



Thesis for the Degree of Master of Engineering

Analysis and Detection of Red Tide Sea Areas Using Machine Learning-Based Clustering

by

Mi So Park

Division of Earth and Environmental System Sciences Major of Spatial Information Engineering The Graduate School Pukyong National University

February, 2023

Analysis and Detection of Red Tide Sea Areas Using Machine Learning-Based Clustering (머신러닝 기반 클러스터링을 적용한 적조 해역 분석 및 탐지 연구)

Advisor: Prof. Hong Joo Yoon

by Mi So Park

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Engineering

in Division of Earth and Environmental System Sciences Major of Spatial Information Engineering, The Graduate School, Pukyong National University

February 2023

Analysis and Detection of Red Tide Sea Areas Using Machine Learning-Based Clustering



February 17, 2023

CONTENTS

CONTENTS i
LIST OF FIGURES ii
LIST OF TABLES iii
ABSTRACT iv
I. Introduction 1
II. Data and Methods7
2.1 Study area and date7
2.1.1 Study area ······7
2.1.2 Reesarch data ······9
2.2 Research method15
2.2.1 Unsupervised Learning Clustering15
2.2.2. Cluster Validity Assessment
III. Theoretical Background of Red Tide waters Analysis 23
3.1. optical properties23
3.2. biological properties27
IV. Result28
4.1. Analysis of clustered red tide waters
V.Conclusions and considerations43
References 45

LIST OF FIGURES

Fig. 1. Harmful algal blooms caused by C.polykrikoides
Fig. 2. Cochlodinium polykrikoides observed under a microscope
Fig. 3. South Sea of Korea
Fig. 4. Map of red tide occurrence on August 14, 2015 provided by the
National Institute of Fisheries Scienc14
Fig. 5. gaussian mixture model17
Fig. 6. schematization of the general Gaussian mixture model
Fig. 7. schematization of bayesian gaussian mixture model
Fig. 8. Absorption spectrum and scattering coefficient spectrum by red
tide species ······25
Fig. 9. Absorption spectrum and scattering coefficient spectrum by red
tide species (Ahn et al, 2000) 25
Fig. 10. Scattering coefficient spectrum of suspended solids according to
concentration (Menon et al. 2018)26
Fig. 11. AIC and BIC were calculated to evaluate the effectiveness of
the Gaussian mixture model
Fig. 12. The red tide sea area is clustered to K
Fig. 13. The red tide sea area is clustered to K
Fig. 14 result of clustering sea areas that are not red tides
Fig. 15 AIC and BIC are calculated about the sea area which is not
the red tide
Fig. 16 The data expressing the cluster result on map
Fig. 17 Map of red tide occurrence on August 11, 2013 provided by

the	NIF	S			••••••	•••••	•••••				39
Fig.	18.	Data	comparing	the	estimated	red	tide	position	with	the	actual
		red	tide positio	n			•••••				39



LIST OF TABLES

Table. 1. Band composition of GOCI 11								
Table. 2. Normalized water leaving radiance for each band12								
Table. 3. Measurement data of red tide on August 14, 2015 provided by								
the NIFS 13								
Table. 4. AIC and BIC calculated values 29								
Table. 5. Bayesian Gaussian Mixture weights_								
Table. 6. Number of sea area pixels corresponding to the cluster 30								
Table. 7. Data of red tide occurrence on August 11, 2013								
provided by the NIFS40								
RST II OF III								

Analysis and Detection of Red Tide Sea Areas Using Machine Learning-Based Clustering

Mi So Park

in Division of Earth and Environmental System Sciences Major of Spatial Information Engineering,

> The Graduate School, Pukyong National University

ABSTRACT

본 연구에서는 기존의 이분법적인 적조 해역 분류 방식을 개선하고 다양한 해수 환경을 고 려하기 위해 비지도 학습 클러스터링 기법을 사용하였다. 사용한 가우시안 혼합 모델은 다 른 클러스터링에 비해 매우 유연하게 수행되며 다양한 데이터에 적용할 수 있다는 장점이 있다. 클러스터링은 일반적인 지도학습 모델과 달리 실제 정답값이 주어지지 않아 정확도 와 같은 정량적인 값으로 판단하기 어렵다. 클러스터링 모델의 성능은 데이터의 유사성을 기반으로 하여 형성한 클러스터링의 유효성으로 평가된다. 따라서 GMM은 일반적인 클러스 터링 모델과 달리 확률에 기반한 모델이므로 우도를 활용한 AIC와 BIC를 이용하여 평가된 다. 연구 수행에 사용된 Water leaving radiance는 적조가 빈번히 일어난 2015년 8~9월에 해당하는 GOCI의 Level 1B 자료 중 , 실제로 적조 발생한 날에 대해 수집하고 이를 대기보 정하여 Level 2A 자료로 변환한 후 추출한 값이다. AIC와 BIC가 각각 -99584, -97798로 가 장 작은 값을 나타내는 K는 5로 산출되었으며 이는 K가 5일 때(클러스터 형성이 5개 일 때) 적조 해역 데이터들을 가장 잘 설명한다고 볼 수 있다. 이에 대하여 가우시안 혼합 모 델 클러스터링을 수행하였으며 이를 분광 프로파일 형태로 나타내었다. 클러스터링 된 결 과는 적조와 비적조에 대한 이분적 분류가 아닌 실제 해수의 환경을 고려하여 형성되었으 며 시각적으로도 뚜렷한 차이를 보였다. 특히 외해와 연안에 대하여 확실한 차이를 나타내 었다. 또한 해당 픽셀을 실제 위경도로 변환하여 지도에 매핑한 결과 각각의 클러스터가 서로 외해와 연안에 다르게 위치한 것을 알 수 있었다. 이에 대하여 실제 적조 위치 자료 와 적조로 추정되는 클러스터의 위치 값을 비교한 결과 거의 유사한 정도를 보였다. 이는 클러스터링 된 값으로 적조 픽셀을 구분하여 적조를 탐지 할 수 있음을 나타낸다.



