J. Appl. Environ. Biol. Sci., 4(11)83-90, 2014
© 2014, TextRoad Publication

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Modeling and Prediction of Changes in Anzali Pond Using Linear Regression and GMDH Neural Network

Farshad Parhizkar Miandehi[1], Asadollah Shahbahrami[2]

[1]Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran
[2]Faculty of Engineering, Department of Computer Engineering, University of Guilan, Rasht, Iran

Received: July 22, 2014
Accepted: October 3, 2014

ABSTRACT

Iranian ponds and water ecosystems are valuable assets which play decisive roles in economic, social, security and political affairs. Within the past few years, many Iranian water ecosystems like Urmia Lake, Karoun River and Anzali Pond have been under disappearance threat. Ponds are habitats which cannot be replaced and this makes it necessary to investigate their changes so that we are able to save these valuable ecosystems. The present research aims to investigate and evaluate the trend of variations in Anzali Pond using meteorological data between 1991-2010 by means of GMDH, which is based upon genetic algorithm and is a powerful technique in modeling complex dynamic non-linear systems, and linear regression technique. Input variables of both methodsinclude all factors (inside system and outside system factors) which affect variations in Anzali Pond. Exactness of linear regression method was 78% and exactness of GMDH neural network method was more than 97%. As it can be seen, exactness of GMDH neural network method is significantly better than regression model.

KEYWORDS: Anzali Pond, Regression analysis, modeling, error measurement criterion, GMDH neural network

INTRODUCTION

Investigation of conditions of natural ecosystems like jungles, range, lakes and ponds is of great importance in every country (Jamalzad, F, 1998). At present, Iran uses its natural resources 3.6% more than its normal use.Iranian environment will be disappeared if it goes on like this (Zebardast, L and Jafari, H, 2001). Within the past few years, many Iranian natural ecosystems like Urmia Lake, Arasbaran jungles and Anzali Pond have received irreparable harms and are prone to complete disappearance. Ponds are important natural ecosystems which cannot be replaced and they cannot be revived if they are not safeguarded. This makes it necessary to investigate the trend of their changes (Ghahraman, A and Attar, F). One of the uninvestigated points about ponds is absence of attention to non-linear changes and behavioral nature of them, which can be affected by many factors (Gaoning, J, 2007). Therefore, the present research tries to model the trend of ponds changes using linear regression and GMDH neural network methods and compare their prediction exactness. It is necessary to understand and model relationship between input-output data in order to model any system. Fuzzy logic, neural networks and genetic algorithm are good techniques in solving complex non-linear systems (Ivakhenco, G, 1995-Ivakhenco, G, 1996-Abbaspour and NazariDoust, 2007). Numerous studies have been conducted on prediction of ntural ecosystems changes in different spots of the world (Ozsemi, S and Baer, M, 2002-Yang, J, 2008-Ivakhenco, G, 1996). Most of them have used aerial images or satellite images for evaluation. One of the main studies in this case is titled "trend of ecosystems changes in general and ponds changes in particular" (Dirac, Y and Jones, K, 2008). In this research, the author believes that understanding of the trend of changes in natural ecosystems and especially ponds can be useful in prediction of future status of them. Another research tried to investigate the trend of changes in South African ponds and then identify factors which affect these changes and interactions between the factors using satellite images and geographical information systems (GIS) (Dirac, Y and Jones, K, 2008). Prediction of natural ecosystems changes in Iran started when Urmia Lake went under crisis and many studies dealt with the reduction of the volume and area of Urmialake using visual analysis of satellite images and meteorological data (Abbaspour and NazariDoust, 2007-global environmental warning system, 2012). In a similar study (Ahmadi et al, 2011), researchers used image processing and recognition of Urmia Lake textures and identification of salt sections and calculation of increase in these sections around Urmialake to investigate bioenvironmental threats on this lake and then used linear regression to evaluate the present status of the lake. In the present research, table of factors affecting area and depth changes was created first of all. Then, we analyzed the trend of changes in area and depth of Anzali Pond using linear regression. In the next step, we predicted a time series for changes in Anzali pond using GMDH neural network based on genetic algorithm and used all factors affecting changes in the pond. 70% of data were used as input and 30% were used as test. Results showed that exactness of prediction of area and depth in

