A Hierarchical Clustering Approach to Analyze Similarities between Sea Surface Temperature Patterns in the Caribbean

Marc Boumedine
Computational and Computer Sciences Department
College of Science and Mathematics
University of the Virgin Islands
U.S. Virgin Islands, St. Thomas, 00802
mboumed@uvi.edu

***Abstract*—This study presents a clustering approach to analyze similarities between sea surface temperature patterns in the Caribbean. Our goal is to supplement existing predictive systems with data mining techniques by automatically extracting new patterns and ultimately increase the predictive accuracy of coral reef monitoring systems. The approach presented analyze times series sea temperature data from 2000 until 2014 collected the National Oceanographic and Atmospheric Administration National Environmental Satellite, Data and Information Science (NOAA/NESDIS) Coral Reef Watch(CRW) monitoring system. Unsupervised techniques (cluster analysis) are performed on Twenty three virtual stations in the Caribbean in an attempt to discover similarities and stressing patterns that might affect coral reef health in the region. The approach follows main three steps: (1) raw data selection and processing, (2) discretization and dimensionality reduction of time series using the Symbolic Aggregate Approximation (SAX), and (3) determination of sequence similarities/dissimilarities and hierarchical clustering of virtual weather stations according to sea surface temperature patterns.**

Keywords—Times Series Similarity; Discretization; SAX; Data Mining; Hierarchical Clustering; Coral Reel Ecological Systems (key words)

I. Introduction

Sea warming acts as an environmental stressor on coral health and may have a devastating impact on Caribbean economies. Environmental agencies and decision makers are strongly committed in assessing the impact of climate change and land use on coral health [1-3]. In order to support this type of assessment, intensive data analysis is required to better understand this phenomena. Due to the vast amount of data to process, this analysis can be aided by automated or semi-automated computational tools in order to discover interesting patterns, anomalies or trends in various data sources available such as atmospheric, oceanographic, biologic etc. Coral reef organisms are very extremely sensitive to changes (increase or decrease) in water temperatures. Ocean excessive warming causes coral polyps to expel the symbiotic algae (called zooxanthellae), essential for its survival. Once the algae is expelled, coral polyps look white or bleached. If stressing conditions persist, the coral will likely die. In order to monitor ocean warming and impacts on ecosystems, complex sensor networks and satellite imaging systems have been deployed. These systems collect large temporal data sets that can be exploited for discovering any hidden structure or useful patterns leading to stressing conditions.

This work investigates relationships between sea surface temperature patterns and stressing episodes that might threaten coral reef ecosystems. We specifically focus on similarities (dissimilarities) between SST patterns occurring in the Caribbean. Data sets have been obtained from the National Oceanographic and Atmospheric Administration National Environmental Satellite, Data and Information Science (NOAA/NESDIS) Coral Reef Watch(CRW) monitoring system. We analyze times series sea surface temperatures (SST) from January 2000 until October 2014 sampled from twenty three virtual stations in the Caribbean. We use unsupervised techniques (cluster analysis) of time series in an attempt to discover stressing patterns and trends that might affect coral reef health in the region. Data transformations are carried out to allow a lower dimensionality representation, analysis and visualization.

The major contribution of this paper is the application of SAX algorithm in an attempt to discover association pattern between thermal stress and coral bleaching alerts in the Caribbean. This new approach offers a new way of analyzing data collected from (NOAA/NESDIS) Coral Reef Watch(CRW) monitoring system and providing new insights on analyzing globally the effect of SST on fragile ecosystems such as coral reefs.

The approach described in this study presents the three steps: (1) raw data selection and processing, (2) using the Symbolic Aggregate Approximation (SAX), (3) determination of sequence similarities using hierarchical clustering and (4) Discussion and validation of the results. The remainder of this paper is organized as follows: section II describes the background and previous work, section III describes the methodology, section IV reviews SAX method , section V presents the clustering approach, and finally section VI presents the resulting clusters .

