Applying Macro-Grammatical Evolution to Salinity Estimation Using MODIS Data on Taiwan Strait

Li Chen[1] Basmah Alabbadi[1]

[1] Department of Civil Engineering, Chung Hua University

Hsinchu 707, Taiwan, ROC

lichen@chu.edu.tw

basmaalabadi@yahoo.com

Received 22 December 2014; Revised 17 February 2015; Accepted 13 March 2015

Abstract. Strait water quality is traditionally monitored and estimated based on in-situ data. Collecting and analyzing in-situ water quality data are expensive, time consuming and with large parts of the water body never sampled. In this study we utilize MODIS data to estimate the water quality of Taiwan Strait, and propose a nonlinear model which incorporates improved real-coded grammatical evolution (GE) with a genetic algorithm (GA). The GE, an evolutionary automatic programming type system, automatically discovers complex nonlinear mathematical relationships among observed salinity concentrations and remote sensed imageries. The algorithm discovers significant input variables and combines them to form mathematical equations automatically. Utilizing GA with GE optimizes an appropriate type of function and its associated coefficients. To enhance searching efficiency and genetic diversity during GA optimization, the macro-evolutionary algorithm (MA) is processed as a selection operator. The results of this study indicate that the proposed GEMA yields an efficient optimal solution. GEMA has the advantages of its ability to learn relationships hidden in data and express them automatically in a mathematical manner. Compared with linear regression (LR_1), LN transform of linear regression (LR_2), and back-propagation neural network (BPN), the performance of GEMA was found better than LR_1, LR_2 and BPN.

Keywords: water quality, Taiwan Strait, MODIS, grammatical evolution, macro-evolutionary algorithm

1 Introduction

Salinity refers to the fresh water inputs from various resources and processes such as precipitation and river runoff, is used as an indicator of the sea and ocean density. It is one of the major factors affecting ocean water quality, which can produce changes in numerous physical and biochemical processes [1]. In addition, salinity has important role in circulation patterns and affects the distribution of several marine organisms [2]. The instability in ocean salinity may appear as a result of increased evaporation (resulted from increased temperature) and changes in ocean circulation or induced by climate change [3]. In the ocean researches, several scientists and researchers have employed in situ salinity data from buoys or commercial ships. These data can be used to quantify temporal changes in sea surface salinity at specific points However, they remain sparse, irregular, expensive, time-consuming, and large parts of the global oceans have yet to be sampled. In recent years, satellite remote sensing provides the potential of estimating sea surface salinity because of its advantages of large spatial coverage within a short time [4, 5]. Satellite remote sensing provides the potential of estimating sea surface salinity across entire water bodies at the frequency of satellite overpass [2]. Effective efforts to estimate sea surface salinity by applying remote sensing have included Landsat Thematic Mapper (TM) data [6], Landsat Multispectral Scanner (MSS) [7] and electronically scanned thinned array radiometer (ESTAR) [8].

Water quality assessment applying satellite remote sensing data has been performed since the first remote sensing satellite, the Landsat Multispectral Scanner (MSS) became operational [7, 9]. Distinct remote sensing data, such as data obtained using the Landsat TM [7], electronically scanned thinned array radiometer (ESTAR) [8], and microwave sensors [10] have been used for water quality assessment including salinity. The most commonly used data are obtained using the Landsat Thematic Mapper, a multispectral imaging sensor. This system supplies a highly continuous dataset of high-spatial-resolution images of global land and water surfaces and enables the synoptic monitoring of water quality problems. However, using this system to obtain quantitative results is difficult [11]. The Landsat TM sensor is calibrated for land use; thus, its signal-to-noise ratio for a low-reflectance seawater surface is unsuitable for obtaining substantial data [12]. Recently, the Moderate Resolution Imaging Spectroradiometer (MODIS) has been recognized as major challenge for sea surface water quality assessment including salinity. Notably, Moderate Resolution Imaging Spectroradiometer (MODIS) sensors aboard

the Terra and Aqua satellites are multispectral sensors with several wavebands designed for monitoring the Earth's environment, including atmosphere, land, and ocean. Their data have been used for estimating water quality assessment including salinity. For example, Hu et al. [13] used MODIS data for estimating the water quality and proposed that the colored dissolved organic matter concentration (CDOM) is the only constituent with a linear and inverse relationship with sea surface salinity. Barbini et al. [14] used the light detection and ranging (LIDAR) fluorosensor for estimation chlorophyll-a concentration in transects between New Zealand and Italy and obtained a favourable agreement between MODIS and SeaWiFS datasets. Wong et al. [12] used Aqua/MODIS data for estimating suspended solids and salinity in marine Hong Kong, where monitoring stations are few, and determined significant correlations between MODIS data and in situ data. Application of Terra/MODIS with 500 m images could be considerably useful because it frequently achieves high accuracies [12] and might help capture the variation in sea surface salinity of Taiwan Strait. Another valuable feature of Terra/MODIS is free archive access.

In recent years research efforts have been focused on optimization and predicting technique for solving large-scale problems of various research fields [15, 16, 17, 18, 19]. For example in oceans research, statistical applications are traditionally used to establish algorithms for predicting various water quality variables. Regression analysis is widely used in formulating predictive models [20, 12]. However, nonlinear transfer functions are frequently observed when relating water quality variables to satellite imagery data [21]. In addition, the linear regression is too simple, it may generate inaccurate results. Generally, sophisticated regression models must go through time-consuming trial and error procedures so that the correct regression type can be obtained. Khorram [22] used satellite remote sensing data for estimating sea surface salinity and developed a multiple linear relationship between Landsat MSS bands and sea surface salinity. Xie, Zhang, and Berry [23] used Landsat TM data for sea surface salinity monitoring in Florida Bay by applying the geographically weighted regression (GWR) approach. Wong et al. [12] have developed multilinear retrieval algorithms to estimate sea surface salinity based on data of MODIS sensor. In recent studies, Urquhart et al. [2], Alabbadi et al. [24] and Geiger et al. [25] have used MODIS sensor data and applied different statistical methods to predict sea surface salinity. Urquhart et al. [2] developed eight statistical methods for predicting sea surface salinity in the Chesapeake Bay. Alabbadi et al. [24] used genetic algorithm combining operation tree to estimate sea surface salinity in Taiwan Strait. Geiger et al. [25] used neural network models to predict sea surface salinity in the Atlantic coastal. Therefore, sea surface salinity can be expressed as a function of remote sensing reflectance. Qing et al. [4] developed a simple multilinear regression model for sea surface salinity by using in situ measurements and medium-resolution imaging spectrometer (MERIS) visible band remote sensing reflectance along with sea surface salinity data in the Bohai Sea.