I. Introduction

Red tide generally refers to a phenomenon in which phytoplankton proliferates in large quantities at a time or is physically integrated to change the color of water to red or reddish brown. However, it is called the meaning of HABs (Harmful Algal Blooms) due to the subsequent damage such as the death of marine life due to red tide and the destruction of marine ecosystem. Red tide species that appear in the waters of the Korean peninsula are Cochlodinium polykrikoides, Karenia mikimotoi, Alexandrium sp., and Chattonella marina, among which Cochlodinium polykrikoides (C. polykrikoides), a type of dinoflagellate, is considered to be the representative species causing damage.

C. polykrikoides is a dinoflagellate, a single-celled eukaryote that photosynthesizes, and generally has the property of rapidly dividing and diffusing when creating an environment that meets growth conditions. In addition, in an environment that does not meet the growth conditions, such as a decrease in illumination, it does not disappear but sinks into the surface sediments with resting cysts and enters a dormant state (*Matsuoka et al., 2010; Yoon and Shin., 2013*). This is to increase the survival rate of the community, and it is known to spread in the floating ecosystem (surface) after re-emergence by external environmental stimuli. In addition, C. polykrikoides are photothermal and photochromic organisms that can live in a wide range of temperatures and a wide range of salinity changes. They can survive and grow in most parts of the Korean peninsula, It is possible to ingest abundant nutrients or swim in high light conditions (*Oh et al., 2010*). Due to these same characteristics, C. polykrikoides have a better chance of

survival than other red tide species and are more likely to spread long-term after red tide.





Figure 2. Harmful algal blooms caused by C.polykrikoides.



Figure 3. Cochlodinium polykrikoides observed under a microscope.

The National Fisheries Research and Development Institute defines red tide as toxic, harmful, and harmless depending on the damage caused by red tide. Red tide species that cause only changes in water color and do not damage aquatic organisms are classified as harmless, red tide species that cause suffocation of fish due to mucus secretion are classified as harmful, and red tide species that cause paralytic shellfish toxins due to toxic substance secretion are classified as toxic . C. polykrikoides secrete large amounts of mucinous material, which interferes with the respiration of fish and causes them to suffocate, which is classified as harmful.

C. polykrikoides red tide has occurred mainly in the southern coast of summer since the 1990s, causing various damages to fisheries and tourism. Especially in 1995, it is reported that record damage of about 76.4 billion won per year occurred (*Ministry of Maritime Affairs and Fisheries, 1999*). In addition, the amount of damage that occurred in the last 10 years is about 50.6 billion won, and the red tide that occurred in 2013 is known to have caused a great deal of damage to the fisheries industry by showing a very wide range of mechanisms such as spreading to the east coast and the west coast as well as the south coast (*Kim et al, 2014; Ministry of Maritime Affairs and Fisheries, 2019*). As the damage caused by the red tide phenomenon occurs every year, there is a need for continuous red tide research and monitoring.

Currently, red tide detection and monitoring is carried out by dividing into ship, land, and air. In the case of a ship, which is the main means of forecasting, it is a method to directly take a ship and measure the population of red tide creatures using a microscope method and judge whether or not red tide occurs. This method requires a lot of manpower, time, and cost, and there is a limit due to the characteristics of red tide that occurs diffusely and broadly in an unspecified space only by field survey. Therefore, research using remote sensing techniques has been actively conducted for efficient red tide detection and monitoring. Conventional remote sensing red tide monitoring was a method of distinguishing red tides by the concentration of chlorophyll-a estimated from satellite images, but this is calculated based on clear waters belonging to CASE - 1, and is likely to be overestimated in the Korean peninsula where water signals are very complex (Ahn et al., 2006). The waters of the Korean Peninsula correspond to CASE - 2, and not only phytoplankton but also dissolved organic matter and suspended solids have more complex characteristics because they affect the water signal. In addition, there is a difficulty in distinguishing between problems such as harmless red tide or increase of non-red tide creatures that do not cause red tide damage. In order to overcome these limitations, research has been conducted to analyze the optical characteristics of water and to identify the species, existence, concentration, and distribution of red tide species. (2007) proposed an algorithm to detect red tides using normalized water leaving radiance and sea surface temperature calculated from MODIS, a polar orbiting satellite. (2011) and Son et al. (2012) classified the red tide and non-red tide waters using the band ratio of MODIS and GOCI, respectively. (2016) presents a method for classifying red tide pixels by reducing the complex computational procedure required in previous photonic red tide detection techniques. This is a method of analyzing the spectral characteristics of the red tide area and performing the operation by expressing it as a formula. It has the advantage of classifying the red tide using a relatively simple equation. However, due to the diverse water backgrounds in the Korean peninsula, it is difficult to classify red tides clearly by simple equations. Recently, research has been conducted using machine learning techniques to detect red tides considering complex water signals. (2018), *Enkhjargal et al.*, (2020) attempted to detect red tide pixels in satellite images using logistic regression models and decision tree models. As a teaching and learning method based on the correct answer value given in advance, we tried to distinguish the cause and effect of red tide caused by complex water background.

