Hyperspectral Image Segmentation, Feature Reduction and Clustering using *k*-means

1st Hitenkumar Motiyani
Department of Civil Engineering
Institute of Infrastructure, Technology, Research And Management, Ahmedabad
hitenkumar.motiyani.19c@iitram.ac.in

2nd Prashant Kumar Mali
Department of Civil Engineering
Institute of Infrastructure, Technology, Research And Management, Ahmedabad,
prashant.mali.19c@iitram.ac.in

3rd Anand Mehta
Department of Civil Engineering
Institute of Infrastructure, Technology, Research And Management, Ahmedabad,
anandmehta@iitram.ac.in

***Abstract*—In this study, a novel clustering methodology is proposed, which utilizes *k*-means sequentially for performing feature reduction, segmentation, and clustering on hyperspectral imagery, to extract information. The proposed methodology is a multi-stage framework. Initially, *k*-means is utilized to perform feature reduction. In the next stage, *k*-means is again deployed to perform hyperspectral image segmentation, using new feature set obtained from the first stage. Finally, *k*-means clustering is carried out on segmented hyperspectral image by making use of reduced feature set. To evaluate the performance of the proposed methodology, experiments are conducted over three sets of hyperspectral images. For evaluation purpose Normalized Mutual Information (NMI) score and purity are used. The experimental results prove that the proposed methodology has an edge over the other compared clustering methodologies. Results show incorporation of feature reduction and image segmentation techniques leads to significant improvement in accuracies.**

***Index Terms*—Hyperspectral image, clustering, feature reduction, segmentation**

I. Introduction

The multidimensional dataset that makes up hyperspectral data can be displayed as a collection of images. A sufficient amount of spectral information is provided by hyperspectral images to identify and distinguish various materials [1]. Hyperspectral imaging is one of the dataset from which we can find and recognize things in a variety of images. Red, green, and blue are the main colour ranges on which human vision is based, but spectroscopy breaks vision into numerous more bands. The electromagnetic spectrum's hundreds of adjacent spectral bands make up the visual data that hyperspectral sensors collect [2]. The most crucial techniques in hyperspectral imaging that aids in information extraction are cluster analysis, dimensionality reduction, and image segmentation [3].

One important data analysis method that is frequently used for numerous empirical applications in new areas is cluster analysis [4]. The process of identifying groups of things is called clustering. This ensures that the objects in a group will be similar to one another and distinct from the objects in other groups, and that they will generally share the same properties [5].

Some of the clustering work on hyperspectral image are based on the concept of nearest neighbor [6]. While, it is often observed that incorporating spatial regularization into a multiscale clustering framework corresponds to smoother and more coherent clusters when subjected to hyperspectral imaging data and leads to more accurate clustering labels [7]. Also, it is noticed that Sparse Subspace Clustering (SSC) ignores the spatial information in the hyperspectral images, its discrimination capability is limited, hampering the clustering results' spatial homogeneity [8].

Dimensions, or the number of bands, attributes, or variables that make up a dataset, are reduced by dimensionality reduction. The aim is to reduce the number of sizes, which are represented as bands [9]. Since these bands are frequently connected, there is some extra information that adds to the dataset's noise. The learning outcomes of our model are negatively impacted by this extraneous input, hence it is essential to employ feature reduction approaches.

In image segmentation, the image is divided into various regions so that the pixels have a high difference between spaces and a high similarity in each area [10]. It is a useful tool for a variety of tasks, including hyperspectral image processing, and there are various methods for image segmentation, including neural networks and thresholds. Utilizing the clustering method is one of the other techniques that is quick and simple [5].

k-means is one of the widely used clustering methods [11]. Compared to a hierarchical clustering, it is simpler and faster in terms of computation. Additionally, it works well with a lot of different factors. For a different number of groups, however, it produces a different cluster outcome. Consequently, it is crucial to establish the proper amount of clusters. Moreover, a different set of initial centroid values will provide a different set.

In this paper, the main contribution is that a clustering methodology is proposed in which clustering, feature reduction and segmentation are performed by utilizing *k*-means only. Rest of the paper is organised as follows. Section II describes the methodology. Section III presents datasets considered and functions utilised. Section IV presents the results of the study. Section V concludes this paper.

