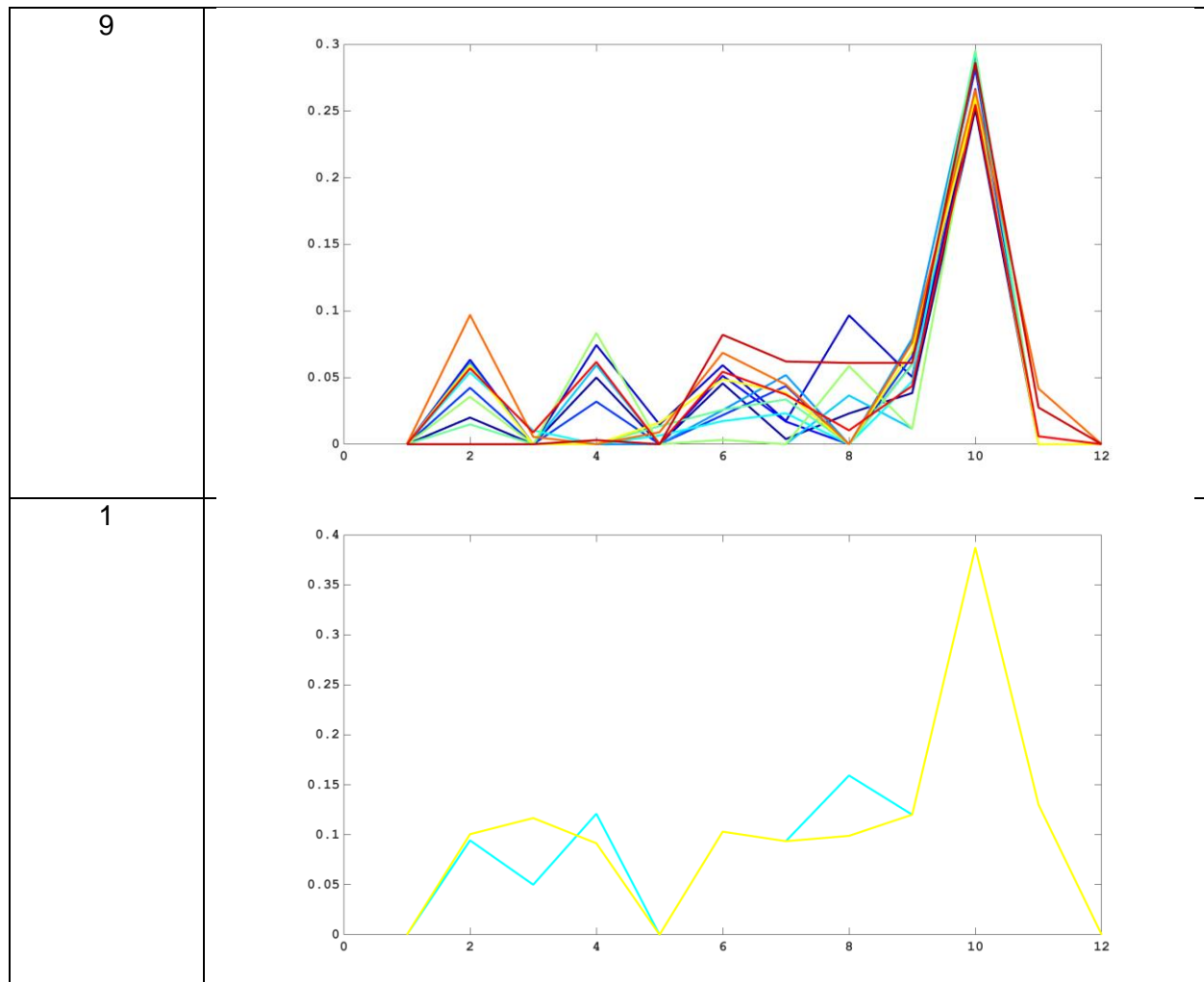


Table 24: Hier-Fréchet – continuum removed spectra – signatures of clusters containing granitoid pixels. Each distinct spectral signature within a cluster is represented by a unique color.

It seems that the Hier-Fréchet algorithm demonstrates improved outcomes compared to K-means algorithm in terms of the similarity observed within the clusters when examining their spectral signatures. Nevertheless, it remains evident that the clusters do not exhibit identical signature patterns. For instance, in cluster label 1, despite the algorithm grouping pixels from the eastern granitoid area into a single cluster, these signatures do not share the same pattern. The explanation of potentially more than one granitoid intrusions is to be examined in the future.

