ESTIMATION OF OCEAN WATER CHLOROPHYLL-A CONCENTRATION USING FUZZY C-MEANS CLUSTERING AND ARTIFICIAL NEURAL NETWORKS

By

Kevin Michael Turner

B.S. Bucknell University, 2004

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Electrical Engineering)

The Graduate School
The University of Maine
August, 2007

Advisory Committee:

Habtom W. Ressom, Assistant Professor of Biostatistics, Georgetown U., Co-Advisor
Mohamad T. Musavi, Professor of Electrical and Computer Engineering, Co-Advisor
Richard Eason, Associate Professor of Electrical and Computer Engineering
Bruce Segee, Associate Professor of Electrical and Computer Engineering
ESTIMATION OF OCEAN WATER CHLOROPHYLL-A CONCENTRATION USING FUZZY C-MEANS CLUSTERING AND ARTIFICIAL NEURAL NETWORKS

By
Kevin Michael Turner
B.S. Bucknell University, 2004

A THESIS
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science
(in Electrical Engineering)

The Graduate School
The University of Maine
August, 2007

Advisory Committee:
Habtom W. Ressom, Assistant Professor of Biostatistics, Georgetown U., Co-Advisor
Mohamad T. Musavi, Professor of Electrical and Computer Engineering, Co-Advisor
Richard Eason, Associate Professor of Electrical and Computer Engineering
Bruce Segee, Associate Professor of Electrical and Computer Engineering
LIBRARY RIGHTS STATEMENT

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at The University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for "fair use" copying of this thesis for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature:

Date:
ESTIMATION OF OCEAN WATER CHLOROPHYLL-a CONCENTRATION USING FUZZY C-MEANS CLUSTERING AND ARTIFICIAL NEURAL NETWORKS

By Kevin Michael Turner

Thesis Advisors: Dr Habtom W. Ressom and Dr. Mohamad T. Musavi

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(in Electrical Engineering)
August, 2007

A system incorporating a fuzzy c-means clustering and an ensemble of artificial neural networks (ANNs) is proposed to estimate chlorophyll-a (Chl a) concentration from remotely sensed reflectance (Rrs) measurements. Fuzzy c-means is used to measure and define multiple spectral clusters from a pre-specified training set. A radial basis function (RBF) neural network is used to emulate the function of the fuzzy c-means clustering to determine the cluster and grade of membership for previously unseen spectral measurements. Next, a feed forward multi-layer perceptron (MLP) neural network is incorporated and used for Chl a estimation. The proposed method can be used to estimate Chl a concentration from Rrs measured at various global oceanic locations representing heterogeneous water types. The performance of the proposed method is presented in two experiments representing a proof of concept and a potential global Chl a prediction model. The two experiments are compared to the traditional approach, where a single ANN is used for all water types. It is shown that the cluster-based approach has the potential to build a more global Chl a prediction model.
A system incorporating a fuzzy c-means clustering and an ensemble of artificial neural networks (ANNs) is proposed to estimate chlorophyll-a (Chl a) concentration from remotely sensed reflectance (Rrs) measurements. Fuzzy c-means is used to measure and define multiple spectral clusters from a pre-specified training set. A radial basis function (RBF) neural network is used to emulate the function of the fuzzy c-means clustering to determine the cluster and grade of membership for previously unseen spectral measurements. Next, a feed forward multi-layer perceptron (MLP) neural network is incorporated and used for Chl a estimation. The proposed method can be used to estimate Chl a concentration from Rrs measured at various global oceanic locations representing heterogeneous water types. The performance of the proposed method is presented in two experiments representing a proof of concept and a potential global Chl a prediction model. The two experiments are compared to the traditional approach, where a single ANN is used for all water types. It is shown that the cluster-based approach has the potential to build a more global Chl a prediction model.
ACKNOWLEDGEMENTS

This work was supported by the NASA/Maine Space Grant Consortium Grant awarded to H.W. Ressom (EP-04-08).

I would like to thank my parents for their support in whatever form it may take.
# Table of Contents

Acknowledgements........................................................................................................................................... ii  
List of Tables ................................................................................................................................................... iv  
List of Figures .................................................................................................................................................. v  
1. Introduction ........................................................................................................................................... 1  
   1.1. Background ........................................................................................................................................ 1  
   1.2. Objective ........................................................................................................................................... 4  
   1.3. Organization of the Thesis ................................................................................................................ 5  
2. Data ....................................................................................................................................................... 7  
   2.1. Experiment 1....................................................................................................................................... 7  
   2.2. Experiment 2....................................................................................................................................... 8  
3. Methods and System Development ................................................................................................. 11  
   3.1. Fuzzy C-Means ................................................................................................................................. 11  
   3.2. Radial Basis Function Network ....................................................................................................... 13  
   3.3. Multilayer Perceptrons ....................................................................................................................... 19  
      3.3.1. Experiment 1 .............................................................................................................................. 19  
      3.3.2. Experiment 2 .............................................................................................................................. 26  
   3.4. System Performance .......................................................................................................................... 29  
      3.4.1. Experiment 1 .............................................................................................................................. 29  
      3.4.2. Experiment 2 .............................................................................................................................. 33  
4. Comparison of Cluster-Based and Single MLP Chlorophyll a Estimation ......................... 35  
   4.1. Experiment 1....................................................................................................................................... 35  
   4.2. Experiment 2....................................................................................................................................... 40  
5. Conclusion ............................................................................................................................................. 43  
References .................................................................................................................................................... 45  
Biography of the Author ......................................................................................................................... 47
LIST OF TABLES

Table 1: Summary of $S_{X_B,m}$ values for Exp2.................................................................12

Table 2: Network statistics for the individual clusters and the training results for the seven MLP networks used for Chl $a$ estimation from a previously clustered input ..19

Table 3: Exp1 validation results using the NOMAD derived testing set .........................23

Table 4: Summary of performance for all systems............................................................40
LIST OF FIGURES