Evolutionary algorithms, such as genetic programming, have been used with much success for the automatic generation of programs or equations between the inputs and outputs. It has an advantage over traditional statistical methods because it is distribution free, i.e., no prior knowledge is needed about the statistical distribution of the data like the back-propagation network (BPN) [26] and its abilities to learn relationships hidden in data and expresses them automatically in a mathematical manner [27]. Nevertheless, it is well known that the BPN is considered as a nonlinear black-box model, and it is not unusual for it to be criticized as not enhancing our understanding of the physical mechanisms because of its complex weighting coefficients and numerous other parameters.

Chen [28] pointed out that constructing a tree-type data structure genetic programming is a difficult task for computer programming, because it is hard to choose the suitable size of a tree that can express a meaningful equation in advance. Recently, the newly developed grammatical evolution (GE) technique is a biologically plausible approach that performs evolutionary processes on a variable-length binary string. This new data structure is flexible and allows researchers to exploit the benefits of genetic algorithms (GAs) [29]. A mapping process generates programs in any language using the binary strings to select production rules in a Backus–Naur form (BNF) grammar definition [30]. The result constructs a syntactically correct program or equation from a binary string, which can then be evaluated by a fitness function [30].

This paper is intended to improve the monitoring techniques of using remote sensing data to estimate strait salinity. Because of the complex nonlinear relationship between sensor bands and salinity concentration in a strait, a new system identified method called GEMA is used for salinity for the first time.

2 Grammatical Evolution

Grammar evolution (GE) [31] has been applied to all manner of automatic programming problems, from symbolic regression, to C programs, or generation of graphical objects. The common view of GE is that, given a particular problem statement, a program that satisfied the fitness function is to be generated. Grammatical evolution (GE) is an evolutionary automatic programming type system that combines of a variable length binary string genome and a Backus-Naur Form (BNF) grammar to evolve interesting structures. It presents a unique

method which exploits grammars in the process of automatic programming. Harrison [32] defined the grammar as a process to produce strings sets. Variable-length binary string genomes are used with several codons representing integer values where codons are consecutive8-bit groups. The integer values are used in a mapping function to select an appropriate production rule from the BNF definition; the numbers generated always representing one of the rules that can be used at that time [29].

2.1 Backus-Naur Form

BNF is a notation for expressing the grammar of a language in the form of production rules [33], was originally developed by Niklaus Wirth [34]. BNF grammars consist of terminals, which are items that can appear in the language, e.g., +, -, etc., and nonterminals, which can be expanded into one or more terminals and nonterminals. A grammar can be represented by the tuple {N, T, P, S}, which N is the set of nonterminals, T is the set of terminals, P is a set of production rules mapping the elements from N to T, and S is a start symbol that is a member of N. When there are a number of productions that can be applied to one particular N, the choice is delimited with the '|' symbol.
Below is an example BNF, where
N = { expr, op, pre_op }
T = { Sin, Cos, +, -, *, /, Variable , Constant }
S = <expr>
And P can be represented as

(1) <expr> :: = <expr><op><expr> …….rule 0
| (<expr><op><expr>)……....rule 1
|<pre-op> (<expr>)……...rule 2
| <var>……..…rule 3
(2) <op> :: = + ……...rule 0
| - ……..…….rule 1
| / ……..…….rule 2
|* ……..…….rule 3
(3) <pre-op> :: = Sin ……..... rule 0
| Cos ……..……rule 1
| Log ……..…….... rule 2
(4) <var> :: = X ……..…….rule 0
| 1.0 ……..…….rule 1

2.2 Mapping Process

The genotype is used to map the start symbol onto terminals by reading codons of 8 bits to generate a corresponding integer value from which an appropriate production rule is selected by using the following mapping function:

$$\text{Rule} = (\text{codon integer value}) \text{ MOD } (\text{number of rules for the current nonterminal}) \quad (1)$$

Considering the following rules, i.e., giving the nonterminal op, there are four production rules to be selected from:

(2) <op> :: = + ……..…….rule 0
| - ……rule 1
| / ……rule 2
|* ……..…….... rule 3

If we assume that the codon which being read produces the integer 6, then

6 MOD 4 = 2

selects <op> as rule 2: /. Each time a production rule has to be selected to map from a nonterminal, another codon is read. In this way, the system traverses the genome. An equation Sin(X)*Cos(X) +1.0, including two pre_ops; Sin and Cos, two ops: * and +, one Variable: X and one Constant: 1.0, is as a simple example. Fourteen 8-bit binary codons in a string can represent this equation using the BNF defined above. Each 8-bit binary codon

in the GE represents 256 distinct integer values. The following describes the decoding process and summarized in (Table 1) [35].

First, concentrating on the start symbol <expr>, we can see that there are four productions to choose. To make a choice, we read the first codon from the chromosome "11001000" and use it to generate a number "200". Because the standard decode of the binary 11001000 is

$1\times 2^7 + 1\times 2^6 + 0\times 2^5 + 0\times 2^4 + 1\times 2^3 + 0\times 2^2 + 0\times 2^1 + 0\times 2^0$, which equals to 200. This value will then decide which production rule to be used according to Equation (1) in BNF. Thus, we have 200 MOD 4 = 0, meaning we must take the zeroth production, rule (0), so that <expr> is now replaced with

<expr><op><expr>.

Second, continuing with the first <expr>, i.e., always starting from the leftmost nonterminal, a similar choice must be made by reading the next codon value 160 and again using the given formula we get 160 MOD 4 = 0， i.e., rule 0. The leftmost <expr> will now be replaced with <expr><op><expr> to give

<expr><op><expr><op><expr>.