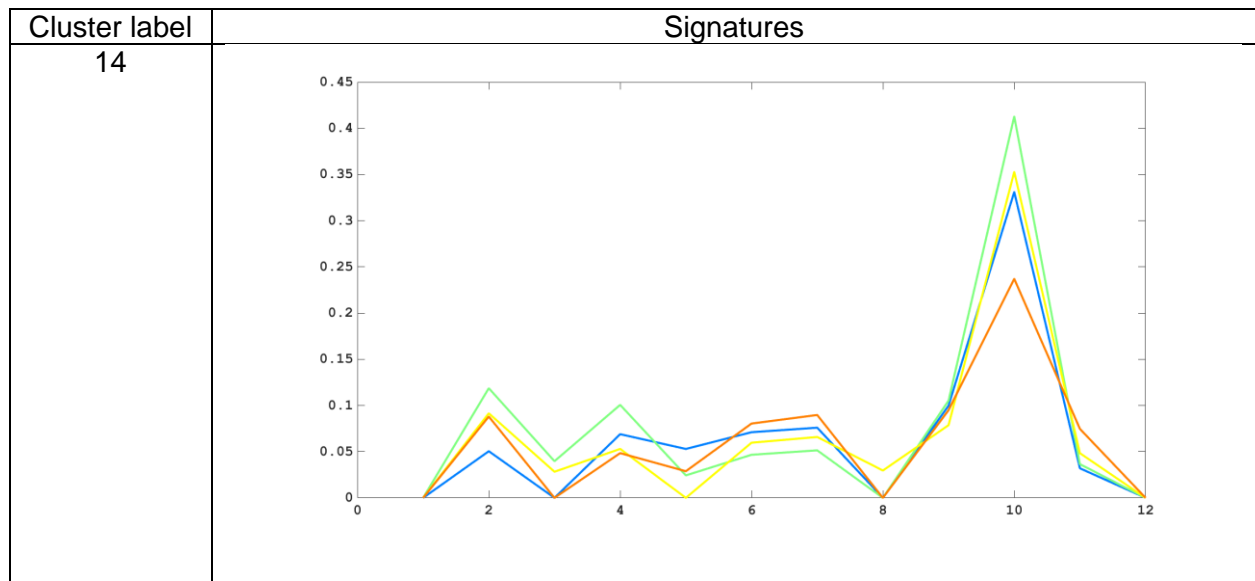
SHC algorithm

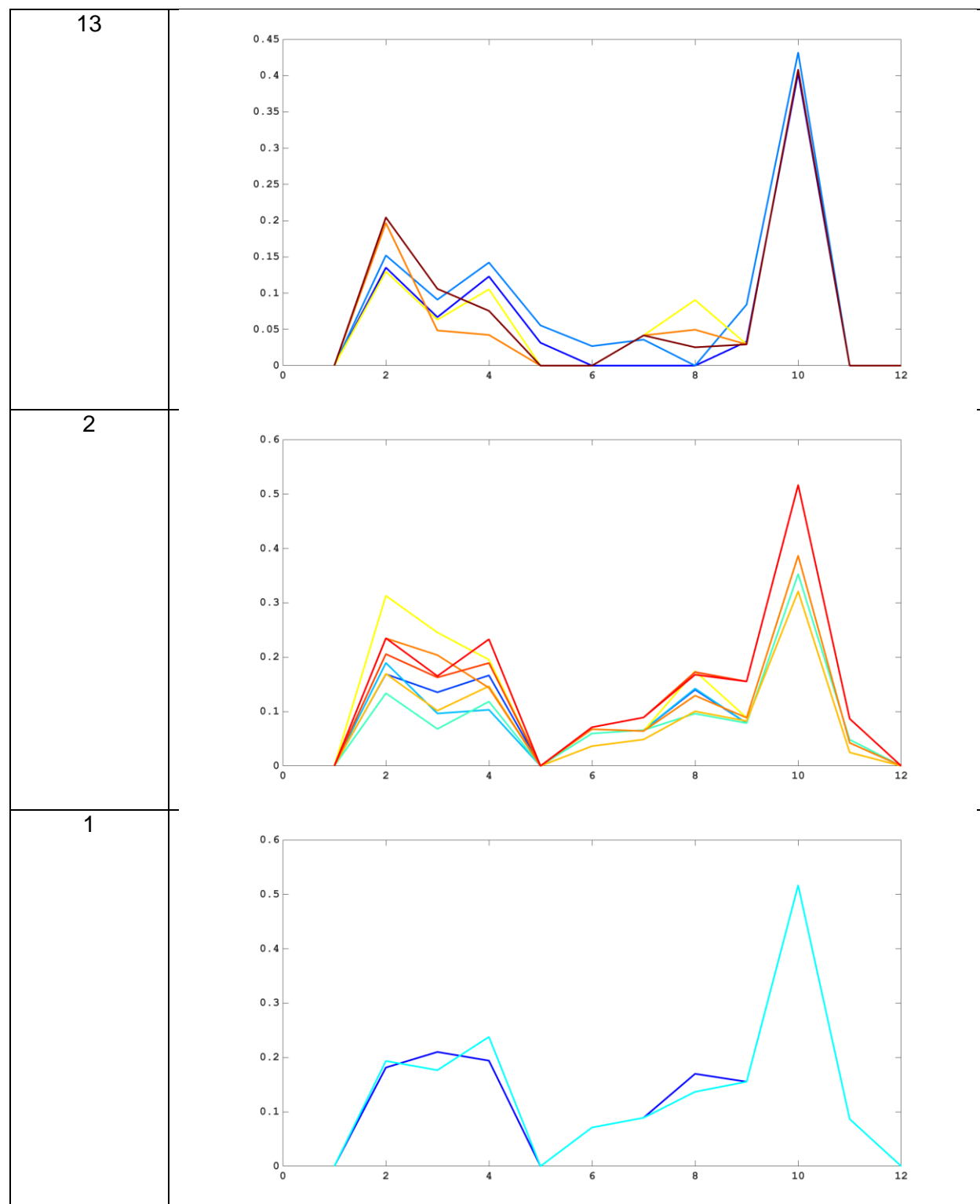
Number of granitoid pixels in the cluster	Number of non- granitoid pixels in the cluster
3	2
2	6
1	13
1	1
1	3
2	7
1	1

