Jurnal Elektronika dan Telekomunikasi (JET), Vol. 22, No. 2, December 2022, pp. 78-86
Accredited by KEMDIKBUDRISTEK, Decree No: **158/E/KPT/2021**
doi: 10.55981/jet.507

Mineral Mapping on Hyperspectral Imageries Using Cohesion-based Self Merging Algorithm

Afnindar Fakhrurrozi[a, *], Izzul Qudsi[b], Mochamad Rifat Noor[c], Anggun Mayang Sari[d]

[a]Research Center for Data and Information Sciences
National Research and Innovation Agency
Komplek BRIN Kawasan Bandung, Jl. Sangkuriang No. 21
Bandung, Indonesia
[b]Geovartha
Jl. Dr. Nurdin 1 no. 21, Grogol
Jakarta, Indonesia
[c]Research Center for Mining Technology
National Research and Innovation Agency
Jl. Kawasan PUSPIPTEK Serpong, Tangerang Selatan, Banten
Serpong, Indonesia
[d]Research Center for Geological Disaster
National Research and Innovation Agency
Komplek BRIN Kawasan Bandung, Jl. Sangkuriang No. 21
Bandung, Indonesia

Abstract

Recently, hybrid clustering algorithms gained much research attention due to better clustering results and are computationally efficient. Hyperspectral image classification studies should be no exception, including mineral mapping. This study aims to tackle the biggest challenge of mapping the mineralogy of drill core samples, which consumes a lot of time. In this paper, we present the investigation using a hybrid clustering algorithm, cohesion-based self-merging (CSM), for mineral mapping to determine the number and location of minerals that formed the rock. The CSM clustering performance was then compared to its classical counterpart, K-means plus-plus (K-means++). We conducted experiments using hyperspectral images from multiple rock samples to understand how well the clustering algorithm segmented minerals that exist in the rock. The samples in this study contain minerals with identical absorption features in certain locations that increase the complexity. The elbow method and silhouette analysis did not perform well in deciding the optimum cluster size due to slight variance and high dimensionality of the datasets. Thus, iterations to the various numbers of k-clusters and m-subclusters of each rock were performed to get the mineral cluster. Both algorithms were able to distinguish slight variations of absorption features of any mineral. The spectral variation within a single mineral found by our algorithm might be studied further to understand any possible unidentified group of clusters. The spatial consideration of the CSM algorithm induced several misclassified pixels. Hence, the mineral maps produced in this study are not expected to be precisely similar to ground truths.

Keywords: clustering, hyperspectral, mineral mapping, cohesion-based self-merging.

I. Introduction

Close-range hyperspectral imagery (HSI) on mineral mapping from rock samples has become popular in the past few years. Several studies on utilizing the availability of more bands compared to the multispectral imagery have proven to be more effective in identifying minerals from the rock samples [1]–[3]. Geologists can take advantage of the higher spectral resolution captured by the hyperspectral sensor. It allows us to identify more mineral variations and to distinguish the degree of crystallinity by the subtle wavelength shift of the absorption features [4].

As demonstrated in the other fields, there are various existing algorithms to classify the pixels from the HSI dataset. In mineral mapping, one of the most popular methods is examining the location of the absorption features for the given mineral [5], [6], along with the Spectral Angle Mapper (SAM) [7]–[9]. Both methods require introducing prior knowledge (endmember and absorption feature of each mineral). The first mentioned method, defining the absorption features for each mineral targeted on every sample, would take time, especially for a large dataset with many mineral class variations. Prior information given, mineralogy from geologist interpretation or regional studies, to select end members of SAM is excellent. Still, with this method, we could probably lose some information that the human eyes might not capture.

Some previous studies proved that various clustering algorithms efficiently extract new information about particular classes from HSI datasets [10]–[12]. However, there are still problems that the previous algorithms could

* Corresponding Author.
Email: afnindar.fakhrurrozi@brin.go.id
Received: October 14, 2022 ; Revised: December 12, 2022
Accepted: December 14, 2022 ; Published: December 31, 2022

not capture some minerals with slight variation and small cluster sizes. Thus, we tried to investigate a solution that might be able to tackle the previously addressed problems with an improved clustering algorithm.

Hybrid clustering algorithms gained much research attention due to better clustering results, and this method is computationally efficient [13], [14], [23-24]. The CSM is a hybrid clustering algorithm that serializes two-stage clustering using conventional partitioning-based and hierarchical-based clustering through cohesion-based self-merging. The cohesion-based self-merging is an approach that minimizes the probability density function of points intra-clusters and inter-clusters and also considers the size of the cluster to compute the distance matrix to reconstruct the clusters. Therefore, the CSM is theoretically robust to outliers, and we hypothesized that CSM could segment the minerals with slight distinct wavelength variance in HSI datasets. In this paper, we investigated the use of conventional partitioning-based clustering algorithms, namely K-means++ alongside hierarchical-based clustering algorithms, with agglomerative clustering that being merged using the CSM algorithm. It should be no exception for any HSI classification studies, including mineral mapping. In this paper, the result and performance of hybrid clustering were compared to a conventional single clustering algorithm.

II. Material and Method

A. Material

A total of three hyperspectral images of milled pebbles samples were obtained from the previous research. The hyperspectral images were scanned by ITC University of Twente using SWIR the SisuCHEMA spectral imagery. The spectral SWIR camera produces high spatial resolution images with 0,2mm/pixels and 1000 - 2500 nm spectral range. This camera has 384 spatial pixels and 288 spectral bands. The complete specification of the sensor can be seen in Table 1.

We used the mineral maps from previous research [15], [16] to validate our algorithm classification result. We decided to refer our results to these maps due to our unavailability to access the actual samples. The research we mentioned has their mineral maps validated using several approaches such as XRF, XRD, thin section, and direct observation with the samples. Since we have access to these datasets, we did not perform any pre-processing and directly used the pre-processed reflectance imagery of the samples.

