Morphological and Physicochemical Characterization of Seagrass Species Using Unsupervised Learning Algorithm

Emmylou A. Borja
Monalee A. dela Cerna
Surigao State College of Technology
Surigao City, Philippines

Abstract
Seagrasses are the only true plants that live completely submerged in the sea and are particularly abundant and diverse in tropical waters. It is necessary for filtering and holding sediments and thus keeping the water over the coral reefs clearer. The aim of this work is to integrate and compare the type of parameters that can influence the characteristics of a certain species by analyzing their morphometric and physicochemical variables. This study uses the unsupervised learning algorithms such as \( k \)-Means and Principal Component Analysis in determining the morphological and physicochemical characterization of four seagrass species. Among all the species observed *Enhalus acoroides* had the most morphological variation regarding of the length of longest leaf which has the highest value. This means that this species of seagrass is the most sensitive to the environmental conditions prevailing in the study site. Hence, seagrass conservation is hereby encouraged to promote more seagrass areas in the country.

Keywords: \( k \)-Means; Morphological; Physicochemical; Principal component analysis; Seagrass

Introduction
Seagrasses are important providers of ecological services such as wave protection, reduced water flow, fishing ground, and oxygen production in areas along the shore (Ahmad-Kamil et al., 2013). They are highly diverse marine angiosperms and found along shallow parts of the sea and proximity to the land/sea interface causes them to be sensitive to changes in their environment (Hackney & Durako, 2004). However, they grow fast and can tolerate a wide range of environmental conditions (Orth, 2006). Thus, their adaptation to the marine environment could lead to variation in their morphology.

Out of the 60 known seagrass species in the world, 14 of them have been recorded in Philippine waters (Green & Short, 2003). Due to the high diversity, there has recently been an expanding interest in evaluating various morphometric structural and dynamic parameters of seagrasses. The study of (Ahmad-Kamil et al., 2013) proved that seagrass population is sensitive to monthly variations in the waters physicochemical parameters. The seagrasses are also affected by weather changes and availability of nutrients. Moreover, the amount of light that reaches the sediment would limit the distribution of seagrasses. Therefore, human activities could affect their growth and abundance (Papenbrock, 2012).

Changes in structural characteristics, such as leaf length, width, and shoot-specific leaf area, reflected responses of seagrasses to
environmental conditions. This response will indicate whether the environment is favorable for their growth. Furthermore, morphological changes may make seagrass more vulnerable to light stress. Thus the data gathered will give information for proper management and utilization of seagrass species.

In Surigao City, Brgy. Day-asan is perceived to have the highest density of seagrasses. Four species of seagrasses were identified along its intertidal zone namely: *Cymodocea rotundata* (Ribbon seagrass), *Enhalus acoroides* (Tape seagrass), *Syringodium isoetifolium* (Noodle seagrass) and *Halodule uninervis*. There was no study conducted yet on the morphological and physicochemical characterization of these seagrass species in Surigao City. In response, the study aimed to assess morphological variation and physicochemical parameters among four seagrass species in Brgy. Day asan, Surigao City.

### Materials and Methods

#### Study Site

Sampling sites were established at Brgy. Day-asan, Surigao City. This barangay of Surigao City was chosen as the experimental unit for it is perceived to have diverse seagrass ecosystems. The study was conducted from December 2017 to May 2018.

Barangay Day-asan is seven (7) kilometers Northeast of the city proper. It has a vast mangrove ecosystem which is the niche of around seven (7) different species. It has a dazzling underwater life including seagrasses that could give important life to other marine resources (Green & Short, 2003).

Three stations with three transects each were made, and the following data provided the exact location. Shown in Table 1 were the exact locations of the study area using Global Positioning System (GPS) HOLUX-GPS GM-305 WT. This was done to locate the seagrass beds in the area. Further, the site of the sampling stations is pictorially shown in Figure 1.