Table 25: SHC algorithm - continuum removed spectra - granitoid pixels vs non granitoid pixels at the clusters.

Cluster label	Granitoid pixels in the cluster
14	(14,10) (14,11) (14,12)
13	(16,22) (16,23)
2	(16,24)
1	(16,25)
3	(17,22)
27	(17,23) (17,24)
15	(17,25)

Table 26: SHC algorithm - continuum removed spectra - granitoid pixels positions at the clusters.





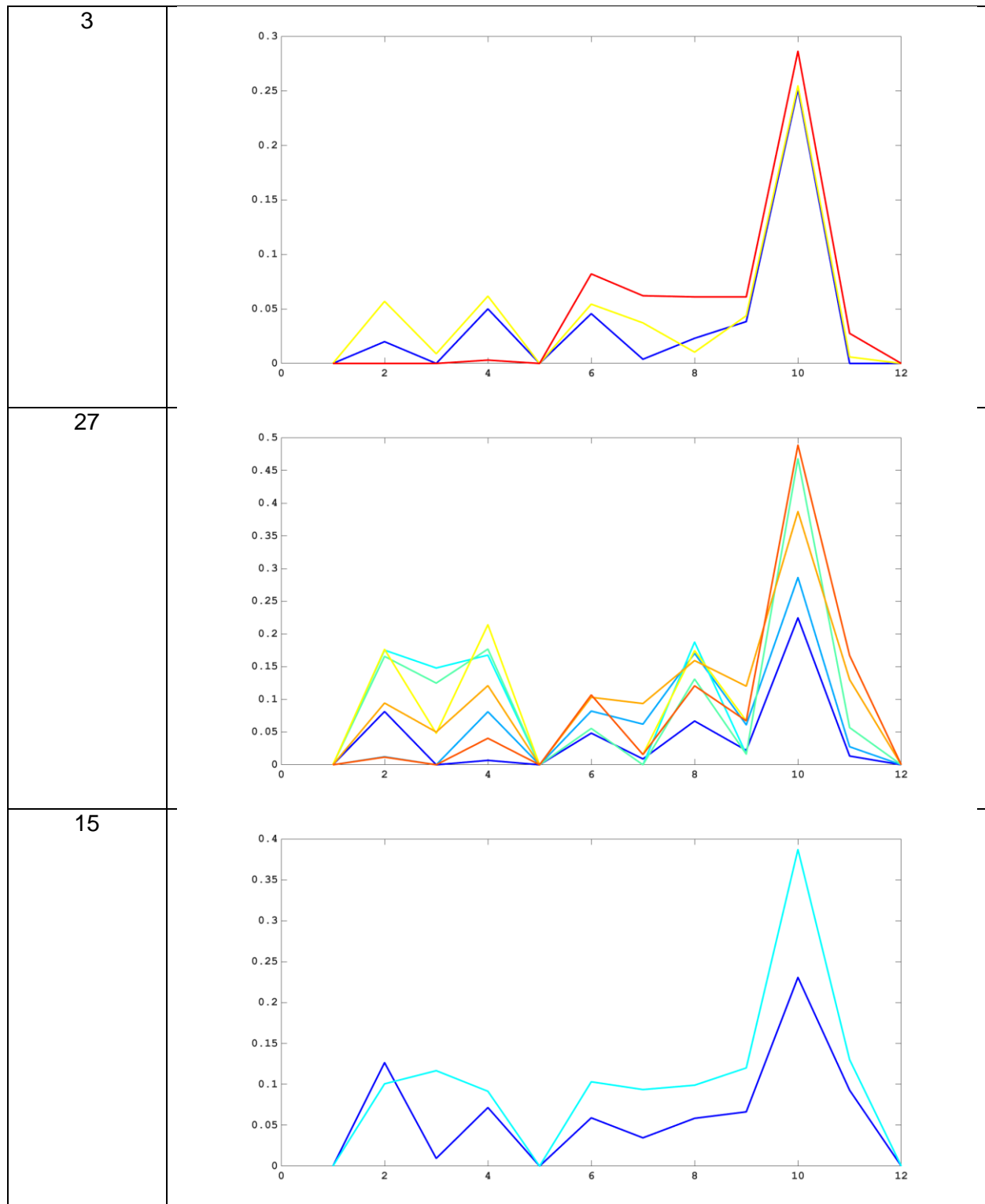


Table 27: SHC algorithm – continuum removed spectra- signatures of clusters containing granitoid pixels. Each distinct spectral signature within a cluster is represented by a unique color.