TABLE 1
SisuCHEMA Sensor Detail Information

Optical Characteristic	Typical Specification
Spectral range	1000-2500 nm
Spectral resolution FWHM	12 nm (30µm)
Spectral sampling	5.6 nm
Spectral resolution	rms spot radius < 15 µm
F/#	F/2.0
Slith width	30 µm (50 or 80 µm optional)
Effective slit length	9.2 mm
Electrical characteristic	
Detector	Cyrogenically cooled MCT detector
Spatial pixels	384
Spectral bands	288
Pixel size	24 x 24 µm
Camera output	16 bit CL
SNR	1050:1 (at max signal level)

We used three dominant minerals in the milled pebbled samples: Muscovite, Tourmaline, and Illite, including high-crystallinity Illite, known as HX Illite. The most intriguing case of these mineral variations is the similarity between the white mica minerals, Muscovite and Illite. The occurrence of the water absorption feature on wavelength 1900 nm (feature number 145-150) is the determinant that differentiates Illite from Muscovite [17].

Spectral libraries from USGS were used to identify each mineralogy class and anticipate any non-identified mineral from the previous studies. One of the potential benefits of our clustering method is the possibility of unravelling a minor amount of mineral abundance in the rock samples. Figure 1 shows the stacked spectral for each mineral according to USGS.

B. Method

The hyperspectral images used in this study contains 230 features of spectral reflectance (a subset from the original images within the wavelength of 1043-2486 nm). This subset was range selected to target the absorption features of the AlOH, FeOH, and MgOH mineral groups. As shown in Table 2, the sample size and dimension of HSI datasets were used in this study. The n-dimension of datasets was transformed into two-dimensional space in the data preparation stage. In the next stage, data exploration, we tried to find the ideal number of clusters.

The optimum number of clusters in the datasets was identified using the elbow and the silhouette method. On the one hand, the number of k clusters was chosen randomly in the elbow method. The sum of the squared distance between the centroid and points of each k cluster was computed to obtain the within-cluster sum of square value (WCCS). The WCCS was then plotted against k clusters. The elbow point was determined using graph analysis. On the other hand, the silhouette method uses average intra-cluster distance and mean of inter-cluster

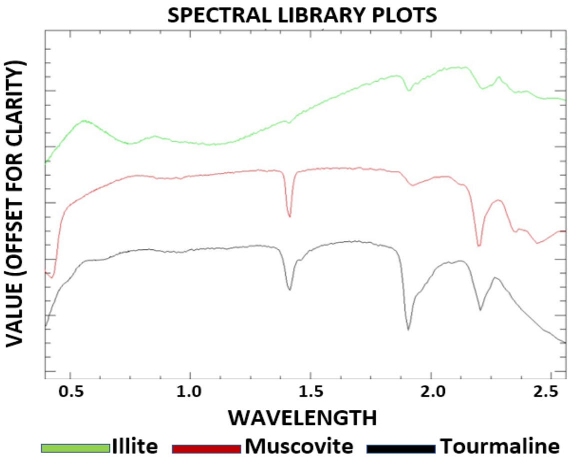


Figure 1. Reflectance Spectra of Muscovite, Illite, and Tourmaline from USGS Spectral library.

TABLE 2
HSI SAMPLE OF MILLED PEBBLED

No	Sample name	Width	Height	Features
1	10a_101012-120551	285	200	230
2	62a_101012-113848	346	220	230
3	76a_101012-114750	255	246	230

distance to obtain the optimum cluster number [18]. The silhouette scoring metrics can be formulated as follows:

$$S = \frac{(b-a)}{(a,b)} \tag{1}$$

where a is the mean of intra-cluster distance, and b is the average of inter-cluster distance. The silhouette metrics (S) scoring ranges from $-1 < S < 1$.

To understand the benefit of the hybrid partitional and hierarchical clustering algorithm, we use the traditional K-means++. After extracting the cluster, we identified the mineral of each class using the USGS spectral library and then later compared it to the previous studies using a confusion matrix.

1) *K-means++ Clustering*

K-means++ is an improved version of simple K-means. Despite the improvement, it still uses simple K-means kernels to cluster the data due to simplicity and speed [19]. The K-means++ method uses seeding from the shortest distance of each data point to initialize the centroids, presumably leading to convergence speeds compared to random centroids initialization on traditional K-means [20]. The K-means++ algorithm can be explained as in Algorithm 1.

2) *Hierarchical Clustering*

In this study, we used the agglomerative clustering technique. This technique is typically a bottom-up approach. It means, in the beginning, each data point of spectral reflectance value of each pixel has its own cluster, and then pair of clusters progressively merged as one to establish the hierarchy. It is necessary to define a certain distance and similarity threshold, known as the linkage criterion, to determine that the pair of clusters can be merged. The linkage criterion ward was used to cluster the datasets. Ward linkage criterion merges pair of clusters with minimum inter-cluster variance after merging [21]. The ward linkage method can be summarized at the beginning, and each n point has its own cluster. Then the initial distance between n points must be squared Euclidean distance. Mathematically it can be formulated as follow:

$$d_{ij} = d(\{X_i\},\{X_j\}) = ||X_i - X_j||^2 \tag{2}$$

where d_{ij} is the distance between clusters i and j, $X_i = \{x_1, x_2, x_3, \dots x_i\}$ set of i data points and $X_j = \{x_1, x_2, x_3, \dots x_j\}$ set of j data points.

3) *CSM*

CSM is a hybrid clustering algorithm that uses a partitioning-based clustering algorithm and a hierarchical-based clustering algorithm [13], [22]. However, CSM has hard constraints as the number of sub-clusters must be greater than the number of objective k clusters, or it cannot work as expected. The detailed hybrid clustering algorithms that were used in this study are described in Algorithm 2.