II. METHODOLOGY

The description of proposed methodology is followed by deploying clustering on hyperspectral image initially by utilizing *k*-means in order to perform feature reduction. Secondly,

ISBN:978-1-6654-6200-6/22/$31.00 ©2022 IEEE

it associates *k*-means with image segmentation but using the new feature set obtained from the previous strategy. Lastly, *k*-means clustering is performed on segmented hyperspectral image by employing reduced feature set. Fig. 1 shows visual representation of proposed methodology.

A. *k-means Clustering Algorithm*

k-means is a popular partition-based clustering technique that seeks to locate for the user a predetermined number of *k* clusters, which are represented by their centroids, by lowering the function of square error [11]. Despite being straightforward, *k*-means can be used for a variety of data sets. One clustering technique based on division and non-hierarchical clustering techniques is the *k*-means algorithm. The *k*-means algorithm searches for a cluster between the set of numerical items X and the integer *k*, which lowers the total square errors within clusters.

Starting with the assembly of *k* cluster centres, the *k*-means method is used. The input data points (pixels) are then dispersed among the clusters that already exist in accordance with their Euclidean boundaries, and the closest cluster is chosen. The centre of each cluster is then updated by computing the average (centroid) for each cluster [12]. The membership of each cluster has changed as a result of this update. In order to prevent additional changes in the value of any cluster centre, the pixel vectors are reset repeatedly, and the cluster centres are updated [13]. The steps for performing *k*-means clustering are shown below (Algorithm 1)

Algorithm 1 *k*-means based clustering

Input: $X = \{x_1, x_2, , \ldots, x_n\}//$ Set of *n* pixel vectors.
Output: Set of *k* clusters

- Select *k* pixel vectors randomly to configure the cluster.
- For each pixel vector, find the center of the nearest cluster and set that pixel vector to the appropriate cluster.
- Update the centers of each cluster by using the centroid for the pixel vectors assigned to that cluster.
- Repeat procedure until there is no more change in the value of the means.

B. *Feature Reduction*

The amount of pixels and bands in a hyperspectral image give it an extremely high spectral resolution [1]. Feature reduction is used to reduce the number of features in a dataset without significantly lowering the quantity of information retained [9]. The analysis will proceed more quickly for the smaller dataset. The procedure for accomplishing feature reduction entails arranging the provided dataset into a 2D matrix, where the rows represent pixels and the columns represent bands. Here, *k*-means is performed on bands instead of pixels in order to obtain representative bands from each set of cluster.

C. *Image Segmentation*

The method of image segmentation involves assigning labels to each pixel in an image so that pixels with the same label have certain characteristics [14]. A series of sections that together make up the full image, or a set of features that are retrieved from the image, are the results of image segmentation [10]. Each pixel in a given area has the same computed or defined attributes, such as texture, density, or colour. In relation to the same property, neighbouring communities can differ significantly. The steps for performing image segmentation are shown as, by organizing the dataset in a matrix and applying *k*-means. Further, defining labels to each clusters and obtaining cluster map. Lastly, converting the cluster map into segmentation map using connected components labelling procedure [5].

III. Experimental Setup

The aforementioned clustering methodology was coded in Spyder, Python 3.9.7 on a computer device with intel i5 10 gen processor running at 4.30 GHz using 8 GB of RAM, running on Windows 11 Operating System. This algorithm was tested on different hyperspectral data followed by comparing three clustering methodologies for each strategy, namely *k*-means and *k*-means with feature reduction. In first strategy only *k*-means clustering is performed while in second strategy, dimensionality of dataset is reduced using feature reduction and then *k*-means clustering is carried out.

Furthermore, for comparison few more methods were utilized as followed. *k*-means with Principal Component Analysis (PCA) was advanced where the data was standardized and covariance was found to determine the correlation between them. Moreover, the principal components (PC) of the datapoints were obtained to find out the percentage of variance each PC constitutes. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [15], the algorithm is constituted to count how many samples are located within a small distance ε (epsilon) from it. The values of epsilon and minimum samples were perpetually altered to calculate the best accuracy. Additionally, the proposed method is compared with three existing hyperspectral clustering methodologies which are as follows; Segmentation assisted Nearest Neighbor based Clustering (SNNC), which uses segmentation and nearest neighbor concept [6]. Hierarchical clustering based on rank-two non-negative matrix factorization (H2NMF) [16] and clustering by fast search and find of density peaks algorithm (CFSFDP) [17]. The segmentation is preceded by CFSFDP, similarly to the first step of the proposed methodology.