Like the reflectance spectra case the SHC algorithm seems to have the best results in returning clusters with common spectral signature patterns than the previous algorithms. The SHC algorithm was the only algorithm that was able to successfully distinguish the western granitoid area, even though some non-granitoid pixels were present, sharing a similar signature pattern. Compared with the reflectance case, this can be explained by the fact that the normalized reflectance values of these granitoid pixels share the same pattern, whereas in the reflectance data, they differ at band 1 (the first band and last band in the continuum removal procedure have 0 value). This explains why these pixels are within the same cluster from the output of the sequential algorithm (first step) where the shape of the spectral forms is compared. The appearance of the non-granitoid pixels into this cluster could happen due to the low spatial resolution Sentinel-2 data (10m), which has as an effect the formation of a mixed common signature pattern for a large area ($100m^2$). Another observation is that all the pixels in cluster 2 exhibit the same signature pattern; however only one granitoid pixel belongs to the cluster. This reflects again the effect of the low spatial resolution into the clustering results. Finally, in some clusters (cluster labels: 1, 13, 15) the signatures differ only in one band, which could be explained by the fact that in the second step of the algorithm the pixels are clustered based on the sum of the absolute differences of all the corresponding coordinates of the derivative signatures between two pixels, which is minimal between these pixels.

6.1.2 WorldView-3 VNIR dataset

Due to the significantly higher spatial resolution offered by the WorldView-3 VNIR image compared to the Sentinel-2 image, the validation of clusters is achieved primarily through visualization rather than pixel-by-pixel verification (Fig.15). To capture the signature patterns in the granitoid areas, a subset of pixels from these regions is utilized. This sample aids in capturing and analyzing the spectral patterns present in the granitoid areas within the higher-resolution WorldView-3 VNIR dataset. Lastly, a sample of pixels in non-granitoid areas is utilized as well to compare the signature patterns with the granitoid areas. (Fig.16) From figure 16, it is evident that the granitoid areas exhibit a similar signature pattern, with minor differences primarily observed in bands 6, 7, and 8. However, these small variations in bands 6,7 and 8 are small indicating that both granitoid areas share some common mineralogical compositions. Conversely, the non- granitoid pixels exhibit minimal variations across all bands, and, despite apparent similarities in signature patterns with the granitoid pixels across many bands, the noticeable differences in the reflectance values indicate a differentiation in their pattern.



Figure 15: WorldView-3 VNIR masked RGB image where the red squares represent the granitoid areas.