There are three stages in the CSM algorithm. The first stage is to obtain the values of the mean vector (μ), and covariance matrix of each cluster (ψ) using the maximum likelihood estimator, as the values are unknown. Given the location of n points, $V = (v_1, v_2, \dots, v_n)$, it was estimated by using the following formulas:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} v_i \tag{3}$$

and

$$\psi = \frac{1}{N}\sum_{i=1}^{n} (v_i - \hat{\mu})\,(v_i - \hat{\mu})^T \tag{4}$$

The location of a point and mean vector are d-variate vectors, and the covariance matrix is an absolute definite $d\ x\ d$ matrix. Moreover, it is assumed that the location of n points in each cluster follows a multivariate normal distribution. In the second stage, it computes the values of the probability density function of each pixel $pdf\ f(v)$ by using the following formula:

$$f(v) = (2\pi)^{-\frac{d}{2}}\,(det\ det\ \psi\,)^{-\frac{1}{2}}\ exp\ exp\left[-\frac{1}{2}\,\Delta^2\,(v)\right] \tag{5}$$

where

$$\Delta^2 = (v_i - \mu)^T \psi^{-1} (v - \mu) \tag{6}$$

Algorithm 1. Let $X = \{x_1, x_2, x_3, \dots x_n\}$ set of n data points, the spectral reflectance value of each pixel in cluster $C = \{\,C_1,\ C_2,\ C_3\,, \dots\ C_n\}$, which is a mineral name cluster with centroid k and $D_{(x)}$is the shortest distance of data points to centroid k.

Step 1	:	Cluster centre C_1 initialization randomly at X
Step 2	:	Get a new cluster centre C_n, pick data points x_n with probability $D_{(x)}$weighting function
Step 3	:	Repeat Step 2 until all centroids obtained
Step 4	:	Assign the x_n data points to the closest centroid of C_n
Step 5	:	Get a new k centroid using the Euclidean distance of x_n in the cluster C_n
Step 6	:	Repeat Step 4 and 5 until convergences

Algorithm 2. Let $X = \{x_1, x_2, x_3, \dots x_n\}$ set of n data points, the number of clusters k, the number of sub-cluster of m and $m > k$

Step 1	:	Apply K-means++ to obtain the number of m sub-cluster
Step 2	:	Apply the CSM algorithm to obtain the similarity matrix of each m sub-cluster obtained in Step 1
Step 3	:	Apply the complete link clustering algorithm on m sub-cluster obtained in Step 1 with cohesion as similarity matrix obtained in Step 2 and stop when k clusters are obtained

where d is the space dimension. The final stage is to compute the joinability of each cluster based on the existence of grouped pixels. The basic rule of CSM related to constructing the joinability:

1. Pixels data located closest to the boundary of two clusters are important
2. The way clusters merge should not be because there are only a few pixels.

It can be formulated as follows:

$$join(p, C_i, C_j) = (f_i(v),\ \ f_j(v)) \tag{7}$$

where f_i and f_j, are probability density functions of cluster C_i and cluster C_j. Thus the cohesion of two clusters C_i and C_j, can be formulated as follows:

$$cohesion(C_i, C_j) = \frac{\sum_{p \epsilon\, C_i, C_j} \quad join\ (p, C_i, C_j)}{|C_i| + |C_j|} \tag{8}$$

where $|C_i| + |C_j|$ is the total sum of the size cluster C_i and C_j. The intuitive illustration of joinability between two clusters (C_i, C_j) applied to mineral mapping, as shown in Figure 2. The cohesion or similarity matrix of m sub-cluster is then applied to agglomerative clustering to obtain desired k cluster.

4) Confusion Matrix

Comparative studies were also done to evaluate the mineral distribution of the clustering algorithm. We compared our result and the mineral maps from the previous study through visual comparison and confusion matrix. It was conducted to assess the results qualitatively and quantitatively.

This matrix has four basic values that represent the classification result, as shown in Figure 3. True Positive (TP) represent the number of positive prediction that is actually correct, and True False (TN) is a negative prediction that is actually correct. False Positive (FP) is a positive prediction when the actual value is negative, means an incorrect prediction, and False Negative (FN) where the prediction is negative and the actual value is positive that also means incorrect prediction.

From those basic values, we then calculate Recall, Precision, and Accuracy using these formulas:

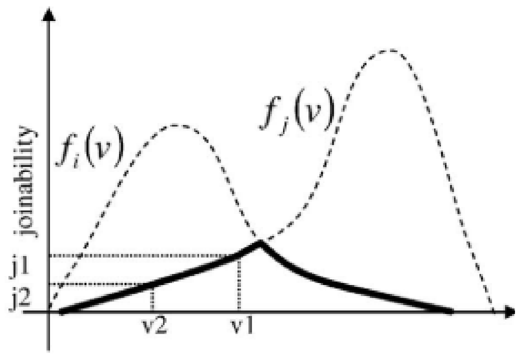
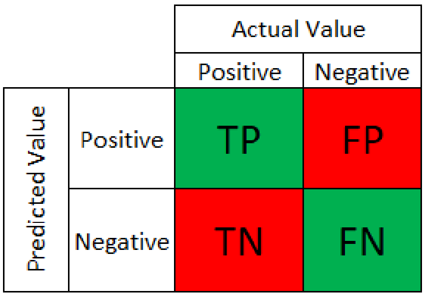


Figure 2. The joinability of cluster C_i and C_j occurs at location pixels v_1 and pixels v_2.



		Actual Value	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	TN	FN

Figure 3. Illustration of the matrix of the classification result.

$$Recall \ = \frac{TP}{TP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Accuracy \ = \frac{TP \ + TN}{TP \ + \ TN \ + \ FP \ + FN} \tag{11}$$

Since we are anticipating other minerals that might be found using this methodology, we can assess the recall of the existing mineral instead of only focusing on the overall accuracy that would potentially be low due to the addition of a new mineral class.

III. Results and Discussion

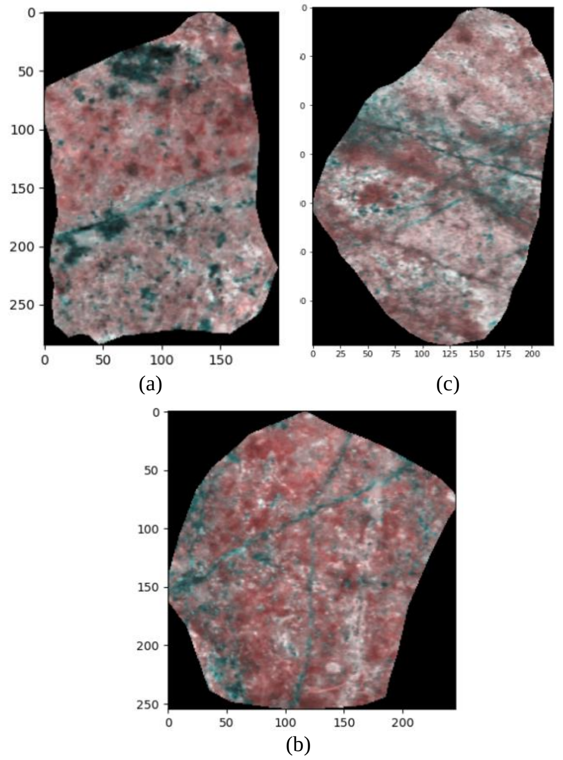
The CSM clustering result was compared to their vanilla clustering algorithm counterpart, K-means++.

