IADIS International Journal on Computer Science and Information Systems
Vol. 12, No. 2, pp. 102-114
ISSN: 1646-3692

MINING DIATOM ALGAE FOSSIL DATA FOR DISCOVERING PAST LAKE SALINITY

Ray R. Hashemi[1], Azita A. Bahrami[2], Jeffrey A. Young[1], Nicholas R. Tyler[3] and Jay Y. S. Hodgson[4]
[1]*Department of Computer Science, Armstrong State University, Savannah, GA, USA*
[2]*IT Consultation, Savannah, GA, USA*
[3]*School of Pharmacy, University of Georgia, Athens, GA, USA*
[4]*Department of Biology, Armstrong State University, Savannah, GA, USA*

ABSTRACT

Climate changes around a large body of water have an intertwined relationship with the salinity of the water and diatom algae growing within it. One may use the diatom algae fossils obtained from bottom of an inland lake to conclude the historical climate changes around the lake and by extension the historical salinity of the water. The discovery of the historical quantified salinity of inland lakes is extremely important to understanding climate change, carbon dioxide levels, and global warming. In this research effort, the past salinity levels for Santa Fe Lake located in New Mexico, USA, were discovered by mining the data of diatom algae fossils. Modified Rough Sets as the first component of the proposed hybrid system were used to establish the relationships between diatom algae data, expressed in linguistic values, and the climate changes. The established relationships were extended to embrace the linguistic values of water salinity. The outcome was a set of fuzzy patterns. Fuzzy Logic as the second component of the proposed hybrid system was employed to: (i) provide the membership functions for the different linguistic values of the salinity and (ii) produce a crisp value for the salinity of the water related to each slice of diatom fossil using the crisp values of algae abundance indices in each slice. The validity of the findings was tested which revealed 72% of accuracy for the produced results.

KEYWORDS

Data Mining, Past Salinity Levels Extraction, Diatom Algae Fossils, Modified Rough Sets, Fuzzy Logic, and Feature Extraction

1. INTRODUCTION

For a given inland body of water, any two of the three elements of the *climate change*, *diatom algae growth*, and *salinity level* of the water have a highly intertwined relationship. To explain it further, diatom algae are considered as one of the most robust climate proxies collected from

lakes (Battarbee 1986). Since they are extremely fastidious to climate changes, their emergence, decline, presence, and absence in lakes are highly indicative of climate changes. When there is a drought, the evaporation of the water accelerates and the salinity level of the water concomitantly increases. An extended drought causes a decline in some species of the diatom algae and a bloom of some other species in the water. Many species of diatom algae have a very narrow range of salinity tolerance and simply cannot live if values become too high or too low.

Climate change entails more than temperature differences alone. For example, since the atmosphere is coupled with the oceans, their combined circulation systems have caused some areas to warm, others to not warm, some to experience drought, and a few to receive more precipitation than before (Jones et al. 2001, Cook et al. 2004). Deciphering these patterns has cast uncertainty on models predicting future climate (Von Storch et al. 2004). Therefore, records of historical climate from areas sensitive to changes are needed to understand potential impacts of future global warming (Mann 2002).

The American southwest is such an area very prone to climate change, particularly drought (Cook et al. 2004). Tree rings (Grissino-Mayer 1996) and lakes studies have suggested that the climate in this region has gone through drought cycles in the past and will likely continue into the future (Hodgson et al. 2013). Using data from inland lakes in this area will be useful in linking the diatom algae to climate and climate to lake salinity and setting the template for our data mining approach.

As we mentioned, algae are extremely sensitive to changes in climate (Smol and Cumming 2000) and, therefore, diatom algae fossils retrieved from the lake's bottom inherently carry the indications for historical climate changes around the lake. The abundance of different types of algae in a given slice of fossil may be expressed in *linguistic* values of "Low", "High", and "Medium". (A linguistic value, in contrast with a *crisp* value, is *fuzzy*. Therefore, we use linguistic value and fuzzy value interchangeably in this paper). The abundance indices may be used by experts to calculate a climate value of "Rain" or "Drought" for the period of time (identified by Carbon 14 dating) represented by the slice of diatom fossil. Observing the generated data, we are intrigued by the following question. Can the same data be used to discover the past salinity level of the Lake's water? An answer to this question is the goal of this research effort. To be more specific, the goal is to mine a crisp salinity value for each slice of a Diatom algae fossil for which the algae abundance is expressed in fuzzy values.