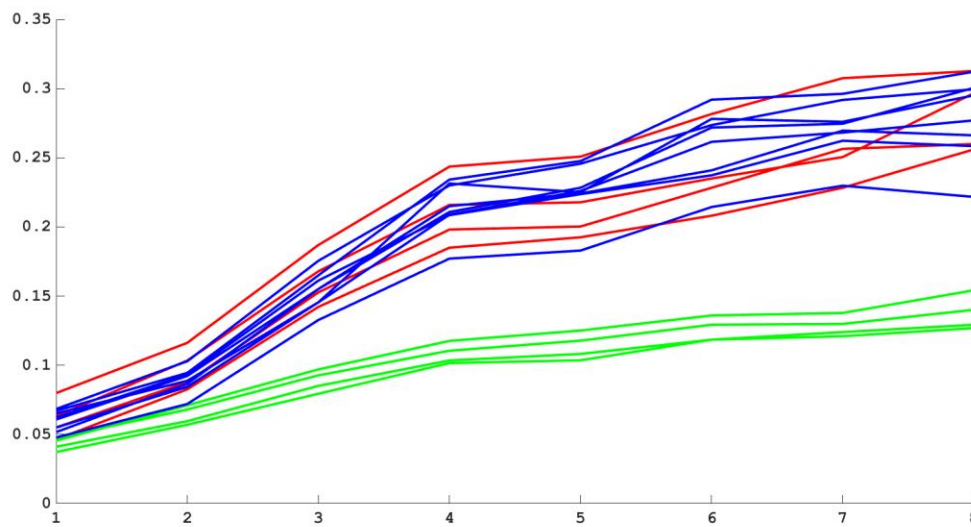
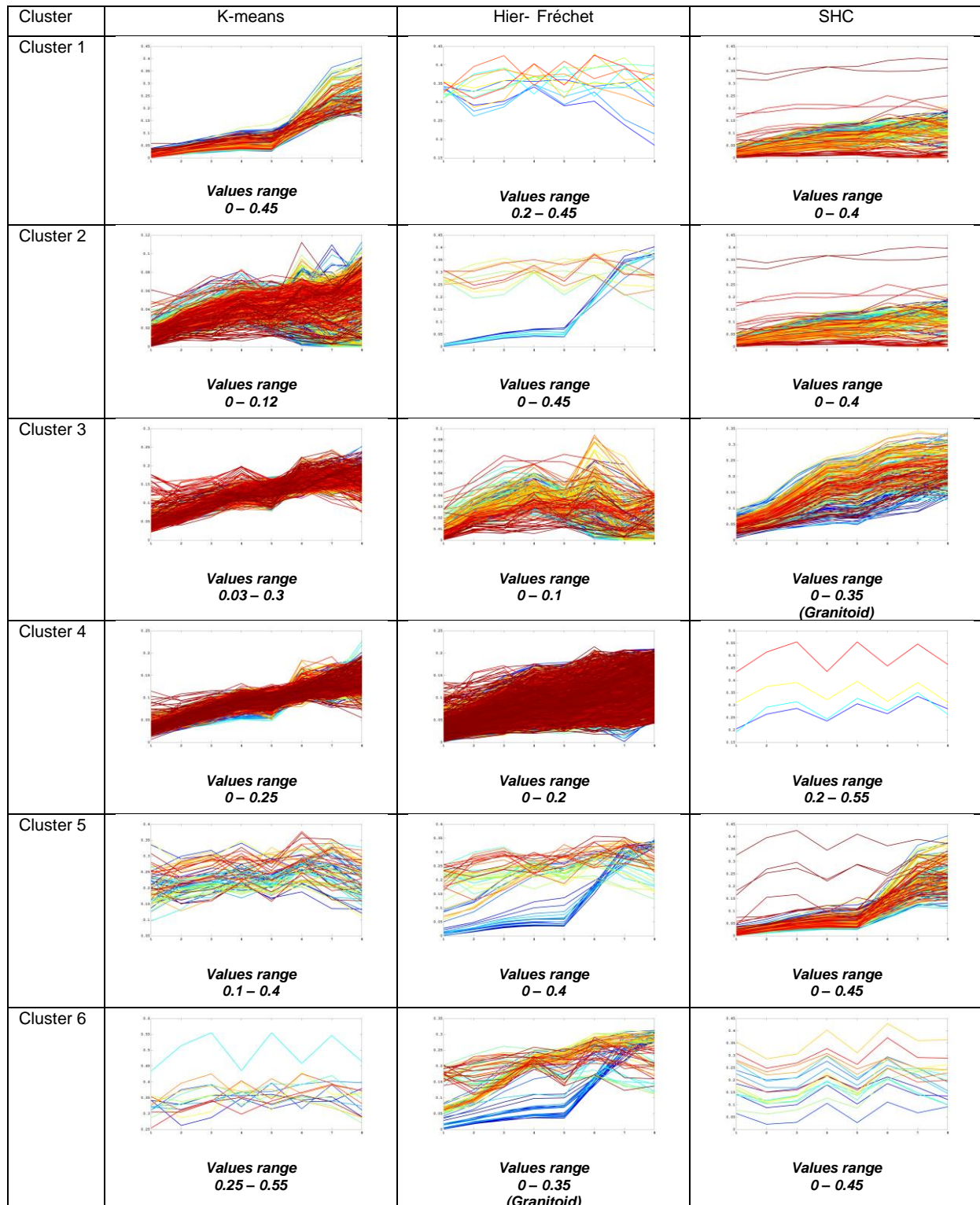


Figure 16: Signatures of granitoid and non- granitoid pixels of the WorldView-3 VNIR image. The red lines represent the signatures of the granitoid pixels in the western area of the islet whereas the blue lines the signatures of the granitoid pixels in the eastern area. The green lines represent pixels in the center of the island where no granitoid intrusions were detected.

From the information presented in table 5, all the algorithms successfully identified the granitoid area. Here below, some more specific observations are provided for each algorithm:

- K-means algorithm: Associates the granitoid area with cluster label 7, along with some pixels located near the sea, possibly indicating some misclassifications or mixed pixels.
- Hierarchical Hier-Fréchet-based algorithm: Associates the granitoid area with two separate clusters (cluster label 6, cluster label 7). This split of the area into two clusters is probably due to spectral variations within the granitoid area. Additionally, some pixels across the island were included in these clusters.
- SHC algorithm: Provides a more accurate delineation of the granitoid area, compared with the previous algorithms, covering the entire area (cluster label 3). However, there are still some pixels across the island, which are grouped within the same cluster as the granitoid area because they share the same signature pattern with the granitoid pixels.