Therefore, in this study, we classify the red tide area and analyze each characteristic by using the clustering technique which does not require a separate labeling operation and forms a cluster based on similarity. Therefore, we will classify the red tide occurrence area considering various water environments rather than simple dichotomous classification and utilize clustering evaluation index for validity verification.

$I\!\!I$. Data and methods

2.1. Research Areas and Data

2.1.1 Research Areas

C. polykrikoides red tide was first recorded in Jinhae Bay in the south coast in 1982, and it has been spreading from the southern coast to the east coast and the west coast every year from late spring to early autumn. In particular, the frequency of occurrence in the central sea area of the South Sea is the highest, and the duration and occurrence scale are also large.

Therefore, in this study, the damage caused by C. polykrikoides red tide is the most, and the south coast area, which is the starting point of C. polykrikoides red tide, was selected as the research area. For the convenience of analysis, the latitude $34.34 \approx 34.98$ and longitude $127.012 \approx 128.6$, including Goheung, Yeosu, Namhae, Goseong and Geoje, were designated as the study range.



Figure 4. South Sea of Korea.

2.1.2 Research data

In this study, in order to perform the clustering of the red tide area, the GOCI Level 1 data of July and August 2013 were collected based on the red tide breaking data and the red tide breaking data provided by the National Institute of Science and Technology (NIFS) of the Ministry of Maritime Affairs and Fisheries. In addition, C. polykrikoides had a characteristic of vertical movement, so data of the time zone with the highest illuminance were collected (ref. 3.2). In addition, data from July and August 2017 when little red tide occurred were collected to compare the results, in order to consider the seasonality of seawater reflection. In addition, in order to reduce the influence of weather such as clouds and precipitation, we collected images of clear days with less than 2 clouds among the images to be collected by referring to the past observation data provided by the Korea Meteorological Administration.

GOCI (Geostationary Ocean Color Imager) is the world's first geostationary orbital marine payload on the Communication, Ocean and Meteorological Satellite (COMS), which is operated by the Korea Ocean Research & Development Institute's Ocean Satellite Center. It consists of six visible light band and two near-infrared band (Table 2), with a spatial resolution of 500m x 500m and a time resolution of 8 times a day, 1 hour (UTC). In addition, it has a narrow band width of 10 \sim compared to the conventional sea color sensor, 40nm and the signal-to-noise ratio (SNR) is over 1000, which is high performance. for semi-real-time GOCI is responsible monitoring of marine ecosystems around the Korean peninsula, production of marine / fishery information, and monitoring of coastal and marine environments.

Approximately 90% of the signals observed in satellites are

signals received through the scattering and absorption process of atmospheric trace gas molecules and aerosols as solar radiation passes through the Earth's atmosphere, that is, atmospheric radiation energy. Therefore, in this study, the collected Level 1B images were converted to Level 2A data. For the convenience of the study, each vertex coordinate of latitude $34.34 \sim 34.98$ and longitude 127.012 ~ 128.6, which are the study areas, were converted into GOCI grid numbers and extracted for the necessary sections. The collected GOCI Level 1B data is 5185 ?? 4967, and the extracted reference grid numbers are 2240, 2916, the upper left to the lower right, 2526. 3065 from and respectively. The converted Level 2A data is a data that performs atmospheric correction to extract less than 10% of the unique seawater signal by removing more than 90% of TOA (Top of Atmosphere) radiance. The software used for data conversion is the GOCI data processing system (GDPS), which extracts seawater signals by applying spectral shape matching (SSMM) and solar reflection point correction. The normalized water leaving radiance for each band was extracted from the data, and about 40,000 pixel values were obtained by removing pixels with abnormal values due to incorrect atmospheric correction and clouds. Water leaving radiation, Lw, refers to the optical energy that passes through the sea surface layer among the optical energy reflected in seawater. Table 2. is a table showing some of the extracted normalized water leaving radiance values.

Band	Central Wavelenghts	Bandwidth	Primary use
1	412 nm	20 nm	Yellow substance and turbidity
2	443 nm	20 nm	Chlorophyll absorption maximum
3	490 nm	20 nm	Chlorophyll and other pigments
4	555 nm	20 nm	Turbidity, suspended sediment
5	660 nm	20 nm	Baseline of fluorescence signal, Chlorophyll, suspended sediment
6	680 nm	10 nm	Atmospheric correction and fluorescence signal
7	745 nm	20 nm	Atmospheric correction and baseline of fluorescence signal
8	865 nm	40 nm	Aerosol optical thickness, vegetation, water vapor reference over the ocean

Table 1. Band composition of GOCI.

pixel number	Band1	Band2	Band3	Band4	Band5	Band6	Band7	Band8
276	13.62318	12,25507	10.09052	5.265652	0.386387	0.393007	0.040944	0.015842
277	12.20409	11.03412	9.411535	4.857144	0.361771	0.399512	0.037856	0.014481
19679	25.98858	22.29384	20.85236	11.54399	4.131459	4.894295	0.527104	0.235297
19680	25.64547	22.84419	18.87983	11.84108	4.702428	4.002549	0.607235	0.271415

Table 2. Normalized water leaving radiance for each band.