A. *Data Set*

The first dataset to be taken into consideration is Salinas (contains 204 bands), which was collected by the 224-band AVIRIS (Airborne Visible Infrared Imaging Spectrometer) sensor over the Salinas Valley in California. It is distinguished by a high spatial resolution of 3.7-m pixels [18]. There are 512 lines and 217 samples in the covered region. While, Salinas ground truth has 16 kinds of land use and cover.

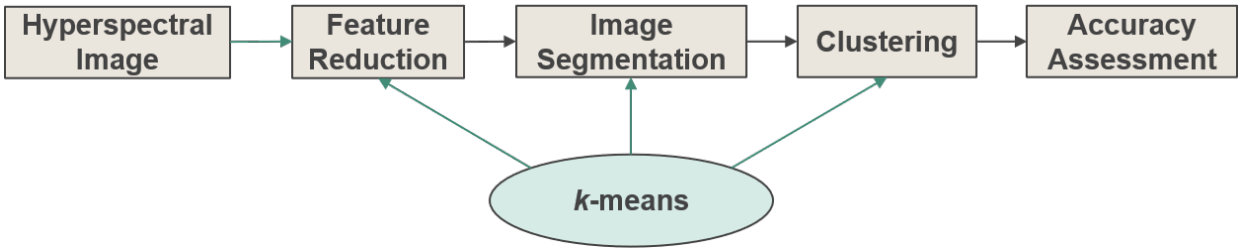

Fig. 1: Proposed Methodology

Pavia Center and Pavia University are the second and third datasets taken into account in the trials. The ROSIS (Reflective Optics System Imaging Spectrometer) sensor photographed these two landscapes during a flight campaign over Pavia, a city in northern Italy. Pavia University has 103 spectral bands and contains nine classes on the other hand Pavia Center has 102 spectral bands and consists nine classes while it's subset has eight. Here, 400 $\times$ 400 pixels and 610 $\times$ 340 pixels are the taken as subset of Pavia Centre and Pavia University, from their original datasets, which are 1096 $\times$ 1096 pixels and 610 $\times$ 610 pixels, respectively. The two dataset's geometric resolution is 1.3 metres.

B. Accuracy Assessment

The clustering evaluation is done using the functions Normalized Mutual Information (NMI) and purity, respectively [19]. The degree to which clusters comprise a single class is referred to as purity. It is a real number between [0, 1] in the range. The performance of clustering will improve as purity increases.

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j| \tag{1}$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ is the set of clusters, $\mathbb{C} = \{c_1, c_2, \ldots, c_j\}$ is the set of classes and N = total number of test pixels. The function known as normalised mutual information assesses how well the two assignments agree with one another. The NMI ranges from 0 to 1. NMI $= 1$ if clustering perfectly recreates the class and NMI $= 0$ if the clustering is random with respect to class membership.

Let $\mathbb{C}$ be the set of classes that are obtained from the ground reference information and Ω is the set of the clusters obtained from the algorithm. Their mutual information, $\text{MI}(\Omega, \mathbb{C})$ can be obtained as follows:

$$\text{MI}(\Omega, \mathbb{C}) = \sum_{k} \sum_{j} p(\omega_k \cap c_j) \log_2 \frac{p(\omega_k \cap c_j)}{p(\omega_k) \cdot p(c_j)} \tag{2}$$

where $p(\omega_k)$, $p(c_j)$ and $p(\omega_k \cap c_j)$ are the probabilities of an arbitrarily selected pixel belonging to cluster ω_k, class c_j and cluster ω_k as well as class c_j at the same time, respectively. Then the NMI can be obtained as follows:

$$\text{NMI}(\Omega, \mathbb{C}) = \frac{\text{MI}(\Omega, \mathbb{C})}{\max(\text{H}(\Omega), \text{H}(\mathbb{C}))} \tag{3}$$

where $\text{H}(\Omega) = -\sum_k p(\omega_k) \log_2 p(\omega_k)$ and $\text{H}(\mathbb{C}) = -\sum_j p(c_j) \log_2 p(c_j)$ are the entropies of Ω and $\mathbb{C}$, respectively.

IV. Results And Discussion

Hyperspectral image classification is carried out using *k*-means with the help of feature reduction and image segmentation. It can be observed from Table IV that classification accuracies (NMI) is significantly improved.