***Corresponding Author:** Farshad Parhizkar Miandehi, Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran. Email: Farshad.parhizkar@gmail.com, Tel: (+98)24-34221003.

regression analysis was 78% and in GMDH neural network method was 98%. General structure of this paper is as follows:

In the second part, we review definitions and methods. In the third section, factors influencing on changes in Anzali Pond are introduced. In the fourth section, the influence of the factors on the trend of pond changes is investigated and in the fifth section, we will investigate the implementation and evaluation of the trend of changes using linear regression and GMDH neural network method. In the sixth section, we present conclusions and recommendations.

1. Definitions and methods

1.1. Linear regression

Existence of large volumes of data requires use of data mining science on this subject. Regression is one of the prediction methods in datamining. By prediction, we aim to determine output structure using present behavior. In other words, it means achieving a result using data. Outputs can be both numeral and analogical. In comparison with data mining strategies, this strategy is of great importance and has a more general concept with respect to other methods. Regression helps obtain the followings:

1. Identification of the behavior of the variable y using the variable x, i.e. how y changes as x changes. For instance, this behavior is linear in one sample and is curved in another sample.
2. Prediction based on data for future samples which is the main target in datamining through statistical methods.

Modeling targets for explaining relationship between x and y depends on use of model results for prediction of applications of estimation but inference is a more exact area.
Linear regression is the most applicable method in prediction of variations trend.

1.2. Introduction of GMDH neural network

Evolutionary methods like genetic algorithm have many applications in different stages of designing neural networks (Sharzei, Q et al, 2008) because they are capable of finding optimal values and enable us to search in unpredictable spaces (Ranjbar, J. 2008). Therefore, we used genetic algorithm in this research in order to design neural network and determine its coefficients because GMDH neural network contains a collection of neurons which is formed as a result of different pair links and via a quadratic polynomial. Assume a set "m" containing x_1, x_2, …x_m and a variable "y" and assume data for x_is and y are present for a particular period of time. In other words, every variable is in the form of a vector which includes time series numbers corresponding to that variable (Howland, J, 2003). Primary data which must be collected for constructing GMDH algorithm include:

We need two things for starting algorithm work: identification of a relationship which produces output variable based on input variables x_is and prediction of y using x_is. in other words, we need to identify the model and relationship between variables (modeling) so that we can predict future values of target variable (output variable) (Abrishami, H et al, 2008).

GMDH algorithm is a process for building a high-degree polynomial which is called Voltra function series and is as follows: (this polynomial is also called Ivakhenco).

Therefore, in GMDH algorithm, we first decompose Voltra functions series into two-variable quadratic polynomials.

$$G(x_i,x_j)=a_0+a_1x_i+a_2x_j+a_3x_i^2+a_4x_j^2+a_5x_ix_j \quad (1)$$

In this decomposition, Voltra series is converted into several chain iterative equations so that Voltra series can be reconstructed if we replace all of the iterative relations in the relationship using algebraic replacement.

$$y_i=f(x_{i1},x_{i2},x_{i3},\ldots,x_{im}) \qquad i=1,2,3,\ldots,m \quad (2)$$

It is approximated by f function:

$$\hat{y}_i=\hat{f}(x_{i1},x_{i2},x_{i3},\ldots,x_{im}) \qquad i=1,2,3,\ldots,m \quad (3)$$

And if we rewrite f function as follows:

$$\hat{y}=a_0+\sum_{i=1}^{m}a_ix_i+\sum_{i=1}^{m}\sum_{j=1}^{m}a_{ij}x_ix_j+\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}a_{ijk}x_ix_jx_k+\ldots \quad (4)$$