The red tide data of the red tide breaking system is field survey data such as red tide species density, red tide occurrence coordinates, and red tide sea map centered on the peak (Table 3) (Figure 4).



Table	З.	Measurement	data	of	red	tide	on	August	14,	2015	provided	by
the NI	FS.											

Date		2015-08-14	
	Sinji-myeon, Wando-gun ~Yaksan-Today,	Seomyeon (Janghang) in	
Location	Iower Deukryangman Bay, Jangheung-gun, Geumsan-myeon, Goheung-gun [~] Doy ang-eup	Namnae-gun Nammyeon (Longguo)Sangju-m yeon (Nodo).Mijo-myeon (secondary)	Geoje City Yulpo~ Jeogu~Hansan Gokryongpo~
KYOM	Yeosu City's Gaedo, Geumodo, Wolho, Hwatae, Uhak-ri, Dolsan	Goseong-gun, Yokji in Tongyeong-si, Sanyang Suwol [~] Yeongun, Hansan Yongpo [~] Jukdo [~] Chub ong	Ulsan City (Seosaeng-myeon), Pohang City (Kuryongpo)
Species		Cochlodinium polykrikoides	Z
Density	4 ~ 130	50 ~ 1800	980 ~ 8400
(sell/mL)	10 ~ 3000	60 ~ 6120	1 ~ 2765
Water	23.8 ~ 26.8	23.5 ~ 26	22.5 ~ 23.2
Temperature(℃)	22.4 ~ 24.3	23.1 ~ 24.2	24.5 ~ 26
Coordinates (Lon, Lat)	[127.59583,34.40944 3],[127.75,34.40889], [127.76583,34.46611] 	[127.748886,34.582222],[127.82389,34.60416 8],[127.80194,34.6775] 	[127.605835,34.27],[12 7.56389,34.245556],[1 27.25278,34.129166]



Figure 5. Map of red tide occurrence on August 14, 2015 provided by the NIFS.

2.2. Research Method

2.2.1 Unsupervised Learning Clustering

In this study, a clustering technique was used to classify and analyze the red tide area considering the complex environment of the Korean peninsula. Clustering is a non-supervised learning technique that groups data based on similarity between data, unlike supervised learning, which is classified based on correct answers given in advance. As a kind of analytical method for understanding the intrinsic structure and characteristics of data, the data belonging to one group are very similar and have characteristics distinguished from other groups. In general, the similarity of clustering is represented by the distance between data and can be divided into hierarchical clustering and non-hierarchical clustering depending on the cluster formation method.

Hierarchical clustering refers to a method of forming a cluster by bundling data with the closest distance. As a result, it is represented by a tree-shaped hierarchy and can be represented by a diagram such as a dendrogram. Since clusters are formed sequentially, it is easy to grasp the formation process and clustering is possible without pre-determining the number of clusters. However, it is difficult to process large amounts of data and the operation speed is slow.

Hierarchical clustering refers to a method of forming a cluster by bundling data with the closest distance. As a result, it is represented by a tree-shaped hierarchy and can be represented by a diagram such as a dendrogram. Since clusters are formed sequentially, it is easy to grasp the formation process and clustering is possible without pre-determining the number of clusters. However, it is difficult to process large amounts of data and the operation speed is slow.

Non-Hierarchical clustering is a method of partially distinguishing the entire area where data exists, and clusters are formed by bundling data of similar attributes into K pieces. It is also called Partitioning clustering depending how the cluster on is formed.Non-hierarchical clustering is divided into K-Means. K-Medoid, DBSCAN, etc. according to the representative value setting method for cluster formation.

In this study, it is assumed that the data can be represented by K Gaussian distributions, and the red tide area is classified by using GMM (Gaussian Mixture Model) which forms a cluster with data having a high probability of belonging to each distribution. In the case of K-Means, which is a representative technique of clusters, clusters are formed by defining similarity in data as distance. However, this shows a rapid performance difference depending on the center point of the cluster, and it is not well explained in the case of complex data distribution.GMM is based on the probability that the data belongs to the cluster represented by the distribution, so it can be applied to more flexible and complex data than K-Means.

GMM is determined by three parameters: the mean and variance of each distribution, and the probability corresponding to the distribution (π). It is assumed that the characteristics of the data are expressed by mixing several Gaussian distributions, which can be explained as shown in the following figure.



Figure 6. gaussian mixture model.

In this case, the probability for each distribution for the data points x_i may be expressed by the following equation.

$$P(x_i) = \pi_{i,1} N(x_i \mid \mu_1, \sigma_1^2) + \pi_{i,2} N(x_i \mid \mu_2, \sigma_2^2) + \cdots$$
 (2)

 $\pi_{i,1}$ and $\pi_{i,2}$ are the probabilities that x_i corresponds to each distribution, which is called mixing coefficient or weight. $P(z=k) = \pi_k$, but the distribution to which the actual x_i is to be applied is an unknown value and is called a latent variable. In general, the likelihood function is obtained to estimate the

parameter and can be expressed as follows.

$$L(\pi,\mu,\sigma) = \prod_{i=1}^{N} P(x_i) = \prod_{i=1}^{N} \sum_{z_i} P(x_i,z_i) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k N(x_i + \mu_k,\sigma_k)$$
(3)

On the other hand, we take logs on both sides and obtain a parameter whose derivative value is zero.

$$\log P(x) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \sigma_k) \right) \tag{4}$$

However, GMM uses the Expectation–Maximization (EM) algorithm because there are actually unknown values (potential variables) such as z_i .EM is a method of finding parameters and clusters that maximize the log likelihood function by repeating Expectation Step and Maximization Step.

