2014 Hawaii University International Conferences
Science, Technology, Engineering, Math & Education
June 16, 17, & 18 2014
Ala Moana Hotel, Honolulu, Hawaii

How can topology, self-organizing map (SOM) and geographical information system (GIS) enhance the study of Asian green mussels (Perna viridis)?

Navarra-Madsen, Junalyn
Texas Woman's University
Dept. of Mathematics & Computer Science

Junalyn Navarra-Madsen
Dept. of Mathematics & Computer Science
Texas Woman's University

HOW CAN TOPOLOGY, SELF-ORGANIZING MAP (SOM) AND GEOGRAPHICAL INFORMATION SYSTEM (GIS) ENHANCE THE STUDY OF ASIAN GREEN MUSSELS (PERNA VIRIDIS)?

Synopsis:

How can topology and geographical information systems enhance the study of Asian green mussel (Perna viridis)? In this paper, the author explains the application of basic topological rules in enhancing geodatabases applied to oceanographic studies.

How can topology, self-organizing map (SOM) and geographical information system (GIS) enhance the study of Asian green mussels (*Perna viridis*)?

Junalyn Navarra-Madsen

Associate Professor of Mathematics

Department of Mathematics and Computer Science

Texas Woman's University

HOW CAN TOPOLOGY, SELF-ORGANIZING MAP (SOM) AND GEOGRAPHICAL INFORMATION SYSTEM (GIS) ENHANCE THE STUDY OF ASIAN GREEN MUSSELS *(PERNA VIRIDIS)?*

JUNALYN NAVARRA-MADSEN
ASSOCIATE PROFESSOR OF MATHEMATICS
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

Abstract. Topology is often described in layman's terms as rubber-sheet geometry. Why? Topologists cannot distinguish their coffee cups from their doughnuts. Topology is crucial in understanding geographical information systems (GIS). Topological data structures are central to GIS. What are the advantages of topological data structures? Storage for polygonal elements is reduced because boundaries between adjacent polygons are not stored twice; relations between features are maintained; and maps are improved by research-based methods. Topology is not needed for every GIS project, but, topology is valuable in finding the optimal path between a set of points, classes and features. Topologists have certain skills and training in filling gaps found in some of these information systems by applying mathematical research-based methods. In this paper, the author explains the application of basic topological rules in enhancing geodatabases applied to oceanographic studies. The use of self-organizing maps (SOM) method is explained. SOM, a powerful type of articial neural network, can mine patterns from a variety of quantitative data. Using SOM, the author describes the seasonal (monthly, inter-annual) pattern variability of spatfall settlement of *Perna viridis* (Asian Green Mussel), and links the output to ancillary variables by using a number of techniques in processing data of high dimensionality and complexity.

1. Introduction

Many mathematical papers establish mathematical and physical insights highly beneficial to mankind. Unfortunately, these papers are too often untapped, overlooked and underappreciated resources hibernating on dusty library shelves. Except for specialists in the field, this pantheon of thousands of man-hours of work is rarely referenced due to its esoteric language. The language of mathematics can be overwhelming to those who are not studying it. This paper will minimize the emphasis of mathematical formulae and focus more on making concepts more understandable by providing more illustrations, because research has shown that most of us are visual learners. The major objectives of this paper are to:

(1) illustrate the importance of topology in geographical information systems (GIS);
(2) define the not-so-friendly mathematical tool called self-organizing map (SOM) and explain some of the important aspects of its output plots and graphs;

The author wishes to acknowledge the generous support coming from the Research Development Fund of the College of Arts and Sciences, Texas Woman's University (TWU) and the 4-month visiting faculty research award from the National Science Foundation via the Institute for Mathematics and its Applications (IMA).

(3) demonstrate the ability of SOM to classify large data sets and still preserve topology.

1.1. **What is a GIS?.** Worboys and Duckham's definition is, "A geographic information system (GIS) is a computer–based information system that enables capture, modeling, storage, retrieval, sharing, manipulation, analysis, and presentation of geographically referenced data [1, p.2]." A GIS allows us to view, understand, question, interpret, and visualize data in many ways that reveal relationships, patterns, and trends in the form of maps, globes, reports, and charts. A GIS helps us answer questions and solve problems by looking at our data in a way that is quickly understood and easily shared [ESRI]."

1.2. **What is topology?** Topology or "rubber-sheet geometry" is the study of properties of space that are preserved under continuous stretching, twisting and bending without ripping or making holes. Because continuous deformations without tearing are allowed, topology is valuable in finding the optimal path between a set of points, classes and features.

1.3. **GIS and topology.** The utility of topology in GIS can be grouped into two broad areas.

The first area is the support of database development. Knowledge of the topological conditions in a data set can be used to discover structural problems with the feature database, e.g. polygons that are not closed or are overlapping. It can also be used to automate feature creation and ensure feature integration.

The second area is the provision of spatial analyses. These can include using connectivity for network analysis, area definition to determine containment, and contiguity for neighborhood analysis.

