J. Agr. Sci. Tech. (2009) Vol. 11: 111-119

Investigation of the Efficiency of the Fuzzy Regression Method in Reconstructing Monthly Discharge Data of Hydrometric Stations in Great Karoon River Basin

S. J. Sadatinejad[1], M. Shayannejad[2]*, and A. Honarbakhsh[2]

ABSTRACT

There are different methods of reconstructing hydrologic data. Depending on the conditions of the station a particular method can produce the best results. Generally, in order to estimate the lost data in a station and its surrounding stations, hydrologic, climatologic and/or physiolographic similarities are used. Recently, the fuzzy regression method has been used to reconstract the hydrologic data. In this research, the efficiency of this method in reconstructing the montly discharge data of hydrometric stations in comparison to other methods was investigated. The credited omission method was used in this investigation, then by omitting the observed data deliberately, their values were estimated using the different methods. Afterwards, by the use of the statistical index of root mean squared error (RMSE) the best method of reconstruction was determined. The results showed that the best methods of reconstructing monthly discharge data for the hydrometric stations in the great Karoon River basin in order of accuracy are artificial neural network, simple linear regression, multiple linear regression, normal ratio, fuzzy regression, autoregresive and graphical methods.

Keywords: Data reconstruction, Fuzzy regression, Karoon river basin, Monthly discharge.

INTRODUCTION

Hydrology studies are based on correct and complete hydrometric data. However, for many reasons such as the lack of data registration, omitting the incorrect data and being out of order and/or destroying the measuring instruments, there are no complete data available. Therefore, data should be reconstructed using a suitable method. By the study of different references, it was found that substantial research has attempted to reconstruct the hydrologic and climatic data. Some of these research studies are mentioned here. Alizadeh and Salvitabar (1990) using a thirty-year period of discharge data from Shirgah Station in Talar River, reconstructed these data using autoregressive models. Finally, two fifty- and hundred-year data sets have been produced and statistical tests have been shown that there is no significant difference between the results of the model and the observed data. Sadatinejad (1997) has compared the reference station, normal ratio, graphical, simple linear and multiple linear regression and autoregressive methods for reconstruction of yearly precipitation data in Esfahan Province. He introduced the normal ratio metod for arid and Mediterranean climates and the linear regression method for semi-dried climates as the best methods. Jamab Consulting Engineers Company (1999), has used the linear and logarithmic regression for reconstruction of montly and yearly dischage data for fifty-five stations

[1] Department of Natural Resources, College of Agriculture, University of Shahrekord, Shahrekord, Islamic Republic of Iran.
[2] Department of Irrigation, College of Agriculture, University of Shahrekord, Shahrekord, Islamic Republic of Iran.
* Corresponding author, e-mail:shayannejad@yahoo.com

[DOR: 20.1001.1.16807073.2010.12.1.5.1]

situated in Karoon and Dez River basins in order to prepare a comprehensive water plan of Iran.

Lookzadeh (2004) has compared the normal ratio, reversed squared distance, geostatistics, simple linear and multiple linear regression methods for reconstruction of monthly, seasonal and yearly precipitation data in the central Alborz region. In this investigation, the RMSE index has been used for evaluating the results. The results show that the normal ratio method is the best method in 69.2% of cases. In addition, this method provides overestimates and underestimates in wet and dry years, respectively. MacCulloch and Booth (1970) used the regression method for reconstruction of monthly and yearly precipitation data in the Pershen Station of Ontario basin in Canada. The monthly and yearly results had a 17% and 4% error, respectively. Selvam *et al.* (2002) in India and China (1995) and in the Netherland accepted the time series method for reconstruction of discharge and precipitation data. Abeb *et al.* (2000) has compared the methods of normal ratio, artificial neural networks and fuzzy logic for reconstruction of precipitation data in the North of Italy. The results showed that the fuzzy logic made fewer errors in comparison with the two other methods. Lohani *et al.* (2006) determined the stage-discharge relationship using fuzzy logic and its results was more accurate than with traditional methods.