- 18 -

The EM algorithm execute phase is as follows.

step 1) .Select an arbitrary initial value for the necessary π,μ,σ
step 2) The probability that data point is included in the specific distribution is calculated about the chosen parameter.
step 3) Using the calculated probability, we estimate π,μ,σ again step 4) For this, the above process is repeated and optimization is performed until the log likelihood function shows the maximum value.

Step 2, Expectation Step, fixes the parameter selected by step1 and calculates the probability that the data point is included in a particular distribution, which can be expressed as:

$$\gamma(z_k) = P(z_k = 1 \mid x_i) = \frac{P(z_k = 1)P(xvertz_k = 1)}{\sum_{j=1}^{K} P(z_j = 1)P(xvertz_j = 1)} = \frac{\pi_k N(x \mid \mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j N(x \mid \mu_j, \sigma_k)}$$

This is an equation that expresses the probability x_i that z_k belongs to the cluster k in the form of a product of likelihood functions. In addition, the latent variable z selects one cluster for K clusters and expresses it as 0 and 1, which satisfies $\sum z_k = 1$.

Step 3, Maximization Step estimates the parameter π, μ, σ using the probability calculated in step 2, and can be expressed as follows.

$$\mu_k = \frac{1}{N} \sum_{n=1}^N \gamma(z_{nk}) x_n \tag{6}$$

$$\sigma_{k} = \frac{1}{N_{k}} \sum_{n=1}^{N} \gamma(z_{nk}) (x_{n} - \mu_{k}) (x_{n} - \mu_{k})^{T}$$
(7)

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} \tag{8}$$

2.2.2. Cluster Validity Assessment

In the case of supervised learning models, there are quantitative evaluation methods such as Accuracy and Precisin, but clustering is performed with data that does not have a designated correct answer label, and it is somewhat difficult to evaluate the performance of the model accordingly. However, the internal validity of the clustering model is verified by using appropriate parameter determination and indicators to evaluate the similarity of clusters.

In the case of K-Means, another clustering technique that sets the number of clusters in advance, such as GMM used in this study, the optimal number of clusters is selected and the degree of cluster cohesion is evaluated using indicators such as inertia and Silhouette score. However, since GMM is a probability-based model, we use AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) to determine the number of clusters and evaluate their performance.AIC and BIC can be expressed as follows.

$$AIC = -2^* \log likelihood + 2s \tag{9}$$

$$BIC = -2^* \log likelihood + s \log(N) \tag{10}$$

At this time, the number of parameters is represented by s, and the number of data is represented by n. Is to give a penalty due to the use of unnecessary variables for the model. The likelihood is a probability value between 0 and 1, which indicates that the value increases as the validity of the model increases, so we select the number of clusters with minimum AIC and BIC to evaluate the validity of the model.



Figure 7. schematization of the general Gaussian mixture model.

Figure 8. schematization of bayesian gaussian mixture model.

The validity of cluster formation is represented by the number of clusters that can best describe the total data, which means that the optimal number of clusters must be known in advance. However, the data for clustering is low-dimensional and relatively simple, and the Gaussian mixture model also has the hassle of setting the optimal number of clusters, which is similar to K-Means. The Gaussian mixture model determines and optimizes the parameters according to the probability that the data points belong to the corresponding cluster (Gaussian distribution). Accordingly, the number of clusters can be determined by utilizing the Bayesian Gaussian Mixture, which makes the weight of unnecessary or weakly explanatory clusters close to zero.

Therefore, in this study, we evaluate the cluster formation effectiveness of the Gaussian mixture model using AIC, BIC, and Bayesian Gaussian Mixture and find the optimal number of clusters.



III. Theoretical Background of Red Tide waters Analysis

2.1. optical properties

C.polykrikoides absorb blue and red wavelengths with chlorophyll, a photosynthetic pigment, and show particularly large absorbance near 440 to 450 nm. At this time, a part of the absorbed blue wavelength band energy is emitted again in the red region of 675 to 690 nm, which is called fluorescence. The emitted fluorescence energy is about 1.3% of the absorbed energy. According to Ahn et al. (2004), the absorbance coefficient of chlorophyll at 440 nm is less than 0.05 m^{-1} and more than 0.6 m^{-1} for waters where the concentration of chlorophyll is 60 ~ 100 m^{-1} or more. The seawater also showed a strong absorbance at the blue wavelength band, but the absorbance coefficient was about 0.01 m^{-1} , so the absorbance of the chlorophyll was more strongly affected. As a result, the most scattering occurs at wavelengths of 570 to 580 nm in the vellow region, which is the non-absorbing zone of seawater and chlorophyll. As the concentration of C. polykrikoides increases, the scattering coefficient near 550~600nm increases and the absorbance near 440~450nm also increases. This causes the wavelength of the blue region and the slope of 555 nm to be gentle based on the band of GOCI when red tide occurs, with a high reflection value at 555 nm (Ahn et al, 2000; Sin et al, 2017).