Third, Again, we have the same choice for the first <expr> by reading the next codon value 206, the result being the application of rule 2 to give

<pre-op>(<expr>)<op><expr><op><expr>.

Fourth, now the leftmost <pre-op> will be determined by the codon value 96 that gives us rule 0, which is <pre-op> becomes Sin. We have the following:

Sin(<expr>)<op><expr><op><expr>

Steps fifth to thirteenth are shown in (Table 1)

Step fourteenth mapping continues until eventually we are left with the following expression:

Sin(X)*Cos(X) +1.0

Table 1. Example of each codon converted into corresponding BNF grammar

No.	8-bit binary codon	Mapping function	BNF grammars
1	11001000	200 MOD 4 = 0 from <expr>	<expr><op><expr>
2	10100000	160 MOD 4 = 0 from <expr>	<expr><op><expr><op><expr>
3	11001110	206 MOD 4 = 2 from <expr>	<pre-op>(<expr>)<op><expr><op><expr>
4	01100000	96 MOD 3 = 0 from <pre-op>	Sin(<expr>)<op><expr><op><expr>
5	00011011	27 MOD 4 = 3 from <expr>	Sin(<var>)<op><expr><op><expr>
6	01001000	72 MOD 2 = 0 from <var>	Sin(X)<op><expr><op><expr>
7	01101011	107 MOD 4 = 3 from <op>	Sin(X)*<expr><op><expr>
8	00111110	62 MOD 4 = 2 from <expr>	Sin(X)*<pre-op>(<expr>)<op><expr>
9	00010110	22 MOD 3 = 1 from <pre-op>	Sin(X)*Cos(<expr>)<op><expr>
10	00110111	55 MOD 4 = 3 from <expr>	Sin(X)*Cos(<var>)<op><expr>
11	01011000	88 MOD 2 = 0 from <pre-op>	Sin(X)*Cos(X)<op><expr>
12	01100100	100 MOD 4 = 0 from <op>	Sin(X)*Cos(X)+<ex pr>
13	11001011	203 MOD 4 = 3 from <expr>	Sin(X)*Cos(X)+<var>
14	00101001	41 MOD 2 = 1 from <var>	Sin(X)*Cos(X)+1.0

Notice that if any extra codons exist, they shall be ignored during the genotype-to-phenotype mapping process. It is possible for individuals to run out of codons and, in this case, we wrap the individual and reuse the codons. This technique of wrapping the individual draws inspiration from the gene-overlapping phenomenon, which has been observed in many organisms [29]. It is possible that an incomplete mapping could occur even after several wrapping events, and in this case the individual in our question gives the lowest fitness value [36].

2.3 Improving GE: Real Coding Representations

Since there is a problem that only integers can be presented by the binary coding scheme mentioned above, we revised it as a real-coded representation. The real numbers which imply that, each chromosome is a real-valued vector, as opposed to binary-coded GA, where chromosomes are 0-1 vectors. It is very useful and efficient to generate the real-number constants and coefficients shown in these output equations. When a codon is decoded as a constant, the real value of real-coded genome can be generated directly. It is the main difference between the improved real coded GE and original binary coded GE. Whereas mapping a codon to the BNF rule, just need to round it off as a non-negative integer within the range between 0~255 (8 bits) then choose one corresponding BNF rule. The same procedure was used to mapping a codon into a corresponding BNF rule via equation (1) as binary coded GE does described in section 2.2. This revision makes it easy for the GE to combine with a real-coded GA and it is described as follows.

3 GE Combined with the Macro-Evolutionary Algorithm

3.1 Genetic Algorithm

The genetic algorithm (GA), originated in the mid-1970s [37], is an iterative procedure, which includes a population of individuals that are candidate solutions to specific domain. During each generation, the individuals in the current population are related to their effective evaluations, and a new population of candidate solutions is formed by specific genetic operators like reproduction, crossover, and mutation. These steps are repeated until the convergence criterion is satisfied or a predetermined number of generations are achieved. A macro-evolutionary algorithm (MA) is presented as a selection scheme [28] which is introduced as follows. Blend crossover (BLX-α) uniformly picks values that lie between two points contain the two parents, but may extend equally on either side determined by a user specified GA-parameter α [38].

The use of MA improves the capability of searching global optimum solutions and avoids premature convergence because the genetic diversity can be maintained. The model exploits the presence of links between species that represent candidate solutions to the optimization problem. Because of the connection matrix, the whole population is able to obtain a rather accurate map of the relative importance of the solutions being explored in the landscape [39].

3.2 Algorithms of MA

For connection matrix, each individual gathers information about the rest of the population through the strength and sign of its couplings $W_{i,j}$ as

$$W_{i,j} = \frac{f(p_i) - f(p_j)}{dis(p_i, p_j)} \tag{2}$$

Where p_i = are the input parameters of the ith individual, and $dis(p_i, p_j)$ means the Euclidean distance between p_i and p_j.

The selection operator allows calculating the surviving individuals through their relations, i.e., as a sum of penalties and benefits. The state of a given individual S_i will be given by

$$S_i(t+1) = \begin{cases} 1, if \sum_{j=1}^{N} W_{i,j}(t) \geq 0, \ alive \\ 0, otherwise, \quad extinct \end{cases} \tag{3}$$

Where t is generation number and $W_{ij} = W(p_i, p_j)$ is calculated according to (2).

3.3 GEMA

Fig. 1 shows a combination of the GE and MA, called the GEMA, which is able to generate the optimal relationship among inputs and outputs automatically. First, a GE was employed to transfer the real-coded string to mathematical function mapping the input onto output. Several ahead steps of inflow data were implemented as inputs in the GE to forecast current inflow. Furthermore, a real-coded GA including blend crossover and uniform mutation was incorporated with the GE in order to optimize objective value of the functions. Blend crossover (BLX-α) uniformly picks values that lie between two points that contain the two parents, but may extend equally on either side determined by a user specified GA-parameter α. Recently, the non-uniform mutation is usually used to produce offspring for the real-coded GA. The GA was regarded as a search strategy to determine the most proper relationship among the salinity data. Moreover, MA was applied to improve the searching efficiency and prevent the premature convergence during the period of the optimization. The basic algorithm begins to choose an initial population randomly. Then it continuously runs from one generation to the next.