II. Background and Previous Work

The past two decades, intense efforts have been developed to monitor sea surface temperatures via remote sensing and in situ technologies. As a result, an increasing number of applications and opportunities are becoming available to drill into these data sets and contribute to developing ecological forecasting system in the Caribbean and globally . Building empirical models is time consuming and requires very specific knowledge of the domain. As the number of environmental variables increases, it is imperative to derive models assisted with machine learning and data mining techniques [4]. Machine learning techniques have been successfully used in many knowledge discovery applications [5-6]. However, despite the availability of NOAA's products and services, there has been very little time series data mining research conducted on NOAA produced SST data sets [7]. NOAA/NESDIS Coral Reef Watch(CRW) monitoring system provides five thermal stress alert levels: *no stress*, *bleaching watch*, *bleaching warning*, *bleaching alert 1* and *bleaching alert 2 (see Fig. 3)*. These levels are calculated based on NOAA's cumulative sum Degree Heating Week model [8]. Using unsupervised techniques (cluster analysis) on SST time series our ultimate goal is to discover association patterns between SST and thermal stress that might affect coral reef health in the Caribbean region.

Table 1. Partial SST St. Croix U.S.Virgin Islands-
http://coralreefwatch.noaa.gov/satellite/index.php

		SST	HoT			
Date	SST	ANO.	SPOT	DHW	Lat	Lon
11/28/2000	27.1	-0.5	0	0	18	-65
12/2/2000	27	-0.6	0	0	18	-65
12/5/2000	26.9	-0.6	0	0	18	-65
12/9/2000	26.9	-0.5	0	0	18	-65
12/12/2000	26.8	-0.5	0	0	18	-65
…						

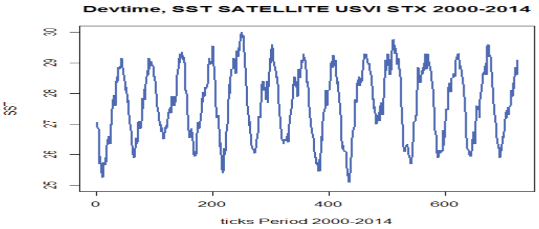

Fig. 1: SST times series for the US Virgin Islands(USVI) St. Croix Station January 2000-October 2014

Most algorithms reduce time series dimensionality using different representation in order to manage computational cost [9]. This is usually accomplished by preserving the general trends of the data using techniques such as single value decomposition, discrete Fourier transformation, piecewise aggregate approximation, adaptive piecewise constant approximation and symbolic aggregate approximation (SAX).

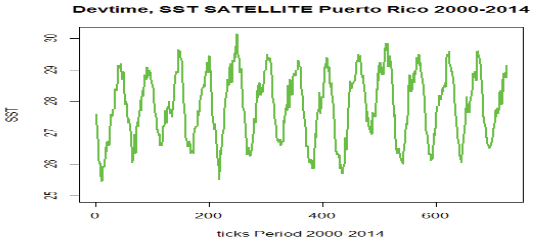

Fig. 2: SST times series for Puerto Rico Station January 2000-October 2014

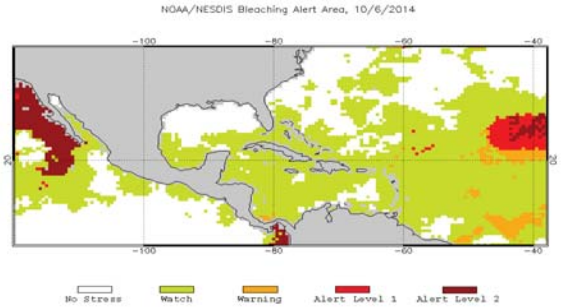

Fig. 3: SST times series for US Virgin Islands St. Croix Station January 2000-October 2014

Clustering algorithms seek to group object that are the most similar in the same cluster while minimizing similarity between clusters. Hierarchical clustering algorithms produce a nested representation represented graphically as a dendrogram which is easier to interpret and validate by domain experts.

Similarity (dissimilarity) measures can be expressed using a variety of approaches such as Euclidean distance, Dynamic Time Warping, distance based on Longest Common Subsequence.

III. Metholology

The overall methodology is shown on Fig. 4. As mentioned previously, our goal is to supplement existing models by automatically extracting new knowledge from NOAA data sets. In particular, this study focuses on finding similarities/dissimilarities between virtual stations in order to detect SST patterns as precursor of thermal stress leading to coral bleaching episodes in the Caribbean. The methodology is summarized below.