In addition to phytoplankton, the factors that determine the color of seawater in CASE-2 waters are dissolved organic matter (CDOM) and

suspended solid (SS). CDOM shows a high absorbance of the wavelength of the blue region, especially at 412 nm. As a result, as the concentration of CDOM increases in the water mass, the wavelength of the blue region is absorbed and becomes relatively black (*Sharpless, C. and Blough, N., 2014; Li, J. et al., 2017*). SS has a much greater effect of scattering than absorption due to this, and as the concentration increases, the peak shifts to the long wavelength band (*Doxaran, D. et al., 2002; Bright et al., 2018*). In the clear waters, the concentration of CDOM and SS is relatively small, and the scattering effect of the blue wavelength of the seawater mass itself is large, whereas in the relatively cloudy waters, the short wavelength absorption due to CDOM and the scattering due to SS are increased (*Menon et al.2018; Liew et al., 2001*).





Figure 9. Absorption spectra for chlorophyll, colour dissolved organic matter, TSM, and water ingots(Pavlov et al., 2014).



Figure 10. Absorption spectrum and scattering coefficient spectrum by red tide species(Ahn et al., 2000).

- 25 -



Figure 11. Scattering coefficient spectrum of suspended solids according to concentration(Menon et al. 2018).



2.1. biological properties

C.polykrikoides are known to be present in the surface layer when the environment is favorable for growth and in the bottom layer when the environment is unfavorable for growth. C. polykrikoides, which show a wide range of survival patterns for salinity and water temperature, are eventually stimulated by light and then spread to the surface layer, which can be referred to as the correct answer for analysis of red tide waters (*Oh et al., 2010; Park et al., 2001*). In addition, the factor that causes a large change in the chlorophyll of water within a short time range corresponding to one week corresponds to the vertical movement of C. polykrikoides (*Kim et al., 2010*). Therefore, among the collected GOCI data, chlorophyll at 14 o'clock with high illuminance and chlorophyll at 9 o'clock with relatively low illuminance were compared with the same date, and the pixels showing a large change were referred to red tide sea area analysis.

म व्यं म

IV. Result

4.1. Analysis of clustered red tide waters

In this study, AIC and BIC were calculated to evaluate the effectiveness of the Gaussian mixture model. AIC and BIC give penalties as the number of variables increases and the complexity of the model increases.



Figure 12. AIC and BIC were calculated to evaluate the effectiveness of the Gaussian mixture model.

Table 4. AIC and BIC calculated values.

K	1	2	3	4	5	6	7	8
BIC	39020	-8457	-40960	-71602.	-97798	-98262	-99280	-99023
AIC	38668	-9167	-42029	-73029	-99584	-100406	-101783	-101884

Figures 11 and 4 show the BIC and AIC values according to the number of clusters when GMM was performed on red tide area data, and there was no noticeable change from the time when K was 5 as the number of clusters increased.

Also calculated the weights for 8 arbitrary clusters using Bayesian Gaussian mixing. The calculated weight ratio is shown in the following table.

Table 5.	Bayesian	Gaussian	Mixture N	weights.		TIS		
bay_gmm.weights								
0.2181	0.1082	0.1527	0.1579	0.1861	0.002	0.0681	0.035	

The value of 0.1 or more for the weight was calculated as 5. Weights below 0.1 were judged to have low explanatory power. Therefore, it can be seen that Bayesian Gaussian Mixture also aims for the number of clusters of 5.

Therefore, this model can be considered to best explain the characteristics of the data when the number of clusters is 5.

In this case, the number of sea pixels corresponding to each cluster is shown in the following table.

gmm_cluster	Number of clustered data
0	13424
1	3758
2	1967
3	11571
4	9280

Table 6. Number of sea area pixels corresponding to the cluster

For a total of 40,000 sea pixels, 13424 were assigned to cluster 0, 3758 to cluster 1, 1967 to cluster 2, 11571 to cluster 3, and 9280 to cluster 4. The following is a picture in the form of a spectral profile to perform GMM on the number 5 of the selected clusters and analyze the characteristics of each cluster. The water leaving radiance for the 8 bands of GOCI is shown, which is a graph corresponding to a total of 40,000 sea pixels.



Figure 13. The red tide sea area is clustered to K.



Clusters 1 and 4 showed smaller overall signals than other clusters, especially the band 6 values of clusters 1 and 3 showed smaller values than other clusters. This value is due to the fluorescence action of chlorophyll, and a small value indicates that the value is less affected by chlorophyll. As a result of identifying the value of the individual pixels assigned to the cluster, it was a pixel corresponding to the relatively clear seawater, indicating that the influence on the scattering of the seawater mass itself was greatly affected. In the case of cluster 1, the blue region wavelength band and the slope with Band 4 are gentle compared to cluster 4, which means that the scattering of the yellow region is increased due to the influence of non-chlorophyll particles. Therefore, it can be indirectly seen that Cluster 1 is closer to the coast than Cluster 4, which is the largest value of Band 1, which is considered to have less absorbance due to low CDOM concentration. CDOM is not only produced on land, but also produced by photolysis and microbial degradation of photosynthetic organisms, which means that the concentration of red tide, which is a photosynthetic organism, is low and the CDOM concentration is also low. Cluster 2, Cluster 3, and Cluster 5 generally showed peak values in Band 4, and Band 1 and Band 2 showed different slopes from Cluster 1 and 4. It is estimated that the band 1 value is smaller than the band 2 value due to the absorption of CDOM, and the large overall value of each band indicates that the scattering due to suspended particles such as chlorophyll and bichlorophyll particles is increased. Therefore, it can be seen that the values corresponding to clusters 2, 3, and 5 are the areas of turbidity. This indicates that the increase of red tide organisms increases the absorption of blue wavelengths and the scattering of yellow regions, so the value corresponding to cluster 2 is a very high concentration of red tide organisms. Clusters 2 and 5 also show high values in Band 6; this is due to fluorescence, which indicates an increase in chlorophyll.Since red tide grows mainly in the area of turbid water such as the coast, this result is considered to represent the relationship between red tide, turbid water, and fresh water.