As it can be seen in the above equations, the order of the above relations from top to bottom is a view of the process of decomposition relation (3) into several quadratic polynomials. On the other hand, the order of these relations from bottom to top indicates completion of relation (4) by iterative relations. In fact, this algorithm aims to find alpha unknown coefficients in Voltra functions series. It must be mentioned that all partial models have a similar structure like this:

$$\hat{f}(x_i,x_j)=v_0+v_1x_i+v_2x_j+v_3x_i^2+v_4x_j^2+v_5x_ix_j \quad (5)$$

Considering the fact that we try to reach a primary system modeling, we can combine partial systems model and repeat this action in order to reach the main system model which is presented in relation (6).

$$\hat{y}=v_0+\sum_{i=1}^{m}v_i x_i+\sum_{i=1}^{m}\sum_{j=1}^{m}v_{ij}x_i x_j+\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}v_{ijk}x_i x_j x_k+\ldots \quad (6)$$

After decomposition the main system into C_m partial systems, we calculate a model with two input variables for each of them. Then, we combine the partial models mutually and the number of new partial models or systems is $\frac{C_m(C_m+1)}{2}$ with at least three or four input variables. Of course, the number of variables dependent on the model or the number of system inputs is not important and only exactness of real estimation of the main system by the created models is important. Therefore, we select partial models which are of high estimation in comparison with other models and eliminate others.

1.3. Criteria for prediction power measurement

Different criteria have been introduced for measuring prediction power of different models. The followings are several of these criteria:

Root square mean error (RSME):

$$\text{RMSE}=\sqrt{\sum_{t=T+1}^{T+h}\left(\hat{y_t}+y_t\right)^2} \quad (6)$$

Mean absolute error (MAE):

$$\text{MAE}=\sum_{t=T+1}^{T+h}\left|\hat{y_t}+y_t\right|/h \quad (7)$$

Mean absolute prediction error (MAPE):

$$\text{MAPE}:\sum_{t=T+1}^{T+h}\left|\frac{\hat{y_t}+y_t}{y_t^a}\right|/h \quad (8)$$

We used the first criterion in this research.

2. factors affecting the trend of changes in Anzali Pond

Te present research aims to model and predict the trend of changes in area and depth of Anzali pond. Table 1 shows factors affecting changes in area of the pond and table 2 shows factors affecting changes in the depth of the pond.

Table 1: independent and dependent variables for modeling and prediction of changes in area of the pond (Jamalzad, F, 2008)

Intended atmospheric parameters	constants	variables
Precipitation-independent	B_1	X_1
Water discharged in river-independent	B_2	X_2
Temperature-independent	B_4	X_4
Pond surface area-dependent	Y	

Table 2: dependent and independent variables for modeling and prediction of depth of water in the pond (Jamalzad, F, 2008)

Intended atmospheric parameters	constants	variables
Precipitation-independent	B_1	X_1
Water discharged in river-independent	B_2	X_2
Temperature-independent	B_3	X_3
Debris-independent	B_4	X_4
Pond depth-dependent	Y	

Investigation of factors affecting the trend of changes in lake

Precipitation in the pond region

Considering figure 1, it can be seen that precipitation level has increased slightly (Y=10.902X – 20552, R^2=0.5088). It shows that precipitation level plays a positive role in pond changes.

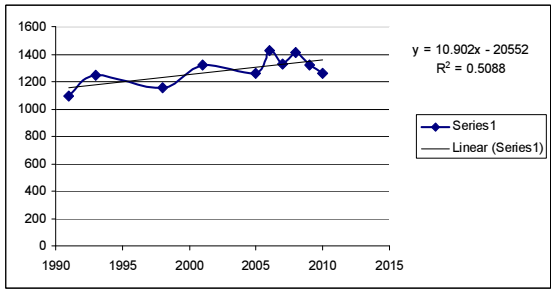

Figure 1: trend of changes in precipitation level in 1991-2010

Level of waters discharging the pond

As it can be seen in figure 2, the lowest level of discharged water belongs to 1991 because of a serious drought at that time and the highest level of discharged water belongs to 1998 due to tide in Caspian Sea, which increased the pond's water level considerably. Since this factor took place only once in the time interval, we did not assume Caspian Sea tide as an independent factor in this research ($Y=-6.7055X + 15350$, $R^2=0.5088$). this shows that the level of discharged water has had a severe descending trend after 1988 and especially within the recent years.