Figure 1: Depiction of how the cluster-based MLPs and the trained RBF work together to estimate the Chl a from new Rrs measurement.............................................................5

Figure 2: Exp1 flow-chart for training RBF and MLP networks using separated IOCCG and NOMAD data. ........................................................................................................9

Figure 3: Exp2 flow-chart for training RBF and MLP networks using combined and randomized IOCCG and NOMAD data.................................................................10

Figure 4: Exp2 RBF Gaussian centers for each of the clusters. .........................................16

Figure 5: Exp1 spectral data clusters of the entire IOCCG and NOMAD dataset as determined by the RBF network trained from 150 NOMAD and 200 IOCCG data points. ..........................................................................................................................17

Figure 6: Exp2 spectral data clusters formed by the RBF network for the training data set \((n=450)\). The numbers of samples in clusters 1 to 7 are 96, 44, 24, 99, 15, 60, and 112, respectively. ........................................................................................................18

Figure 7: Clustered training set from the NOMAD dataset \((n=296)\). ........................................21

Figure 8: Exp1 scatter plots of the network output vs. desired target for the training set for each cluster and for the entire system using the NOMAD training set \((n=296)\) ...22

Figure 9: Exp1 clustered spectra for the validation set. NOMAD \(n=132\). .................24

Figure 10: Exp1 scatter plots of the network output vs. desired target for each cluster and the entire network over the NOMAD validation dataset. \((n=132)\). .................25

Figure 11: Exp2 scatter plots of the MLP outputs vs. desired targets simulated by the cluster-based approach for the training data set \((n=450)\). Scatter plots are shown for each of the seven clusters and the entire network ......................................................27

Figure 12: Exp2 scatter plots of the MLP outputs vs. desired targets simulated by the cluster-based approach for the validation data set \((n=250)\). Scatter plots are shown for each of the seven clusters and the entire network ......................................................28

Figure 13: Exp1 scatter plots of the network output vs. desired targets for each cluster and the entire network over the entire IOCCG dataset \((n=500)\). ........................................31

Figure 14: Exp1 scatter plots of the network output vs. desired targets for each cluster and the entire network for a subset of the IOCCG dataset \((n=300)\) .................................32
Figure 15: Exp2 spectral data clusters formed by the RBF network for testing data set 
(n=228). The numbers of samples in clusters 1 to 7 are 58, 15, 17, 51, 4, 28, and 55, 
respectively .................................................................33

Figure 16: Exp2 scatter plots of the MLP outputs vs. desired targets simulated by the 
cluster-based approach for the testing data set. Scatter plots are shown for each of the 
seven clusters and the entire network (n=228).................................34

Figure 17: Exp1 scatter plot of a single MLP network output (A) vs. target (T) on 
validation dataset (n=128).................................................................36

Figure 18: Exp1 scatter plot of the entire IOCCG dataset simulated by the single MLP 
(n=500)......................................................................................................37

Figure 19: Exp1 the IOCCG dataset simulated by the single MLP separated into clusters 
in order to compare the results to the cluster-based approach (n=500) ..............38

Figure 20: Exp1 scatter plots for a subset of the IOCCG dataset simulated by the single 
MLP separated into clusters in order to compare the results to the cluster-based 
approach (n=300).........................................................................................39

Figure 21: Exp2 scatter plots for the testing data set simulated by the single MLP and 
separated into clusters in order to compare the results to the cluster-based approach 
(n=228).........................................................................................................41

Figure 22: Scatter plots for the Exp2 testing data set simulated by the OC4 algorithm and 
separated into clusters in order to compare the results to the cluster-based approach 
(n=228).........................................................................................................42
1. INTRODUCTION

Ocean color analysis aims to determine the concentration of seawater constituents from remotely sensed measurements of the reflected sunlight at specific wavelengths in the visible and near-infrared wavelengths in ocean water. The water leaving radiance, measured in the visible and near-infrared bands, contains information about the concentrations of chlorophyll \( a \) (Chl \( a \)), color dissolved organic matter (CDOM), and suspended sediments [1].

1.1. Background

The time and effort involved with ship born measurement practices for Chl \( a \) and CDOM concentrations is immense. Given remotely sensed reflectance measurements, the remote sensing of ocean color for the purpose of estimating Chl \( a \) and CDOM concentrations, for a point in the ocean, has proven a viable and effective method. The increasing number of advanced satellite sensors such as NASA’s Sea-viewing Wide Field of view Sensor (SeaWiFS) and the Medium Resolution Imaging Spectrometer (MERIS) on board the ESA-ENVISAT satellite makes it possible to exploit large amounts of optical data to support ocean color studies and applications, including monitoring of upwelling phenomena, algae blooms, and marine environmental changes.

The reflectance spectra are made up of multiple wavelengths between 400 and 700 nm and are calculated as the ratio of the upwelling radiance over the downwelling irradiance of a given sample of water at the surface. From \textit{in-situ} data sets, simple n-order band
ratios have been found to accurately estimate chlorophyll concentration in the water such as Gordon and Morel [2] and O’Reilly et al. [3] [4]. Chl \( a \) can be estimated in this way since chlorophyll has a distinct reflectance signature. These previously developed algorithms have been statistical and empirical in origin. The statistical algorithms are strictly used to derive Chl \( a \) concentration as an index of biomass where the empirical methods are used to derive Chl \( a \) as an index of spectral absorption and backscattering characteristics. Although most of these algorithms are applied globally, it is recognized that they are strictly applicable only to Case I waters [5]. Case I water refers to open ocean areas where the principal optical properties are affected by the presence of phytoplankton and chlorophyll pigments. Case II water refers to coastal and inland waters where optical properties are affected by CDOM and other suspended sediments in addition to chlorophyll pigments [1]. Therefore, it is widely accepted that a universal bio-optical algorithm for all water types is not feasible. In ocean color image scenes, there are often multiple water types with different optical properties, which in turn require different algorithms. Case I algorithms simply fail in Case II waters because they do not account for color absorption and scattering characteristics of CDOM and suspended sediments [6]. Among Case II waters, however there are different substances found globally that effect the optical properties of the water. A single parameterized Case II algorithm might be able to account for variability of one or more substances, but not variability due to wide variety of substances found globally [7, 8].