2. Self-Organizing Map (SOM)

SOM is an artificial neural network (ANN) that is trained using unsupervised learning to produce a two-dimensional, discretized representation of the input space of the training samples by using a neighborhood function to preserve the topological properties of the input space. A Finnish professor, Teuvo Kohonen, was the first to describe SOM [2]. SOM can extract patterns and learn via the training algorithm [3, 4]. Input data are presented successively to the network and after this iterative process, the nodes converge to positions that represent the input data.

Bação and others [5] reported in their paper that SOM is less prone to local optima than k-means and that the search space is better explored by SOM.

2.1. **How is SOM utilized in oceanographic and ecological studies?** Chon [6] has surveyed several ecological studies using SOM. Richardson and others [7] have used SOM to extract patterns in satellite imagery. Hardman-Mountford et al. [8] have applied SOM to altimeter data and related it to Namibian sardine recruitment. Liu et al. have studied ocean current patterns [9] and sea surface temperature patterns [10] using SOM.

2.2. **Why use GIS and SOM together?** Interpretations of self-organizing maps are not easy. There is a need to improve visualization for better interpretation of results [11, 12]. Fincke et al. [11] have made use of GIS to visualize SOM results and shown a way to import traditionally-created SOM into GIS so that 3-D spatial analysis can be performed.

3. The SOM Algorithm

SOM is usually classified as unsupervised learning with the goal of discovering some underlying structures of the input data. SOM as a topology-preserving map maintains neighborhood relations. This is classically written as

$$y_i \sim y_j$$

whenever

$$x_i \sim x_j$$

where x_i, x_j are input data which can be multidimensional, y_i, y_j are nodes from the output layer which is usually 2-dimensional, and $\sim$ is read as "related to". As a general rule, the connections within a group of similar or related input data is greater than the connections outside this group. Figure 1 provides the schematic diagram of this mapping.

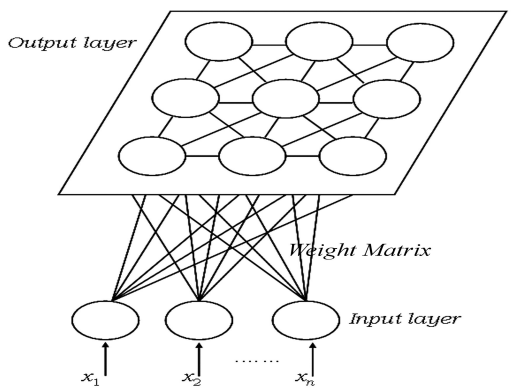

Figure 1. Schematic Diagram of SOM

There are basic assumptions for this algorithm. They are: (1) Output nodes are connected as 1-dimensional or 2-dimensional array and (2) the whole network is fully connected, i.e. all input nodes are connected to all nodes in the output layer. Each output node is weighted. Given a randomly selected input vector x, the winning output node i is denoted by $W_i(x) \geq W_k(x)$ whenever the weights are normalized

$$|w_i - x| \leq |w_k - x| \ \forall k.$$

Given the winning output node i, the weight update is

$$W_k(new) = W_k(old) + \mu \mathcal{N}(i,k)(x - w_k)$$

where $\mathcal{N}(i,k)$ is a neighborhood function.

4. Asian Green Mussels

Asian green mussels (*Perna viridis*) are members of the family *Mytilidae* (Linnaeus 1758). The family Mytilidae is characterized by two narrow, fan-shaped, thin valves, the absence of prominent hinge teeth and often the presence of byssal threads for anchoring to hard substrates. *Perna viridis* can grow between 80 to 100 mm with a maximum length of about 160 mm [13].

4.1. **Mussel Life Cycle.** These mollusks have separate sexes and fertilization occurs externally. Sperm and egg are released; after fertilization, 7 or 8 hours, free-swimming larvae can remain free-swimming for 2-3 weeks [14]. Using their byssus, 3-week old larvae settle onto seaweeds and later attach themselves on rocky subtidal or intertidal flats [15, 16]. During their planktonic period, larvae will be widely dispersed by physical processes, but may aggregate periodically at certain depths through a variety of biological processes, most notably diel vertical migration. Juveniles may be able to reach sexual maturity within 2-3 months, and may live as long as 3 years. Adult populations may reach densities of 35,000 individuals per square meter [17].

4.2. **Reproductive Cycle.** *Perna viridis* become sexually mature after 12 weeks at 20-30mm shell length. The life span of *Perna viridis* is typically 3 years. Growth rates are influenced by environmental factors such as temperature, food availability and water movement [18].

A number of studies have been done on the closely related species *Perna canaliculus* [19, 20, 21, 22]. Figure 2 shows the life cycle of *Perna canaliculus* taken from University of Waikato webpage, `www.biotechlearn.org.nz/focus_stories/farming_green_lipped_mussels/images/mussel_life_cycle`.