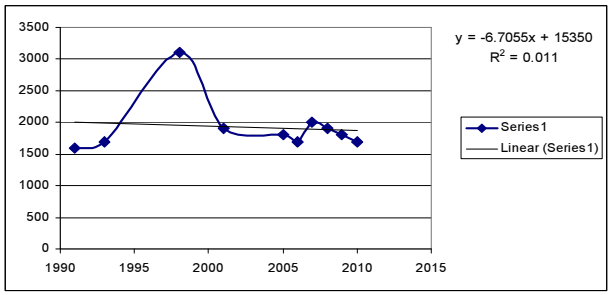

Figure 2.trend of changes in the volume of water discharged in Anzali Pond in 1991-2010.

Temperature changes

Investigation of minimum, maximum and average annual temperature in stations of Anzali Pond shows that temperature variation is about mean value of statistical period until 2010. In other words, temperature mean value is more than mean value of total period in some years and is smaller than mean value of total mean value in some other years. The ascending trend of temperature starts in 2005.

Table 3: data needed for GMDH neural network

Year	ɛ of discharged debris (tons)	Volume of discharged water (million cubic meters)	rain (mm)	ration(MM)	Pond surface area (Km2)
1991	974.44	1600	1095	1100	57.84
1993	990.759	1700	1246	950	58
1998	1073.616	3100	1154	1020	81.87
2001	1175.728	1900	1324	900	66.9
2005	1273.572	1800	1257	850	66.5
2006	1273.189	1700	1425	800	66
2007	1273.158	2000	1326	900	64.5
2008	1272.214	1900	1411	800	62.09
2009	1271.144	1800	1324	1000	60.39
2010	1291.52	1700	1264	900	56.91

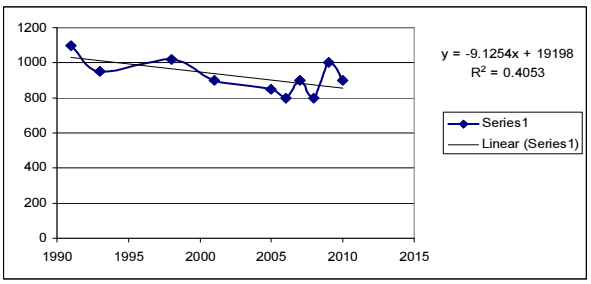

Figure 3:trend of temperature changes in 1991-2010

2.1. Inserted debris

2.2. Considering the fact that many industrial, urban and sewage debris discharge Anzali at present, the volume of these debris has increased a lot within the past few years. Figure 4 shows the discharged debris volume ($Y=36.978X + 983.56$, $R^2=0.7818$) and shows that discharged debris has had a considerable impact on reduction in Anzali Pond depth.

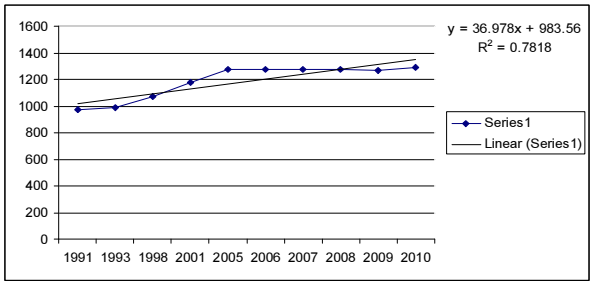

Figure 4: level of debris inserted into Anzali Pond in 1991-2010

2.2. Changes in surface area of Anzali pond

Data for changes in the pond's surface area can be seen in table 3. Regression line equation for these data is y= -0.0701x + 204.58. according to this equation, the descending trend of pond's surface area becomes serious after 1998.