There has been considerable effort devoted to developing Case II algorithms [8, 9], but very little effort has been made to address the problem of merging algorithms of different
water types. Zhan et al. [10] used genetic algorithms to retrieve apparent optical properties from reflectance measurements. Due to the non-linear nature of the problem there are several instances where optimization algorithms such as multilayer perceptron (MLP) neural networks and radial basis function (RBF) networks have been used [11, 12] to retrieve optical properties of water. Zhang et al. [13] used a knowledge-guided segmentation and labeling approach based on unsupervised fuzzy clustering algorithm in conjunction with image processing techniques and applied to coastal zone color scanner (CZCS) images to determine the boundaries of ocean zones with different water types in a single image.

Moore et al. [14] used an unsupervised clustering based on the fuzzy c-means (FCM) cluster algorithm. FCM was applied to the in-situ remotely-sensed reflectance (Rrs) data obtained from different water types. Rrs measurements made in the Gulf of Maine, Middle Atlantic Bight, Sargasso Sea and Tokyo Bay were used to determine the different clusters. Once the clusters were identified, they used the clusters to develop class-specific bio-optical algorithms. Cococcioni and Corsini [15] used a FCM clustering algorithm to cluster the data and then instead of using bio-optical algorithms as in the earlier case, they used fuzzy set rules to estimate ocean color for Case II waters. FCM allows a reflectance measurement to be assigned partial or graded memberships to different water types with which they share spectral characteristics. This is done by using a fuzzy membership function that expresses the likelihood that a reflectance measurement vector belongs to a class or cluster with a known reflectance distribution.
1.2. Objective

In this project, FCM is used to determine spectral clusters and to develop a possibility distribution matrix $U$ from a training set. The latter is used to train an RBF network which is used for the purpose of determining cluster membership for a testing set. The RBF network is then used to estimate the cluster possibility distribution of new spectra that it has not seen before. Class specific algorithms were developed for the individual clusters formed using multilayer perceptrons (MLPs). For any new reflectance measurement for which Chl $a$ is to be retrieved, its cluster membership belonging to the clusters formed is determined via the RBF network. The final estimate of Chl $a$ concentration is the estimation from an individual MLP which has been trained to respond to the reflectance measurements in that cluster. The clusters are created using a subset of the data synthesized by the International Ocean Colour Coordinating Group (IOCCG) based on inherent optical properties of water and a subset of the NASA bio-optical Marine Algorithm Dataset (NOMAD). MLPs are then trained and tested to estimate Chl $a$ from the corresponding Rrs measurements. Figure 1 is the depiction of how the cluster-based MLPs and the trained RBF work together to estimate the Chl $a$ from new Rrs measurement. As shown, a new Rrs measurement is introduced; the RBF network calculates the cluster possibility distribution of the Rrs measurement. Based on the highest cluster possibility the Rrs measurement is input to the particular MLP that has been trained to respond to similar measurements and the Chl $a$ is output.
Two experiments of MLP development are presented and compared to the traditional single MLP approach. The first experiment (Exp1) was developed as a proof of concept for the system structure. In Exp1 the NOMAD and IOCCG data sets are kept separate during the training, training validation and testing phase. The second experiment (Exp2) was developed as a potential global Chl $a$ prediction model where the IOCCG and NOMAD sets were combined, randomized and separated into training, training validation and testing sets. The proof of concept was presented at the International Joint Conference on Neural Networks 2006 [16] and Exp2 was presented at Oceans Boston 2006 [17].

1.3. Organization of the Thesis

Following the Introduction, the Thesis will be presented in four additional parts. The second chapter will present the source for the datasets. The NOMAD and IOCCG
datasets will be discussed including which portions of the datasets were incorporated into the Thesis and how the data points used were derived from those datasets. The Data section will also introduce the two experiments, Exp1 and Exp2. The individual experiments will discuss how the datasets were used and will present the hierarchy for each experiment which is used for the duration of the Thesis. Next, the Methods and System Development chapter will introduce and explain the methods and algorithms used to develop the separate experiments. The chapter will present the results of each experiment in conjunction with the mathematical concepts. Finally the performance of each system over an independent testing set will be presented in the chapter. The chapter will be separated into individual experiments when needed to provide emphasis. After the systems from the two experiments have been developed and system performance presented, Chapter 4 will develop and compare each of the experiments to a traditional single MLP approach. Chapter 5 will conclude the Thesis with a reiteration of the results and main topics and will present topics that may offer additional improvements to the results of the two experiments presented.
2. **Data**

Remotely sensed reflectance measurements (Rrs in sr\(^{-1}\)) from the IOCCG dataset that was derived for a 60-degree zenith angle of the sun were used in this study. The dataset consists of 500 data points calculated for every 10nm from 400nm to 800nm. The 500 data points from 400nm to 800nm from the IOCCG data set were linearly interpolated into seven wavelengths previously used by NASA SeaWiFS project \(\lambda = [411\ 443\ 489\ 510\ 555\ 670\ 683](\text{nm})\). Additional remotely sensed reflectance data were obtained from the NOMAD dataset by selecting points containing data at the selected wavelengths. Rrs was then calculated for the dataset by dividing the spectral upwelling radiance (W/m\(^2\)sr) by the above surface down welling irradiance (W/m\(^2\)). In total there were 428 spectra with values at the seven wavelengths used by the NASA SeaWiFS project. The data points at the given wavelengths (\(\lambda\)) originating from the IOCCG and NOMAD datasets were combined. Hence, a total of 928 spectra were available for this study.

This project will present two experiments that are based on the combination of the two data sets, NOMAD and IOCCG.