The significance of such mining stems from the fact that the salinities of bodies of water are closely related to the global climate system (Adkins et al. 2002, Durack and Wijffels 2010, Mendelsohn et al. 1994, Durack et al. 2012). Specifically, climate changes such as freezing (ice ages), melting (global warming), and evaporation (droughts) that change water levels will alter the concentration of dissolved substances within the water (Verschuren et al. 2000) including salinity level of the water. In turn, changes in salinity will affect how well atmospheric carbon dioxide dissolves in water (Sigman and Boyle 2000). Carbon dioxide is undoubtedly one of the leading causes of global warming because it is a very powerful greenhouse agent, trapping large quantities of heat on the planet's surface (Barry and Chorley 1999). As salinity increases from reductions in water level, the ability of the waterbody to dissolve carbon dioxide decreases and the atmospheric carbon dioxide increases, resulting in greenhouse gas accumulation and global warming (Adkins et al. 2002). As temperatures rise, the lake begins to evaporate, which increases salinity and will then release more carbon dioxide to the atmosphere. Subsequently, the added carbon dioxide will cause additional warming, driving more lake evaporation. Furthermore, since climate change affects

precipitation and consequently lake levels, discovering historical lake salinity is paramount to understanding climate. Ultimately, understanding historical salinities over millennia are important when reconstructing past climates for predicting future changes.

To meet the goal, we (i) identify a set of patterns representing the relationship between linguistic values of the algae indices and climate using Modified Rough Sets, (ii) extend the patterns to represent the relationship between linguistic values of the algae and linguistic values of the water salinity level, and (iii) find a crisp value for the salinity level of each slice of the fossil using Fuzzy Logic.

The organization of the rest of the paper is as follows. The Previous Works is the subject of Section two. The Methodology is presented in Section three. The Empirical Results are discussed in Section four. The Conclusions and Future Research are covered in Section five.

2. PREVIOUS WORKS

The salinity of a body of water and climate changes have an intertwined relationship. The investigation of the historical quantified salinity level of a body of water definitely will shed light on the history of the climate change and provides for a more precise analysis of global warming and human contribution to the problem.

However, past salinities cannot be directly measured and must be interpreted from proxy data. Much attention has been paid to oceanic studies using chemical isotopes (Adkins et al. 2002, Henderson 2002, Emiliani 1955), but researchers are now increasingly using inland lakes as sources of historical salinity (Laird et al. 1998, Wilson et al. 1994). Majority of lakes have greater levels of carbon dioxide than the atmosphere, thus becoming potential major sources of greenhouse gasses (Cole et al. 1994, Jansson et al. 2012). Algal fossils collected from the bottom of lakes have emerged as reliable indicators of past salinity (Wilson et al. 1994). Many of these types of studies have relied on constructing weighted-averaging models, maximum likelihood, bootstrapping, and transfer functions from large datasets of multiple lakes.

One common obstacle challenging all these methodologies is the inherent uncertainty in the data. Here, we present a new method of extracting the historical salinity using a hybrid system of Modified Rough Sets and Fuzzy Logic for a single lake. This hybrid system is more adoptive to the uncertainty nature of data. To the best of our knowledge such a hybrid system has not been reported in the literature.

3. METHODOLOGY

A dataset, F, is given and each record, R_i, of the dataset represents a slice of diatom algae fossils retrieved from bottom of a lake, Figure 1. R_i contains the abundance indices for a set of algae in linguistic values of “Low”, “Medium” and “High”. R_i also contains an age value and a climate value representing the age of the fossil slice and climate around the lake for the slice. Subtraction of ages of the n-th slice and (n-1)-th slice of the fossils represents the number of years that algae of the slice n survived the climate assigned to the slice. The climate values are “Rain” and “Drought”.

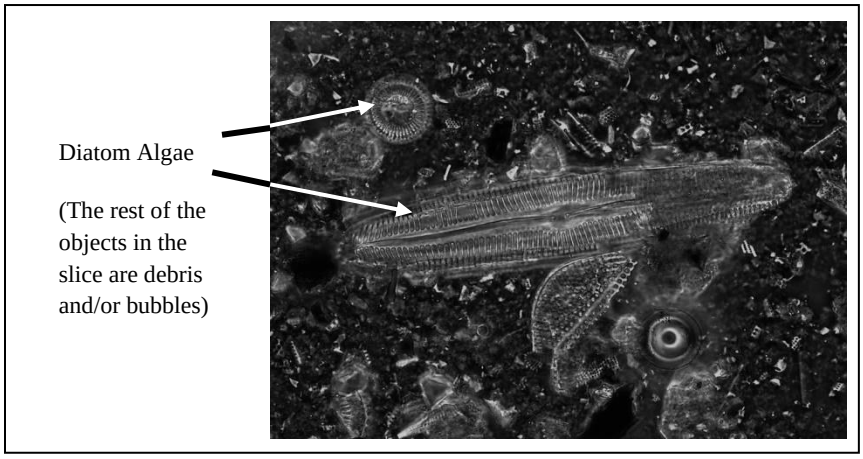

Figure 1. A slice of diatom algae fossil

We take the following three steps to achieve our goal.