### Table 1. GPS of the sampling sites

<table>
<thead>
<tr>
<th>Sampling Sites</th>
<th>GPS Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Station 1</td>
<td>9° 46' 03.1&quot; N</td>
</tr>
<tr>
<td></td>
<td>125° 33' 38.2&quot; E</td>
</tr>
<tr>
<td>Station 2</td>
<td>9° 46' 02.3&quot; N</td>
</tr>
<tr>
<td></td>
<td>125° 33' 38.1&quot; E</td>
</tr>
<tr>
<td>Station 3</td>
<td>9° 46' 02.3&quot; N</td>
</tr>
<tr>
<td></td>
<td>125° 33' 38.9&quot;E</td>
</tr>
</tbody>
</table>
The study utilized quadrat sampling using 100 m transect and ten quadrats per transect starting from the inner margin of the seagrass beds and ending at the outer margin. Samples were rinsed carefully to remove sediments using seawater ensuring the short-shoots remained attached to the rhizomes and they were placed in labeled plastic bags and kept in the fridge for succeeding morphometry. Measurements of the morphological variables were taken: length of rhizomes, length of upright shoot, length of longest leaf (cm), number of leaves, number of nodes, and length of roots. Figure 2 shows some samples of seagrass species.

**Sampling Procedure**

The study utilized quadrat sampling using 100 m transect and ten quadrats per transect starting from the inner margin of the seagrass beds and ending at the outer margin. Samples were rinsed carefully to remove sediments using seawater ensuring the short-shoots remained attached to the rhizomes and they were placed in labeled plastic bags and kept in the fridge for succeeding morphometry. Measurements of the morphological variables were taken: length of rhizomes, length of upright shoot, length of longest leaf (cm), number of leaves, number of nodes, and length of roots. Figure 2 shows some samples of seagrass species.

**System Architecture**

The study used unsupervised learning algorithms such as k-Means and Principal Component Analysis (PCA) for cluster analysis. k-Means algorithm is applied since it is an efficient method in mapping social-ecological systems (Hamann et al., 2015), hazard zonation (dela Cerna, 2017), species distribution and environmental conditions (Jackson, 2010). Likewise, PCA was used as decision support tool in mapping demographic areas for development and strategic planning purposes (Petrișor, 2012) and analyze ecological niches (Janžekovič & Novak, 2012). To investigate the morphological and physicochemical characterization of seagrass species, these two methods were used in the study. The six morphological parameters against the five physicochemical parameters were used as datasets. Morphological variables were taken such as length of rhizomes (LOR), length of upright shoot (LUS), length of the longest leaf (cm) (LLL), number of leaves (NL), number of nodes (NN) and length of roots (LR). The physicochemical parameters such as salinity (parts per thousand), pH, water temperature, type of substrate and water depth are also determined. Figure 3 presents the system architecture of the study.

**Unsupervised Learning Algorithm**

The unsupervised learning is a type of machine-learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping of data. Data clustering is used as part of several machine-learning algorithms, and can also be used to perform ad-hoc data analysis. The most common method for clustering numeric data is called the k-Means algorithm and Principal Component Analysis. k-Means was used to represent all data vectors via a small number of cluster centroids to represent them as linear combinations of a small number of cluster centroid vectors. The PCA was also used to characterize all n data vectors as linear combinations of a small number of eigenvectors and does it
Principal Component Analysis

Principal Component Analysis is useful in reducing and interpreting large multivariate data sets with underlying linear structures, and for discovering previously unknown relationships. PCA is used to maximize the separation of the features. It has a way of rotating the axis which determines the variance of \( x_1 \) and \( x_2 \) axis of the variable to ensure that the dataset will pass the origin of the axis. So the PCA is the new axis of data points. It also produces a low-dimensional representation of a dataset and also serves as a tool for data visualization. Using PCA, we will examine the relationship between morphological variables and physicochemical parameters of seagrass species. PCA is based on the statistical representation of a random variable. Suppose we have a random vector population \( x \), where:

\[
x = (x_1, ..., x_n)^T
\]

and the mean of that population is denoted by

\[
\mu_x = E\{x\}
\]

and the covariance matrix of the same data set is

\[
C_x = E\{(x - \mu_x)(x - \mu_x)^T\}
\]

\( k \)-Means Algorithm

\( k \)-Means algorithm is one of the unsupervised learning and clustering algorithms in neural networks. The algorithm classifies a given dataset by finding a certain number of clusters (K). Their clusters differentiate the groups. The choice is to place them as much as possible far away from each other. The...
algorithm is highly sensitive to the initial placement of the cluster centers. The algorithm consists of the following steps:

i. Input the number of clusters

ii. Set the initial value of centroid points for each cluster.

iii. Assign each input example \( x \) to the cluster \( c(x) \) with the nearest corresponding weight vector:

\[
c(x) = \text{argmin}_j ||x - w_j(n)||
\]  

(4)

iv. Update the weights:

\[
w_j(n+1) = \sum_{x \text{ such that } c(x) = j} x / n_j
\]

(5)

with \( n_j \) the number of examples assigned to cluster \( j \)

v. Increment \( n \) by 1 and go until no noticeable changes of weight vectors occur. The whole process is passed out repetitively until the centroid values become constant.