2.1. **Experiment 1**

Exp1 is based on the flow chart depicted in Figure 2. The Fuzzy-MLP model was developed with the IOCCG and NOMAD data sets kept separate as a proof of concept for the system. In Exp1, the RBF network was trained using a combination of IOCCG (n=200) and NOMAD (n=150). Figure 2 illustrates the combination of the NOMAD
(n=150) and IOCCG (n=200) datasets for use in the FCM to yield the cluster possibility distribution matrix U and the cluster centers. An RBF network is then trained using the combined data and the desired outputs (U) from the FCM. The cluster centers and the trained RBF weight matrix are used in the fuzzy inference system to designate which cluster a data point belongs to. The NOMAD dataset was then clustered and separated into training (n=300) and validation (n=128) datasets. The MLP networks were developed using the NOMAD dataset. The system is tested with the entire IOCCG data set and independently testing using an n=300 IOCCG subset which was not used for any other system process.

2.2. Experiment 2

Exp2 is depicted in Figure 3. The Fuzzy-MLP model was developed by using the combination of the NOMAD and IOCCG data sets. The combined data set was randomly split into training (n= 450), validation (n= 250) and testing (n= 228) sets. Exp2 uses the training set (n=450) and validation set (n=250) for the Fuzzy inference system and MLP development. The testing set (n=228) was used as an independent test set. Figure 3 illustrates the combination of the NOMAD and IOCCG datasets for use in the FCM to yield the cluster possibility distribution matrix U and the cluster centers. An RBF network is trained using the combined data and the desired outputs (U) from the FCM. The cluster centers and the trained RBF weight matrix are used in the fuzzy inference system to designate which cluster a data point belongs to. The training and validation set are then clustered by the system for use in training the individual MLP’s. The entire system is independently tested using the testing set.
Figure 2: Exp1 flow-chart for training RBF and MLP networks using separated IOCCG and NOMAD data.
Figure 3: Exp2 flow-chart for training RBF and MLP networks using combined and randomized IOCCG and NOMAD data.
3. METHODS AND SYSTEM DEVELOPMENT

3.1. Fuzzy C-Means

The Fuzzy c-means (FCM) algorithm [18] is designed to minimize the objective function $J_m$ defined in Eq. (1):

$$J_m = \sum_{k=1}^{c} \sum_{i=1}^{N} (u_{ik})^m \left| \bar{x}_i - \bar{v}_k \right|^2$$  \hspace{1cm} (1)

where $u_{ik}$ is the membership of the $i$th observation to the $k$th cluster; $\left| \bar{x}_i - \bar{v}_k \right|$ is the Euclidean norm between the vectors $\bar{x}_i$ and $\bar{v}_k$; $m$ is the weighting exponent that can be any real number greater than 1 (usually set to 2); $c$ is the number of clusters; and $N$ is the number of samples or observations. The FCM algorithm chooses clusters that minimize the distance between the data points and the prototype cluster centers ($\bar{v}_k$). Cluster centers are iteratively adjusted until the optimization criterion is met (could be the number of iterations or the minimum change residual). The membership function used here is based on the reciprocal of the distance between points.

We used the FCM algorithm to cluster the IOCCG and NOMAD datasets. A number of iterations of the algorithm were done while varying $m$ from 1.2 to 3.0 and the number of clusters $c$ from 4 to 12. Using the cluster validity function $S_{XB,m}$ shown in Eq. (2); a measure of the variance of the clusters over the minimum distance between cluster centers, the optimal number of clusters for Exp1 and Exp2 was estimated to be 6 or 7.
with $S_{XB,m} = 0.22 - 0.20$ where the fuzzy partition coefficient, $v_{pc}$ (Eq. 3) equaled 1.0. Moore et al. [9] used both $S_{XB,m}$ and the fuzzy partition coefficient in their determination of the number of clusters to use. A summary of $S_{XB,m}$ for the number of clusters vs. $m$ for Exp2 is shown in Table 1. Note only $m = 1.2 - 2.2$ and number of clusters = 4 – 10 are shown; results outside these values were inconsistent and extraneous when computed over several iterations. The objective of using the $S_{XB,m}$ algorithm is to find a local minimal based on the cluster centers created through the FCM, where a small $S_{XB,m}$ indicates compact, separated clusters.

**TABLE 1: Summary of $S_{XB,m}$ values for Exp2**

<table>
<thead>
<tr>
<th># Clusters</th>
<th>$m=1.2$</th>
<th>1.4</th>
<th>1.6</th>
<th>1.8</th>
<th>2</th>
<th>2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.34134</td>
<td>0.3244</td>
<td>0.30611</td>
<td>0.27833</td>
<td>0.24443</td>
<td>0.3302</td>
</tr>
<tr>
<td>5</td>
<td>0.68186</td>
<td>0.46408</td>
<td>0.588</td>
<td>0.36774</td>
<td>0.36874</td>
<td>0.25736</td>
</tr>
<tr>
<td>6</td>
<td>0.44459</td>
<td>0.40803</td>
<td>0.33781</td>
<td>0.39166</td>
<td>0.22581</td>
<td>0.31536</td>
</tr>
<tr>
<td>7</td>
<td>0.64998</td>
<td>0.31617</td>
<td>0.31545</td>
<td>0.32487</td>
<td>0.20636</td>
<td>0.26015</td>
</tr>
<tr>
<td>8</td>
<td>0.56954</td>
<td>0.43202</td>
<td>0.36113</td>
<td>0.43879</td>
<td>0.28867</td>
<td>0.25762</td>
</tr>
<tr>
<td>9</td>
<td>0.57277</td>
<td>0.43097</td>
<td>0.43234</td>
<td>0.39272</td>
<td>0.34826</td>
<td>0.27438</td>
</tr>
<tr>
<td>10</td>
<td>0.53872</td>
<td>0.5671</td>
<td>0.44713</td>
<td>0.34853</td>
<td>0.34711</td>
<td>0.36983</td>
</tr>
</tbody>
</table>