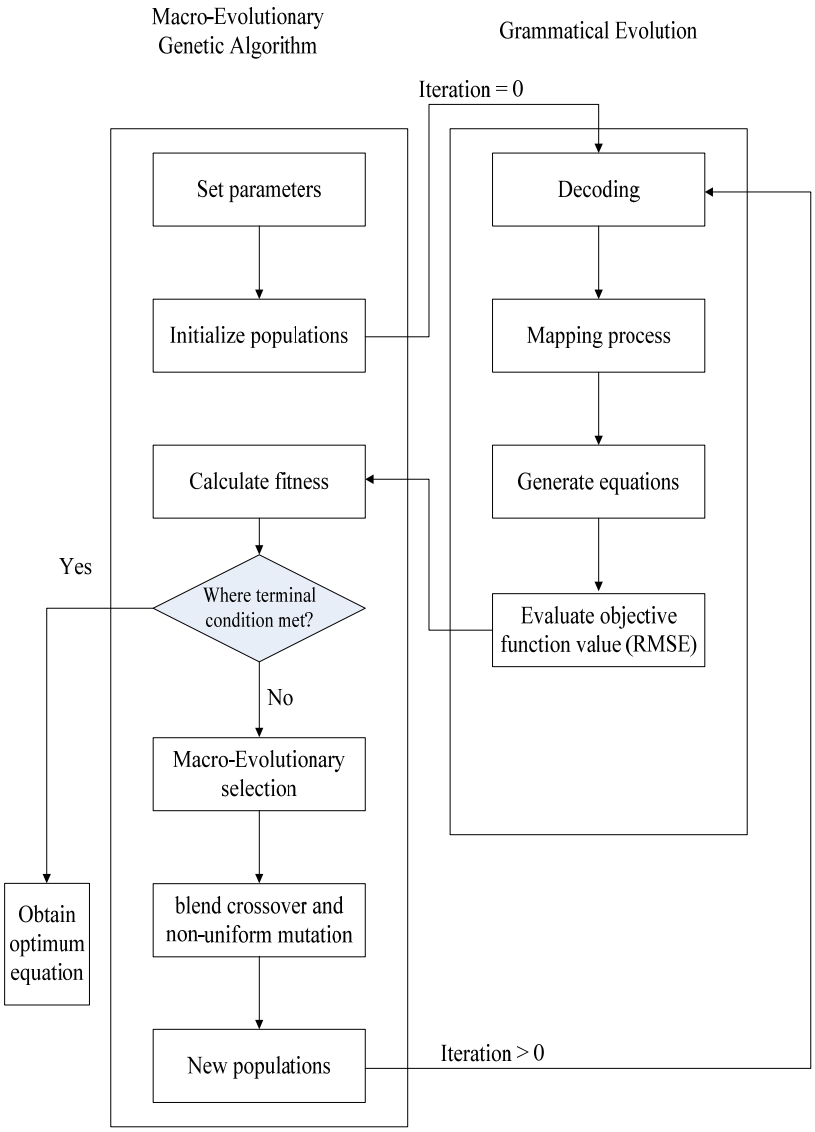

Fig. 1. The flowchart of GE combined with MA

These steps are repeated until the terminal condition is satisfied; an optimal equation which is capable of minimizing the objective function (root mean square error, RMSE) will be obtained.

4 Salinity Estimation Using GEMA

The purpose of this application was to utilize MODIS sensor data to evaluate sea surface salinity concentration in a strait. Whenever the relationship was made, the sea surface salinity concentration in strait may be computed in time. Whereas the relationship between salinity concentration in straits and corresponding image data was constructed through the GEMA. This system identification problem may be viewed as a search for a function type, which maps input values of MODIS onto an output value of sea surface salinity. The correlation coefficient, CC, value and root mean squared errors (RMSEs) are used as the criteria in this study.

4.1 Study Area Taiwan Strait

The Taiwan Strait is a shallow passage, about 350 km long, 180 km wide and 60 m -deep, between the China mainland and the Taiwan Island connecting the South China Sea to the East China Sea in the western North Pacific and its orientation is approximately southwest to northeast as shown in Fig. 2. The strait lies in monsoonal regions: Southwest monsoon (June-August), Northeast monsoon (December- February), fall inter-monsoon (September-November) and spring inter-monsoon (March-May) [40]. The three major water masses identified in the Taiwan Strait are the Kuroshio Branch Water with high temperature and high salinity, the China Coastal Water with low temperature and low salinity, and the South China Sea Water with intermediate temperature and salinity [41, 42]. This variation in salinity affects the temporal and spatial variations of chemical and biological factors.

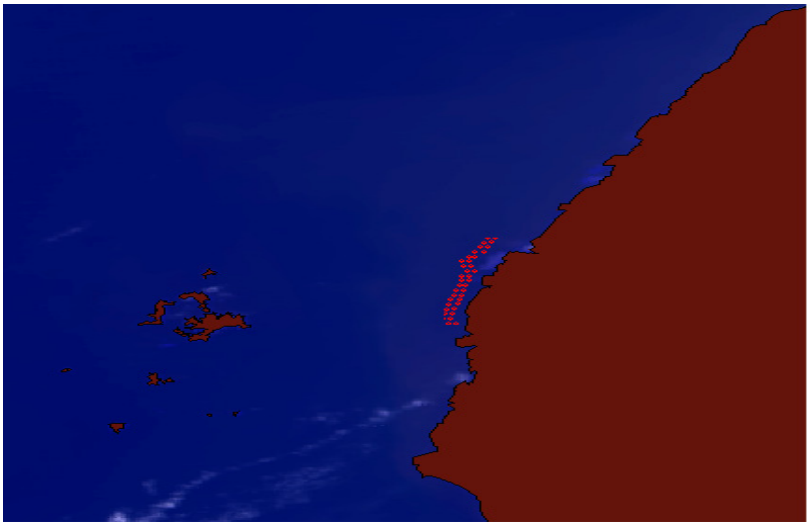

Fig. 2. Taiwan Strait with sampling locations

4.2 Salinity Data Set

In this study, the actual data of sea surface salinity for 13/8/2009, 13/8/2010 and 13/8/2011 as shown in Fig. 2, are used and it is expected to have a high correlation with MODIS reflectance data. These data are used to validate the salinity distributions which derived from MODIS data. Actual data are selected at same time of MODIS/Terra Satellite data overpass.

