Optimum Index Factor and Cloud Removal on the Landsat Imagery Data Processing

M S Ziliwu¹, Tulus¹,², Sutarman³, M Zarlis¹, Z Situmorang¹,³, A S Harahap²

¹Postgraduate School, Department of Technical Information, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
²Department of Mathematics, Faculty of Mathematic and Natural Sciences, Universitas Sumatera Utara, Medan, Indonesia
³Faculty of Computer Science, Universitas Katolik Santo Thomas, Medan, Indonesia

*E-mail : tulus@usu.ac.id

Abstract. The research by using landsat imagery data which is a remote observation and recording technology equipped with Enhanced Thematic Mapper Plus (ETM +) sensor has been conducted. Landsat-7 has a sensor that produces images with 7 bands of 1, 2, 3, 4, 5, and 7 with spatial resolution of 30x30 m and band 6 with a spatial resolution of 120x120 m. The image processing begins with obtaining the best combination of Optimum Index Factor (OIF) is one of the methods used to get the proper combination. Landsat data imagery also contains cloud cover which can interfere and hamper in the process of obtaining more accurate landsat image data. Cloud removal by using K-Means Clustering process can separate and clusterize each object in the image based on the closest color and distance so that clouds and other land cover can be clustered and freed from the cloud visually.

1. Introduction
Landsat-7 is a remote viewing and recording technology equipped with Enhanced Thematic Mapper Plus (ETM +) sensor which is a continuation of the Thematic Mapper (TM) program that has been carried since Landsat-5. Landsat-7 which is equipped with ETM + sensor has advantages that are by using channel or band in observation and remote image recording. Landsat-7 has bands of 1, 2, 3, 4, 5, 7 with spatial resolution of 30 x 30 m and band 6 with spatial resolution of 120 x 120 m. The Landsat data image has the ability to cover the same area on the surface of the earth for every 16 days, at an orbit altitude of 705 km [1]. Another advantage of Landsat data imagery techniques is to produce digital data which can then be processed quantitatively with the help of computers so that the information can be produced quickly and more accurately [2]. However, the image of landsat data turned out to produce standard information in 7 bands so that in the process of image processing requires inefficient time, for it needed an appropriate combination to get the most optimal information. The statistical value can be used to select the optimal combination of the three bands on the landsat image. The resulting band combination level can be done using the Optimum Index Factor (OIF) method.

The OIF value is calculated by a simple formula that uses the standard deviation of the band and the correlation coefficient between pairs of bands. By using three bands, the highest value of OIF will make the best band combinations. The highest OIF values will provide the greatest information (largest
diversity) with the smallest duplication, thus providing more information\[3\]. The analysis of a study always aims to produce an information that comes from data. Data is defined as a representation of the real world object such as humans, animals, events, concepts, circumstances and etc. which are recorded in the form of numbers, letters, symbols, texts, images, sounds or combinations. In other words, the data is a material or raw material that has no meaning or has no direct effect on the user so it needs to be processed to produce something more meaningful. One form of a data is image\[4\]. Land Satellite (Landsat) is the oldest program in earth observation devices.

Landsat began in 1972 under the name of Earth Resources Technology Satellite (ERTS-1). Since its launch in 1972, landsat technology has been widely used, and is successful in monitoring natural resources. In addition, a very important ability of it is able to provide input data and information for the management of natural resources on earth\[5\]. $OIF$ indicates the best band combinations may differ depending on the recording time and the selected image area \[6\]. In the image of landsat data, cloud is a noise that until now, there are still frequent researches conducted to eliminate the cloud. A thin cloud cover can affect the digital value of the pixels below and around the cloud, while the thick cloud cover can remove all the Earth’s surface object information that lies beneath the cloud \[7\].

Methods to remove clouds and cloud shadows (cloud removal) have been widely developed. In general, the initial phase of the cloud removal method used is to perform cloud and cloud shadow detection. The difficulty level for determining clouds and shadows is very high, indicating that clouds and shadows are detected resulting in 30-40 % errors of reflectance observed for affected pixels\[8\]. Landsat data imagery often produces image objects that are less obvious or not visible due to noise or disturbance. One of the frequent noises is the image covered by the cloud. So that the actual object image is vague or invisible. With existence of centroid, each objects may exhibit respective differences according to color indexation which are counter from the nearest distance to the centroid. Centroid could minimalize imbalance within the cluster \[9\]. To remove clouds in the image can be used method of cloud removal. In implementing the cloud removal process, K-Means Clustering is expected to cluster cloud and not cloud.