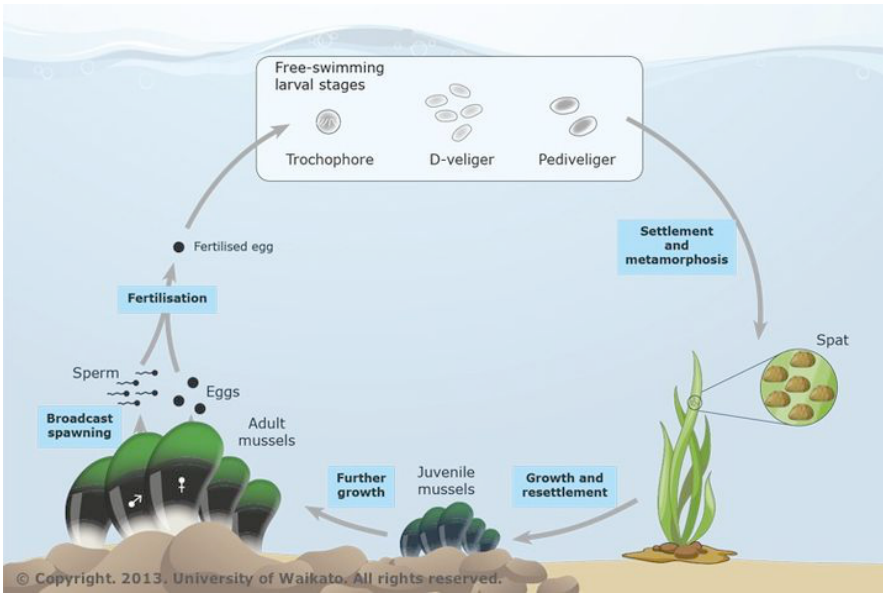

Figure 2. Life Cycle of *Perna canaliculus*

4.3. **Larval Transport, Recruitment and Settlement.** A number of researchers [23, 24, 25] have reported that transport of the planktonic larvae of benthic invertebrates is mostly influenced by a number of physical and biological factors. One of the important questions in most mussel ecological studies for aquaculture application is how to determine the location of adult broodstock mussels and the accompanying macroalgae assemblages that serve as initial habitat for recruitment and settlement. Although a number of studies had been published about this topic, knowledge gaps still exist. For instance, more studies need to be done to describe the mechanisms involved in the transport, recruitment and settlement of mussel larvae to bottom drifting macroalgal materials, e.g. the free-swimming larvae which are approximately $.3mm$ in length undergo morphological changes and start attaching themselves on seaweeds. The morphologically transformed larvae accumulate. This biomass, collectively called "spat',' can sometimes engulf large area the seaweeds assemblages. More mechanistic questions arise as to how these seaweeds containing spat arrive near certain surf zones in different parts of the world at specific times of the year. The next section explains the temporal variability of spat arrival using SOM.

5. Describing Mussel Spat Arrival Variability via SOM

In Alfaro et al.'s paper [22], certain statistical analyses describing the relationship between temporal patterns of arrival of *Perna canaliculus* spat in the Ninety Mile Beach in Northern New Zealand were presented. Various hydrodynamics and oceanographic factors such as wind speed and direction, tidal range, water temperature and swell height and direction were the explanatory variables utilized. Three models were generated and several hypotheses tested to study the 1990-1999 data set. The first logistic regression model fitted on the probability of spatfall events (any amount of algae and spat collected on a given day) contain the following parameters (1) mean wind speed (m/s), wind direction (onshore and offshore), tidal range (mm), and date (January 1990 to December 1998); water temperature (degrees C) was added to the first model; and swell height (m) in the onshore direction was added to the model. This study led to the conclusion that spatfall events and the amount of spatfall increased with strong offshore winds.

5.1. **Modeling and Simulation via SOM and Hierarchical Clustering.** How do we utilize SOM to gain insight about the variability of spat arrival of *Perna viridis* in surf zones of several islands in the Philippines?

We start with the simplest model using three parameters and increase its complexity by adding a parameter at a time. Model 1 contains 200 randomly selected set of days in a 10-year period; has input vectors $\vec{x}_i$, having three dimensions (physical factors: mean wind speed (ms^{-1}), wind direction (onshore or offshore) and tidal range (mm). Figure 3 shows the "Training Progress and SOM Neighbor Distances" describing Model 1. The output layer is chosen to be a 5×5 hexagonal 2-dimensional lattice.

We use the packages "kohonen" with "hdc" [26, 27] provided via software R [28], version R-3.0.3, and implement the simulations using the graphical user interface (GUI), *R Studio*®, [29].

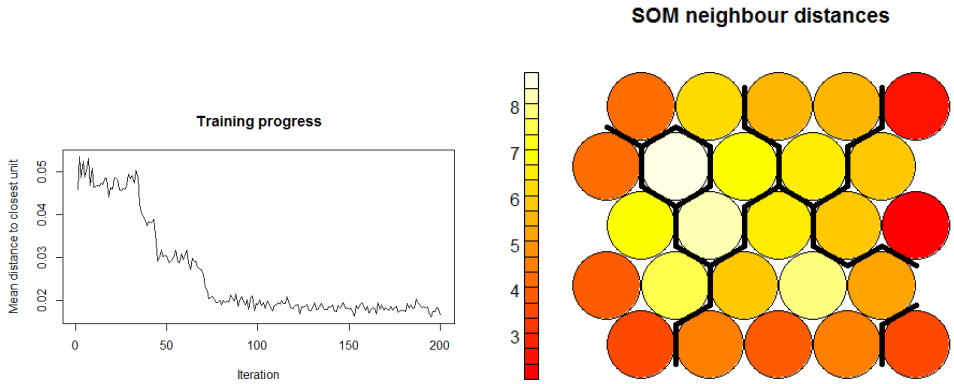

FIGURE 3. Training Progress and SOM Neighbor Distances Describing Model 1

We then add mean sea surface temperature ($°C$) to model 1 to obtain Model 2. Figure 4 shows the "Training Progress and SOM Neighbor Distances" describing Model 2.