A. HSI Data Exploration

The HSI dataset sample consists of three samples, as shown in Figure 4. These samples have been pre-processed and calibrated to obtain the reflectance value of each pixel location in the image.

We also discovered the occurrence of four different mineral clusters using visual analysis according to their different colour, tone, and spectral wavelength, as shown in Figure 5. These wavelength reflectance differences result from each mineral's different levels of light absorbance. However, the segmentation process of each mineral becomes puzzling when there is an identical spectral wavelength to two distinct minerals, such as high-crystalline illite and illite itself.

Hence, there was a challenge to determine the optimum number of clusters using the elbow method and the silhouette method, as the plotted curve in the elbow method shown in Figure 6 is considerably smooth and becomes ambiguous. The ambiguity appeared due to low variance between points of each cluster despite the large gap of identified spectral wavelength starting from number 0 until 190, as shown in Figure 5. As for the silhouette analysis results, as shown in Table 3, the separation distance value of each cluster is getting lowered as the number of clusters arose. Moreover, the presence of unequal size of each cluster appeared when the number of clusters is more than four clusters, although it satisfies the minimum silhouette score, as shown in Figure 7.



(a)

(c)

(b)

Figure 4. Sample picture of HSI milled pebbled: (a) 10a_101012-120551; (b) 76a_101012-114750; (c) 62a_101012-113848.

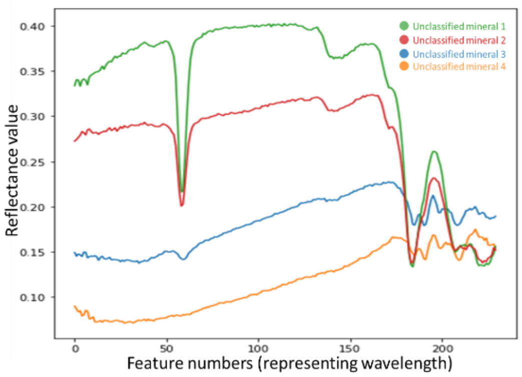


Figure 5. Spectral reflectance of unclassified minerals of 62a_101012-113848 sample.

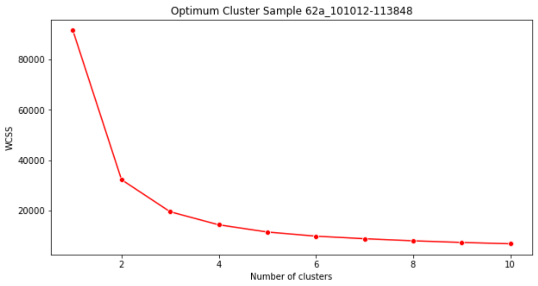


Figure 6. Shows the optimum cluster number selection using the elbow method.

B. Clustering Result

Since the elbow method and silhouette analysis gave ambiguous information regarding the optimum cluster number, we decided to cluster the data into six clusters. The other two additional clusters were intended to capture the anticipated of other unidentified minerals other than the four previously classified minerals. Details about the clustering algorithm results on each sample are discussed below.

TABLE 3
SILHOUETTE SCORE OF EACH SAMPLE DATASET

No	Sample name	C = 3	C = 4	C = 5	C = 6
1	10a_101012-120551	0.42	0.39	0.37	0.33
2	62a_101012-113848	0.44	0.38	0.36	0.33
3	76a_101012-114750	0.40	0.36	0.33	0.30

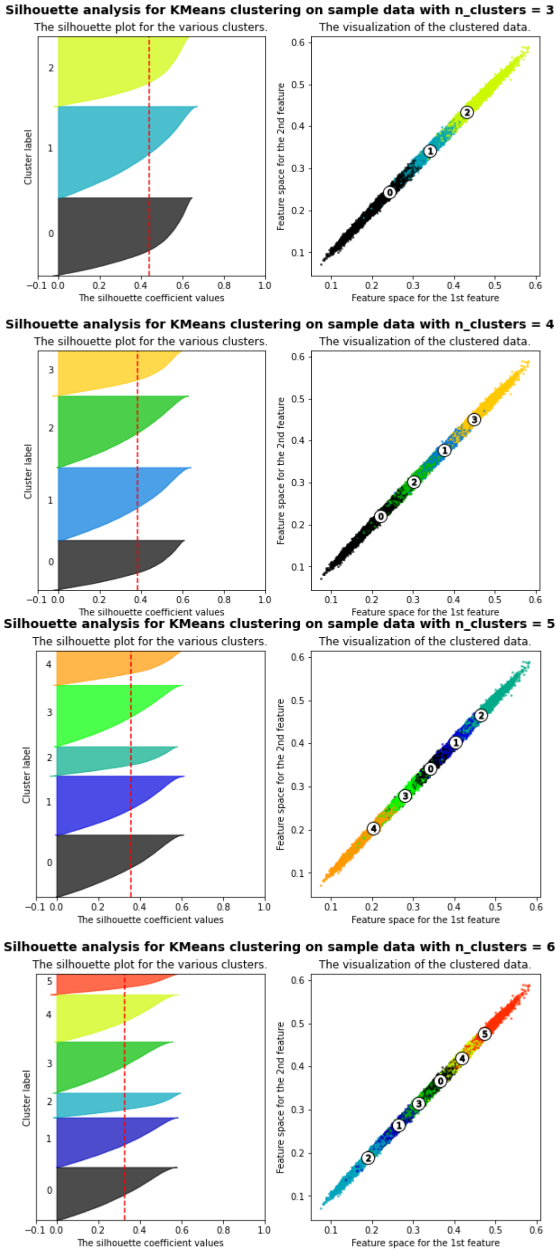


Figure 7. Silhouette analysis with various k numbers of the cluster on sample 62a_101012-113848.