The fuzzy partition coefficient where $0 \leq v_{pc} \leq 1$, is an indication of the fuzzyness of the partition space. A $v_{pc} = 1.0$ indicates hard partitions while $v_{pc} < 1$ is a measure of how fuzzy the partitions are. Obtaining a high $v_{pc}$ is a positive indication as the fuzziness of the partitions decreases as $v_{pc} \to 1.0$. In our study, for all experiments regardless of the number clusters, $v_{pc}$ was equal to 1. Although the partitions used the study are hard, it is still possible and/or likely for a vector to have a partial cluster membership in an FCM.
The optimal number of clusters for Exp1 was determined by using 150 data points from the NOMAD dataset and 200 data points from the IOCCG dataset. The optimal number of clusters for Exp2 was determined by using the training set. Spectral from both experiments were clustered into seven groups based on highest possibility of their membership using the possibility distribution $U$. It is important to note that the possibility distribution matrix $U$ from the FCM function summed to one for all spectra.

$$S_{XB,m} = \sum_{k=1}^{c} \sum_{i=1}^{N} (u_{ik})^m |\bar{x}_i - \bar{v}_k|^2$$
$$N(\min_{i,j} d_{wij}^2)$$

$$v_{pc} = \frac{1}{N} \sum_{k=1}^{c} \sum_{i=1}^{N} (u_{ik})$$

### 3.2 Radial Basis Function Network

For a radial basis function network, an input is passed directly to a hidden layer without weighting. The hidden nodes represent the radial basis function. The output of the hidden node is then weighted and linearly summed at the output. The Gaussian function (Eq. 4) is the most common function used in RBF networks and was used throughout this project; where $x$ is an n-dimensional, in this project seven-dimensional, input vector $m_i$ is the Gaussian center with the same dimension as $x$ and $\sigma$ is the Gaussian width.
To build an RBF model that estimates the possibility distribution U matrix for Exp1, the 150 NOMAD and 200 IOCCG spectra were divided into training and testing datasets. The RBF network had seven inputs (size of the reflectance measurement vector) and seven outputs (number of clusters). In Exp2 the training and validation sets were used.

The reason for building an RBF network is to be able to determine which cluster an unknown (previously unseen) reflectance measurement vector (spectra) belongs to. The Chl \( a \) estimation process is depicted in Figure 1. Without a model like this, one would have to determine the cluster of an unknown vector on the basis of its proximity to the cluster centroids. The drawback of this approach is that it does not take direct advantage of the U matrix. Another option would be to re-cluster the combined (previous and new) vectors and obtain a new U-matrix. This approach may lead to reorganization of the vectors, which would require the need to train new MLP networks for each cluster. By using an RBF model of the U matrix, one is able to determine which cluster a previously unseen Rrs vector belongs to, thereby allowing a previously trained MLP for that cluster to estimate the corresponding Chl \( a \) concentration. The use of the RBF network is advantageous due to the fact that it is contains the cluster centers and vector weighting information that is critical to determining cluster memberships as some Rrs values at different wavelengths may have more impact (weight) in determining vector cluster membership.
The seven cluster centroids found by FCM were used as the Gaussian function centers of the RBF network. The RBF weight matrix that represents the heights of the Gaussian functions was determined using the least squares method. A learning rate of 0.0003 was found to yield good results and would converge. The Gaussian center widths were varied and a final center width of .0001 was used since it produced a low mean-squared error (MSE) between the original and estimated U matrices. Note that the individual reflectance values can vary from $R_{rs} = 0.0001 - 0.03 \text{ sr}^{-1}$. Hence, the choice of a center width is consistent with the data range. The network was trained until a local minimum was found; a zero change in error was used as the stopping criterion. For each wavelength ($\lambda$), the position of each center of Exp2 and its Gaussian function are shown in Figure 4. This plot serves to indicate that a width of 0.0001 for the functions was a good choice since some of the centers are located very close to each other for the given values of $\lambda$. It is also important to note that some of the centers are spaced quite far apart; this is due to using the same width for each Gaussian function. An attempt was made using different widths for different $\lambda$’s using the r-nearest neighbor heuristic developed by Moody and Darken [11]. The heuristic attempt did not produce any useful results.
Figure 4: Exp2 RBF Gaussian centers for each of the clusters.
After the RBF was trained for the possibility distributions, the error between the RBF output and the target FCM U matrix was not considered important as long as the outputs produced the correct clustering; where the RBF cluster membership maximum was the same as the FCM membership max. As illustrated in Figure 5 and 6, visual inspection indicates the RBF network reasonably clustered spectra for all training points for both Exp1 and Exp2.

Figure 5: Exp1 spectral data clusters of the entire IOCCG and NOMAD dataset as determined by the RBF network trained from 150 NOMAD and 200 IOCCG data points.
Figure 6: Exp2 spectral data clusters formed by the RBF network for the training data set \( (n=450) \). The numbers of samples in clusters 1 to 7 are 96, 44, 24, 99, 15, 60, and 112, respectively.
3.3. Multilayer Perceptrons

3.3.1. Experiment 1

The 428 NOMAD spectra were clustered into seven groups via the RBF network for Exp1. Table 2 shows the number of data points in each of the seven clusters (Figure 7). The spectra were then divided up into testing and validation data for each cluster. Approximately 2/3 of the data per cluster was used for network training and the remaining 1/3 was used for network validation. The training data was transformed to have zero mean and unit variance.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Input Nodes</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>R train coef</td>
<td>0.817</td>
<td>0.98</td>
<td>0.941</td>
<td>0.999</td>
<td>0.958</td>
<td>0.932</td>
<td>0.933</td>
</tr>
<tr>
<td>Linear Reg. Slope</td>
<td>0.668</td>
<td>0.957</td>
<td>0.855</td>
<td>0.998</td>
<td>0.906</td>
<td>0.834</td>
<td>0.809</td>
</tr>
<tr>
<td>MSE</td>
<td>0.0204</td>
<td>0.0004</td>
<td>0.0142</td>
<td>0.0024</td>
<td>0.023</td>
<td>0.0134</td>
<td>0.0034</td>
</tr>
<tr>
<td>Training Points</td>
<td>18</td>
<td>30</td>
<td>24</td>
<td>22</td>
<td>123</td>
<td>63</td>
<td>16</td>
</tr>
<tr>
<td>Validation Points</td>
<td>6</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>60</td>
<td>30</td>
<td>6</td>
</tr>
</tbody>
</table>