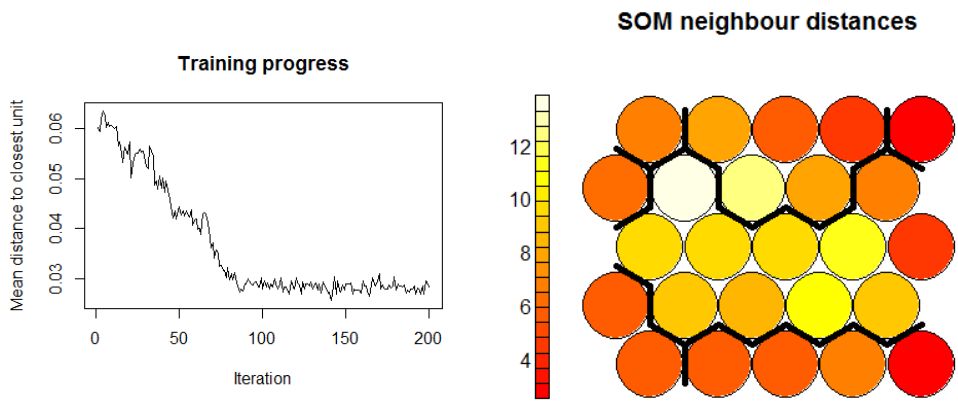

FIGURE 4. Training Progress and SOM Neighbor Distances Describing Model 2

To obtain Model 3, we add mean swell height (m) to Model 2. Figure 5 shows the "Training Progress and SOM Neighbor Distances" describing Model 3.

5.2. **Insights Gained.** Interpretation of SOM images resulting from simulation is not an easy task.

The three graphs in Figures 3, 4 and 5 showing the rates of SOM training progress of these three models look similar at first glance. The number of iterations utilized is 200 and there is a fast decrease of mean distance closest to unit between the first and ninetieth iterations. As soon as the hundredth iteration is reached, all three graphs oscillate. Model 1 has the smallest oscillation rate in terms of mean distance to unit while Model 3 has the largest.

Examining the vertical legend describing "SOM Neighbor Distances" of the three maps in Figures 3, 4 and 5, the distances between neighbors increase as the number of parameters increases. Model 1 has the smallest distance range of only 5 between neighbors while Model 3 has the largest (10). The presence of more whitish-yellowish

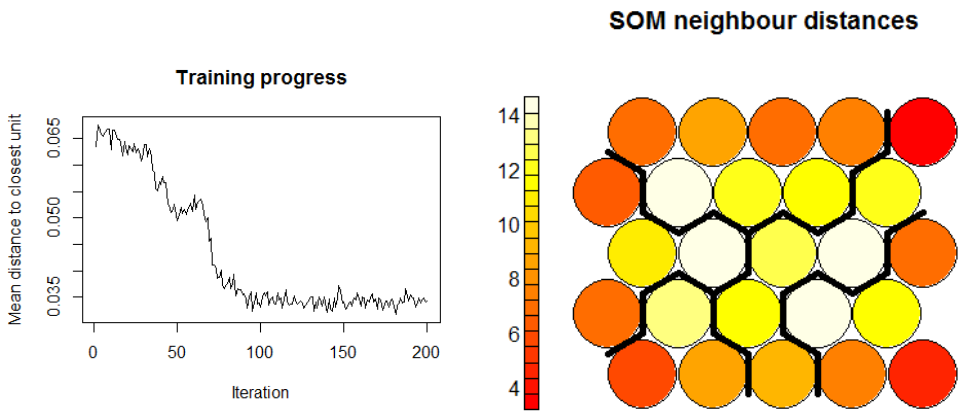

FIGURE 5. Training Progress and SOM Neighbor Distances Describing Model 3

color in Model 3 map can be interpreted as a better visual aid in distinguishing one cluster from another. The variability of these three colored lattices demonstrates the need for additional input parameters to improve the delineation of different clusters.

5.3. **Dendograms.** Using hierarchical clustering [27], the SOM simulation generated graphical outputs such as dendograms. Reading and interpreting dendograms as output graphs of certain model simulation containing large number of inputs can sometimes be complicated. Figure 6 illustrates this scenario. Given a large data set to cluster, the bottom branches of the dendograms can be unreadable.