TABLE I: Accuracy values for Salinas.

	k	*k*_fr	*k*_seg	NMI	Purity
k-means	16	-	-	0.7242	0.6734
k-means + FR	16	8	-	0.7254	0.6748
*p*roposed	16	8	14	0.8115	0.6835

Table I contains different values of *k* that are initialised while performing *k*-means clustering for Salinas Dataset. While, *k*_fr represents number of features in bands and *k*_seg shows number of regions. Optimum number of clusters for Salinas was found to be 16. For feature reduction, a set of values of *k*_fr are taken. Starting with *k*_fr = 2 to *k*_fr = 20 multiple value of corresponding NMI are generated and maximum NMI value is marked (as shown in Fig. 2). As there is no significant change in values of NMI so any value in above range can be considered.

Similarly, for image segmentation a set of values of *k*_seg are taken. Initiating with *k*_seg = 2 to *k*_seg = 20 corresponding values of NMI are plotted and maximum NMI value is marked (as shown in Fig. 2). From the optimum values of *k*_fr and *k*_seg NMI is calculated and it's analysed that NMI is significantly improved from 72.42% to 81.15%. Using feature reduction and image segmentation accuracy is increased by 01.20% and 08.61% respectively.

TABLE II: Accuracy values for Pavia Center.

	k	*k*_fr	*k*_seg	NMI	Purity
k-means	8	-	-	0.7694	0.7961
k-means + FR	8	8	-	0.7871	0.8080
*p*roposed	8	8	5	0.8097	0.8155

Table II contains different values of *k* that are initialised while performing *k*-means clustering for Pavia Center Dataset.

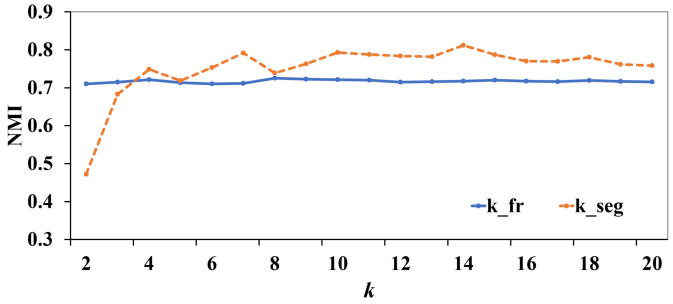

Fig. 2: Optimum value of NMI for Salinas using feature reduction and image segmentation.

For feature reduction, a set of values of k_fr are taken. Starting with k_fr = 2 to k_fr = 20 multiple value of corresponding NMI are generated and maximum NMI value is marked (as shown in Fig. 3).

Similarly, for image segmentation a set of values of k_seg are taken. Initiating with k_seg = 2 to k_seg = 16 corresponding values of NMI are plotted and maximum NMI value is marked (as shown in Fig. 3). From the optimum values of k_fr and k_seg NMI is calculated and it's analysed that NMI is improved from 76.94% to 80.97%. Using feature reduction and image segmentation accuracy is increased by 01.77% and 02.26% respectively.

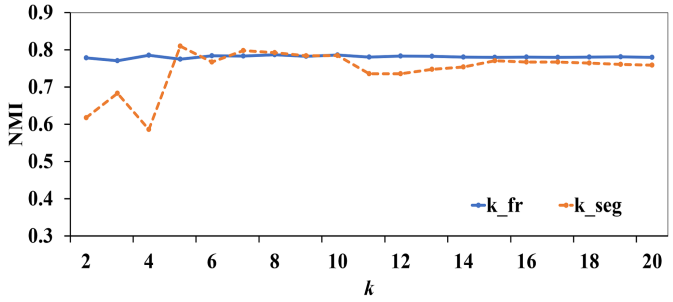

Fig. 3: Optimum value of NMI for Pavia Center using feature reduction and image segmentation.

TABLE III: Accuracy values for Pavia University.

	k	k_fr	k_seg	NMI	Purity
k-means	9	-	-	0.5330	0.6696
k-means + FR	9	5	-	0.5562	0.6902
proposed	9	5	6	0.6166	0.7362

Table III contains different values of k that are initialised while performing k-means clustering for Pavia University dataset. For feature reduction, a set of values of k_fr are taken. Starting with k_fr = 2 to k_fr = 20 multiple value of corresponding NMI are generated and maximum NMI value is marked (as shown in Fig. 4).

Similarly, for image segmentation a set of values of k_seg are taken. Initiating with k_seg = 2 to k_seg = 20 corresponding values of NMI are plotted and maximum NMI value is marked (as shown in Fig. 4). From the optimum values of k_fr and k_seg, NMI is calculated and it's analysed that NMI is improved from 53.30% to 61.66 %. Using feature reduction and image segmentation accuracy is increased by 02.32% and 06.04% respectively.