Results and Discussions

Principal Component Analysis

This part presented the PCA results. The scree plot in Fig. 4 was used as a guide to determine how many components to consider. A scree plot is a useful visual aid for determining the appropriate number of principal components. It shows significant contribution of the 1st components, 2nd components and the next components which also denotes major component to the system. As we look for an "elbow," the 2nd point explains 3.0963 eigenvalue result in the eigenanalysis of the correlation matrix so that we will keep two main components.

There are only ten (10) variables tested which reveal two major components which are depicted in Table 2. The pH was not included since the values in the four species is nearly constant. Here, the PCA is the linear combinations of the original variables that explained for the divergence in the data. The most number of components extracted always equals the number of variables.
The eigenvectors, which are comprised of coefficients corresponding to each variable, are used to calculate the principal component scores. The coefficients indicate the relative weight of each variable in the component.

As shown in Table 2, the first principal component is strongly indirectly correlated with five of the original variables. In the first principal component NN increases as LUS, LLL, NL and LR scores decrease. This suggests that these five are related factors which will respond to the amount of light and the kind of substrate (Hamsiah et al., 2016). Moreover, seagrass growth for most species can be determined by marking leaves and roots and measuring tissue morphology (Short & Duarte, 2001).

In the second principal component, the two variables are indirectly correlated since their values are of different direction. It implies that as the salinity increases, substrate and water depth scores decrease. This suggests that these three criteria inversely vary together and could affect the morphology of seagrasses.

Figure 5 presents the loading plot for the variables to show their collinearity. The loading plot is the scheme of the correlation between the primary variables and the subspace measurement. It is used to interpret relationships between the variables in the space of the first two components. In the loading plot, we can see that LR, LLL, LUS and NL have large negative loadings on principal component 1, while LOR, Temperature and NN have large positive loadings, so these variables strongly influence the component on the variances of the morphology of seagrasses.

Similarly, the salinity have large positive loading in component 2, while the substrate and water depth, however, large negative loadings for the said component.

In Fig. 6 loading plot presents the variables using covariance matrix wherein LLL and water depth variables stand out. Since the values of these four species are too large.

Also, Fig. 7 presents the score plot that gives a projection of data onto subspace. It is used to interpret relationships between observations. PCA simulation in score plot describes the morphological and physicochemical parameters which display the first component and second component. In the first component, it is characterized by the species Cymodocea rotundata, Halodule uninervis and Syringodium isoetifolium while in second principal component it is characterized by the species Enhalus acoroides.

**k-Means**

This section presents the clustering derived using the k-Means algorithm. The k-Means algorithm is utilized, and the data are clustered. The k-Means presents the numerical output data to interpret the similarity and dissimilarity within groups. The eleven (11) variables are tested which reveal four major clusters is depicted in Table 3. It can be shown in cluster 2 that the higher value of the length of rhizomes will contribute to the greater number of nodes. It implies that the number of nodes will result to a longer rhizome. However, considering the length of upright shoot and number of leaves, cluster 2 indicates that the length of rhizomes and the said variables inversely vary. It follows that those with a higher value of the length of rhizomes has the lesser value of the length of upright shoot and which also have a tendency
of the low value of the number of leaves. This result is evident in seagrasses with dimorphic rhizomes (Hemminga & Duarte, 2000) wherein the leaves produced on the horizontal rhizomes are shorter and are rapidly shed.