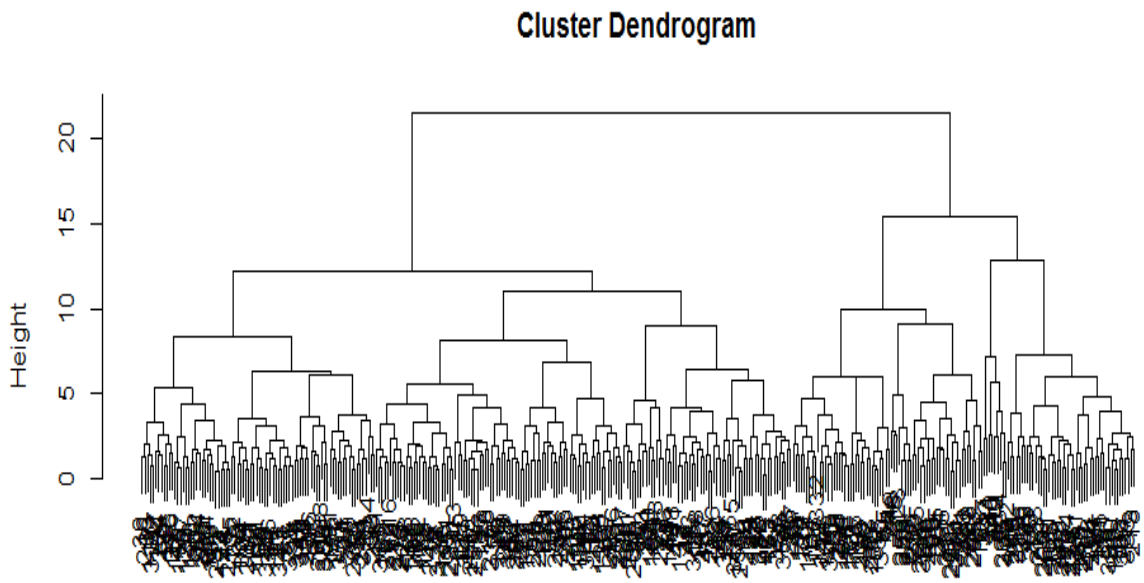

FIGURE 6. A Not-So-Readable and Not-So-Easy to Interpret Cluster Dendogram

For the sake of making the explanation simpler and constructing more readable graphs, let us take a sample of 50 input vectors and via hierarchical clustering we obtain its corresponding cluster dendogram shown in Figure 7. It is then easy to check the

smallest branches of the dendogram and find out to which cluster a certain input vector of interest belongs.

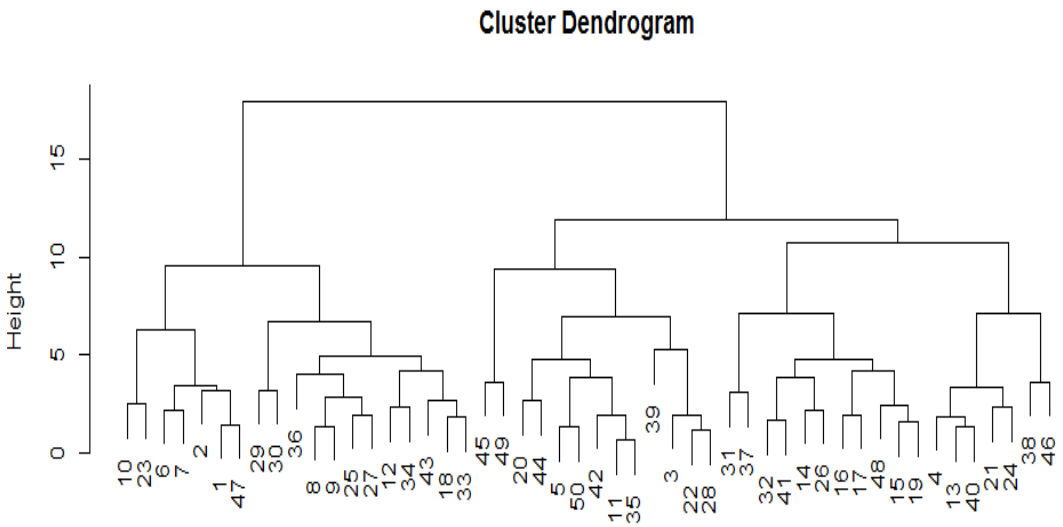

FIGURE 7. A More Readable and Easier to Interpret Cluster Dendogram

5.4. **Clustering via Mapping Plots.** How do we know that we have truly mapped each input vector of the raw data set to each node in the output layer?

5.4.1. *Large Sample - Larger 2-Dimensional Output Lattice.* Figure 10 shows two hexagonal mapping plots where each node of the 25 output layer nodes (except Node 13) contains classified input vectors. The grid used for this is 5×5 hexagonal grid because the sample is large enough $n = 365$ to cover all 25 nodes in the output layer for better clustering. The two background colors signify two seasons: gray for "dry season" and pink for "rainy season". Although these plots look similar in terms of the predicted background colors of gray and pink, if we check closely each node memberships, we can visually determine that these two plots have differences.

5.4.2. *Smaller Sample Size - Smaller 2-Dimensional Output Lattice.* The choice of the the size of the 2-dimensional output lattice depends on the input data complexity and bulk. An interdisciplinary research team with a good understanding of the issues could use exploratory data analysis to decide the shape (hexagonal or rectangular) and the $n \times n$ 2-dimensional lattice. Figure 9 shows two rectangular-shaped mapping plots where each node of the 9 output layer nodes contains 3 or more input vectors. The grid used for this is 3×3 hexagonal grid because the sample size is only 60. The node memberships are visually tractable and hence provide a better approach to and the zoomed-in box illustrating mussel landing and study sites in three islands of interest. Spat (millions of tiny green mussel larvae attached to seaweeds and other debris) arrive daily in the shores of these islands. This large collection can sometimes go unnoticed and untapped. With some initiative, some fishermen harvest the spat to start a mussel farm.