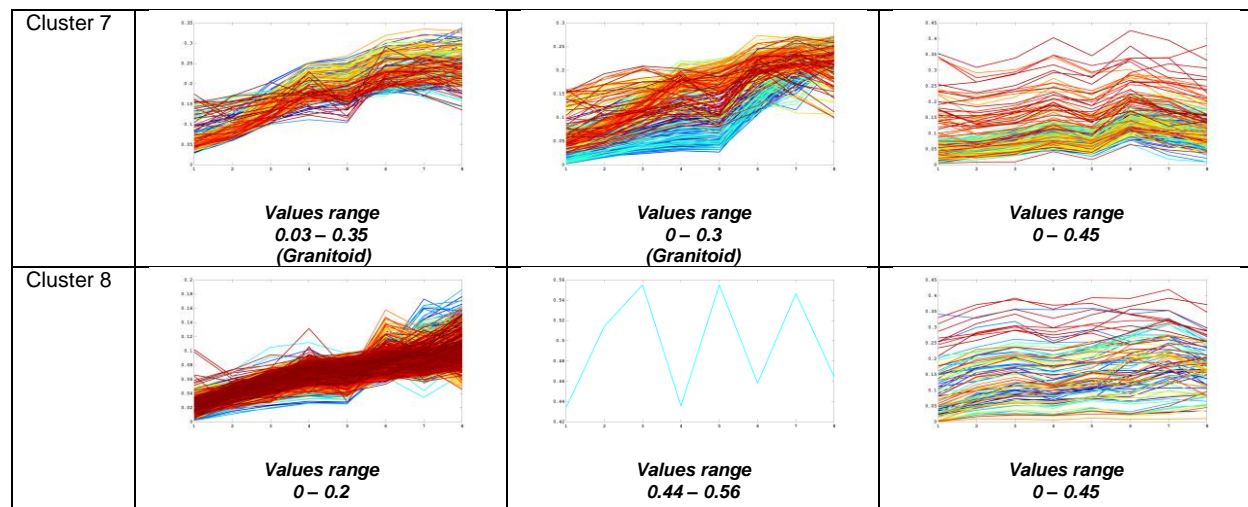


Table 28: Signatures of clusters of all the algorithms in the WorldView-3 VNIR dataset.

Comparing the signatures obtained from the clusters formed by the algorithms (table 28):

- SHC algorithm: Shows the most consistent and common patterns across clusters, indicating a higher level of similarity among the signatures of pixels within each one of them. This suggests a better overall performance in capturing shared spectral patterns within clusters, due to the two step SHC algorithm and the fact that the derivatives of the spectral signatures of each pixel are utilized.
- K-means algorithm: Exhibits good performance in preserving common patterns among clusters, but noticeable errors are evident in cluster 5, indicating some discrepancies or misclassification in this particular cluster.
- Hierarchical Hier-Fréchet-based algorithm: Displays the least favorable performance among the algorithms in terms of signature patterns. In clusters like 2, 5, and 6, the signatures do not exhibit the same pattern, indicating challenges or limitations in accurately delineating clusters based on spectral similarities.

It is important to mention that the SHC algorithm was executed with a higher threshold value (*threshold_1* set to 0.04) in order to achieve satisfactory results in the WorldView-3 VNIR dataset. This adjustment from the significantly lower value threshold used in the Sentinel-2 dataset suggests that the choice of the *threshold_1* plays a crucial role in effectively differentiating pixels into distinct clusters and it should be adjusted based on the distribution of the reflectance values on the corresponding dataset.

6.2 Detection for alteration minerals (Spectral unmixing approach)

The spectral unmixing technique was used to correlate specific locations on the island with various minerals. To validate these findings, information from previous research regarding the distribution of each mineral across the island was employed. [1].

Phases	Abbr.	schist	Granodiorite-1	Granodiorite-2	contact
micas	micas	+	+	+	+
quartz	qz	+	+	+	+
chlorite	chl	+	+	+	
K-feldspars	fsp		+	+	
calcite	cal	+		+	+
plagioclase	plg	+		+	
Mn-oxides	Mn-ox	+			+
goethite	goe				+
halite	hal				+

Table 29: Mineralogy of the lithologies present in the study area according to XRD analysis results on the four samples collected in the field.(Source: [1]).

The observations from the provided information indicate the presence of muscovite (micas) across the entire island. Chlorite is widespread throughout the island except in the areas where schist contacts with granodiorites occur. Additionally, the minerals goethite and Mn-oxides (representing pyrochroite) are exclusively found at the boundaries where schist meets granodiorites.

6.2.1 Reflectance spectra

Relating the results obtained from various spectral unmixing methods with mineral presence, the following observations can be extracted:

- The U-LS method accurately identified muscovite and chlorite across the entire island, although with some pixels being near the sea. However, it unexpectedly found goethite spread throughout the island rather than in contacts, which requires

further investigation. Pyrochroite was not detected despite its anticipated presence at the schist contacts with granodiorites.