An MLP network was initialized to two hidden nodes with hyperbolic tangent sigmoidal activation function and one linear output node. The network was trained using the Levenberg-Marquardt algorithm for 50 epochs or until a stopping criterion was met. The initial stopping criterion was a correlation coefficient $R_{val} \geq 0.920$. $R_{val}$ was found by simulating the validation data in the trained network and finding the linear regression of the network output versus the desired output. The linear regression output is also
graphed to determine the slope of the linear regression, which is expected to be close to one. If there was a high R\textsubscript{val} but a poor linear fit the training process was repeated. If a reasonable solution was found, the desired R\textsubscript{val} value was increased and the process was repeated in an attempt for a better-correlated solution. If, after the data was randomized and a network was trained for 50 epochs, the desired R\textsubscript{val} value was not reached, the same process was repeated using the Bayesian regularization method [12,13]. The Bayesian regularization allows the weights and biases of the network to have small values. Thus, it enables the network to be less likely to overfit. If neither the Levenberg-Marquardt nor the Bayesian regularization technique yielded the desired results over the 50 repetitions for a given network configuration then the number of input nodes was increased incrementally and the number of training epochs was decreased. The number of epochs was decreased because the cluster training sets contained less than 50 data points in five of the seven clusters and it was very easy to over-train the networks. Determining whether a network was over-trained was done by looking at the mean squared error over the training set. If the training set had an MSE of less than $10^{-4}$, the network was considered to be over-trained and the results were ignored and the training process repeated. This entire process was repeated for all seven clusters until a good solution was found or until changes in the network size did not change the output performance. The complete process yielded the networks listed in Table 2. The training data scatter plot is shown in Figure 8.
Figure 7: Clustered training set from the NOMAD dataset (n=296).
Figure 8: Exp1 scatter plots of the network output vs. desired target for the training set for each cluster and for the entire system using the NOMAD training set ($n=296$).
From Table 2 and Figure 8, it can be seen that a linear regression analysis of the training output vs. the target values yields slopes that are close to one for most of the clusters and high $R^2_{\text{train}}$ values have been obtained. Aside from the error in cluster 1, the other clusters have very good training outputs with the best being Cluster 4. All of the plots have been log$_{10}$ transformed. Also, the linear regression analysis of the entire system is plotted in Figure 8 which yielded an $R^2_{\text{train}} = 0.964$ and $\text{MSE}_{\text{train}} = 0.010$.

To validate the system of neural networks developed for Exp1, the validation spectra shown in Figure 9 were used. The performance of the MLP networks in predicting the Chl $a$ for the validation dataset is shown in Table 3 and Figure 10. It can be seen from the results, our method was able to model the validation data nearly as well as the training data. Overall, $R^2_{\text{val}} = 0.955$ and $\text{MSE}_{\text{val}} = 0.025$ for the entire training validation (n=132).

**Table 3: Exp1 validation results using the NOMAD derived testing set**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2_{\text{val.}}$</td>
<td>0.833</td>
<td>0.827</td>
<td>0.921</td>
<td>0.818</td>
<td>0.919</td>
<td>0.897</td>
<td>0.826</td>
</tr>
<tr>
<td>Linear Reg. Slope</td>
<td>0.68</td>
<td>0.98</td>
<td>1.09</td>
<td>1</td>
<td>1.04</td>
<td>0.87</td>
<td>1.01</td>
</tr>
<tr>
<td>$p$-values</td>
<td>0.017</td>
<td>0.015</td>
<td>0.0094</td>
<td>0.037</td>
<td>0.028</td>
<td>0.008</td>
<td>0.004</td>
</tr>
<tr>
<td>MSE</td>
<td>0.07</td>
<td>0.004</td>
<td>0.011</td>
<td>0.12</td>
<td>0.032</td>
<td>0.009</td>
<td>0.004</td>
</tr>
</tbody>
</table>
Figure 9: Exp1 clustered spectra for the validation set. NOMAD $n=132$. 
Figure 10: Exp1 scatter plots of the network output vs. desired target for each cluster and the entire network over the NOMAD validation dataset. (n=132).
3.3.2. Experiment 2

The training (Figure 6) and validation data sets for Exp2 were clustered into seven groups via the RBF network. For each cluster, an MLP network was trained using the Levenberg-Marquardt algorithm. Each network had at most three hidden nodes with hyperbolic tangent sigmoidal activation function and one linear output node. The training data set was used to determine the weights and biases of the MLP network. The validation data set was used as a stopping criterion to prevent overfitting. Figure 11 depicts scatter plots of MLP outputs vs. desired targets (Chl $a$ concentrations) for each cluster in the training data set. From Figure 11, it can be seen that a scatter plot of the training output vs. the target values yields slopes that are close to one for all seven clusters. All of the plots have been log$_{10}$ transformed. The scatter plot of the entire system is plotted in Figure 11. The entire system yielded a correlation coefficient, $R^2_{\text{train}} = 0.978$, and a mean squared error, $\text{MSE}_{\text{traing}} = 0.0150$. Similar performance was observed in the validation data set, where the $R^2_{\text{val}} = 0.964$ and $\text{MSE}_{\text{val}} = 0.0241$. Figure 12 is a representation of the system output for the training validation.
Figure 11: Exp2 scatter plots of the MLP outputs vs. desired targets simulated by the cluster-based approach for the training data set ($n=450$). Scatter plots are shown for each of the seven clusters and the entire network.
Figure 12: Exp2 scatter plots of the MLP outputs vs. desired targets simulated by the cluster-based approach for the validation data set \((n=250)\). Scatter plots are shown for each of the seven clusters and the entire network.
3.4. System Performance

3.4.1. Experiment 1

In order to evaluate if the approach represents a good generalized solution, Exp1 was tested with the IOCCG dataset. Note that a subset of the IOCCG dataset was used to develop the clustering system but none of the spectra was used during MLP training for Chl \( a \) estimation.