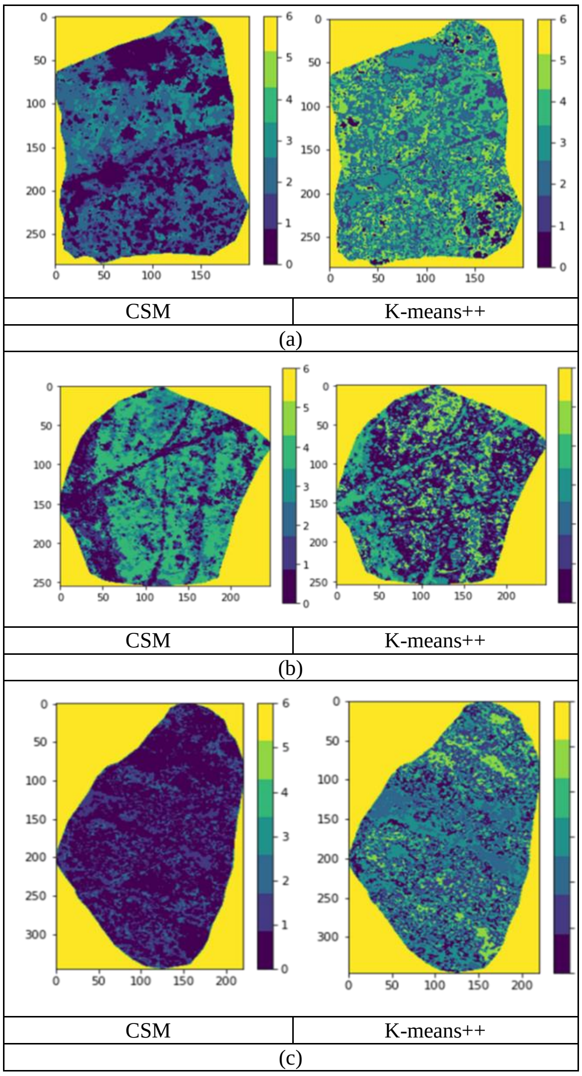
K-means++ was able to extract the expected six clusters. Cconversely, the CSM algorithm can only obtain four to five clusters from these samples. It is highly anticipated that the CSM algorithm merges the clusters from 10 subclusters used in this experiment. CSM algorithm tried to connect clusters with the closest distance in terms of the points located inside the two clusters or intra-cluster. Moreover, it also tried to connect clusters by considering the distance points location between two clusters or inter-cluster. The CSM also consider the size of each cluster in the merging process.

Clustering performance using the CSM algorithm visually has better separation and distinct boundary between clusters than K-means++, as shown in Figure 8.

The mineral map and the spectra of each cluster of sample 10a_101012-120551, as depicted in Figure 9, show an apparent disparity between clusters. Cluster k_1, k_2, k_3 have discrepancy form of wavelength to other clusters despite spectral mixing occurrences. The spectral mixing also occurred in k_0. It means that several pixels are clustered in the wrong label. While k_4 has distinct cluster form and boundary.

Nonetheless, there is also the occurrence of resemblance spectral found between cluster k_0 and k_4 but the CSM algorithm is able to capture both of these clusters. These events are applied to the other samples.

The mineralogy of each class in both mineral maps was determined by comparing them with the USGS spectral library as the baseline, as depicted in Figure 1. The comparison result shows that the mineralogy of each class is defined as depicted in the Table 4A and 4B



CSM | K-means++

(a)

CSM | K-means++

(b)

CSM | K-means++

(c)

Figure 8. Minerals segmentation using CSM and K-means++ on milled pebbled samples: (a) 10a_101012-120551; (b) 76a_101012-114750; (c) 62a_101012-113848.

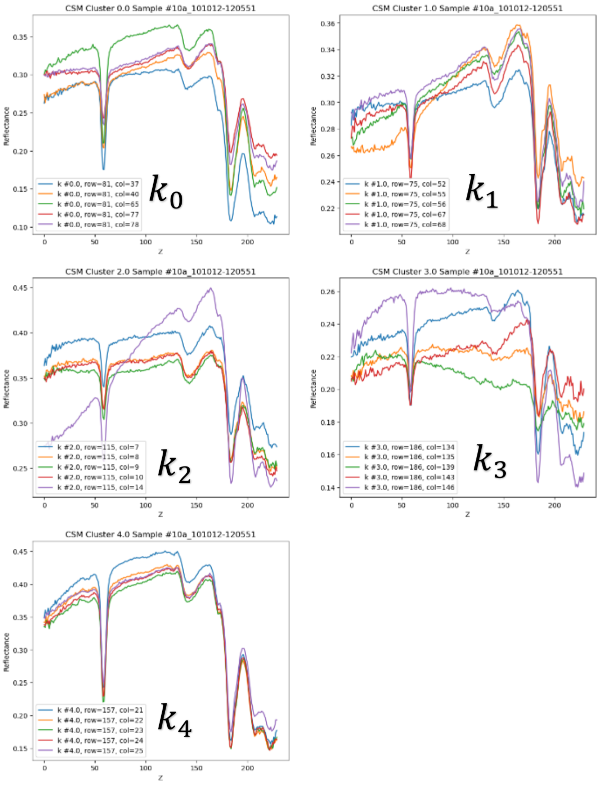


Figure 9. The spectral wavelength of each cluster (k_n) of sample 10a_101012-120551 were segmented using the CSM algorithm.