- We analyze SST times series obtained from (NOAA/NESDIS) Coral Reef Watch Monitoring systems. from January 2000-October 2014 sampled from virtual stations in the Caribbean (see Figures 1 and 2). Sea Surface Temperatures (SST) observations are sampled twice weekly at night-time at 0.5-degree (50-km) resolution by infrared radiometers**.** SST Times series are accessible from NOAA Coral Reef Watch portal (see Table 1).

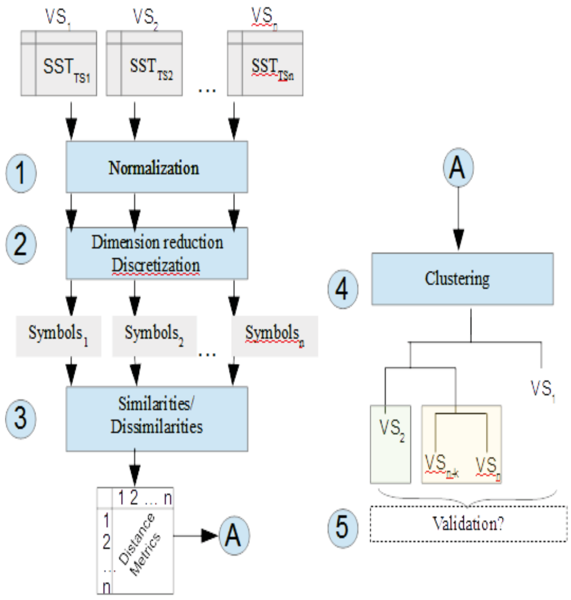

Fig.4 : Overall approach for constructing the clusters between virtual stations

- Time series are then normalized and transformations are carried out using Symbolic Aggregate Approximation [10] and to allow a lower dimensionality representation for analysis efficiency and visualization purposes (Fig. 4 steps 1 & 2). Partial samples observations are shown in Table I. Figs. 2 and 3 show NOAA/NESDIS SSTs observed at nighttime for two stations (U.S. Virgin Islands and Puerto Rico) in order to reduce variability due to solar glare.
- Hierarchical clustering is applied based on a pairwise distance matrix (see Fig.4,steps 3&4).
- The resulting dendograms are validated the the approximately unbiased and bootstrap probability Value [12]

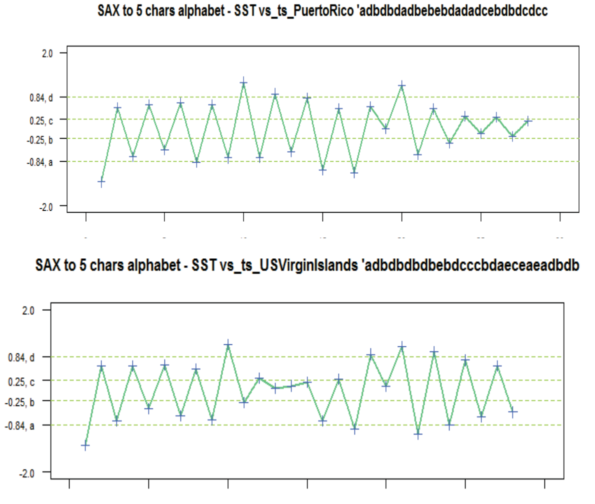

Fig. 5: SAX symbols representing USVI and Puerto Rico SST

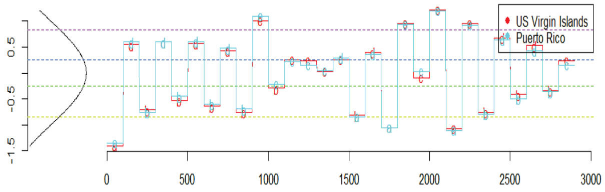

Fig. 6: SAX transformation of SST for USVI and Puerto Rico

IV. Symbolic Aggregate Approximation (SAX)

Sea Surface Temperatures (SST) observations are sampled twice weekly at night-time at 0.5-degree (50-km) resolution by infrared radiometers**.** SST Times series are accessible from NOAA Coral Reef Watch portal. Since our approach focus on SST trends (increase or decrease), we are particularly interested in high-level representation representation of the data. The piecewise aggregate approximation (PAA) and Symbolic Aggregate Approximation (SAX) have been widely used to compare sequence similarities for their interesting properties. SAX is briefly reviewed in the following section.