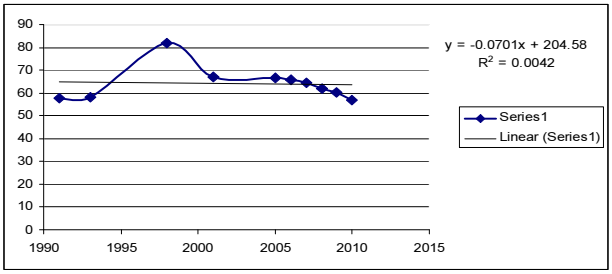

Figure 5: changes in the pond's surface area in 1991-2010

3. Implementation and evaluation

3.1. Modeling and prediction of changes in Anzali Pond using GMDH neural network

Data listed in table 3 were used as GMDH neural network input. The present research aims to model and predict changes in the depth and area of the pond as output variable. Statistics and data were collected from Anzalimeteorological station and Guilan Province fishery website over 1991-2010.

Primary assumptions in GMDH neural network analysis are as follows:

- The number of latent layers is equal to 3.
- Percentage of the samples considered for test is equal to 30%.

First, we obtain independent variables influencing on the surface area and depth of the pond according to tables (1) and (2) as neural network inputs.

After neural network analysis, the number of population produced for prediction of depth is 1540 and the number of population produced for estimation of surface area is 231. These samples have been obtained by combining input variables in two layers. Sample production in any layer corresponds to combination of

variables in the previous layers. In other words, combination of the three variables using $\frac{C_m(C_m+1)}{2}$ equation in the first layer yields 6 answers and combination of these 6 answers in the second layer yields 21 answers and in the third layer, we obtain 231 layers. For the case of 4 variables for pond depth calculation, we obtain 1540 answers. However, it is necessary to select the best answers out of all answers in order to avoid neural network's divergence. Therefore, training error and prediction error was calculated for all final combinations. Selection of optimal answers seeks two targets: minimization of modeling error and prediction. Another point in selection of optimal final input is observation of the order of selected variables to avoid scattering. As it can be seen in figures (4) and (5), 10 samples were selected for estimation of surface area of the pond and 17 samples were selected for estimation of pond depth using genetic algorithm. Rows 5 and 8 were considered for calculation of the trend of changes due to maintaining the order of input variables.

Table 4: table of selection of variables for depth

Row	Variables index								trainingerror	Prediction error
6	3	4	5	3	5	7	5	2	0.01524698	0.003695478
7	6	6	4	6	3	3	4	5	0.01325698	0.00548793
8	4	4	5	5	6	1	2	3	0.0212121	0.00666254
17	2	7	4	5	6	2	4	6	0.06598421	0.008970471

Table 5: table of selection of variables for surface area

	Variables index								training error	Prediction error
3	2	2	4	4	4	2	5	5	0.0248712	0.00168541
4	4	4	2	6	6	2	5	6	0.0123548	0.00231548
5	4	5	6	7	1	2	3	3	0.0136589	0.00652141
6	5	6	4	4	2	4	3	3	0.0258326	0.00725001
10	5	6	4	0	1	2	3	1	0.0369855	0.01595293

Inputs were classified in two categories: training data (including at least 70% of data) and test data (30% of data). tables 6 and 7 indicate the exactness of depth and surface area evaluation.