The purpose of this research was the examination of the ability of the fuzzy regression method to reconstruct the monthly discharge data in the great Karoon River basin in comparison with the others including, regressional , normal ratio, graphical, time series and artificial neural networks methods.

Fuzzy systems can be used well to model phenomena with two main kinds of uncertainty. The first one is the uncertainty resulting from the weakness of human knowledge about the phenomena process. The second one is the uncertainty related to the ambiguity of the phenomena. One of the subjects in fuzzy systems is fuzzy regression. On the whole, it is possible to use the fuzzy regression in the following conditions:

1) Insufficiency in the number of observed data;

2) Ambiguity of the relationship between dependent and independent variables;

3) Inaccuracy of linear theories.

In a classic regression, a certain amount for the output variable is calculated for each series of their input variables, while the fuzzy regression estimates the output variables as amounts inside two boundaries. The distribution of these amounts is determined as the membership function. Numerous researchers have presented different methods for solving fuzzy regressional problem. Tanaka *et al.* (1982) investigated this problem for the first time and, thereafter, some other researchers such as Selmins (1987), Bardossy (1990), Bardossy *et al.* (1990), Savic *et al.* (1991), Ishibuchi (1992), Chang *et al.* (2001) and Sanchez *et al.* (2003) continued working on this subject.

MATERIALS AND METHODS

Introducing the River Basin under Investigation

Since the purpose of this research is to compare different methods for the reconstruction of the discharge data, it was attempted to consider the different climatic and physiographic conditions. Therefore, the great Karoon River basin, including the Karoon and Dez basins (Figure 1) was chosen. These basins are situated inside the middle Zagros mountain range in Iran and are limited to 48° , 10' to 52° , 30' eastern longitude and 30° , 20' to 34° , 5' northern latitude.

There are 83 hydrometric stations in the branches of the rivers in the great Karoon River basin. In these stations, discharge, sediment concentration and qualitative parameters are measured. The greatest

[DOR: 20.1001.1.16807073.2010.12.1.5.1]

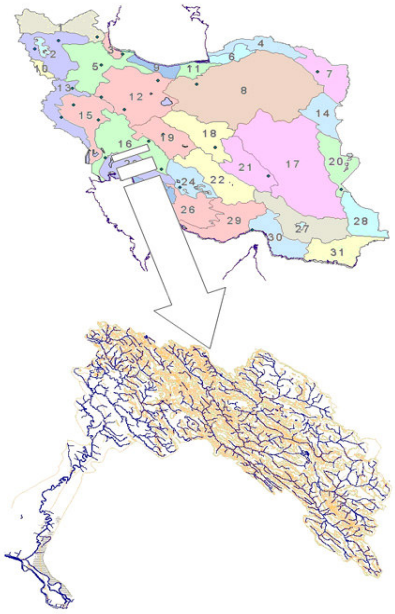

Figure 1. Map of the great Karoon River basin.

number of data in this basin is related to Ahvaz Station from where data is available since 1950. Among these 83 stations, 20 stations were closed and three new stations established in 1990. Karoon and Dez River basins were divided to 8 sub–basins on the basis of their topographic and hydrologic properties. These 8 sub-basins are listed in Table 1.

Determining the Reconstruction Groups

For the reconstruction of discharge data, finding hydrologic, physiographic and climatologic similarities between stations under investigation is necessary. For any station, at least four surrounding stations having a common statistical period and without missing data are needed. So, it is required to divide the basin into some areas

Table 1. Properties of 8 sub-basins in great Karoon River basin.