- The LASSO method failed to detect muscovite, chlorite, and pyrochroite, with only small amounts of goethite found, which is in partial agreement with [1].
- The NS-LS algorithm identified muscovite across the island and chlorite only on the west side. It also found goethite mainly in pixels near the sea, conflicting with the expected distribution. Finally, pyrochroite was detected across the entire island, contradicting prior research.
- Both the S-LS and the N-LS methods yielded results like the NS-LS approach. Considering the previous research, the U-LS method provides the most plausible outcomes by correctly identifying muscovite and chlorite across the island. However, it failed to identify pyrochroite, possibly due to its small spatial appearance. The unexpected presence of goethite beyond the schist contacts with granodiorites warrants further investigation.

6.2.2 Continuum-removed spectra

In the continuum-removed spectra case, the results differ notably from the reflectance spectra case across various algorithms:

- The U-LS method identified muscovite across the entire island, but with higher abundance values compared to the reflectance case. However, as in the previous case some pixels near the sea are shown to contain muscovite. Chlorite is not detected on the island. Goethite and pyrochroite results seem reasonable, observed with low abundances near schist contacts and at smaller appearances elsewhere.
- The LASSO method shows similar outcomes with the U-LS method, with differences observed in the abundance value of muscovite in each pixel. Moreover, fewer pixels containing pyrochroite were found compared to the U-LS method.
- The NS-LS method identifies muscovite across the island. Small portions of chlorite are found on the west side, but with very low abundance values. The abundance values of goethite and pyrochroite are even smaller than chlorite in

pixels where these minerals are detected.

- The S-LS method detects muscovite and chlorite across the entire island, whereas goethite is observed near the sea. Pyrochroite was found on east side but not at pixels near the eastern granitoid area as expected. On the west side of the island the pyrochroite was found in reasonable areas with the previous research but with higher abundance values than chlorite which contradicts the samples analysis in the previous research [1].
- The NS-LS method found muscovite and chlorite widely across the island with higher abundances than every other method. Muscovite and pyrochroite were found mostly at pixels near the sea but in very small abundances.

Considering the continuum-removed case, the NS-LS method appears to provide the most reasonable results in terms of previous chemical analysis [1]. Muscovite and chlorite are found across the island as expected, although goethite and pyrochroite are identified with small portions at unexpected locations, necessitating further investigation.

Considering the results obtained from both the reflectance and continuum removed cases, it is evident that none of the methods succeeds to accurately reproduce the expected results based on the previous research (table 28). A possible explanation of this problem is the low spatial and spectral resolution of the Sentinel-2 image for this type of studies, which poses challenges in accurately identifying these minerals. Additionally, the variations in results among the different methods further emphasize the complexity and difficulty of accurately resolving this problem with the available data and methods. However, it is important to mention that the minerals that exhibit high abundances in the island (muscovite, chlorite) are successfully identified by most of the algorithms.

7. Conclusion

The aim of this study was (a) to identify granitoid and schist formations in it and (b) to detect alteration minerals on the island of Koutala using EO data. Two different datasets (Sentinel-2, WorldView-3 VNIR) on the Koutala island were utilized, with different spatial and spectral resolutions. Two different machine learning methods (clustering, spectral unmixing) were used to extract the results from the two datasets. Additionally, the continuum removal procedure was also applied to compare spectral patterns (e.g. absorptions) from a common baseline.

Clustering was applied to both datasets to delineate regions with similar spectral signatures, aiming to identify granitoid intrusions and schist formations, as is referred on previous research insights [1] (aim (a)). A new novel clustering algorithm named SHC designed especially for spectral data was introduced due to the inability of common off-the-shelf clustering algorithms (K-means, hierarchical methods) to provide accurate results. In general, the SHC algorithm yielded to better results than the other algorithms, based on visual comparisons of pixel spectral signature patterns of pixels within the clusters. This underlines the significance of spectral analysis using the derivative of a pixel within both two steps of SHC. The SHC algorithm was the only algorithm that was able to identify one out of the two granitoid areas into the Sentinel-2 dataset, while none of the algorithms accurately detected the granitoid clusters. The difficulty is probably due to the low spatial resolution of this dataset, which results to mixed pixel signatures. In the case of World-View-3 VNIR dataset, all the algorithms successfully identified the granitoid areas, highlighting the importance of high spatial resolution. However, the SHC algorithm identified more accurately the granitoid areas than the other algorithms when comparing the spectral signatures of pixels in the granitoid clusters. In general, the SHC algorithm exhibited more "coherent" clusters compared with the clusters produced by other algorithms.

Spectral unmixing was used to detect alteration minerals on Koutala island. It was applied exclusively on the Sentinel-2 dataset, given its larger number of spectral bands (compared to WorldView-3 VNIR dataset).

Various linear unmixing methods were employed for spectral unmixing by applying or not various constraints, such as the sum-to-one constraint, the non-negativity constraint, or other constraints (e.g., as in the Lasso case). The results indicated that despite the low spatial resolution of the Sentinel-2 dataset, the alteration minerals with high degree of appearance in the island (such as muscovite and chlorite) were identified quite accurately by most algorithms. This can be attributed to the fact that mixed signatures in a dataset with low spatial resolution are primarily influenced by minerals with high abundance.