First, the entire IOCCG dataset is used for system testing. The dataset is made up of 500 spectra over the same wavelengths as the NOMAD set and was created through calculations based on water optical properties. The 500 spectra were clustered and the MLPs that correspond to the respective clusters were used to estimate Chl \( a \). The scatterplots for each cluster and the entire network are result shown in Figure 13. Cluster networks 3, 5, 6 and 7 were able to obtain correlations greater than 0.900 and the overall \( R^2_{\text{test}} = 0.913 \). Although some cluster-based MLPs were not able to obtain good correlation, the MSE between the target and the estimated Chl \( a \) was relatively small. What is interesting is that all of the points were estimated within an order of magnitude for each cluster. Having a low MSE_{\text{test}} = 0.13 for the entire system and having estimated a dataset that the network has never seen is an indication that a good general solution has been found. It should be noted that the target outputs for cluster 4 were outside of the training output values so the network could not be expected to accurately model that data.
Then, we looked into a subset of the IOCCG dataset, which was used neither in the cluster analysis nor in the training of MLPs. This subset is made up of 300 spectra, whose clusters were determined by using the RBF network. The MLPs that correspond to the respective clusters estimated Chl \( a \) with an \( R^2_{\text{test}} = 0.841 \) and \( \text{MSE}_{\text{test}} = 0.12 \). The scatter plots are shown in Figure 14.
Figure 13: Exp1 scatter plots of the network output vs. desired targets for each cluster and the entire network over the entire IOCCG dataset ($n=500$).
Figure 14: Exp1 scatter plots of the network output vs. desired targets for each cluster and the entire network for a subset of the IOCCG dataset (n=300).
3.4.2. Experiment 2

In order to evaluate Exp2 of the proposed approach, a testing data set that was not involved the training of the RBF or the MLP networks was used. Figure 15 shows the testing spectra clustered using the previously trained RBF network. Figure 16 depicts the scatter plot for each cluster in the testing data set and for the entire system. The latter yielded $R^2_{\text{test}} = 0.943$ and $\text{MSE}_{\text{test}} = 0.0388$.

![Figure 15: Exp2 spectral data clusters formed by the RBF network for testing data set ($n=228$). The numbers of samples in clusters 1 to 7 are 58, 15, 17, 51, 4, 28, and 55, respectively](image-url)
Figure 16: Exp2 scatter plots of the MLP outputs vs. desired targets simulated by the cluster-based approach for the testing data set. Scatter plots are shown for each of the seven clusters and the entire network ($n=228$).
4. COMPARISON OF CLUSTER-BASED AND SINGLE MLP CHLOROPHYLL A ESTIMATION

The performance of the cluster-based method developed in Exp1 and Exp2 was compared to a single MLP. The single MLP was developed in a consistent manner with the development of both Exp1 and Exp2.

4.1. Experiment 1

Using the 300 NOMAD spectra, a single MLP was trained after the spectra were transformed to have zero mean and unit variance. The network had five input nodes with hyperbolic tangent sigmoidal activation functions and a single linear output node. The resulting network was validated with 128 spectra yielding $R^2_{\text{val}} = 0.927$ and $\text{MSE}_{\text{val}} = 0.037$ (Figure 17). The network size was also varied up to a 7-7-1 MLP with sigmoidal hidden layer nodes with a marginal increase $< 0.010$ for the correlation coefficient and $< 0.0010$ for the mean squared error. Comparing the results of the single network solution to the clustering method developed in Exp1, both the clustering method and the single network method were able to achieve high correlation with the clustering approach having a slightly higher correlation of $R^2_{\text{val}} = 0.955$ versus $R^2_{\text{val}} = 0.927$ for the single network method. The estimation error of the clustering approach ($\text{MSE}_{\text{val}} = 0.025$) is smaller than the error of the single network approach ($\text{MSE}_{\text{val}} = 0.037$).
The performance of the single MLP was tested over the entire IOCCG dataset (Figure 18). The single MLP predicted the Chl $a$ that correspond to the 500 spectra with $R^2_{\text{test}} = 0.951$ and $\text{MSE}_{\text{test}} = 0.094$. The single MLP output of the IOCCG dataset was separated into the “clustered” response for comparison against the cluster-based MLP (Figure 19). The cluster-based approach and the single network based approach succeeded in some clusters, but both performed below expectations over similar data in clusters 1, 2 and 4.

The single MLP was also tested with the 300 IOCCG spectra, which were used for testing the cluster-based approach. The single MLP estimated Chl $a$ that correspond to these spectra with $R^2_{\text{test}} = 0.827$ and $\text{MSE}_{\text{test}} = 0.095$. Figure 20 illustrates the single MLP output for the 300 IOCCG spectra separated into the “clustered” response for comparison.
against the cluster-based MLP. The results for all systems are summarized in Table 4.

Figure 18: Exp1 scatter plot of the entire IOCCG dataset simulated by the single MLP (n=500).
Figure 19: Exp1 the IOCCG dataset simulated by the single MLP separated into clusters in order to compare the results to the cluster-based approach (n=500).
Figure 20: Exp1 scatter plots for a subset of the IOCCG dataset simulated by the single MLP separated into clusters in order to compare the results to the cluster-based approach ($n=300$).
4.2. Experiment 2

The performance of the cluster-based method was compared to a single MLP in estimating Chl $a$ concentration. Using the training data set from Exp2, a single MLP was trained. The MLP had five input nodes with hyperbolic tangent sigmoidal activation functions and a single linear output node. The resulting network was tested on the 228 previously unseen testing spectra, yielding $R^2_{\text{test}} = 0.933$ and $\text{MSE}_{\text{test}} = 0.0462$. Comparing the results of the single network solution to the clustering method developed here, both the clustering method and the single network method were able to achieve high correlation with the clustering approach having a slightly higher correlation of $R^2_{\text{test}} = 0.943$ (Figure 16) versus $R^2_{\text{test}} = 0.933$ (Figure 21) for the single network method. The estimation error of the clustering approach ($\text{MSE}_{\text{test}} = 0.0388$) is smaller than the error of the single network approach ($\text{MSE}_{\text{test}} = 0.0462$). The results for all systems are summarized in Table 4.