2. **Research Methodology**

To conduct this research, one area of the city located on Nias Island was chosen as a sample and adjusted to the observation time and image recording by Landsat. The distance and area of this research is up to 500 hectares. For data processing purposes, Personal Computer (PC) compatible with Microsoft Windows operating system and relevant software were used. The plot of the research design can be seen in figure 1.
3. Results and Discussion

This study used Landsat imagery data taken in 2013. The size of data is 512 x 445 pixels. The bands used were bands 1, 2, 3, 4, 5, and 7 since its similarity of wavelength. The image of the processed Landsat data can be seen in Figure 2.

To obtain the best combination of Landsat image data bands according to the method (OIF) is by evaluating the result of the standard deviation of each band with the correlation coefficient between bands. There are 6 bands to be evaluated; each band performed mapping to calculate the standard deviation for each band. Then, the correlation between two different bands is determined so that the upper or lower triangular matrix is formed. To get a standard deviation result each band first calculated the average of each band by calculating the gray level of each band distributed with the number of pixels of each band. After the average of each band is revealed, the calculating process of the variant of each band with the number of gray band level reduced by the average divided by the number of bands minus one band. Standard deviation is obtained from the square root of the number of variants of each band. The results of the calculation of the mean, variant, and standard deviation of each band can be seen in Table 1.
Figure 2. The landsat imagery data of Band 1, 2, 3, 4, 5 and 7

Table 1. Mean, Variant and Deviation Standard for each band

| Note | Band 1    | Band 2    | Band 3    | Band 4    | Band 5    | Band 7    |
|------|-----------|-----------|-----------|-----------|-----------|-----------|
| Xhi  | 14.424.656| 10.187.480| 8.830.985 | 5.604.163 | 5.766.946 | 4.194.589 |
| N    | 227.840   | 227.840   | 227.840   | 227.840   | 227.840   | 227.840   |
| Mean | 63.310    | 44.713    | 38.760    | 24.597    | 25.311    | 18.410    |
| Var  | 2961.830  | 2214.181  | 2207.544  | 1380.241  | 1567.796  | 803.112   |
| $S_b$ | 54.423    | 47.055    | 46.985    | 37.152    | 39.595    | 28.339    |

After the result of the band combination with the standard deviation and the combination coefficients of the band is determined, it will then be tested in implementing the OIF. Standard deviations are distributed with correlation coefficients due to the band combinations. From the 20 band combinations, there are values in accordance with Table 2.
Table 2. The result of combination of band with OIF method

| Band Combination | OIF   | Rank |
|------------------|-------|------|
| 457              | 6.567 | 1    |
| 347              | 6.205 | 2    |
| 247              | 6.202 | 3    |
| 357              | 6.069 | 4    |
| 257              | 6.066 | 5    |
| 147              | 5.917 | 6    |
| 157              | 5.785 | 7    |
| 237              | 5.719 | 8    |
| 345              | 5.553 | 9    |
| 245              | 5.550 | 10   |
| 137              | 5.443 | 11   |
| 127              | 5.440 | 12   |
| 145              | 5.292 | 13   |
| 234              | 5.240 | 14   |
| 235              | 5.131 | 15   |
| 135              | 5.117 | 16   |
| 134              | 4.992 | 17   |
| 124              | 4.989 | 18   |
| 125              | 4.886 | 19   |
| 123              | 4.615 | 20   |

Where:

\[
OIF = \frac{\sum_{h=1}^{3} S_h}{\sum_{h=1}^{3} \text{abs}(f, g)}
\]

Rank = Rank OIF

Based on Table 2, the greatest number of combinations are bands 4, 5 and 7 with 6.567 while the smallest combination are bands 1, 2 and 3 with 4.615. From the test results, it is indicated that the OIF combination results are sorted based on the biggest of the results to the smallest result with the descending mode resulting in a combination of bands 4, 5 and 7 and is displayed in figure 4.
Figure 3. The result of landsat imagery data combination of band 4, 5 and 7

To apply the cloud removal method, the image used as the data processed in this study is the Landsat data image data which is the partial cutting of the result of the band combination through the OIF method which is processed and displayed in figure 4.