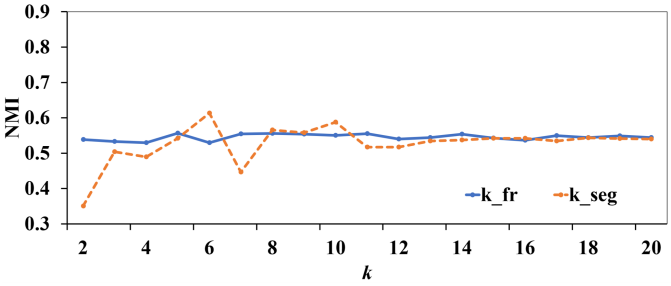

Fig. 4: Optimum value of NMI for Pavia University using feature reduction and image segmentation.

Table IV shows the comparison between different methods. Nonetheless, in k-means + PCA, the proportion of principal components (PC) are selected such that, PC are able to account for at least 99% of the variance in the data set. While, in DBSCAN the optimum epsilon value obtained for Salinas, Pavia Center and Pavia University are 500, 600 and 600, respectively while minimum sample taken are 40, 80 and 60. Fig. 5 shows the comparison between the original image and the improvised images of dataset Salinas.

From Table IV, it can be observed that proposed methodology has performed better than the compared clustering methodologies. Further it can be seen that proposed method resulted in higher values for both NMI and Purity.

V. CONCLUSION

In this paper, several approaches of image classification based on k-means clustering, feature reduction and image segmentation have been discussed. In methodology, experiments were too performed on three datasets by organizing the data into a matrix and finally applying classification methods through k-means. Lastly, the results gathered were compared. Higher accuracy is achieved by image segmentation. Thus, it can be concluded that the proposed method has shown significant improvement over other clustering methods.

VI. ACKNOWLEDGEMENT

The authors would like to thank Prof. David Landgrebe & Prof. P. Gamba for Hyperspectral Images used in this work.

REFERENCES

[1] A. Femenias and S. Marín, "Hyperspectral imaging, chapter 15," in *Electromagnetic Technologies in Food Science*. John Wiley Sons, Ltd, 2021, pp. 363–390.

[2] M. Mateen, J. Wen, D. Nasrullah, and M. Azeem Akbar, "The role of hyperspectral imaging: A literature review," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, 09 2018.

[3] A. Mehta and O. Dikshit, "Comparative study on projected clustering methods for hyperspectral imagery classification," *Geocarto International*, vol. 31, pp. 296–307, 05 2015.

TABLE IV: Accuracy assessment for Salinas, Pavia Center and Pavia University from different methods.

Dataset	Salinas		Pavia Center		Pavia University	
	NMI	Purity	NMI	Purity	NMI	Purity
k-means	0.7742	0.6734	0.7694	0.7961	0.5330	0.6696
k-means + PCA	0.7656	0.6785	0.7750	0.8048	0.5663	0.7150
k-means + FR	0.7254	0.6748	0.7871	0.8080	0.5562	0.6902
DBSCAN	0.7228	0.5846	0.7591	0.7684	0.5059	0.5852
SNNC	0.6860	0.5698	0.7978	0.7080	0.1776	0.4954
H2NMF	0.6372	0.5877	0.7596	0.8592	0.4641	0.6261
CFSFDP	0.6960	0.6703	0.7591	0.8219	0.5772	0.6973
proposed	0.8115	0.6835	0.8097	0.8155	0.6166	0.7362