1. Use Modified Rough Sets (Hashemi et al. 2008) to identify a set of patterns representing the relationships between linguistic values of the algae and climate values.
2. Extend the patterns to embrace the linguistic values of water salinity.
3. Build the membership function for the linguistic values of algae abundance and water salinity level (Zimmermann 2001) using the reported data in literature.
   Apply the process of defuzzification to derive a crisp value for the lake's water salinity level at the time associated with the age of the slice using the crisp values of algae indices along with the extended patterns (Biacino and Gerla 2002).

The details of each step are discussed in the following three sub-sections.

One may ask why the hybrid system of Modified Rough Sets and Fuzzy Logic is chosen to act as the methodology in this research effort. The answer lies on two facts: (i) The presence of the linguistic values in the fossil data and (ii) the expression of the possible tolerance level of salinity for algae, in literature, as a range that also differs from one citation to the next. The Modified Rough Sets approach easily lends itself to extraction of fuzzy patterns from data with linguistic values and the use of Fuzzy Logic is the best way, if not the only way, to extract the crisp salinity levels from the patterns.

3.1 Modified Rough Sets and Pattern Extraction

In a nutshell, member of a rough set (in contrast with a traditional set) is either *totally* inside the set or *partially* inside the set. A modified rough set is a version of rough set for which those members that are partially inside the set are moved to be totally inside the set with an *assigned confidence level (CL)* and a *dominant decision* that both are determined by a Bayesian model. For details consult Hashemi et al. (2008).

As a brief description of rough set and modified rough set, we start with a few essential definitions as follows.

Def. 1: An approximation space P is an ordered pair, P = (U, R) where, U is a set called the universe and R is a binary equivalence relation on U.

Def. 2: Let $A \subseteq U$ and let E^* be a family of equivalence classes of *R*. The set *A* is definable in *P* if for some $G \in E^*$, set *A* is equal to the union of all the sets in *G;* Otherwise, *A* is a *non-definable* or a *rough set*.

Def. 3: A rough set *A* is represented by its upper and lower approximations.

Upper(*A*) = $\{a \in U \mid [a]_R \cap A \notin \emptyset\}$ and

Lower(*A*) = $\{a \in U \mid [a]_R \subseteq A\}$, where, $[a]_R$ is the equivalence class of relation *R* containing a. The set *Ms*(*A*) = *Upper(A) – Lower (A)* is called a *boundary* of *A* in *P*.

To find the rough sets for a given dataset, the dataset is considered as an *information system* (a rough Sets terminology) and it is defined as follows:

Def. 4: An information system, S, is a quadruple (U, Q, V, d) Where,

U is a non-empty finite set of objects, u;

Q is a finite set of attributes, q;

$V = \cup_{q \in Q} V_q$, and V_q is the domain of attribute q.

d is a mapping function such that $d(a,q) \in Vq$ for every $q \in Q$ and $a \in U$.

As an example, Table 1 presents an information system. The information system is reduced vertically and horizontally. In vertical reduction process, all the duplicated records are collapsed into one. This reduced information system has an internal configuration which simply says all the records are unique considering the Q attributes. In horizontal reductions, those condition attributes that their removal does not change the configuration of the system are removed, Table 2.

Table 1. An information system with linguistic values of L: low, M: medium, and H: high

U	Q			
	Condition Attributes			Decision Attribute
	C_1	C_2	C_3	
a1	L	H	L	1
a2	M	L	M	2
a3	L	L	M	2
a4	M	M	H	1
a5	L	H	L	1
a6	M	M	H	2
a7	M	L	M	2
a8	M	L	M	1
a9	L	M	M	2

Table 2. Reduced information system of Table 1

U	Q		
	Condition Attributes		Decision Attribute
	C_1	C_2	
z1={a1, a5}	L	H	1
z2={a2, a7}	M	L	2
z3={a3}	L	L	2
z4={a4}	M	M	1
z5={a6}	M	M	2
z6={a8}	M	L	1
z7={a9}	L	M	2

The partition of the reduced information system using the decision values produce the *elementary sets* of $L_1 = \{z_1, z_4, z_6\}$ and $L_2 = \{z_2, z_3, z_5, z_7\}$ for D = 1 and D = 2, respectively. The elementary sets for the condition attributes in C are: $R_1 = \{z_1\}$, $R_2 = \{z_2, z_6\}$, $R_3 = \{z_3\}$, $R_4 = \{z_4, z_5\}$, and $R_5 = \{z_7\}$.