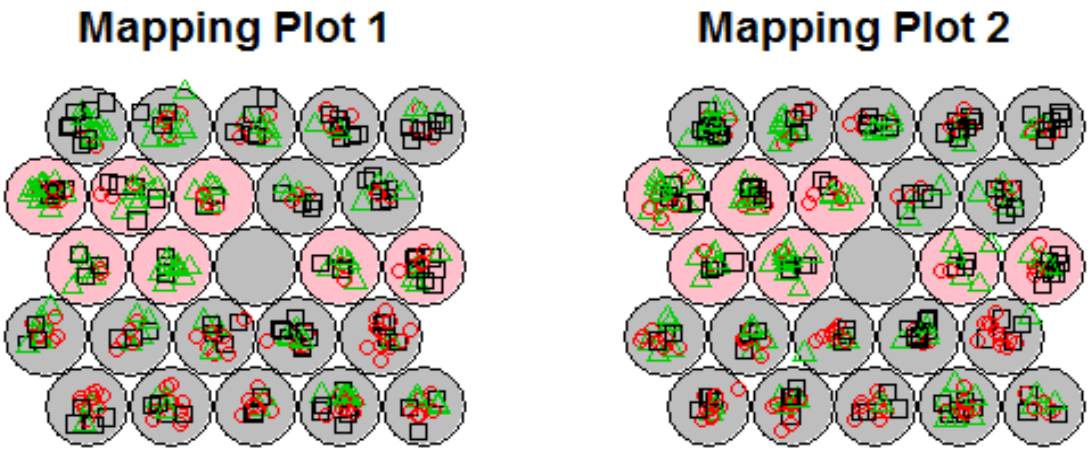

FIGURE 8. Almost Similar Mapping Plots (Quick Glance); Different Plots (Close Inspection)

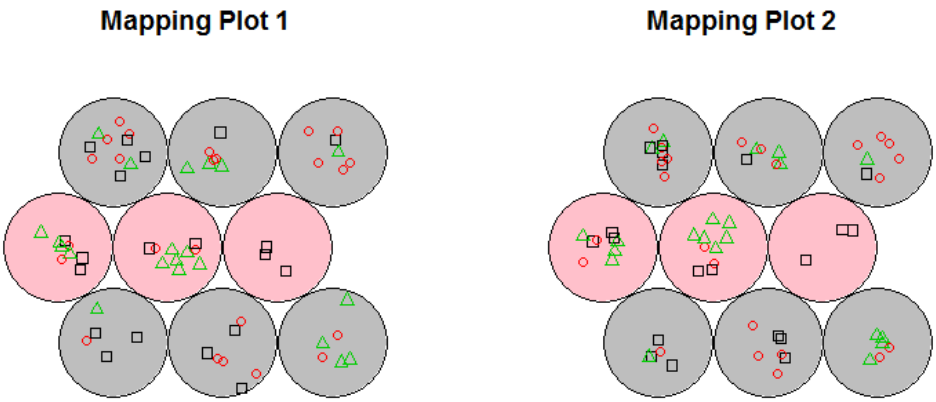

FIGURE 9. Again, Almost Similar Mapping Plots (Quick Glance); Different Plots (Close Inspection)

5.5. **Sample Interpretations.** There are three colored shapes (red circle, green triangle and black square) representing the mussel spatfall amount classification of the input vectors.

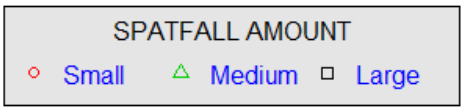

FIGURE 10. Three colored shapes (red circle, green triangle and black square) representing the spatfall amount.

The memberships of each node illustrated as pink or gray circle describe the spatfall amount occurring in that span of time. Two very obvious examples leap out of Figures 10, and 9. First, one of the gray nodes of Figure 10 does not contain any shape. This can explain that there is an instance during the dry season when no spatfall event occurs. Second, one of the pink circles (rightmost) in two mapping plots 1 and 2 of

Figure 9 contains three squares. This is explicitly describing that large spatfall amounts occur during this portion of rainy season.

6. GIS and SOM Combined

There are a number of Asian green mussel landing sites in the Philippines. Using an open source GIS software, QGIS (2014) [30], this paper illustrates a method to study mussel spat arrival. Figures 11 shows the Philippine archipelago with 7,107 islands.