The MODIS images were acquired from the Level 1 and Atmosphere Archive and Distribution System (LAADS Web) for 13/8/2009, 13/8/2010, and 13/8/2011. The level 1 can be obtained every day for all earth parts. It consist: level 1A scans raw radiance measurements, level 1 geolocation, level 1B calibrated radiance (MODIS at 250m, 500m. and 1 km resolution), atmospheric profiles and cloud mask. MODIS 500m resolution data with 7 bands covering the spectral range 459-2155 nm, was selected in this study.

Geometric corrections of the MODIS images data were performed in order to compare the images data with salinity monitoring locations. The geometric correction was applied by using the “Georeferenced MODIS” function in ENVI 4.5. Then data band for each image was obtained by using ERDAS Imagine (2010). Seventy seven entries are used as training data and twenty five as predictive data; the total number of data entries is 102.

4.3 Estimation the Salinity of Taiwan Strait

Using LR_1 and LR_2. To estimate the spatial variation of salinity in the Taiwan Strait using remotely sensed images, empirical relationship between digital numbers of the pre-processed image bands and salinity was established using the linear regression (LR_1) and LN transform of linear regression (LR_2) methods initially. These models, $LR_1$and LR_2, utilized MODIS bands 1 to 7 were given by the following equations, respectively:

$$\text{Salinity}_{LR1} = -0.00022X_1 + 0.00015X_2 + 0.0045X_3 - 0.0040X_4 - 0.000031X_5 + 0.000055X_6 + 0.00072X_7 + 23.40 \quad (4)$$

$$\text{LN (salinity)}_{LR2} = -0.000029X_1 + 0.000012X_2 + 0.00038X_3 - 0.00032X_4 - 0.0000047X_5 + 0.0000064X_6 + 0.000027X_7 + 2.73 \quad (5)$$

where X_1= band$_1$, X_2 = band$_2$, X_3= band$_3$, X_4= band$_4$, X_5= band$_5$, X_6 = band$_6$ and X_7= band$_7$. In equation (4), the weight of X_1 (-0.00022) is similar with those of X_2 (0.00015) and X_7 (0.00072), which are all less than those of X_3 (0.0045), X_4 (-0.0040), and much higher than the weight of X_5(-0.000031) and X_6 (0.000055). In equation (5), the weight of X_1 (-0.000029) is similar with those of X_2 (0.000012) and X_7 (0.000027), which are all less than those of X_3 (0.00038), X_4 (-0.00032), and much higher than the weight of X_5 (-0.0000047) and X_6 (0.0000064). The two models were applied on MODIS image and the results show on Fig. 3 and 4, respectively.

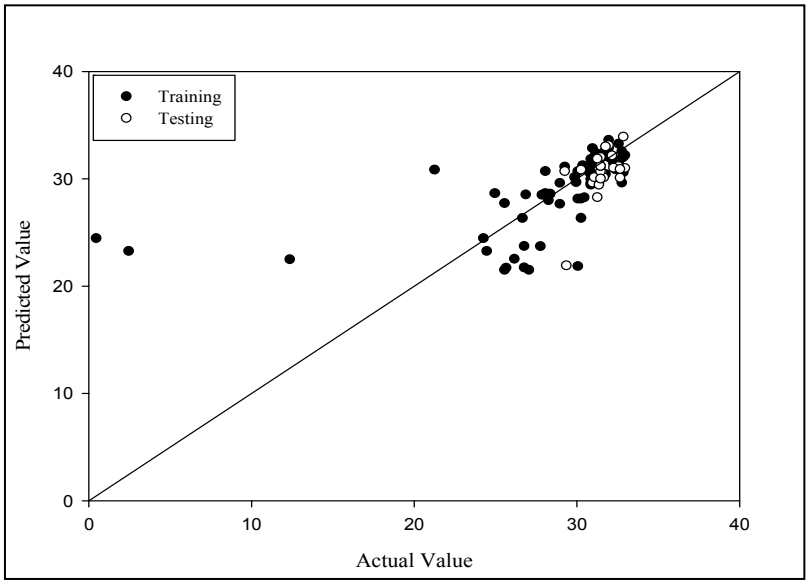

Fig. 3. Taiwan Strait salinity-LR_1

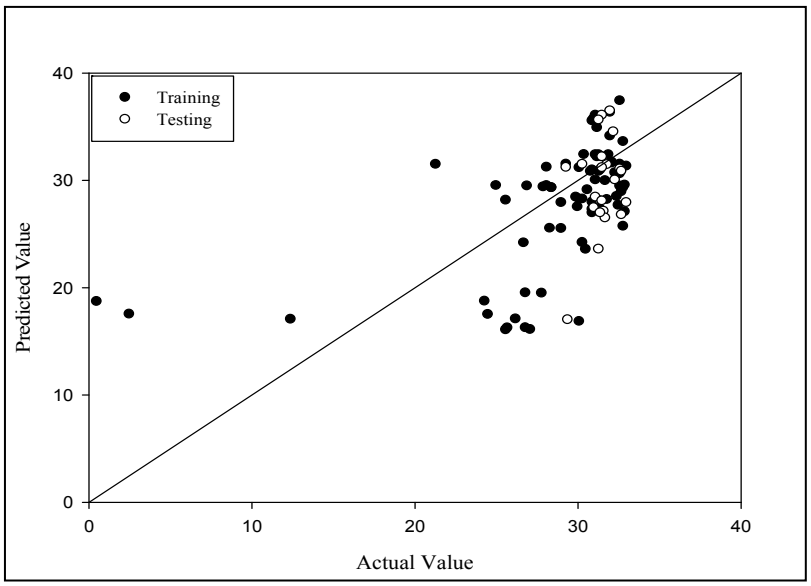

Fig. 4. Taiwan Strait salinity-LR_2

The CCs of equation (4) are 0.59 and 0.56 for training set and testing set, respectively. The CCs of equation (5) are 0. 55 for training set and 0.39 for testing set. The RMSEs of equation (4) are 4.45 and 2.06 for training set and testing set, respectively. The RMSEs of equation (5) are 5.34 for training set and 4.55 for testing set as shown in Table 2. Since nonlinear relationships may exist between the inputs and outputs, it is necessary to use a more advanced automatic programming and optimization model, such as GEMA to fit the complex nonlinear transfer function between the MODIS bands and salinity parameter.