The lower approximation spaces for D = 1 and D =2 are: Lower $(L_1) = \{z_1\}$ and Lower $(L_2) = \{z_3, z_7\}$. The upper approximation spaces for D = 1 and D = 2 are: Upper $(L_1) = \{z_4, z_5, z_2, z_6\}$ and Upper $(L_2) = \{z_4, z_5, z_2, z_6\}$.

We extract patterns from the lower approximation spaces of the decisions in form of *if . . . Then* rules. These patterns are known as the *local certain rules* and they are:

If $C_1 = L \wedge C_2 = H$ *then* $D = 1$, *If* $C_1 = L \wedge C_2 = L$ *then* $D = 2$,
If $C_1 = L \wedge C_2 = M$ *then* $D = 2$.

We also extract patterns from the upper approximation spaces of the decisions and they are known as the *local possible rules*:

If $(C_1 = M \wedge C_2 = M)$ *then* $D = 1$, *If* $(C_2 = M \wedge C_2 = M)$ *then* $D = 2$,
If $(C_1 = M \wedge C_2 = L)$ *then* $D = 2$, *If* $(C_1 = M \wedge C_2 = L)$ *then* $D = 1$.

The set of objects in Upper(L_1) is the same of the set of objects in Upper (L_2). The Bayesian model suggests D = 2 as the *dominant decision* with the *confidence level* of 3/5 = 0.6. One can change the decisions for the objects in Upper (L_1) to the dominant decision by introducing a confidence level to the decision values of all objects in the information system. As a result, the previous local possible rules change into two new local certain rules with confidence level of 60%. Since the local possible rules no longer exist, we refer to local certain rules simply the *Local rules*. The list of all extracted local rules is shown in Table 3.a.

Basically, for each rough set the lower approximation space has been expanded to make sure all the members are totally inside the set (with assigned confidence levels) and the upper approximation space no longer exist. Enforcing the expansion of the lower approximation space and, therefore, generation of local rules constitutes Modified Rough Sets methodology.

Table 3. Extracted patterns using Modified Rough Sets: (a) Local rules and (b) Global rules.

Local Rules:	*Global Rules*
If (C1=L ^ C2=H) then D=1 (CL=100%)	If (C2=H) then D=1 (CL=100%)
If (C1=L ^ C2=L) then D=2 (CL=100%)	If (C1=L ^ C2=L) then D=2 (CL=100%)
If (C1=L ^ C2=M) then D=2 (CL=100%)	If (C1=L ^ C2=M) then D=2 (CL=100%)
If (C1=M^ C2=M) then D=2 (CL=60%)	If (C1=M^ C2=M) then D=2 (CL=60%)
If (C1=M^ C2=L) Then D=2 (CL=60%)	If (C1=M^ C2=L) Then D=2 (CL=60%)
(a)	**(b)**

The set of conditions in each local rule is unique throughout the entire set. The local rules can be turned into the *global rules* by minimizing the number of conditions in each local rule while preserving its uniqueness. The list of global rules produced from the local rules of Table 3 is shown in Table 3.b.

3.2 Extension of Patterns

Let us assume that an attribute of interest (AOI) exist such that a domain expert can generate linguistic values for it using a set of global rules. For example, let the water salinity level be an AOI for the dataset F. If decision D = 1 means “Drought” in a given global rule then, the

confidence level of 100% suggests the water salinity (the AOI) has a linguistic value of "high". Whereas the decision D = 2 (i.e. Rain) with the confidence level of 100% for the same rule suggests the linguistic value of "low" for the AOI. The same suggestions may be extended to the confidence levels of $CL \geq \alpha$ where, α is close to 100% and it is selected by the domain experts. A threshold value for confidence level, T_{CL}, is also selected by the domain experts and any global rule with the confidence level less than or equal to T_{CL} is dismissed.

Those confidence levels that fall in the interval of (T_{CL}, α) are categorized based on the number of selecting subintervals that is dictated by the desired granularity of the linguistic values. For example, if the desired granularity of the linguistic values are *coarse* (e.g. "Low", "Medium", and "High") then, the interval of (T_{CL}, α) may be divided into two subintervals (not necessarily of the same width) to categorize the confidence levels into "low" and "medium. If the desired granularity of the linguistic values are finer (e.g. "low", "medium-low", "medium", "medium-high", and "high") then, the interval of (T_{CL}, α) may be divided into four subintervals to categorize the confidence levels of every global rule. The number of subintervals and the width of each subinterval are decided by the domain experts.