TABLE 4A
THE MINERALOGY IDENTIFICATION OF EACH CLUSTER k

# k	Mineralogy 10a_101012-120551		Mineralogy 62a_101012-113848	
	CSM	K-means++	CSM	K-means++
0	Illite	Muscovite	Muscovite	Muscovite
1	Tourmaline	Muscovite	Tourmaline	HX Illite
2	HX Illite	Tourmaline	Illite	Tourmaline
3	Muscovite	HX Illite	-	Illite
4	Muscovite	HX Illite	-	Muscovite
5	-	Illite	-	Illite

TABLE 4B
THE MINERALOGY IDENTIFICATION OF EACH CLUSTER k

# k	Mineralogy 76a_101012-114750	
	CSM	K-means++
0	Tourmaline	Muscovite
1	Illite	Muscovite
2	HX Illite	Muscovite
3	Muscovite	Illite
4	Muscovite	Tourmaline
5	-	HX Illite

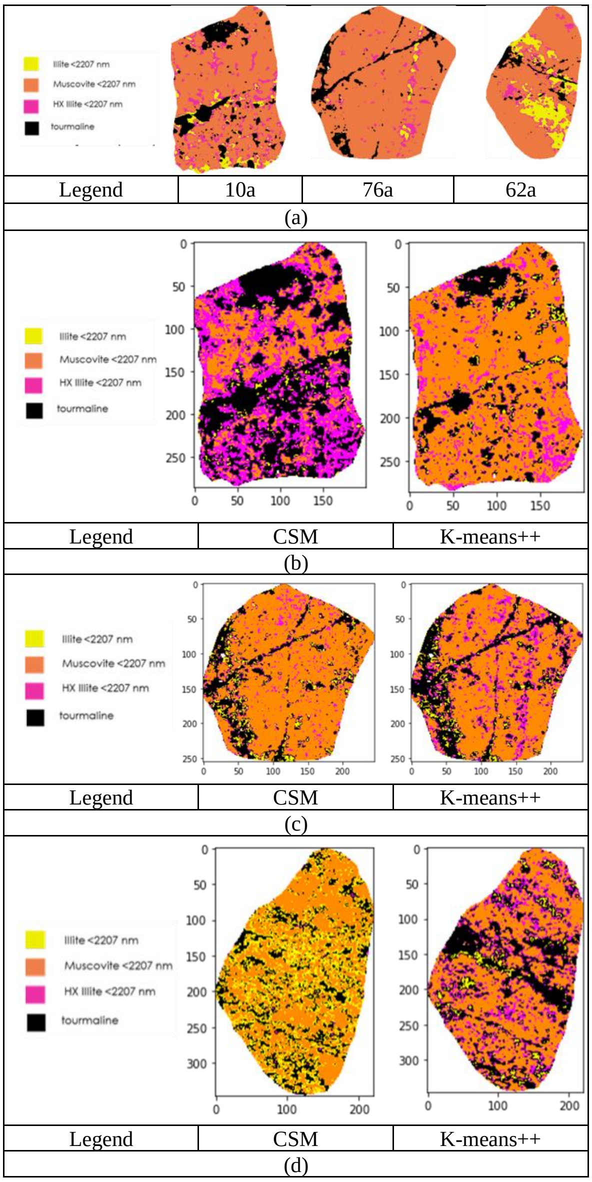
The mineral identification analysis surprisingly discovered slight variations in mineral classes. The spectral wavelength variance between minerals is low. Nonetheless, using the K-means++ algorithm, the minerals extraction result showed that two variations of HX Illite and two of Muscovite were clustered differently.

While using the 62a_101012-113848 sample, the best CSM estimator can only extract three clusters, while K-means++ has six clusters. In this sample, both Illite and Muscovite are clustered into two and three different classes by the K-means++ algorithm. Due to the merged cluster done by CSM, HX Illite could not be captured in this sample. Conversely, in sample 76a_101012-114750, both algorithm able to capture all of the intended minerals despite CSM captures Illite less than K-means++.

C. Comparison with The Previous Study

After identifying the mineralogy of each sample, we compare the new mineral maps and the reference from the previous study. For this purpose, we merge our clusters into four kinds of minerals to match the total classes of the previous study. Minerals were divided into multiple clusters by CSM, and K-means++ was grouped into a single class. We can see the visual comparison of each mineral map as shown in Figure 10.

According to the visual comparison as shown in Figure 10 and confusion matrix assessment as shown in Figure 11, both clustering algorithms produce decent outputs from the samples 10a_101012-120551 and 76a_101012-114750. However, misclassification exists in any minerals. As we can see in these samples, only Muscovite and Tourmaline consistently have good precision and recall values in any algorithm. The



Legend	10a	76a	62a

(a)

Legend	CSM	K-means++

(b)