The species *Cymodocea rotundata* typically occurs in shallow clear waters, in areas that experience minimal exposure at low tide. The substrate in station 2 was rocky and sandy,
Figure 7. Score plot of species distribution using PCA

Table 3. Cluster centroids

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Grand Centroid</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUS</td>
<td>7.61</td>
<td>3.9936</td>
<td>11.86</td>
<td>19.58</td>
<td>6.2493</td>
</tr>
<tr>
<td>LLL</td>
<td>48.53</td>
<td>14.8318</td>
<td>70.11</td>
<td>87.57</td>
<td>28.308</td>
</tr>
<tr>
<td>NN</td>
<td>2.2</td>
<td>7.7667</td>
<td>2</td>
<td>2.2</td>
<td>6.3583</td>
</tr>
<tr>
<td>NL</td>
<td>3.9</td>
<td>2.8889</td>
<td>4</td>
<td>5</td>
<td>3.2417</td>
</tr>
<tr>
<td>LR</td>
<td>9.96</td>
<td>6.1311</td>
<td>19.01</td>
<td>13.01</td>
<td>8.0967</td>
</tr>
<tr>
<td>Substrate</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Salinity</td>
<td>31</td>
<td>31.3333</td>
<td>33</td>
<td>30</td>
<td>31.3333</td>
</tr>
<tr>
<td>pH</td>
<td>8.1</td>
<td>8.1</td>
<td>8.1</td>
<td>8.1</td>
<td>8.1</td>
</tr>
<tr>
<td>Temperature</td>
<td>27.2</td>
<td>28.0344</td>
<td>28.1</td>
<td>26.9</td>
<td>27.8758</td>
</tr>
<tr>
<td>WaterDepth</td>
<td>107</td>
<td>100.6667</td>
<td>65</td>
<td>130</td>
<td>100.6667</td>
</tr>
</tbody>
</table>

and this species was not exposed to low tide thus, they have longer rhizomes and number of nodes than the rest of the stations.

While in cluster 3, salinity, temperature and the length of roots obtained the highest values. This implies that the area with high salinity and temperature are the ones with a higher length of roots.

For the 4th cluster, the length of the longest leaf is related to the length of upright shoot and number of leaves. The longer the shoot, it is expected that it will grow into many leaves and a long leaf. The longer leaves of *Enhalus acoroides* in station 3 indicated that the water in this station is turbid and not clear since plants found in areas with poor water clarity grow longer leaves to maintain optimal light requirements.

In the same table, the water depth obtained the highest mean of 130 which is in cluster 4 and centroid of 100.6667. In the same way, among all variables, water depth is the most diverse from its centroid, and so, it is the primary variable that contributes to the
Figure 8. Seagrass Species Clustering using k-Means, CR (Cymodocea rotundata), EA (Enhalus acoroides), HU (Halodule uninervis), SI (Syringodium isoetifolium)

morphological variation of seagrass species. Moreover, the length of longest leaf obtained the highest mean of 87.57 in that same cluster and centroid of 28.308; this means that the length of the longest leaf is second most dissimilar from its centroid, and so, it showed the most significant variation among the morphological variables. Thus, the more saline the water is, the more diverse species can be observed.

The researchers are interested in identifying parameters for which there is a significant variation of the length of the longest leaf between the four seagrasses. These variables might be useful predictors of seagrass species that stay longer in a particular part of seawater. Substrate types of soil and water depth have a significant relationship between the profiles of seagrass as physicochemical parameters.

Furthermore, Fig. 8 presents the scatter plot that illustrates the simulated data set with 120 observations in 2-dimensional space. The panel shows the results of applying $k$-Means clustering with four values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the $k$-Means clustering algorithm. As shown in the figure Enhalus acoroides dominates all seagrass species in the morphometry which means that it has the highest variability in all factors.

Conclusion

This paper proves that the two clustering techniques, PCA and $k$-Means clustering reveal similar result in predicting variations of
data. The results of PCA and k-Means are in fact analogous since data clustering also is a form of data reduction to classification space. Data points in the same cluster are considered belonging to same class while points in different clusters are considered belonging to different classes. k-Means grand centroids result matches the PCA covariance matrix result as to classification and clustering of variables. These results indicate that unsupervised dimension reduction is closely related to unsupervised learning.

In this study, the two algorithms in unsupervised learning can be useful to predict the variation and clustering of the morphological and physicochemical characterization of seagrass species. Among all the species observed Enhalus acoroides had the most morphological variations. It is also evident that physicochemical parameters contributed much to this variation. The species Cymodocea rotundata, Syringodium isoetifolium, and Halodule uninervis did not show much morphometric variations, therefore, displayed the most uniform adaptation to the prevailing environmental conditions.

Acknowledgment

The authors would like to thank Ms. Irmalyn B. Paymalan for her scholarly comments and suggestions which helped in the refinement of the manuscript.

References Cited