We refer to the above mentioned process as the *extension* of patterns (the global rules) in reference to an AOI. The result of such extension for global rules of Table 3.b and for the AOI of the water salinity level is shown in Table 4.

Table 4. Extended patterns for global rules of Table 3.b and water salinity level(S) as an AOI

Patterns: (For α = 90, coarse linguistic values, and T_{CL} = 57%)	
P1: If (C2 = H) then S= H, P2: If (C1 = L ^ C2 = L) then S= L P3: If (C1 = L ^ C2 = M) then S= L	P4: If (C1 = M^ C2 = M) then S= M P5:If (C1 = M^ C2 = L) then S= M

3.3 Membership Functions and Defuzzification Process

Let P be the set of patterns in Table 4 in which the conditions C_1 and C_2 stand for two algae abundances—$Alga_1$ and $Alga_2$. The $Alga_1$, $Alga_2$, and salinity level have values of "low", "medium", and "high". Each fuzzy value has its own membership function. To define these functions, we have collected the possible tolerance level of salinity for algae from the literature (Hodgson et al. 2013, Borton 2005, Round et al. 1990, Melack et al. 2001) and two domain experts used the data to conclude the membership functions for the fuzzy values of algae indices and the salinity levels. The results are shown in Figure 2.

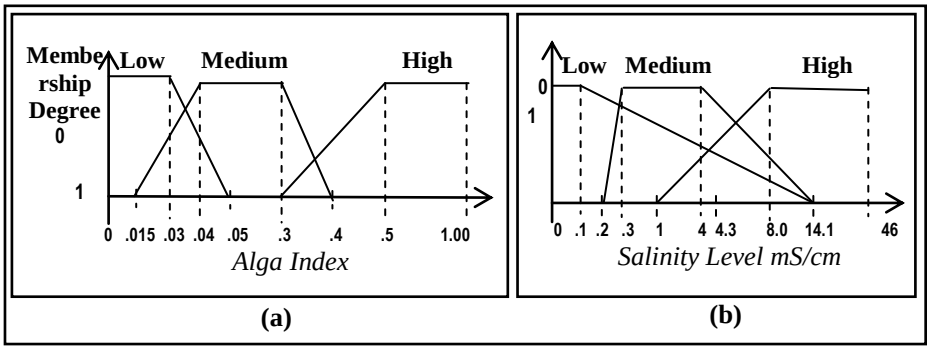

Figure 2. Membership functions for: (a) the alga index values and (b) the salinity levels

There are five patterns in Table 4 and each pattern has a set of conditions and a consequent which are expressed in fuzzy values. Calculation of a crisp value for the consequent given crisp values for the conditions is done through a defuzzification process. The following steps are involved in this process:

Step 1: Calculating degree of membership for conditions' crisp values.
Step 2: Inferring degree of membership for the consequent using P and the results delivered by step 1. The outcome is a set of membership degrees for the consequent (one membership degree per pattern).
Step 3: Grouping the set of membership degrees into m groups in reference to the m fuzzy values for the consequent and find the minimum value in each group as the composed value for the group.
Step 4: Determine the areas under the membership function curves for the fuzzy values of the consequent using the results obtained in step 3. Calculate the coordinates of the overall centroid of the areas. The x coordinate of the overall centroid is the crisp value for the consequent.

The following subsection covers the detailed explanation of each step.

3.3.1 Detailed Explanations

Let us assume that the abundance crisp values of algae for a record, R, are obtained in the lab, and they are: $Alga_1 = 0.03$ and $Alga_2 = 0.35$. The explanation of each step in reference to the crisp values of Aga_1 and $Alga_2$ are cited below.

Step 1: The crisp value of $Alga_1 = 0.03$ has membership degrees of 1, 0.5 and 0 in the "low", "medium", and "high", respectively, Figure 3.a. The degree of membership for $Alga_2 = 0.35$ in the three fuzzy values are 0, 0.4, and 0.2., Figure 3.a.

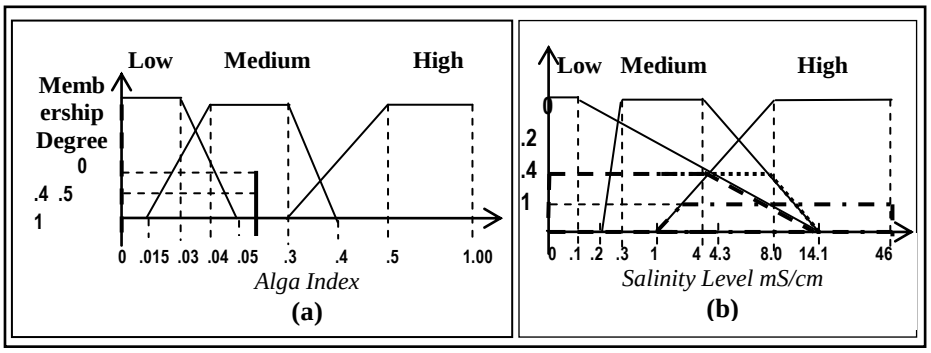

Figure 3. Degree of memberships for: (a) crisp values of Alga1 = 0.03 and Alga2 = 0.35 and (b) The boundary of the area under curves for the record R