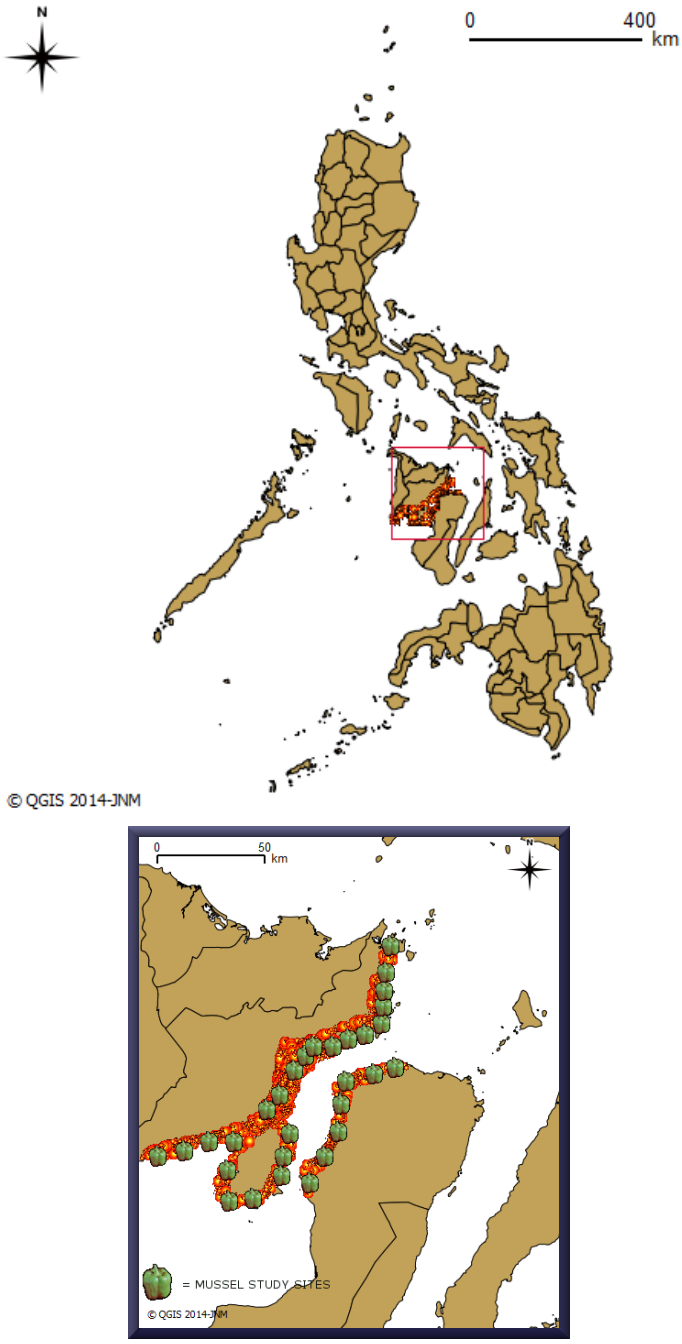

Figure 11. The Philippine Archipelago and the Zoomed-in Mussel Study Sites

SOM can catalog mussel spatfall data. This resulting insight could be used to optimize mussel aquaculture. GIS with SOM could be utilized to find the landing sites with the largest biomass of mussels and other economically-important invertebrates. These two tools could also pinpoint specific areas (rocky intertidal surf zones) to be recommended to stakeholders as favorable marine protected areas for sustainability and conservation purposes.

7. Remarks

Given the accelerated evolution of technological advancements (software and hardware), researchers all over the world gather immense amounts of data every second. The mathematics of "Big data" will be crucial in answering basic scientific questions. The neural network approach in this paper (SOM) can be a powerful technique to find hidden structures (topological or otherwise). SOM and its variants have the ability to "learn" (unsupervised, semi-supervised, or supervised) and can deal with incomplete and noisy data.

We do not claim that the modelling of mussel spatfall amount described above with only five physical factors is comprehensive. Every model has flaws. Ocean scientists believe that there are also biological factors (e.g., species pelagic larval duration and larvae swimming capability) to consider.

Classification and prediction are two principal goals in ecology and other scientific fields. Given the basic and simpler description and interpretation of SOM in this paper, scientific researchers (oceanographers, marine biologists, climatologists, etc.) can integrate and employ GIS with SOM to investigate large data sets for possible patterns and important biological and physical dynamics.

Gaining insights from plots and graphs highlights potential areas for georeferencing and geodatabasing. Some marine biologists and oceanographers will start collaborating with mathematicians and statisticians to improve the number-crunching capability of their research. Although, interdisciplinary work is highly beneficial, it is still crucial that nonmathematician researchers who are employing SOM in their studies have a good grasp of the topic and can explain better the implications of certain results. Given the expertise in their respective fields plus the ability to convey and interpret mathematical results, researchers can provide important stakeholders a certain level of confidence in considering research-based recommendations.

The tools described and simplified in this paper assist us in visualizing and understanding complex data structures.

References

[1] M. F. Worboys and M. Duckham. GIS: A Computing Perspective. 2004.

[2] T. Kohonen. The self-organizing map. *Neurocomputing*, 21:1–6, 1998.

[3] A. Ultsch and H. P. Siemon. Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. In *Widrow, Bernard; Angeniol, Bernard. Proceedings of the International Neural Network Conference*, pages 305–308. 1990.

[4] J Vesanto and E Alhoniemi. Clustering of the self-organizing map. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 11(3):586–600, 2000.