Figure 4. Landsat image with land coverage

Based on Table 2, it shows that the visual appearance of land cover are clouds and cloud shadows. In implementing the cloud removal method, K-means clustering was used for the process. The stages in the K-means clustering process are known to by preparing the image in the color level, classified in RGB to make it easier in determining the number of clusters. Determining the number of clusters K of the image is the first thing to do in case of determining the number of 10 clusters. The cluster center that becomes the initial value is determined by random K data, so that it becomes the reference in the process of determining the distance of each calculation by using euclidian distance. The K-means algorithm is iterated in increasing the variation of value in each cluster, making it easier to place in the closest group calculated from the midpoint of the cluster determined by 20 iterations. The result produces a cluster centroid as shown in Table 3.
Table 3. The centroid of cluster

| Centroid of cluster | X               | Y               |
|---------------------|-----------------|-----------------|
| 1                   | 85.7270826071802 | 158.111039187372|
| 2                   | 101.934957588883 | 142.522107922758|
| 3                   | 108.671988892746 | 97.9392572023603|
| 4                   | 90.037643436878  | 136.599916036944|
| 5                   | 78.3435569285084 | 166.250992939100|
| 6                   | 121.583791895948 | 127.15557788894 |
| 7                   | 105.508571428571 | 135.522932330827|
| 8                   | 97.0615152634170 | 147.434053421960|
| 9                   | 98.1888362983516 | 117.918230781786|
| 10                  | 91.6528981429375 | 152.523804164322|

Where:
X = Rows of Pixel
Y = Column of Pixel

The Landsat data image shows the calculation of the cluster centroid based on the number of pixels present in the column and the rows evaluated. The relationship between the x-axis and the y-axis is the cluster's centroid. Cluster centroid indicates the distance between the cluster centroid in carrying out the clustering process. After all iterations have been executed and processed so that the closest distance to the cluster centroid will be clustered until it reaches the predetermined number of clusters, therefore the cluster centroid is no longer experiencing changes. The results of clustering can be seen in figure 5 below. Based on figure 6, it shows that through the method of cloud removal with the K-means clustering process, it can produce a cloud-free image.

Figure 5. A. Cluster result that are considered cloud, B. The cloud-free image
4. Conclusion

It can be concluded that the result of the band combination of landsat data using Optimum Index Factor (OIF) method according to the sample that has been used as the data which are the combination of band 4, 5 and 7. The result of the combination is then re-processed because of the noise (cloud cover). Finally, it can be generated a cloud-free image (cloud removal) with K-Means Clustering which is indicated by the cluster results that can be seen by visualizing the landsat data image that free from the cloud.

5. Acknowledgement

Authors would like to acknowledge Regent of Nias Utara, Sumatera Utara to provide the research funding.

References

[1] Sitanggang G 2008 Majalah Sains dan Teknologi Dirgantara 3 33
[2] Richards J and Jia X 2006 Remote Sensing Digital Image Analysis: An Introduction (New York: Springer)
[3] Chavez P S, Berlin G L and Sowers L B 1982 J. Appl. Photogr. Eng., 8 23
[4] Yakub 2012 Pengantar Sistem Informasi (Yogyakarta: Graha Ilmu)
[5] Skidmore A K, Oindo B O and Said M Y 2003 Biodiversity Assessment by Remote Sensing Proc. 30th Int. Symp. Remote Sens. Environ. Inf. risk Manag. Sustain. Dev., pp. 1–4, 2003.
[6] Sirait P and Arymurthy A M 2010 Cluster centres determination based on KD tree in K-Means clustering for area change detection International Conference on Distributed Frameworks for Multimedia Applications, DFmA
[7] Trisakti B 2014 J. Penginderaan Jauh 11 88
[8] Simpson J J and Stitt J R 1998 IEEE Trans. Geosci. Remote Sens., 36 880
[9] Sirait K, Tulus and Nababan E B 2017 J. Phys.: Conf. Ser. 930 12016