Step 2: Let us first look at the pattern P_2: *If ($Alga_1 = L \wedge Alga_2 = L$) then S= L.*

The condition ($Alga_1$ = Low) is replaced by the membership degree of $Alga_1$ in the "low" fuzzy value (i.e. 1) and condition ($Alga_2$ = Low) is replaced by value of zero (UNC-Charlotte 2016). The value for the condition (S = Low) is the Min(1, 0) = 0. Following the same process, the inferred degrees of membership for the Salinity Level using the rest of the five patterns of Table 4 are:

P_1: High(S) = Min[High ($Alga_2$)] = 0.2
P_3: Low(S) = Min[Low($Alga_1$), Med($Alga_2$)] = 0.4
P_4: Med(S) = Min[Med($Alga_1$), Med($Alga_2$)] = 0.4
P_5: Med(S) = Min[Med($Alga_1$), Low($Alga_2$)] = 0

Step 3: Compose one crisp value out of the several crisp values produced for the membership degree of salinity level for a given fuzzy value by choosing the maximum crisp value among all the membership degree for the sanity level. As an example, for the salinity level of "low", P_2 delivers the crisp value of zero for Low(S) and P3 delivers the crisp value of 0.4 for the same fuzzy value of salinity. Thus, the composed crisp value for: Low(S) = Max(0, 0.4) = 0.4. Using the same analogy, the rest of the composed values are: Med(S) = Max(0.4, 0) = 0.4 and High(S) = Max(0.2) = 0.2.

Step 4: Find separately the area under the curves for the membership functions of the sanity level "Low" "Medium", and "High" for Low(S) = 0.4, Med(S) = 0.4 and High (S) = 0.2, respectively. The result is three under the curve areas in shape of trapezoids and the boundary of each area under curves, for the record R, is shown by different types of thick broken lines in Figure 3.b.

Finding the overall centroid for the identified areas is done by, first, finding the centroids of the three trapezoids separately and then, finding the centroid of the triangle that its vertices are the centroids of the three trapezoids. To find the centroid of a trapezoid, consider the trapezoid of Figure 4, that its larger parallel side is on X axis (Efunda 2016).

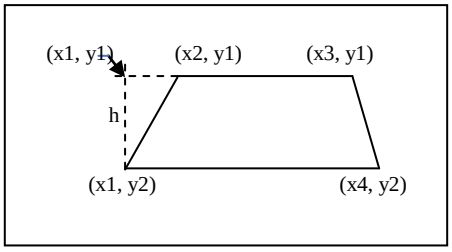

Figure 4. A typical under the curve area

The coordinates of the centroid for a trapezoid are:

$$x = \frac{2ab+b^2+ac+bc+c^2}{3(b+c)} + x1 \text{ and } y = \frac{h(2b+c)}{3(b+c)} \quad (1)$$

Where, a =(x_2-x_1), b=(x_3-x_2), c=(x_4-x_1), and h = y_1.

The coordinates of the centroid for a triangle are:

$$x = \frac{x1+x2+x3}{3} \; and \; y = \frac{y1+y2+y3}{3} \quad (2)$$

Where, x_i and y_i are the x and y coordinates of the i-th vertex of the triangle.

Using the above formulas deliver the x= 10.97 and y = 0.15 for the coordinates of the overall centroid. The value of x is the crisp value for the water salinity level of the given record.

4. EMPIRICAL RESULTS

The given historical dataset is based on the data collected by Hodgson et al. (2013) and it is representative of the diatom algae fossils and climate change for nearly 2000 years in the Santa Fe Lake region, New Mexico, USA. The dataset has 75 records and 84 attributes. One attribute represents the *climate* as the dependent variable with possible values of "Drought" and "Rain". The other 83 attributes represent the abundance of 83 species of algae (algae indices) as independent variables expressed in linguistic values of "Low", "Medium" and "High". Each record represents a slice of the retrieved Diatom fossils (one cm in height) from the bottom of the lake.