Table 2. The results of LR_1, LR_2, BPN and GEMA on the training and testing data

Models	LR_1		LR_2		BPN		GEMA	
	CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE
Training	0.59	4.45	0.55	5.34	0.71	3.94	0.81	3.25
Testing	0.56	2.06	0.39	4.55	0.61	2.11	0.68	0.87

Using BPN. The same data were used to run back-propagation neural network (BPN). The results show on Fig. 5. The CCs of BPN are 0.71 and 0.61 for training set and testing set, respectively, which are better than LR_1 and LR_2 as shown in Table 2. The RMSEs are 3.94 and 2.11 for training set and testing set, respectively, which are lower than RMSEs of LR_1 and LR_2. BPN was found better than the traditional LR_1 and LR_2 models for salinity estimation for both training and testing sets as indicated by the higher CC and lower RMSE.

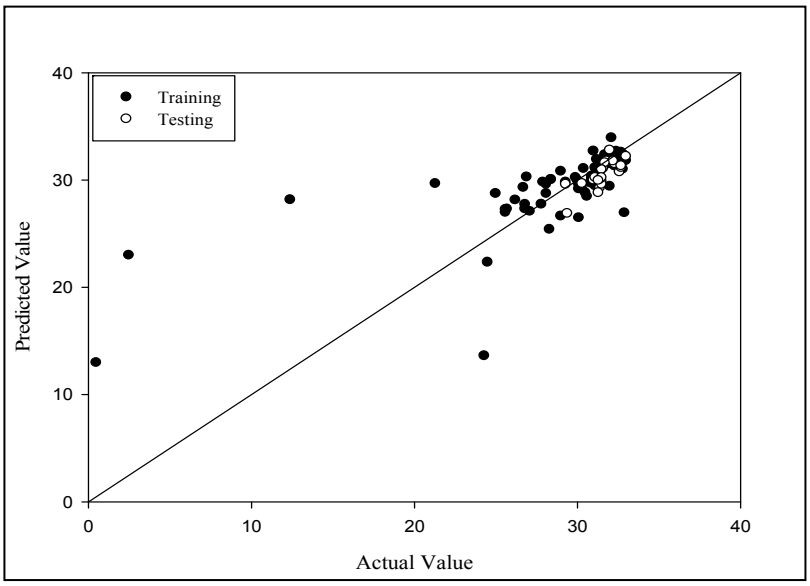

Fig. 5. Taiwan Strait salinity-BPN

Using GEMA. The GEMA model is implemented in C++ Language. The function library types in BNF that apply several sets of the terminals including mathematical operators such as { +, -, ×, /, LN, EXP, POWER}. During the training stage, the GEGA model is built up through predefined 1000 generations with a population size of 100. After 700 generations, the final converged solution obtained from GEMA is shown as equation (6). After ten experiments, the objective values of different final solutions from 3.253 to 3.725 were obtained between 700 and 800 generations.

$$Salinity_{GEMA} = 33 - 4.4 * X_3 / [X_4 + X_2 + X_4 * (X_6 - X_5) / (X_3 + X_6 - X_5)] \quad (6)$$

The result shows that only six input variables X_1, X_2, X_3, X_4, X_5 and X_6 were chosen automatically from total seven input variables by GEMA to form equation (6) through a lot of generations' evolutions and competitions. It shows the six input variables have most strong effects on the predicted salinity. Fig. 6 shows the salinity map using GEMA model.

In Table 2, the result indicates that the CC = 0.81 and RMSE = 3.25 for training data, and CC = 0.68 and RMSE = 0.87 for testing data of GEMA are better than those of LR_1, LR_2 and BPN. GEMA was found better

than the traditional LR_1, LR_2 and BPN for salinity estimation for both training and testing sets as indicated by the higher CC and lower RMSE.

In order to realize the performances of these three models, their diagrams are depicted and compared with each other. The horizontal axis is the actual value salinity, and the vertical axis is the predicted value salinity. Fig. 6 shows that the predicted values for GEMA are closer to ideal line (45°) than LR_1 and LR_2 and BPN, Fig. 3, 4 and 5.