[5] F. Bação, V. S. Lobo, and M. Painho. Self-organizing Maps as Substitutes for K-Means Clustering. In *International Conference on Computational Science*, pages 476–483. 2005.

[6] T-S. Chon. Self-Organizing Maps applied to ecological sciences. *Ecological Informatics*, 6:50–61, 2011.

[7] A. J. Richardson, C. Risien, and F. A. Shillington. Using self-organizing maps to identify patterns in satellite imagery. *Progress in Oceanography*, 59:223–239, 2003.

[8] N. J. Hardman-Mountford, A. J. Richardson, D. C. Boyer, A. Kreiner, and H. J. Boyer. Relating sardine recruitment in the Northern Benguela to satellite-derived sea surface height using a neural network pattern recognition approach. *Progress in Oceanography*, 59:241–255, 2003.

[9] Y. G. Liu and R. H. Weisberg. Patterns of ocean current variability on the West Florida Shelf using the self-organizing map. *Journal of Geophysical Research*, 110, 2005.

[10] Y. G. Liu, R. H. Weisberg, and R. He. Sea surface temperature patterns on the West Florida Shelf using growing hierarchical self-organizing maps. *Journal of Atmospheric and Oceanic Technology*, 23, 2006.

[11] T. Fincke, V. Lobo, and F. Bação. Visualizing self-organizing maps with GIS. In *GI Days*, 2008.

[12] F. Bação, V. S. Lobo, and M. Painho. Geo-Self-Organizing Map (Geo-SOM) for Building and Exploring Homogeneous Regions. In *Geographic Information Science*, pages 22–37, 2004.

[13] S. Rajagopal, V. P. Venugopalan, G. van der Velde, and H. A. Jenner. Greening of the coasts: a review of the Perna viridis success story. *Aquatic Ecology*, 40:273–297, 2006.

[14] K. R. Hayes, R. Cannon, K. Neil, and G. Inglis. Sensitivity and cost considerations for the detection and eradication of marine pests in ports. *Marine Pollution Bulletin*, 50:823–834, 2005.

[15] C. K. Yap, S. G. Tan, A. Ismail, and H. Omar. Genetic Variation of the Green-lipped Mussel Perna viridis (L.) (Mytilidae: Mytiloida: Mytilicae) from the West Coast of Peninsular Malaysia. 2002.

[16] S. Rajagopal, V. P. Venugopalan, G. van der Velde, and H. A. Jenner. Mussel colonization of a high flow artificial benthic habitat: Byssogenesis holds the key. *Marine Environmental Research*, 62:98–115, 2006.

[17] S.Y. Lee. The reproductive cycle and sexuality of the green mussel perna viridis (l.) (bivalvia: Mytilacea) in victoria harbour, hong kong. *Journal of Molluscan Studies*, 54(3):317–323, 1988.

[18] S. Rajagopal, V. P. Venugopalan, K. V. K. Nair, G. Van der Velde, and H. A. Jenner. Settlement and growth of the green mussel Perna viridis (L.) in coastal waters: influence of water velocity. *Aquatic Ecology*, 32:313–322, 1998.

[19] A. C. Alfaro, A. G. Jeffs, and R. G. Creese. Bottom-drifting algal/mussel spat associations along a sandy coastal region in northern New Zealand. *Aquaculture*, 241:269–290, 2004.

[20] A. C. Alfaro. Byssal attachment of juvenile mussels, Perna canaliculus, affected by water motion and air bubbles. *Aquaculture*, 255:357–361, 2006.

[21] A. C. Alfaro. Population dynamics of the green-lipped mussel, Perna canaliculus, at various spatial and temporal scales in northern New Zealand. *Journal of Experimental Marine Biology and Ecology*, 334:294–315, 2006.

[22] A. C. Alfaro, B. McArdle, and A. G. Jeffs. Temporal patterns of arrival of beachcast green-lipped mussel (Perna canaliculus) spat harvested for aquaculture in New Zealand and its relationship with hydrodynamic and meteorological conditions. *Aquaculture*, 302:208–218, 2010.

[23] R. K. Cowen, C. B. Paris, and A. Srinivasan. Scaling of Connectivity in Marine Populations. *Science*, 311:522–527, 2006.

[24] C. B. Paris, L. M. Chrubin, and R. K. Cowen. Surfing, spinning, or diving from reef to reef: effects on population connectivity. *Marine Ecology-progress Series*, 347:285–300, 2007.

[25] M. Roughan, M. S. Macdonald, M. E. Baird, and T. M. Glasby. Modelling coastal connectivity in a Western Boundary Current: Seasonal and inter-annual variability. *Deep-sea Research Part Ii-topical Studies in Oceanography*, 58:628–644, 2011.

[26] R. Wehrens and L.M.C. Buydens. Self- and super-organising maps in R: the kohonen package. *J. Stat. Softw.*, 21(5), 2007.

[27] Laurent Bergé, Charles Bouveyron, and Stéphane Girard. HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, 46(6):1–29, 2012.

[28] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[29] RStudio Team. *RStudio: Integrated development environment for R, version 0.98*. RStudio, Boston, MA. Retrieved Feb 22, 2014.

[30] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2014.