The most favorable outcomes were obtained from the U-LS method for the reflectance spectra case and the NS-LS method for the continuum-removed spectra case ($1 - S_{cr}$), as compared to the results of a previous chemical analysis conducted on the island [1].

In future studies, the potential of spectral unmixing could be further examined in the WorldView-3 VNIR dataset, where clustering results using this dataset were more accurate due to higher spatial resolution compared with the Sentinel-2 corresponding one. Finally, it is worth investigating in the future the relations between the distribution of alteration minerals extracted from the unmixing procedure with the generated clusters from clustering.

8. Data and code availability

The datasets and the related code are available at:

https://github.com/kostsamko/clustering_koutala

References

- [1] O. Sykioti, A. Ganas, C. Vasilatos, Z. Kypridou, “Investigating the Capability of Sentinel-2 and Worldview-3 VNIR Satellite Data to Detect Mineralized Zones at an Igneous Intrusion in The Koutala Islet (Lavreotiki, Greece) Using Laboratory Mineralogical Analysis, Reflectance Spectroscopy and Spectral Indices”, *Bulletin Geological Society of Greece*, vol. 59, no. 1, pp. 175-213, Nov. 2008, <https://doi.org/10.12681/bgsq.31982>
- [2] U.S. Geological Survey, “Remote sensing usage”, <https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used>
- [3] U.S. Government Publishing Office, “Remote sensing data: applications and benefits”, <https://www.govinfo.gov/content/pkg/CHRG-110hhrg41573/html/CHRG-110hhrg41573.htm>
- [4] J. A. Richards, “Remote Sensing Digital Image Analysis: An Introduction”, Springer Berlin Heidelberg, 5th edition 2013, 494 p., Sep. 2012, <https://doi.org/10.1007/978-3-642-30062-2>
- [5] G. Xian, H. Shi, J. Dewitz, Z. Wu, “Performances of WorldView 3, Sentinel 2, and Landsat 8 data in mapping impervious surface”, *Remote Sensing Applications: Society and Environment*, vol. 15, no. 100246, Aug. 2019. <https://doi.org/10.1016/j.rsase.2019.100246>
- [6] S. Theodoridis, K. Koutroumbas, “Pattern Recognition”, Academic Press, 4th edition, 961 p., 2009, <https://doi.org/10.1016/B978-1-59749-272-0.X0001-2>
- [7] S. Liang, “Comprehensive Remote Sensing”, Elsevier, 1st edition, 3134 p., 2017
- [8] S. Gaci, O. Hachay, O. Nicoli, “Methods and Applications in Petroleum and Mineral Exploration and Engineering Geology”, Elsevier, 1st edition, 396 p., 2021, <https://doi.org/10.1016/C2020-0-02594-6>

- [9] European Space Agency, “Sentinel-2 MSI Overview“, <https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/overview>
- [10] European Space Agency, “WorldView-3 Instruments“, <https://earth.esa.int/eogateway/missions/worldview-3>
- [11] European Space Agency, “Spectral signatures“, https://www.esa.int/SPECIALS/Eduspace_EN/SEMPNQ3Z2OF_0.html
- [12] R. N. Clark, T. L. Roush, “Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications”, *Journal of Geophysical Research*, vol. 89, no. B7, pp. 6329-6340, Jul. 1984, [doi:/10.1029/JB089iB07p06329](https://doi.org/10.1029/JB089iB07p06329)
- [13] H. Alt and M. Godau, “Computing the Fréchet distance between two polygonal curves”, *International Journal of Computational Geometry & Applications*, vol.5, no. 01n02, pp. 75-91, Mar. 1995, <https://doi.org/10.1142/S0218195995000064>
- [14] A. Makris, I. Kontopoulos, P. Alimisis, and K. Tserpes, “A Comparison of Trajectory Compression Algorithms Over AIS Data”, *IEEE Access*, vol. 2, pp. 92516-92530, Jun. 2021, <https://doi.org/10.1109/access.2021.3092948>
- [15] J. Wei and X. Wang, “An Overview on Linear Unmixing of Hyperspectral Data,” *Mathematical Problems in Engineering*, vol. 2020, no. 3735403, 12 p., Aug. 2020. <https://doi.org/10.1155/2020/3735403>
- [16] B. Koirala, M. Khodadadzadeh, C. Contreras, Z. Zahiri, R. Gloaguen, and P. Scheunders, “A Supervised Method for Nonlinear Hyperspectral Unmixing”, *Remote Sensing*, vol.11, no. 20, 2458, Oct. 2019. <https://doi.org/10.3390/rs11202458>
- [17] S. Boyd, L. Vandenberghe, “Convex Optimization”, Cambridge University Press, 1st edition, 727 p., 2004, <https://doi.org/10.1017/CBO9780511804441>
- [18] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso”, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp 267–288, Jan. 1996, <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>