SAX is used to transform original time series into a symbolic representation [10-11] while preserving essential trends (see Fig. 5). This approach is based PAA representation and required to normalize SST observations in order to take advantage of Gaussian distribution properties. Normalized SST vectors are reduced using PAA which produces equal sized segments. The segments are converted into a symbolic representation (or a string) using an alphabet of symbols. Because of the five coral reef stressing level we chose to represent the size of the alphabet with the set {a,b,c,d,e}. Each symbols is then assigned to an equal sized interval under the Gaussian curve (see Fig. 6).

Lin and Keogh have shown that the distance measure between two symbolic strings created by SAX is a lower bound of the true distance between the two original time-series [10]. Since the times series for all the virtual stations have the same length the Euclidean distance applied to derive the distance matrix used in the hierarchical clustering process.

V-Clustering Approach

The purpose of the hierarchical clustering process is to reveal any similarities/dissimilarities between SSTs at various Caribbean virtual stations in an attempt to discover relationships between SST and coral reef thermal stress episodes. These episodes are determined using the Degree Heating Weeks (DHW) variable. DHW represents the weekly accumulation of heat exceeding the coral bleaching threshold. DHW values are available for each virtual stations for the period 2000-2014. In order to clusters the set of virtual stations represented by the sequences of symbols (or strings) we are using hierarchical clustering algorithm for both SST and DHW SAX symbols. Basically, the algorithm attempts to group the N virtual stations (N=23) based on the *N*N* similarity/dissimilarity matrix between all sequences of symbols. Fig. 5 shows SAX symbols obtained after PAA transformation for Puerto Rico and US Virgin Islands stations. The length of the original time series have been reduced to twenty-nine (w=29). The alphabet size is five (alpha=5).

Clustering process

1.Initially, each object (sequence of symbols) is assigned to a cluster.

2.The closest (most similar) couple of clusters are identified and merged.

3. The similarity matrix is updated by computing the distance between the new merged cluster and other clusters.

4.Steps 2 and 3 are repeated until all objects are grouped into a single cluster.

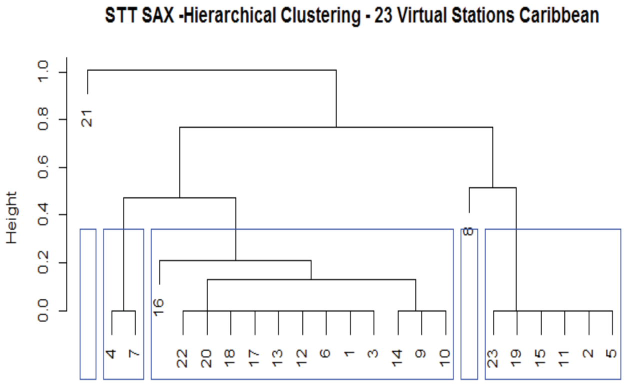

Fig.7: Dendogram for SST obtained with average method, Euclidean distance

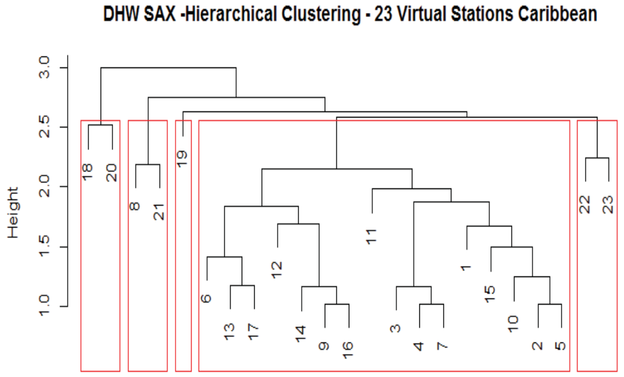

Fig. 8- Dendogram for 23 stations based on Degree Heating Week

7.VI- Preliminary Results

We run experiments with the single, complete and average linkage hierarchical clustering. In [17] Kaufman and Rousseeuw have shown that average linkage is a good technique performs with Euclidean distances but others metrics as well such . The results presented on Figs. 7 & 8 show clusters that were generated by the average-linkage clustering algorithm from the Tclust R package [16] The distance between each pair of clusters is the average distance from any items of one cluster to any items of the other clusters [13]. Table 2 list the names of the virtual stations with their associated numbers used on the dendograms shown on Figs. 7-10.