Altitude (m)		Area Km^2	Major rivers	Sub-basin code
Minimum	Maximum			
1500	4050	6274	Tyreh and Marbareh	3-3-1
500	4082	4729	Sabzeh, Zaz and Sezar	3-3-2
1500	4082	6448	Bakhtyari-Zalaki	3-3-3
15	2585	5799	Deze paeen and Shavver	3-3-4
1790	4409	8976	Beshar, Marbareh and Kharsan	3-4-1
1000	4221	15037	Abvanak, Juneghan, Kyar and Bazoft	3-4-2
200	3701	8531	Karoon	3-4-3
0	1391	12687	Bohlool & Shoor	3-4-4

[DOR: 20.1001.1.16807073.2010.12.1.5.1]

and to choose the best stations located in each area. To do this, the area being studied that including eight sub-basins was divided to 11 reconstruction groups. Based on the area division accomplished, the stations located in a certain sub-basin were allocated to one or several groups (For example the 3-3-1 and 3-3-2 sub-basins have been located in group 1).

Fuzzy Regression

There are three kinds of models that fit a fuzzy linear regression equation: (1) fuzzy posibilistic regression, (2) fuzzy least squares regression and (3) fuzzy regression based on interval analysis. The fuzzy posibilistic regression models represent the best regression model through minimizing the sum of the widths of membership functions of fuzzy regression equation coefficients. In one of these models, the coefficients of regression are fuzzy and observed input and output are not fuzzy. In this research this model has been used as the following equation:

$$\tilde{y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \tilde{A}_2 x_2 + \tilde{A}_3 x_3 + \ldots\ldots\ldots + \tilde{A}_n x_n \quad (1$$

where coefficients of $\tilde{A}_0$, $\tilde{A}_1, \ldots\ldots, \tilde{A}_n$ are fuzzy numbers , x_1 x_2 x_3 و.....و x_n are independent variables as classic numbers, $\tilde{y}$ is a dependent variable as a fuzzy number and n is the number of variables. We suppose there are m rows observed with n input variables and one output variable in each row. For the fuzzy number as a symmetrical triangle (Figure 2), the membership function is written as the following equation:

$$\mu_{\tilde{A}}(a_i) = \begin{cases} 1 - \dfrac{|p_i - a_i|}{c_i} & p_i - c_i \le a_i \le p_i + c \\ 0 & \end{cases} \quad (2$$

where c_i and p_i are the width and center of the fuzzy number, respectively.

$\tilde{A}$ in Equation (2) is used to indicate the "almost equal to p_i" amount and c_i indicates its degree of being fuzzy, which this concept can be shown as $\tilde{A}_i = (p_i, c_i)$. So the fuzzy regression equation is as follows:

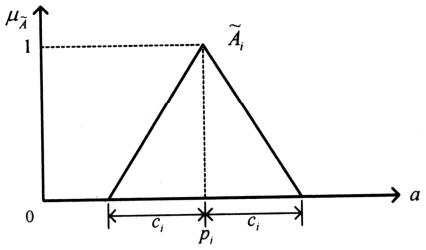

Figure 2. Membership function of a fuzzy number as a symmetrical triangle.

$$\tilde{y} = (p_0, c_0) + (p_1, c_1)x_1 + (p_2, c_2)x_2 + \ldots\ldots\ldots + (p_n, c_n)x_n \quad (3$$

The membership function of the output fuzzy variable is represented as follows:

$$\mu_{\tilde{y}}(y) = \begin{cases} \max\left(\min\left[\mu_{\tilde{A}i}(a_i)\right]\right) & \{a \mid y = f(x,a) \neq \phi \\ 0 & \end{cases} \quad (4$$

Substitution of Equation (2) in (3) gives following equation:

$$\mu_{\tilde{y}}(y) = \begin{cases} 1 - \dfrac{y - p_0 - \sum_1^n p_i x_i}{c_0 + \sum_1^n c_i |x_i|} & x_i \neq 0 \\ 1 & x_i = 0, y = 0 \\ 0 & x_i = 0, y \neq 0 \end{cases} \quad (5$$

There are different algorithms to solve the fuzzy linear regression problem, one of which changes the fuzzy linear regression model to a linear programming problem. In this case, the purpose of the regression model is to determine the optimum amounts for $\tilde{A}$ as the membership degree of the fuzzy output variable which should be greater than a given amount of h which is determined by the user. In other words, the following unequality should be correct for m rows of data (j =1, 2, 3,…, m):

$$\mu_{\tilde{y}j}(y_i) \ge h \quad (6$$

By increasing h, the amount of output fuzziness increases too. The unequality 4 explains that the fuzzy output should be taken between A and B, which have been determined in Figure 3.