Table 6: investigation of predicted exactness for pond depth in 2008-2010

Prediction exactness (percentage)	Prediction error(percentage)	Predicted depth	RMSE	Time period
97.02	2.98	7.3	4.41	**2008**
99.34	0.66	6.8	4.48	**2009**
98.4	1.6	6.2	5.48	**2010**

Table 7: investigation of predicted exactness of pond surface area in 2008-2010

Prediction exactness (percentage)	Prediction error(percentage)	Predicted area	RMSE	Time period
96.87	3.13	60.4	4.50	**2008**
97.21	2.79	58.70	4.84	**2009**
98.80	1.20	56.22	5.53	**2010**

3.2. Introduction of multiple linear regressions

Investigation of pond water level and surface area indicates that there is a strong relationship between these variables and atmospheric conditions.

The question is that: whether we can calculate depth and surface area of the pond out of atmospheric conditions data?

In regression analysis, the depth and surface area of the pond were considered as dependent variables and atmospheric parameters were considered as independent variables and the regression equation is as follows:

$$Y=B0+B1X1+B2X2+B3X3+B4x4+\ldots$$

Xs are atmospheric parameters and Bs are calculated in a way that least squares index is satisfied. The calculated value for B0 indicates predicted values of Y as X values are held constant. For instance, B2 value indicates values predicted by X2 variable in case of holding Xs constant.

On the other hand, comparison of B coefficients reveals the rank and impact size of each of the factors.

Further, B variable sign shows that this parameter has a direct or reverse relationship with dependent variable.

3.2.1. Depth investigation

As it was mentioned in section 3, atmospheric data are independent variables in linear regression and elevation of pond level is as presented in table 2. We reach the following equation after investigation of data and variables. the calculated determination coefficient is equal to 0.72.

(7) Y=124.36+0.0551X1+0.00291X2-0.3658X3

3.2.2. Surface area investigation

Further, atmospheric data are independent variables in linear regression and surface area of pond level is as presented in table 1.

Investigation of the data and variables leads us to the following equation and coefficients. The corresponding determination coefficient is equal to 0.83.

(8) Y=1265.1202.3X1+0.987X2-217.0074X3

Estimations of the coefficients of this model were presented in the previous sections. Prediction of the estimated model by means of linear regression in the time period using equations (7) and (8) reveals that error percentage of the linear regression model for prediction of trend of depth changes is 9.3% and also RMSE for surface area was 7.3%. Real data and predicted data using both methods can be observed in figures (6) and (7).

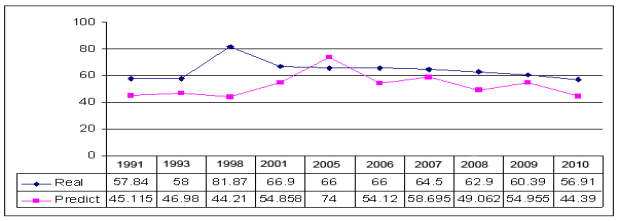

Figure 6: real data graph and linear regression line equation figure for pond surface area

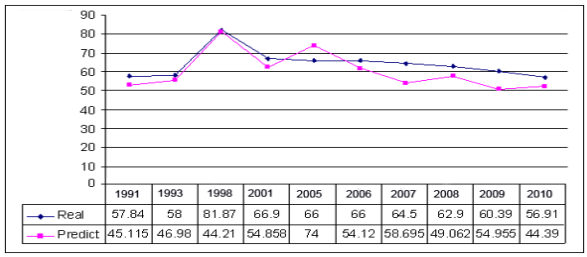

Figure 7: real data figure and GMDH neural network for pond surface area

4. Conclusion

In the present research, we modeled and predicted the trend of changes in Anzali pond using linear regression and GMDH neural network based on genetic algorithm and using table of factors affecting changes in Anzali pond in 1991-2010 and investigated the impact of each of the factors on depth and surface area of Anzali Pond. Results of GMDH neural network modeling analysis on all factors which affect changes in the pond's surface area (as inputs, 10 inputs) proved the serious reduction in surface area (from 82 square kilometers in 1998 to 57 square kilometers in 2010) and its prediction exactness is more than 97%. Using linear regression method, this value was equal to 69 square kilometers. Further, results of analyses conducted on input data (17 inputs) indicated serious reduction in the pond's depth. Further, exactness above 98% for prediction of changes in pond's depth verifies the results of this prediction.