A pre-processing step was completed to identify and remove the redundant independent variables using the entropy approach (Natt et al. 2012). The final cleaned dataset had 75 records and seven attributes: climate and six non-redundant algae species of *Achnanthes lanceolata var. lanceolata*, *Cyclotella meneghiniana*, *Fragilaria construens var. venter*, *Fragilaria pinnata var. pinnata*, *Navicula laevissima, and Pinnularia abaujensis var. subundulata*. (The membership functions of Figure 2 are derived only for the non-redundant algae).

Our hybrid methodology of Modified Rough Sets and Fuzzy Logic was applied on the dataset and salinity level of each slice of the diatom fossils was calculated. The calculation was completed under the following parameters: TCL = 67%, α = 90%, β = 74%, and desired level of granularity = coarse. It is crucial to determine the validity of the calculated historical salinity results. The process for measuring the validity of the results is the subject of the following subsection.

4.1 Validity Measurement

To measure the validity of results produced by the proposed hybrid system, we have generated seven pairs of training and test sets out of the fossil dataset (in linguistic values) such that the records of one test set do not appear in any of the other test sets and each test set contains ten records (roughly 13% of the total records) that are chosen randomly. The crisp values of the algae for the test records along with the climate value of "drought/rain" were borrowed from the data reported in (Hodgson et al. 2013).

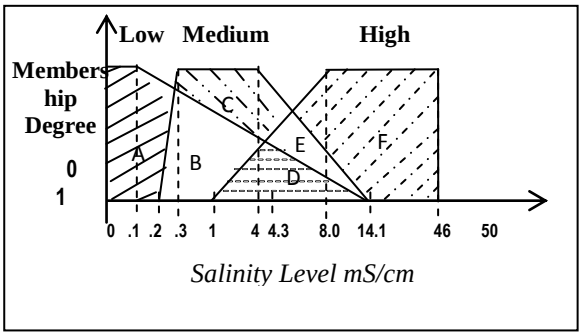

Figure 5. The regions of interest in the membership functions of salinity

The following two steps were applied separately on every training set and the salinity level for each record of the corresponding test set was produced:

1. The proposed hybrid system was applied on the training set to get the global rules and identify the extended patterns in reference to water salinity level.
2. For every record in the test set, the crisp values of the conditions were used to derive a crisp salinity level for the record using the membership functions and defuzzification process.

The produced salinity level for each test record suggests a climate for the test record based on Figure 5. There are six specific regions of interest under the curves of the salinity membership functions of Figure 5, named A, B, C, D, E, and F. If the overall centroid appears in areas of A, B, C, D, E, and F, the suggested climates are "rain, "rain", "unknown", "rain", "drought", and "drought", respectively. For our example, the overall centroid is located in area F which means climate = "drought".

The suggested climate value may be in agreement or disagreement with the actual climate value of the record. In either case the results are recorded. The percentage of the agreement between the suggested climate by the salinity level and the actual climate value for each test set is shown in Table 5 which suggests that the mined historical salinity levels have a high average degree of validity—72%.

Table 5. The results of validity checking of the calculated salinity

	Test Sets							
	1	2	3	4	5	6	7	Average
Validity (%)	60	70	80	82	70	80	60	72

5. CONCLUSIONS AND FUTURE RESEARCH

The data used in this research represented the diatom algae fossils and climate change for nearly 2000 years in an alpine region of New Mexico, USA. The drought/rain cycles of the past for the region were established by Hodgson et al. (2013). During a drought cycle water was evaporated in a larger scale causing increase in the salinity of the water, which in turn eliminated some type of algae while some other types flourished. During a rain cycle, the water level increased causing a decrease in the salinity level of the water, which in turn those algae flourished with lower levels of tolerance for the saltier water. Therefore, it seems logical to use the diatom algae fossils to discover the salinity level of water in which the algae lived. In addition, the expression of the fossil data in linguistic values and disagreement about the reported range of tolerance level of salinity for different algae in literature suggest the use of Modified Rough Sets and Fuzzy Logic as the best way (if not the only way) to extract the historical water salinity levels.

The results revealed that our methodology has a great potential in mining of the historical salinity level for a body of water using diatom algae fossils. Discovery of the historical salinity of inland lakes is extremely important to understanding climate change, carbon dioxide levels, and global warming.

Investigation of the relationships among the three entities of historical salinity, solar intensity, and Human contribution to the global warming is in progress as the future research.

REFERENCES

Adkins, J.F. et al, 2002. The salinity, temperature, and delta18O of the glacial deep ocean. *Science* 298:1769-73.

Barry, R., and R. Chorley. 1999. Atmosphere, weather, and climate. 8th Ed. Routledge Publishing, New York, New York, USA.

Biacino, L., Gerla, G., 2002. Fuzzy logic, continuity and effectiveness. *Archive for Mathematical Logic*, Vol. 41 No.7, pp. 643–667, 2002.