Figure 15. result of clustering sea areas that are not red tides.

Figure 14 shows the clustering of the sea area where red tide did not occur in order to compare and analyze the previously clustered results. Like the above red tide sea clustering, the calculated K was applied to cluster by obtaining the minimum values of BIC and AIC.



Figure 16 AIC and BIC are calculated about the sea area which is not the red tide.

The above results showed that the overall value of all clusters was relatively smaller than in red tide waters, indicating less light scattering effect on chlorophyllous particles such as red tide organisms. In addition, the small value of Band 6 also indicates that chlorophyll-induced fluorescence is low, indicating that red tide organisms do not appear or are very insignificant. In the case of cluster 3, the reflection value due to the wavelength of the blue region is most, and the band 4 value, which is the yellow region, is also small, so the slope is steeper than that of the red sea, which is similar to the characteristics of fresh water corresponding to the general open sea. Cluster 1 and Cluster 2 show smaller values than the red tide area, but basically the southern coast is affected by bichlorophyll particles and CDOM, and these clusters are turbidity-driven.In addition, the slope of the short-wavelength zone is relatively gentle compared to the red tide area, which is considered to be due to the lack of influence on chlorophyll of red tide organisms.

Figure 16 is a picture of the data location of each cluster, and the pixel values of each cluster are mapped to map the geographical distribution of the data values classified into five clusters. At this time, the number of the cluster assigned to represent the pixel in the sea map is labeled and the pixel number of the corresponding value is converted into the actual coordinate. The pixel number is the value corresponding to the grid of GOCI, which was used as the index value of the data during the research process, but the corresponding value was converted into actual latitude and longitude for geographical distribution analysis.



Figure 17. The data expressing the cluster result on map.

As shown in Figure 16, the pixels corresponding to Cluster 1 and Cluster 4 are relatively close to the open sea compared to other pixels, and Cluster 1 is relatively closer to the coast than Cluster 4.Cluster 2, Cluster 3, and Cluster 5 showed a distribution close to the coast.

On August 11, 2013, the latitude and longitude corresponding to the location of red tide were extracted from the red tide data provided by the red tide breaking system and the corresponding values were mapped on the map.In order to compare them, they were shown with cluster 2, which is estimated to be the spectral profile of red tide.

Therefore, the results of clustering the red tide area showed different locations on the sea map and it was possible to distinguish the location from the outer sea and the coast. In addition, it was found that the location of the cluster, which is estimated to correspond to the red tide, and the location of the actual red tide were very similar.





Fig. 18. Data comparing the estimated red tide position with the actual red tide position.



Figure 19. Map of red tide occurrence on August 11, 2013 provided by the NIFS.

Table 7. Data of red tide occurrence on August 11, 2013 provided by the NIFS.

Date	2013-08-11						
	Sacheon Samcheonpo to Seopo, Sindosu, Goseongman Bay, High-myeon to Hail-myeon to Samsan-myeon	Geoje West (Dundeok-Bisan-Jan gsado)	southeastern Geoje (Hakdong ~ Gujora ~)Wahyeon)				
Location	Yokjido Island, Yeonhwado Island, Hansan Island, Yongchodo Island, Gokyongpo, Tongyeong	Tongyeong Sanyang Jeodo, Yeonmyeong~ Daldal, Gonri, Obi.	Western Namhaedo (Janghang-Wolgok), Southern (Nodo-Yugu), Eastern (Changseon-Mijo), Northern (Jangpo-hyang), Geumnam, Hadong				
Species		Cochlodinium polykrikoides	13				
Density	580~6200	150~11000	420~15000				
(sell/mL)	1240~6500	1410~4421	500~8000				
Water	23.1~25.1	19.1~21.2	22.5 ~ 23.2				
Temperature(℃)	18.2~21.8	18.5~21.5	19.5~23.3				
Coordinates (Lon, Lat)	127.960487, 34.962740, 127.944134, 34.952742	128.179431, 34718374, 128.188188, 34.7204290	127.824788, 34.801992, 127.829197, 34.773424 				

V. Conclusions and considerations

In this study, unsupervised learning clustering technique was used to improve the existing dichotomous red tide classification method and to consider various seawater environments. The Gaussian mixture model used is very flexible compared to other clustering and has the advantage of being applicable to various data. The performance of the clustering model is evaluated by the validity of the clustering based on the similarity of the data. Therefore, GMM is evaluated using AIC and BIC using likelihood because it is a probability based model unlike general clustering model. The water leaving radiance used in the study was collected from Level 1B data of GOCI from August to September 2015, when red tide occurred frequently, and it was converted into Level 2A data after atmospheric correction.

AIC and BIC were -99584, -97798, respectively, and K, which represents the smallest value, was calculated as 5, which can be considered to best explain the red tide data when K is 5 (when cluster formation is 5).

For this, Gaussian mixed model clustering was performed and it was presented in the form of a spectroscopic profile.