Table 2: Caribbean Virtual Stations

1) Banco Chinchorro, Mexico	9)Flower Garden Banks, Texas	17)Negril, Jamaica
2)Barbados	10)Glovers Reef, Belize	18) Puerto Morelos, Mexico
3)Bay Islands, Honduras	11) Guadeloupe	19) Puerto Rico
4) Bocas del Toro, Panama	12) Isla de la Juventud, Cuba	20)San Bernardo, Colombia
5)Bucco Reef, Tobago	13)Jardines de la Reina,	21) Santa Marta, Colombia
6) Cayman Islands	14)Lee Stocking Island, Bahamas	22)Turks and Caicos
7)Cayos Miskitos, Nicaragua	15) Los Roques, Venezuela	23) US Virgin Islands
8)Curacao and Aruba	16)Montecristi, Dominican Republic	

In order to assess results, the validation consists of analyzing the results and determine if the partitioning best fits the patterns represented by the SAX strings. Various cluster validity approaches have been proposed in literature such as the Adjusted Rand Index, the Silhouette Width, the Dunn Index [14-16]. Our approach assess the validity through the stability criteria. This criteria will guarentee that the clustering output will be similar for two different time series. In order to measure stability and reduce the uncertainty of the clustering process, *p* values are calculated based on the multiscale bootstrap resampling technique developed by Suzuki and Shimodaira [12]. Two indicators, the approximately unbiased (AU) and Bootstrap Probability Value (BP), provide a level of support into the reliability of the structure obtained with the hierarchical clustering algorithm. Fig. 9 shows p-values greater than 95% which suggest that the clusters were not generated by chance.

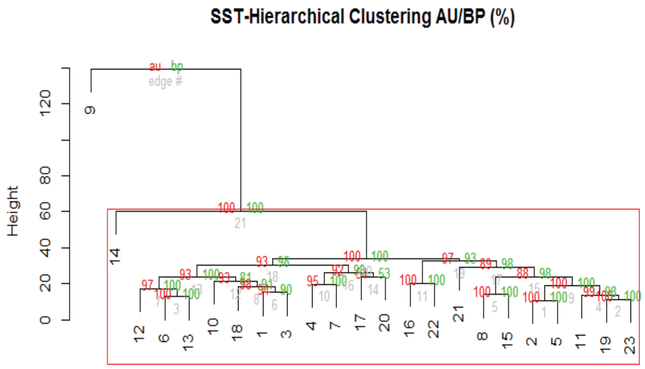

Fig. 9. Dendogram for SST with AU/AP percentage for Caribbean Stations

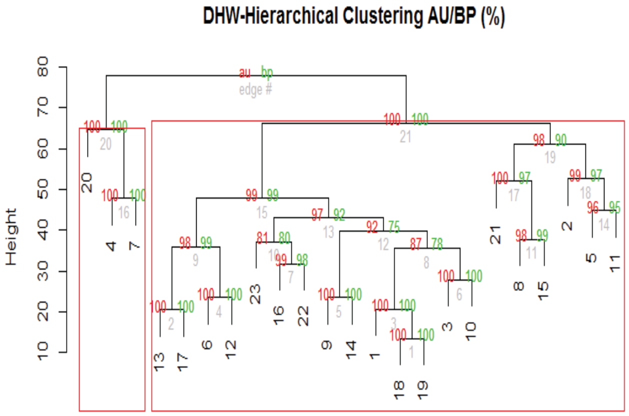

Fig. 10. Dendogram for DHW with AU/AP percentage for Caribbean Stations

VII- CONCLUSIONS AND FUTURE WORK

In this work, we proposed a hierarchical clustering approach in an attempt to reveal relationships between common SST patterns and thermal stress patterns (DHW) that might affect coral reef health. Observations from twenty-three virtual stations from 2000 until 2014 have been transformed using SAX and agglomerated into dendograms for further analysis. Before mapping SST patterns and DHW patterns the stability of the results have been assessed. The mapping between SST and DHW clusters will be presented in future work.