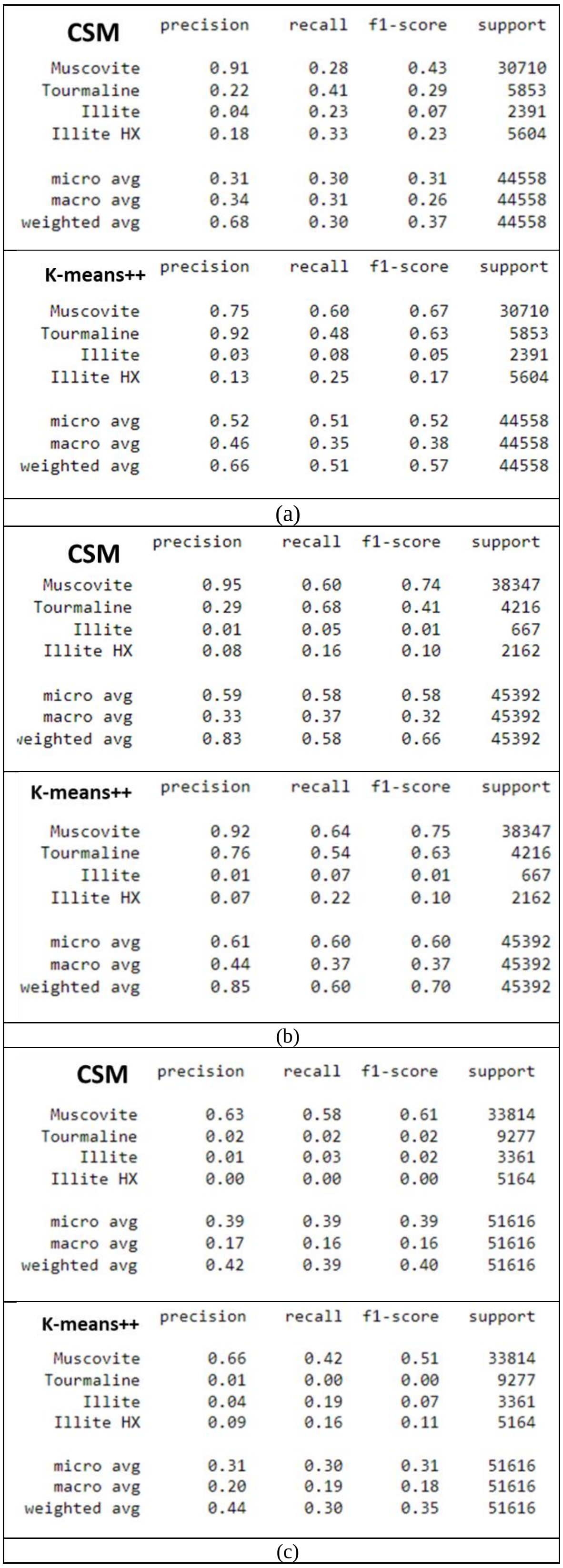
Legend	CSM	K-means++

(c)

Legend	CSM	K-means++

(d)

Figure 10. Visual comparison between the mineralogy after merging with the previous study: (a) Previous Study; (b) 10a_101012-120551; (c) 76a_101012-114750; (d) 62a_101012-1138.



```
CSM          precision    recall  f1-score   support

   Muscovite       0.91      0.28      0.43     30710
  Tourmaline       0.22      0.41      0.29      5853
      Illite       0.04      0.23      0.07      2391
   Illite HX       0.18      0.33      0.23      5604

   micro avg       0.31      0.30      0.31     44558
   macro avg       0.34      0.31      0.26     44558
weighted avg       0.68      0.30      0.37     44558
```

```
K-means++    precision    recall  f1-score   support

   Muscovite       0.75      0.60      0.67     30710
  Tourmaline       0.92      0.48      0.63      5853
      Illite       0.03      0.08      0.05      2391
   Illite HX       0.13      0.25      0.17      5604

   micro avg       0.52      0.51      0.52     44558
   macro avg       0.46      0.35      0.38     44558
weighted avg       0.66      0.51      0.57     44558
```

(a)

```
CSM          precision    recall  f1-score   support

   Muscovite       0.95      0.60      0.74     38347
  Tourmaline       0.29      0.68      0.41      4216
      Illite       0.01      0.05      0.01       667
   Illite HX       0.08      0.16      0.10      2162

   micro avg       0.59      0.58      0.58     45392
   macro avg       0.33      0.37      0.32     45392
veighted avg       0.83      0.58      0.66     45392
```

```
K-means++    precision    recall  f1-score   support

   Muscovite       0.92      0.64      0.75     38347
  Tourmaline       0.76      0.54      0.63      4216
      Illite       0.01      0.07      0.01       667
   Illite HX       0.07      0.22      0.10      2162

   micro avg       0.61      0.60      0.60     45392
   macro avg       0.44      0.37      0.37     45392
weighted avg       0.85      0.60      0.70     45392
```

(b)

```
CSM          precision    recall  f1-score   support

   Muscovite       0.63      0.58      0.61     33814
  Tourmaline       0.02      0.02      0.02      9277
      Illite       0.01      0.03      0.02      3361
   Illite HX       0.00      0.00      0.00      5164

   micro avg       0.39      0.39      0.39     51616
   macro avg       0.17      0.16      0.16     51616
weighted avg       0.42      0.39      0.40     51616
```

```
K-means++    precision    recall  f1-score   support

   Muscovite       0.66      0.42      0.51     33814
  Tourmaline       0.01      0.00      0.00      9277
      Illite       0.04      0.19      0.07      3361
   Illite HX       0.09      0.16      0.11      5164

   micro avg       0.31      0.30      0.31     51616
   macro avg       0.20      0.19      0.18     51616
weighted avg       0.44      0.30      0.35     51616
```

(c)

Figure 11. Confusion matrices and classification reports of the samples: (a) 10a_101012-120551; (b) 76a_101012-114750; (c) 62a_101012-1.

expected problem, the similarity between white mica, caused the mixed cluster between Illite and HX Illite.

Conversely, the clustering results from sample 62a_101012-113848 did not show good numbers in the visual comparison and confusion matrix application. Both algorithm in this sample could not map the mineralogy of the sample close to the mineral map from the previous study and also the spatial distribution of the minerals in this sample.

IV. Conclusion

The mineral maps produced by CSM and K-means++ results were not showing exactly a similar result to the previous study. Especially in the CSM algorithm, several pixels were clustered in the wrong class due to the location of the pixel trapped inside the other class. This different outcome is exactly the intended result of this study, performing data mining to unravel hidden information from the existing mineral maps.

However, both clustering algorithms face the same problem while dealing with white mica minerals. Spectral clustering and density-based spatial clustering combination, along with a little treatment in the pre-clustering process, might be applied to try separating these minerals in a better way.

Despite the limitations demonstrated in this study, both algorithms are able to distinguish the variation of each mineral. Any mineral that was defined as a single unit in previous studies was clustered into two or more classes using this clustering algorithm. Advanced study on what caused these variations might be done to understand if there is any unidentified event, such as other chemical content that is recorded in these mineral variations.

Declarations

Conflict of Interest

The authors have declared that no competing interests exist.

Credit Authorship Contribution

Afnindar Fakrurrozi: Conceptualization, Methodology, Investigation, Software, Formal Analysis, Validation, Writing-Original Draft; Izzul Qudsi: Conceptualization, Methodology, Data Curation, Software, Investigation, Writing-Original Draft; Muhammad Rifat: Resources, Data Curation, Writing-Reviewing & Editing; Anggun Mayang Sari: Visualization, Data Curation, Supervision, Writing – Review & Editing.

Funding

The authors received no financial support for the publication of this article.

Acknowledgment

The authors would like to thanks to Faculty of Geo-information Science and Earth Observation (ITC), University of Twente for supplying these hyperspectral imageries of milled pebbled as the research material in this paper.