In addition, it was found that the clusters were located in the open sea and the coast differently from each other as a result of mapping the pixels to the actual latitude. As a result of comparing the actual red tide location data with the location value of the cluster estimated as red tide, it showed a similar degree, indicating that the red tide can be detected by distinguishing the red tide pixels with the clustered value.

In the future, if the clustered values are used as new variables, it is

expected that they can be used as learning values or parameters for the red tide occurrence prediction model.



REFERENCE

Ahn, Y., Moon, J., Seo, W and Yoon, H., 2009, "Inherent Optical Properties of Red Tide Algal for Ocean Color Remote Sensing Application," J. of the Korean Society for Marine Environmental Engineering, vol. 12, no. 1, pp. 47-54.

Shin, H., MATSUOKA, K., Yoon, Y., 2010, "Response of dinoflagellate cyst assemblages to salinity changes in Yeoja Bay, Korea.", Marine Micropaleontology, vol. 77, no.1-2, pp. 15-24

Yoon, Y. and Shin, H., 2013 "Summary on the Dinoflagellate Cyst Assemblages of Modern Sediments from Korean Coastal Waters and Adjoining Sea.", Korean Society of Environmental Biology, vol. 31, no. 4, pp. 243-274

Oh, S et al., 2010, "Effects of Water Temperature, Salinity and Irradiance on the Growth of Harmful Dinoflagellate Cochlodinium polykrikoides Margelef isolated from South Sea of Korea in 2008.", Korean Journal of Fisheries and Aquatic Sciences, vol. 43, no. 6, pp.715-722

Kim, D. and Yoo, H., 2014, "Analysis of Temporal and Spatial Red Tide Change in the South Sea of Korea Using the GOCI Images of COMS.", Korea Spatial Information Society, vol. 22, no. 3, pp.129-136

Ahn, Y. and Shanmugam, P., 2006, "Detecting the red tide algal blooms from satellite ocean color observations in optically complex Northeast-Asia Coastal waters.", Remote Sensing of Environment, vol. 103, no. 4, pp.419-437 Son, Y. et al., 2011, "Cochlodinium polykrikoides red tide detection in the South Sea of Korea using spectral classification of MODIS data.", Ocean Science Journal, vol. 46, pp. 239-263

Son, Y. et al., 2012, "Monitoring red tide in South Sea of Korea (SSK) using the geostationary ocean color imager (GOCI).", Korean Journal of Remote Sensing, vol. 28, no. 5, pp. 531-548

Hu, C. and Feng, L., 2016, "Modified MODIS fluorescence line height data product to improve image interpretation for red tide monitoring in the eastern Gulf of Mexico.", Journal of Applied Remote Sensing, vol. 11, no. 1, 012003

Kim, Y. et al., 2006, "Detection of Cochlodinium Polykrikoides Red Tide Using MODIS Level 2 Data in Coastal Waters.", Korean Society of Civil Engeneers, vol. 27, no. 4D, pp.535-540

Bak, S. et al., 2018, "Study on Detection Technique for Cochlodinium polykrikoides Red tide using Logistic Regression Model and Decision Tree Model.", Korea Institute of Electronic Communication Science, vol. 13, no. 4, pp.777-786

Unuzaya. E. et al., 2020, "Study on Detection for Cochlodinium polykrikoides Red Tide using the GOCI image and Machine Learning Technique.", Korea Institute of Electronic Communication Science, vol. 15, no. 6, pp.1089-1098

Shin, J., et al., 2017, "A study on red tide surveillance system around the Korean coastal waters using GOCI.", Korean Journal of Remote Sensing, vol. 33, no. 2, pp. 213-230

Ahn, Y., 2000, "Development of Remote Sensing Reflectance and Water Leaving Radiance Models for Ocean Color Remote Sensing Technique.", ournal of the Korean Society of Remote Sensing, vol. 16, no. 3, pp.243-260

Sharpless, C. and Blough, N., 2014, "The importance of charge-transfer interactions in determining chromophoric dissolved organic matter (CDOM) optical and photochemical properties.", Environmental science ,vol. 16, pp. 654-671

Li, J. et al., 2017, "Remote sensing estimation of colored dissolved organic matter (CDOM) in optically shallow waters.", Remote Sensing, vol. 128, pp.98-110

Doxaran, D. et al., 2002, "Spectral signature of highly turbid waters: Application with SPOT data to quantify suspended particulate matter concentrations.", Remote Sensing of Environment, vol. 81, no. 1, pp. 149-161

Bright, C. et al., 2018, "Predicting suspended sediment concentration from nephelometric turbidity in organic-rich waters.", vol. 34, no. 7, pp. 640-648

Menon, H. and Adhikari A., 2018, Remote Sensing of Chlorophyll-A in Case II Waters: A Novel Approach With Improved Accuracy Over Widely Implemented Turbid Water Indices.", Journal of Geophysical Research: Oceans, vol. 123, no. 11, pp. 8138-8158

Liew, S. et al., 2001, "Retrieval of chlorophyll absorption spectra from remote sensing reflectance of turbid coastal waters.", IEEE, DOI:

- 47 -

10.1109/IGARSS.2001.976132

Park, J. et al., 2001, "Diurnal vertical migration of a harmful dinoflagellate, Cochlodinium polykrikoides (Dinophyceae), during a red tide in coastal waters of Namhae Island, Korea.", Phycologia, vol. 40, no. 3, pp.292-297