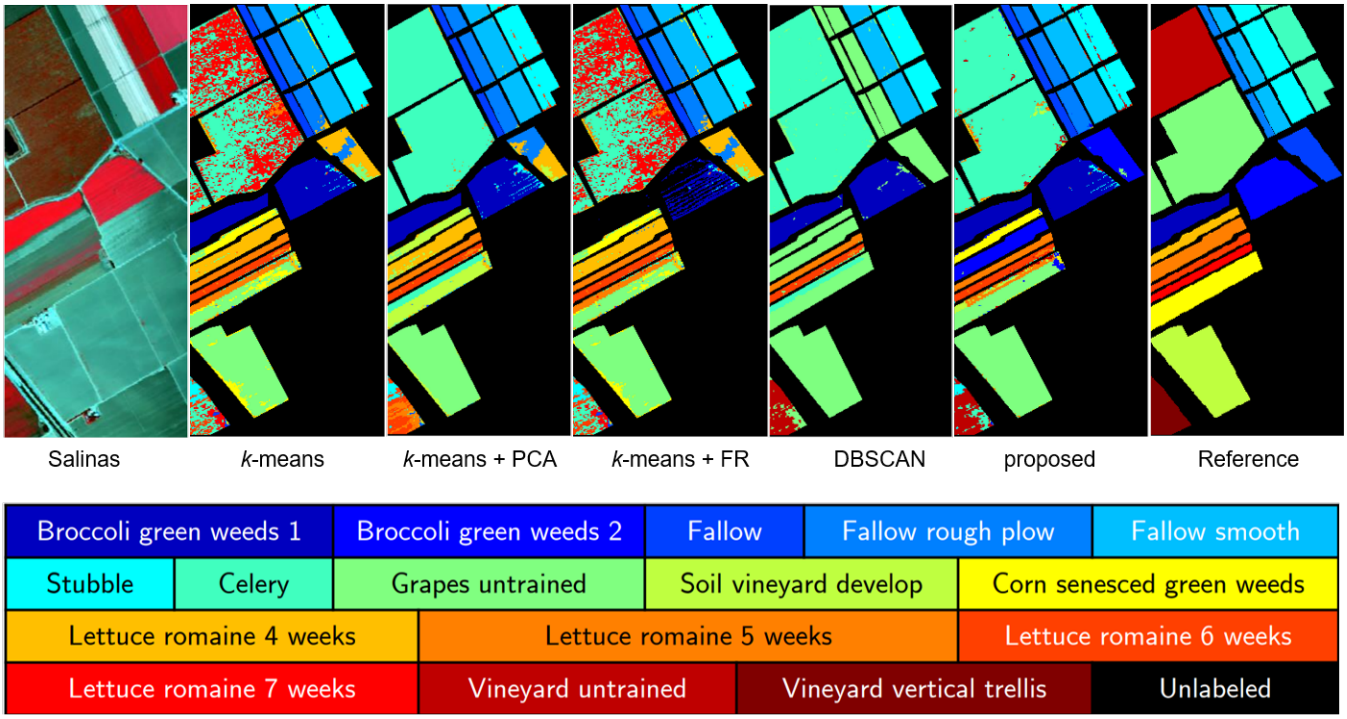

Fig. 5: Clustering maps of Salinas Dataset.

[4] A. F. Alkarkhi and W. A. Alqaraghuli, "Cluster analysis, chapter 11," in *Easy Statistics for Food Science with R*, A. F. Alkarkhi and W. A. Alqaraghuli, Eds. Academic Press, 2019, pp. 177–186.

[5] A. Mehta and O. Dikshit, "Projected clustering of hyperspectral imagery using region merging," *Remote Sensing Letters*, vol. 7, no. 8, pp. 721–730, 2016.

[6] A. Mehta and S. Pasari, "Hyperspectral image clustering using nearest neighbor," in *2021 IEEE International India Geoscience and Remote Sensing Symposium (InGARSS)*, 2021, pp. 194–197.

[7] S. L. Polk and J. M. Murphy, "Multiscale clustering of hyperspectral images through spectral-spatial diffusion geometry," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 4688–4691.

[8] C. Hinojosa, E. Vera, and H. Arguello, "A fast and accurate similarity-constrained subspace clustering algorithm for hyperspectral image," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 773–10 783, 2021.

[9] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag, Berlin, Heidelberg, 5th Edition, 2013, pp. 403–446.

[10] A. Mehta and O. Dikshit, "Segmentation-based projected clustering of hyperspectral images using mutual nearest neighbour," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 12, 2017, pp. 5237–5244.

[11] Y. Li and H. Wu, "A clustering method based on k-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 12 2012.

[12] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[13] A. K. Jain, M. N. Murty, and P. J. Flynn., *Data Clustering: A Review*, 1999, pp. 264–323.

[14] H. Chokshi and A. Agarwal, "Image segmentation," in *Advanced Sensing in Image Processing and IoT 1st Edition*, 02 2022, pp. 43–62.

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[16] N. Gillis, D. Kuang, and H. Park, "Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2066–2078, 2014.

[17] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[18] M Graña, MA Veganzons, B Ayerdi, "Hyperspectral remote sensing scenes," accessed: 2021-12-26. [Online]. Available: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[19] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, cambridge Books.