References

[1] D. Krupnik and S. D. Khan, “High-resolution hyperspectral mineral mapping: Case studies in the Edwards Limestone, Texas, USA and sulfide-rich quartz veins from the Ladakh Batholith, Northern Pakistan,” *Minerals*, vol. 10, no. 11, pp. 1-16, 2020, doi: 10.3390/min10110967.

[2] Á. F. Egaña, F. A. Santibáñez-Leal, C. Vidal, G. Díaz, S. Liberman, and A. Ehrenfeld, “A robust stochastic approach to mineral hyperspectral analysis for geometallurgy,” *Minerals*, vol. 10, no. 12, pp. 1–32, 2020, doi: 10.3390/min10121139.

[3] R. Gore, A. Mishra, and R. Deshmukh, “Exploring the Mineralogy at Lonar Crater with Hyperspectral Remote Sensing,” *Journal of the Geological Society of India*, vol. 97, no. 3, pp. 261-266, Mar. 2021, doi: 10.1007/s12594-021-1676-4.

[4] M. Pineau *et al.*, “Estimating kaolinite crystallinity using near-infrared spectroscopy: Implications for its geology on Earth and Mars,” *American Mineralogist*, vol. 107, no. 8, pp. 1453–1469, Aug. 2022, doi: 10.2138/am-2022-8025.

[5] R. G. Skirrow *et al.*, “Mapping iron oxide Cu-Au (IOCG) mineral potential in Australia using a knowledge-driven mineral systems-based approach,” *Ore Geology Reviews*, vol. 113, p. 103011, Oct. 2019, doi: 10.1016/j.oregeorev.2019.103011.

[6] F. J. A. Van Ruitenbeek *et al.*, “Mapping the wavelength position of deepest absorption features to explore mineral diversity in hyperspectral images,” *Planet. Space Sci.*, vol. 101, pp. 108-117, Oct. 2014, doi: 10.1016/J.PSS.2014.06.009.

[7] J. M. Meyer, R. F. Kokaly, and E. Holley, “Hyperspectral remote sensing of white mica: A review of imaging and point-based spectrometer studies for mineral resources, with spectrometer design considerations,” *Remote Sens. Environ.*, vol. 275, p. 113000, Jun. 2022, doi: 10.1016/J.RSE.2022.113000.

[8] B. B. Sinaice *et al.*, “Spectral Angle Mapping and AI Methods Applied in Automatic Identification of Placer Deposit Magnetite Using Multispectral Camera Mounted on UAV,” *Minerals*, vol. 12, no. 2, p. 268, Feb. 2022, doi: 10.3390/min12020268.

[9] M. Wang, Z. Huang, X. Zhang, Y. Zhang, and M. Chen, “Altered mineral mapping based on ground-airborne hyperspectral data and wavelet spectral angle mapper tri-training model: Case studies from Dehua-Youxi-Yongtai Ore District, Central Fujian, China,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102409, Oct. 2021, doi: 10.1016/j.jag.2021.102409.

[10] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, “Yinyang K-means clustering for hyperspectral image analysis,” in *Proc. 17th Int. Conf. Comput. Math. Methods Sci. Eng*, 2017, pp. 1625-1636, 2017.

[11] C. Ding *et al.*, “Hyperspectral Image Classification Promotion Using Clustering Inspired Active Learning,” *Remote Sens.*, vol. 14, no. 3, 2022, doi: 10.3390/rs14030596.

[12] Y. Chen, S. Ma, X. Chen, and P. Ghamisi, “Hyperspectral data clustering based on density analysis ensemble,” *Remote Sens. Lett.*, vol. 8, no. 2, pp. 194–203, Feb. 2017, doi: 10.1080/2150704X.2016.1249295.

[13] C.-R. Lin and M.-S. Chen, “Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 145-159, Feb. 2005, doi: 10.1109/tkde.2005.21.

[14] S. Koley and A. Majumder, “Brain MRI segmentation for tumor detection using cohesion based self merging algorithm,” in *Proc. 2011 IEEE 3rd International Conference on Communication Software and Networks*, May 2011, doi: 10.1109/iccsn.2011.6015005.

[15] M. R. Noor, “Ore Texture Measurment Using Infrared Hyperspectral Imagery of Porphyry Cu Pebbles for Copper Content Estimation,” 2019.

[16] I. C. Contreras, C. Hecker, and F. van der Meer, “Mapping Epithermal Alteration Mineralogy with High Spatial Resolution Hyperspectral Imaging of Rock Samples,” *GRSG Conf.*, May 2018, pp. 123-128, 2015.

[17] B. Portela, M. D. Sepp, F. J. A. van Ruitenbeek, C. Hecker, and J. H. Dilles, “Using hyperspectral imagery for identification of pyrophyllite-muscovite intergrowths and alunite in the shallow epithermal environment of the Yerington porphyry copper district,” *Ore Geology Reviews*, vol. 131, p. 104012, Apr. 2021, doi: 10.1016/j.oregeorev.2021.104012.

[18] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, “A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, Feb. 2021, doi: 10.1186/s13638-021-01910-w.

[19] M. Shutaywi and N. N. Kachouie, “Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering,” *Entropy*, vol. 23, no. 6, p. 759, Jun. 2021, doi: 10.3390/e23060759.

[20] D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” in *Proc. The Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2006, pp. 1027–1035.

[21] J. H. Ward, “Hierarchical Grouping to Optimize an Objective Function,” *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236-244, Mar. 1963, doi: 10.1080/01621459.1963.10500845.

[22] C. Ye and C. Zhong, “An improved cohesion self-merging clustering algorithm,” in *Proc. 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Jul. 2011, doi: 10.1109/fskd.2011.6019670.

[23] C. Zhong, T. Luo, and X. Yue, “Cluster Ensemble Based on Iteratively Refined Co-Association Matrix,” *IEEE Access*, vol. 6, pp. 69210-69223, 2018, doi: 10.1109/access.2018.2879851.

[24] T. Alqurashi and W. Wang, “Clustering ensemble method,” *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 6, pp. 1227-1246, Jan. 2018, doi: 10.1007/s13042-017-0756-7.