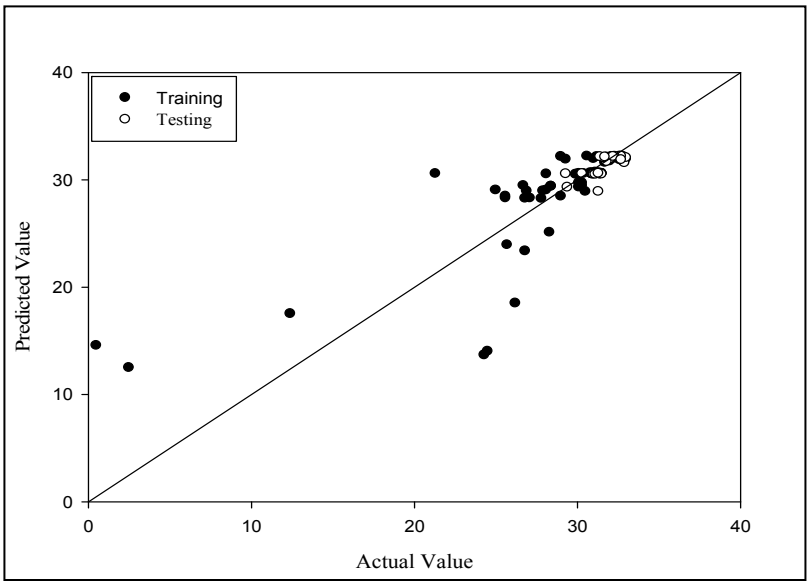

Fig. 6. Taiwan Strait salinity-GEMA

5 Conclusions

This paper provides an improved real-coded grammatical evolution combined with the macroevolution algorithm (GEMA), to predict strait salinity and compared with the conventional linear regression (LR_1), LN transform of linear regression (LR_2) and back-propagation neural network (BPN). GEMA deals with nonlinear transfer problems among several input and output data to generate a fittest mathematical equation. Few significant variables can be chosen from all input variables automatically. The result shows that GEMA used real number coding as an efficient and robust model. Although, using of the GEMA was not as simple as the basic formula, it provided an appropriate model to predict strait salinity using the five input variables. The response figure demonstrates that the relationship between predicted salinity and actual salinity generated by GEMA was reasonable. The result of this case study indicates that the CC = 0.81 and RMSE = 3.25 for training data, and CC = 0.68 and RMSE = 0.87 for testing data of GEMA are better than those of LR_1 (CC = 0.59 and RMSE = 4.45 for training set, and CC = 0.56 and RMSE = 2.06 for testing set), LR_2 (CC = 0.55 and RMSE = 5.34 for training set, and CC = 0.39 and RMSE = 4.55 for testing set), and BPN (CC = 0.71 and RMSE = 3.94 for training set, and CC = 0.61 and RMSE = 2.11 for testing set) as shown in Table 2. The results confirms that GEMA would be the better option than linear regression (LR_1), LN transform of linear regression (LR_2) and back-propagation neural network (BPN), because it models salinity without the limitation of linear property which conventional linear regression (LR_1), LN transform of linear regression (LR_2), and cannot conquer. The current study shows a successful application of GEMA on salinity predicting and can be effectively used for Taiwan Strait salinity predicting. Further researches of water quality parameters of lakes, reservoirs and oceans can be improved to use the real-coded expression of grammatical evolution combined with the macroevolution algorithm.

Acknowledgement

The authors wish to express their gratitude and sincere appreciation to professor Lien-Siang Chou from National Taiwan University, R.O.C., for the data collection.

References

[1] S.Z. Feng, F.Q. Li and S.J. Li, *Introduction to Marine Science*, Higher Education Press, Beijing, 1999.

[2] E.A. Urquhart, B.F. Zaitchik, M.J. Hoffman, S.D. Guikema, E.F. Geiger, "Remotely Sensed Estimates of Surface Salinity in the Chesapeake Bay: A statistical approach," *Remote Sensing of Environment1*, Vol. 23, pp. 522-531, 2012.

[3] N.S. Cooper, "The Effect of Salinity on Tropical Ocean Models," *Journal of Physical Oceanography*, Vol. 18, No. 5, pp. 697-707, 1988.

[4] S. Qing, J. Zhang, T.W. Cui, Y.H. Bao, "Retrieval of Sea Surface Salinity with MERIS and MODIS Data in the Bohai Sea," *Remote Sensing of Environment*, Vol. 136, pp. 117-125, 2013.

[5] L.W. Harding, E.C. Itsweire, W.E. Esaias, "Determination of Phytoplankton Chlorophyll Concentrations in the Chesapeake Bay with Aircraft Remote Sensing," *Remote Sensing of Environment*, Vol. 40, No. 2, pp. 79-100, 1992.

[6] J. McKeon, R. Rogers, "Water Quality Map of Saginaw Bay from Computer Processing of Landsat-2 Data," Special Report to Goddard Space Flight Center, Greenbelt, Maryland. 1976.

[7] S. Khorram, "Development of Water Quality Models Applicable Throughout the Entire San Francisco Bay and Delta," *Photogrammetric Engineering and Remote Sensing*, Vol. 51, No. 1, pp. 53-62, 1985.

[8] D.M. Le Vine, J.B. Zaitzeff, E.J. D'Sa, J.L. Miller, C. Swift and M. Goodberlet, *Sea Surface Salinity: Toward an Operational Remote-Sensing System*, Elsevier Oceanography Series 63, Amsterdam, 2000.

[9] P. Lavery, C. Pattiaratchi, A. Wyllie, P. Hick, "Water Quality Monitoring in Estuarine Waters Using the Landsat Thematic Mapper," *Remote Sensing of Environment*, Vol. 46, No. 3, pp. 268-280, 1993.

[10] W.J. Wilson, S.H. Yueh, F.K. Li, S. Dinardo, C. Yi, C. Koblinsky, G. Lagerloef, S. Howden, "Ocean Surface Salinity Remote Sensing with the JPL Passive/Active L-/S-band (PALS) Microwave Instrument," in *Proceeding of IEEE 2001 International Geoscience and Remote Sensing Symposium*, pp. 937-939., 2001

[11] A.G. Dekker, R.J. Vos, S.W.M. Peters, "Analytical Algorithms for Lake Water TSM Estimation for Retrospective Analyses of TM and SPOT Sensor Data," *International Journal of Remote Sensing*, Vol. 23, No. 1, pp.15-35, 2002.

[12] M.S. Wong, K.H. Lee, Y.J. Kim, J.E. Nichol, Z. Li, N. Emerson, "Modeling of Suspended Solids and Sea Surface Salinity in Hong Kong using Aqua/MODIS Satellite Images," *Korean Journal of Remote Sensing,* Vol. 23, No. 3, pp. 161-169, 2007.

[13] C. Hu, Z. Chen, T.D. Clayton, P. Swarzenski, J.C. Brock, F.E. Muller-Karger, "Assessment of Estuarine Water-Quality Indicators Using MODIS Medium-Resolution Bands: Initial Results from Tampa Bay, FL," *Remote Sensing of Environment*, Vol. 93, No. 3, pp. 423-441, 2004.

[14] R. Barbini, F. Colao, L. De Dominicis, R. Fantoni, L. Fiorani, A. Palucci, E.S. Artamonov, "Analysis of Simultaneous CholorophyII Measurements by Lidar Fluorosensor, MODIS and SeaWiFS," *International Journal of Remote Sensing*, Vol. 25 , No. 11, pp. 2095-2110, 2004.

[15] M.-T. Wu, J.-S. Wu, C.-N. Lee, M.-C. Chen, "A Genetic Algorithm-Fuzzy-Based Voting Mechanism Combined with Hadoop Map-Reduce Technique for Microarray Data Classification," *Journal of Computers*, Vol. 24, No. 3, pp. 40-48, 2013.