REFERENCES

[1] C. M. Eakin, J.M. Lough and S.F. Heron (2009). Climate Variability and Change: Monitoring data and evidence for increased coral bleaching stress. In M. Van Oppen & J.M. Lough [Eds.], Coral Bleaching: Patterns, Processes, Causes and Consequences. Ecological Studies 205, Springer, Berlin. 178 pp.Amigó G., Gonzalo, J. Artiles J., and Verdejo F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints," Inf. Retr. Boston., vol. 12, no. 4, pp. 461–486, 2009

[2] G.M. Wellington, P.W. Glynn, A.E. Strong, S.A. Navarrete, E. Wieters and D. Hubbard (2001). Crisis on coral reefs linked to climate change.EOS 82(1): 1,5.

[3] J. P McWilliams., I.M. Côté, J.H. Gill., W.J. Sutherland, and A. R. Watkinson (2005). Accelerating impacts of temperature-induced coral bleaching in the caribbean. Ecology 86:2055–2060. http://dx.doi.org/10.1890/04-1657.

[4] M. Boumedine. 2008. Mining ICON/CREWS Data Sets for Discovering Relationships Between Environmental Factors and Coral Bleaching. 11th International Coral Reefs Symposium (2008), Fort Lauderdale, USA, unpublished.

[5] Böttcher M, Höppner F, Spiliopoulou M (2008) On exploiting the power of time in data mining. ACM SIGKDD Explorations 10(2):3–11.

[6] Duda, R.O.,Hart, P.E.and Stork, D.G. (2001) Pattern Classification, 2nd ed. John Wiley and Sons Ltd.

[7] NOAA Coral Reef Watch, updated twice-weekly. NOAA Coral Reef Watch Operational 50-km Satellite Coral Bleaching Degree Heating Weeks Product, Jan. 1, 2001-Dec.31, 2010. Silver Spring, Maryland, USA: NOAA Coral Reef Watch. Data set accessed 2014-04-15 at http://coralreefwatch.noaa.gov/satellite/hdf/index.php

[8] Liu, G., A.E. Strong, J.W. Skirving and L.F. Arzayus (2006). Overview of NOAA Coral Reef Watch Program's Near-Real-Time Satellite Global Coral Bleaching Monitoring Activities. Proceedings of the 10th International Coral Reef Symposium, Okinawa: 1783-1793.

[9] Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems, 3(3),263-286.

[10] Lin, J., Keogh, E., Lonardi, S. & Chiu, B. (2003) A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.

[11] Lin, J. Keogh, E. Wei, L. and Lonardi, S. "Experiencing SAX: a novel symbolic representation of time series," Data Mining and Knowledge Discovery, vol. 15, pp. 107–144, 2007.

[12] R. Suzuki, and H. Shimodaira (2004) "An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters?", The Fifteenth International Conference on Genome Informatics 2004, P034

[13] Liao T. W, (2005) Clustering of time series data: a survey. Pattern Recognigtion 38(11):1857–1874

[14] Ben-Hur, A., A. Elisseeff, and I. Guyon (2002). A stability based method for discovering structure in clustered data. In Aetman, R.B. et al. (eds), Pacific Symposium on Biocomputing World Scientific Publishing Co., New Jersey.

[15] Dunn, J. C. (1974). Well separated clusters and fuzzy partitions. J.Cybernet.,4, 95–104.

[16] H. Fritz, L. A. Garcia-Escudero, A. Mayo-Iscar A. (2012). tclust: An R Package for a Trimming Approach to Cluster Analysis. Journal of Statistical Software, 47(12), 1-26.

[17] L. Kaufman and P. Rousseeuw (1990). Finding groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, Inc.