REFERENCES

Abrishami, Hamid and Moeeni, Ali and Mehrara, Mohsen and AHrari, Mahdi and SoleimaniKia, Fatemeh (2008), "modeling and prediction of gasoline price using GMDH neural network", quarterly of Iranian economic studies, 12th year, number 36, pp: 37-58.

Abrishami, Hamid and Mehrara, Mohsen and Ahrari, Mahdi and Mir Ghasemi, Soudeh (2009), "modeling and prediction of Iranian economic growth with a GMDH neural network approach", journal of economic studies, number 88, pp: 1-24.

Abbaspour, M. and Nazaridoust, "Determination of Environmental Water Requirements of Lake Urmia, Iran: an Ecological Approach", International Journal of Environmental Studies, Vol.64, pp.161-169, 2007.

Ahmadi, R., Mohebbi, F., Hagigi, P., Esmailly, L., Salmanzadeh, R. Macro-invertebrates in the Wetlands oftheZarrineh "estuary at the south of Urmia Lake. International Journal of Environmental Restoration", 5(4), 1047-1051. (2011).

De Roeck, E., Jones, K., “Integrating Remote Sensing and Wetland Ecology: a Case Study on South African Wetlands”, pp.1-5, 2008.

Ghahraman, A and Attar, F. Anzali Pond in death coma (an ecological-floristic investigation). Journal of environmental studies: special notes on Anzali Pond: 1 to 38.

Howland. J.C, Voss. M.S. “Natural Gas Prediction Using the Group Method of Data Handling”, ASC. . (2003)

Ivakhnenko.G.A (1995),”The Review of Problems Solvable by Algorithms of the Method of Data Handling (GMDH)”, Pattern Recognition and Image Analysis, Vol.5, No.4, PP 527-535.

Ivakhnenko. G.A and Muller. J.A. (1996). “Recent Development of Self-Organizing Modeling in Prediction and Analysis of Stock Market”, Available in URL Address: http://www.inf.kiev.ua/GMDH Home/Articles.

Jamalzad, F.(2008). determination of the level of sensitivity of different areas of Anzali Pond using GIS, master degree thesis, environment faculty, Tehran University, page 52.

Ozesmi, S. L., E. M., Bauer. “Satellite Remote Sensing of Wetlands. Wetlands Ecology and, Management”, Vol.10, pp.381-402, 2002.

Sharzei, Gholam Ali and AHrari, Mahdi and Fakhraee, Hasan (2008), "structural models, time series and GMDH neural network", journal of economic studies, number 84, pp: 151-175.

Tavakkoli, B and SabetRaftar, K. investigation of the impact of area, population and population compression factors of water basin on rivers discharging Anzali Pond, journal of environmental studies: special notes on Anzali pond: 51 to 57, 2007.

UNEP Global Environmental Alert Service(GEAS) The drying of Iran's Lake Urmiaand its environmental consequences. February(2012)

Van Stappen, G., Bossier, P., Sepehri, H., Lotfi, V., RazaviRouhani, S., Sorgeloos, P., “Effects of Salinity on Survival,Growth, Reproductive and Life Span Characteristics of Artemia Populations from Urmia Lake and Neighboring Lagoons”, Journal of Biological Sciences, Vol.11, pp.164-172, 2008.

Yung, J.L., “Sustainable Wetland Management Strategies under Uncertainties”, the Environmentalist, Vol.19, pp. 67-79, 2008.

Zebardast, L, Jafari, H. R, evaluation of the trend of changes in Anzali Pond using remote sensing and presentation of a managerial solution, journal of environmental studies, 57-64, 2011.

Zhaoning, G., et al. “Using RS and GIS to Monitoring Beijing Wetland Resources Evolution”, IEEE International, Vol.23, pp.4596 – 4599, 2007.