**Table 4: Summary of performance for all systems**

<table>
<thead>
<tr>
<th></th>
<th>$R^2_{\text{training}}$</th>
<th>MSE training</th>
<th>$R^2_{\text{validation}}$</th>
<th>MSE validation</th>
<th>$R^2_{\text{testing}}$</th>
<th>MSE testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp1</td>
<td>0.964</td>
<td>0.01</td>
<td>0.955</td>
<td>0.025</td>
<td>0.841</td>
<td>0.12</td>
</tr>
<tr>
<td>Exp2</td>
<td>0.978</td>
<td>0.015</td>
<td>0.964</td>
<td>0.0241</td>
<td>0.943</td>
<td>0.0388</td>
</tr>
<tr>
<td>MLP Exp1</td>
<td>0.931</td>
<td>0.0427</td>
<td>0.927</td>
<td>0.037</td>
<td>0.827</td>
<td>0.095</td>
</tr>
<tr>
<td>MLP Exp2</td>
<td>0.978</td>
<td>0.155</td>
<td>0.948</td>
<td>0.349</td>
<td>0.933</td>
<td>0.0462</td>
</tr>
<tr>
<td>OC4 Exp2</td>
<td>0.918</td>
<td>0.0898</td>
<td>0.918</td>
<td>0.0779</td>
<td>0.898</td>
<td>0.1015</td>
</tr>
</tbody>
</table>
Exp2 was also compared to the OC4 version 4 formula [23]; the current operational algorithm for the SeaWiFS project. When compared to OC4 v4, Exp2 performed better over the training, validation and testing sets. Figure 22 depicts the Exp2 clustered OC4 algorithm performance over the Exp2 testing set.

Figure 21: Exp2 scatter plots for the testing data set simulated by the single MLP and separated into clusters in order to compare the results to the cluster-based approach (n=228).
Figure 22: Scatter plots for the Exp2 testing data set simulated by the OC4 algorithm and separated into clusters in order to compare the results to the cluster-based approach (n=228).
5. Conclusion

In this project, it has been shown how clusters formed by using the fuzzy c-means algorithm can be used to develop an ensemble of cluster-based artificial neural networks to estimate Chl $a$ concentration for heterogeneous water types. The performance of two different experiments of the proposed method was compared with the traditional approach, where a single neural network was used for all water types. Although a comparison of clustered-based approach to a single neural network performance yielded similar results for the datasets used, the cluster-based approach has the potential to build a more global Chl $a$ prediction model and performs slightly better for the presented experiments. The results offered by Exp2 offer a better indication of the comparison between the cluster-based system and a single MLP since the use of the training, training validation, and testing is more consistent throughout the experiment when compared to the reuse of data points and complete separation of the IOCCG and NOMAD data in Exp1. Exp2 was also compared to the existing working algorithm used by the NASA SeaWiFS project and was able to perform better than the OC4 algorithm over the same data. The comparison of Exp2 and OC4 serves to offer evidence that the cluster based approach is able to perform better than a single MLP network and offer equivalent, if not better performance, than existing standards.

There are a few topics that are not discussed in this study that would be worthwhile investigating. Feedback received on this research has indicated the need to investigate combining the outputs of various cluster-based neural networks based on their degree of membership to accurately estimate Chl $a$; a weighted cluster based approach. It has been
proposed that such an approach may improve the estimation of Chl $a$ when the reflectance measurement is from previously unseen water types or from locations where a transition in water type occurs. The system used in both Exp1 and Exp2 is based on a hard decision in the FIS and does not consider the partial cluster membership of a spectra. If work were done on this topic, one concern that will need to be addressed is how to account for the error in the outputs of the individual cluster MLP’s, specifically how to account for error in the output of an MLP network when the input is not like the training date. An MLP is expected to perform well over data that closely resembles the data that was used to train and validate the training of the network, it cannot be expected to accurately estimate an expected value for an input that does not resemble the training data or data that is associated with an output that is out of the range of what the MLP can produce.

Another topic for future work not explored in this study is the meaning of the individual clusters. The data used represents different heterogeneous waters types from around the globe. The location and water type of each spectra was not tracked or analyzed in this study. It would be worthwhile to investigate if the FCM algorithm was able to cluster the data by geographic location and investigate whether the created clusters had any significance. The investigation of the meaning and geographic location of the clustered spectra could aid in determining like water bodies and local ecosystems.
REFERENCES


BIOGRAPHY OF THE AUTHOR

Kevin Turner was born in Morristown, NJ and moved to Bangor, ME where he graduated from Bangor High School in 2000. After graduating from High School, Kevin moved on to graduate from Bucknell University with a Bachelor of Science Degree in Electrical Engineering in 2004. After his undergraduate experience, Kevin moved back to Bangor and enrolled in the Graduate program in the Electrical and Computer Engineering Department at the University at Maine. While at the University of Maine Kevin was supported by a Grant from the NASA/Maine Space Grant Consortium and conducted research on Ocean Water Chlorophyll-a Concentration utilizing computational intelligence techniques. Working under H.W. Ressom and M.T. Musavi, Kevin published and presented a paper at the 2006 International Joint Conference on Neural Networks and contributed to a paper presented at Oceans 2006. Kevin is currently a student member of The Society of Automotive Engineers and has been a member of the Computational Intelligence Society and IEEE. Kevin is currently employed as a Research Engineer and Project Manager at a non-profit research institute. In his spare time, Kevin enjoys competing in the New England Region SCCA Rally Cross Series. He is a candidate of the Master of Science degree in Electrical Engineering from The University of Maine in August, 2007.