Borton, S.A., (Ed.), 2005. *Estuarine Indicators*, CRC Press, Boca Raton, Florida, USA

Battarbee, R.W. 1986. Diatom analysis. Pages 527-570 *in* B.E. Berglund, editor. *Handbook of Holocene palaeoecology and palaeohydrology*. The Blackburn Press, Caldwell, New Jersey, USA.

Cole, J.J. et al, 1994. Carbon-dioxide supersaturation in the surface waters of lakes. *Science* 265:1568-1570.

Cook, et. al. 2004. Long-term aridity in the western United States. *Science* 306:1015-1018.

Durack, P. J. and Wijffels, S. E., 2010. Fifty-Year Trends in Global Ocean Salinities and Their Relationship to Broad-Scale Warming, *Climate*, DOI: http://dx.doi.org/10.1175/ 2010JCL13377.1

Durack, P. J. et al, 2012. Ocean Salinities Reveal Strong Global Water Cycle Intensification During 1950 to 2000. *Science* 336: 455-458.

Efunda, http://www.efunda.com/math/areas/trapezoid.cfm, cited in April 2016.

Emiliani, C., 1955. Pleistocene temperatures. *Journal of Geology* 63:538-578.

Grissino-Mayer, H.D. 1996. A 2129-year reconstruction of precipitation for northwestern New Mexico, USA. Pages 191-204 *in* J.S. Dean, D.M. Meko, and T.W. Swetnam, editors. Tree rings, environment, and humanity. Radiocarbon 1996, Department of Geosciences, University of Arizona, Tucson, Arizona, USA.

Hashemi, R. R. et al, 2008. Use of Rough Sets as a Data Mining Tool for Experimental Bio-Data, *A book chapter in: "Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Applications"*, Springer-Verlag Publisher, pp. 69-91.

Henderson, G.M. 2002. New oceanic proxies for paleoclimate. *Earth and Planetary Science Letters* 203: 1-13.

Hodgson, J.Y.S. et al., 2013. An independently corroborated diatom-inferred record of long-term drought cycles occurring over the last two millennia in New Mexico USA. *Inland Waters* 3: 459-472

Jansson, M. et al, 2012. Carbon dioxide supersaturation promotes primary production in lakes. *Ecology Letters* DOI:10.1111/j.1461-0248.2012.01762.x.

Jones, et. al, 2001. The evolution of climate over the last millennium. *Science* 292:662-667.

Laird, K.R. et al, 1998. A diatom-based reconstruction of drought intensity, duration, and frequency from Moon Lake, North Dakota: a sub-decadal record of the last 2300 years. *Journal of Paleolimnology*. Vol. 19, pp. 161-179.

Mann, M.E. 2002. The value of multiple proxies. Science 297:1481-1482.

Melack, J.M. et.al (Eds), 2001. Saline Lakes, *Proceedings of the 7the International Conference on Salt Lakes*, Volume 466, 2001.

Mendelsohn, R. et al, 1994. The Impact of Global Warming on Agriculture: A Ricardian Analysis, *The American Economic Review*, Vol. 84, No. 4, pp. 753-771.

Natt, J., et al, 2012, Predicting Future Climate Using Algae Sedimentation", *Proceedings of the 9th International Conference on Information Technology: New Generation*. Las Vegas, Nevada, USA, pp. 560-565.

Round, F.E. et al, 1990. *The Diatoms: Biology & Morphology of the Genera.* Cambridge University Press, Cambridge, UK.

Sigman, D.M., and E.A Boyle, 2000. Glacial/interglacial variations in atmospheric carbon dioxide. *Nature* No. 407:pp. 859-859.

Smol, J. P. and Cumming, B. F., 2000. Tracking long term changes in climate using algal indicators in lake sediments. *Journal of Phycology*. Vol. 36, No. 6, pp 986-1011.

UNC-Charlotte, cited on March 2016. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, coitweb.uncc.edu/~ras/courses/Fuzzy-Sets.ppt, cited on March 2016.

Verschuren, D. et al, 2000. Rainfall and drought in equatorial east Africa during the past 1100 years. *Nature* 403, pp. 410-414.

Von Storch et. al, 2004. Reconstructing past climate from noisy data. *Science* 306:679-682.

Wilson, S.E., et al. 1994. Diatom-salinity relationships in 111 lakes from the Interior Plateau of British Columbia, Canada: the development of diatom-based models for paleosalinity reconstructions. *Journal of Paleolimnology* 12: 197-221.

Zimmermann, H. 2001, *Fuzzy set theory and its applications*. Kluwer Academic Publishers, Boston, USA.