[16] C.-Y. Chang, H.-J. Wang, C.-F. Li, "Image Content Analysis Using Modular RBF Neural Network," *Journal of Computers,* Vol. 21, No. 2, pp.41-54, 2010.

[17] Shih Chi Peng and Chuan Yi Tang ,"Computer-Aided Engineering for Inference of Genetic Regulatory Networks Using Data from DNA Microarrays," *Journal of Computers*, Vol. 21, No. 3, pp. 1-11, 2010.

[18] M.-Y, Cheng, Y.W. Wu, "Evolutionary support vector machine inference system for construction management," *Automation in Construction*, Vol. 18, No. 5, pp. 597-604, 2009.

[19] N. Padhy, R. Panigrahi, "Data Mining: A prediction Technique for the workers in the PR Department of Orissa (Block and Panchayat)," *International Journal of Computer Science, Engineering and Information Technology*, Vol.2, No.5, pp. 19-36, 2012.

[20] R.J. Allee, J.E. Johnson, "Use of Satellite Imagery to Estimate Surface Chlorophyll-a and Secchi Disc Depth of Bull Shoals Reservoir, Arkansas, USA," *International Journal of Remote Sensing,* Vol. 20, No. 6, pp.1057-1072, 1999.

[21] Y. Zhang, J.T. Pulliainen, S.S. Koponen, M.T. Hallikainen, "Water Quality Retrieval from Combined Landsat TM Data and ERS-2 SAR Data in the Gulf Finland," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 41, No. 3, pp. 622-629, 2003.

[22] S. Khorram, "Remote Sensing of Salinity in the San Francisco Bay Delta," *Remote Sensing of Environment*, Vol. 12, pp. 15-22, 1982.

[23] Z. Xie, C. Zhang, L. Berry, "Geographically Weighted Modelling of Surface Salinity in Florida Bay Using Landsat TM Data," *Remote Sensing Letters*, Vol. 4, No. 1, pp.76-84, 2013.

[24] B.M. Alabbadi, L. Chen, "Applying Genetic Algorithm Combining Operation Tree (GAOT) for Estimating Salinity of Taiwan Strait Using MODIS/Terra," in *Proceeding of 2013 Fourth Global Congress on Intelligent Systems (GCIS)*, pp. 16-20., 2013

[25] E.F. Geiger, M.D. Grossi, A.C. Trembanis, J.T. Kohut, M.J. Oliver, "Satellite-derived coastal ocean and estuarine salinity in the mid-atlantic," *Continental Shelf Research,* Vol. 63, pp. 235-242, 2013.

[26] J.K. Kishore, L.M. Patnaik, V. Mani, V.K. Agrawal, "Application of genetic programming for multicategory pattern classification," *IEEE Transactions on Evolutionary Computation,* Vol. 4, No. 3, pp. 242-257, 2000.

[27] Y.H. Chen, F.J. Chang, "A study on reservoir inflow forecasting using genetic programming," *Journal of Taiwan Water Conservancy,* Vol. 58, No. 1, pp.1-9, 2010.

[28] L. Chen, "A study of applying macroevolutionary genetic programming to concrete strength estimation," *Journal of Computing in Civil Engineering*, Vol. 17, No. 4, pp. 290-294, 2003.

[29] G.D. Elseth and K.D. Baumgardner, *Principles of Modern Genetics,* West Publishing Company, Minnesota, 1995.

[30] M. O'Neill, C. Ryan, "Grammatical evolution," *IEEE Transactions on Evolutionary Computation*, Vol. 5, No. 4, pp. 349-358, 2001.

[31] M. O'Neill and C. Ryan, *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*, Kluwer Academic Publishers, Massachusetts, 2003.

[32] M. Harrison, *Introduction to Formal Language Theory*, Addison-Wesley, Massachusetts, 1978.

[33] J.W. Backus, F.L. Bauer, J. Green, C. Katz, J. McCarthy, P. Naur, A.J. Perlis, H. Rutishauser, K. Samelson, B. Vauquois, J. H. Wegstein, A. van Wijngaarden, M. Woodger, "Revised report on the algorithmic language Algol 60," *Communication of the ACM*, Vol. 6, No. 1, pp. 1-17, 1963.

[34] N. Wirth, "What can we do about the unnecessary diversity of notation for syntactic definitions?," *Communications of the Association for Computing Machinery*, Vol. 20, No. 11, pp. 822-823, 1977.

[35] L. Chen, C. -H. Tan, S. -J. Kao, T. -S. Wang, "Improvement of remote monitoring on water quality in a subtropical reservoir by incorporating grammatical evolution with parallel genetic algorithms into satellite imagery," *Water Research*, Vol. 42, No. 1-2, pp. 296-306, 2008.

[36] L. Chen, T. -S. Wang, "Modeling strength of high-performance concrete using an improved grammatical evolution combined with macrogenetic algorithm," *Journal of Computing in Civil Engineering*, Vol. 24, No. 3, pp.281-288, 2010.

[37] J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, A Bradford Book, Massachusetts, 1992.

[38] L. Chen, F.J. Chang, "Applying a real-coded multi-population genetic algorithm to multi-reservoir Operation," *Hydrological Processes*, Vol. 21, pp. 688-698, 2007.

[39] J. Marin, R. Sole, "Macroevolutionary algorithms: A new optimization method on fitness landscapes," *IEEE Transactions on Evolutionary Computation,* Vol. 3, No. 4, pp. 272-285, 1999.

[40] C.-T.A. Chen, "Rare northward flow in the Taiwan strait in winter: A note," C*ontinental Shelf Research*, Vol. 23, pp. 387-391, 2003.

[41] T.-Y. Chu, "Environmental Study of the Surrounding Waters of Taiwan," *Acta Oceanographica Taiwanica*, No. 1, pp. 15-32, 1971.

[42] S. Jan, D.D. Sheu, H.-M. Kuo, "Water mass and throughflow transport variability in the Taiwan strait," *Journal of Geophysical Research*, Vol. 111, No. C12, pp. 1-15, 2006.