[DOR: 20.1001.1.16807073.2010.12.1.5.1]

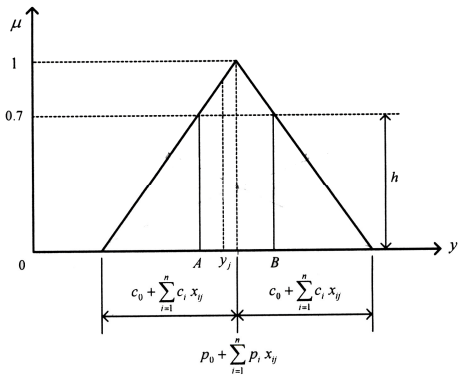

Figure 3. The membership function of fuzzy output.

Considering Equation (5), the center and width of the membership function are equal to $p_0 + \sum_{1}^{n} p_i x_i$ and $c_0 + \sum c_i x_i$ respectively.

For determining the regression coefficients, the fuzzy output width is minimized for all of the data sets. Thus, the subjective function and constraints of linear programming can be represented as in Table 2.

Table 2. Linear programming model to solve the linear regression fuzzy.

Regression equation:
$\tilde{y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \tilde{A}_2 x_2 + ... + \tilde{A}_n x_n$
Subjective function:
$\text{Minimize} : \text{mc}_0 + \sum_{j=1}^{m}\sum_{i=1}^{n} c_i \lvert x_{ij} \rvert$ (7
Constraints:
$p_0 + \sum p_i x_{ij} - (1-h)\left[c_0 + \sum c_i x_{ij}\right] \le y_j$ (8
$p_0 + \sum p_i x_i + (1-h)\left[c_0 + \sum c_i x_{ij}\right] \ge y_j$ (9

The constraints in Table 2 have been obtained by substituting Equation (5) into (6). Thus, in order to solve a linear regression problem with fuzzy coefficients and non fuzzy data, we need to solve a linear programming model based on Table 2 equations. Equations (8) and (9) are written separately for each one of the observed data pairs. So, based on the above mentioned equations, the number of $2m$ unequalities are established. This work is carried out by HYDROGENERATOR software (prepared by the authors). The unequalities set are entered into LINGO software and the coefficients of p_i and c_i are obtained. The fuzzy output change to non fuzzy data using the gravity center method is based on the following equation:

$$A = \frac{\int \mu\ (x).x.dx}{\int \mu\ (x).dx} \quad (10$$

In this research the following six methods have been considered in order to compare their results with the fuzzy regression (FR) method. This comparison is made using correlation coefficients and root mean squared error (RMSE) as the following equation:

$$RMSE = \sqrt{\frac{(Q_m - Q_c)^2}{n}} \quad (11$$

where Q_m is measured discharge, Q_c is calculated discharge and n is the number of data.

Normal Ratio Method (NR)

In this method, the monthly discharge of station X is estimated using around stations A, B,... by following equation:

$$Q_X = \frac{1}{N}\left[\left(\frac{\overline{Q}_X}{\overline{Q}_A}.Q_A\right) + \left(\frac{\overline{Q}_X}{\overline{Q}_B}.Q_B\right) + ...\right] \quad (12$$

where N is the number of stations, Q_X, Q_A, Q_B are the monthly discharge of stations X , A and B respectively and $\overline{Q}_X, \overline{Q}_A, \overline{Q}_B$ are their average values.

Graphic Method (GR)

In this method, the coordinates of station X are assumed (0, 0) and its discharge is estimated using the following equation:

[DOR: 20.1001.1.16807073.2010.12.1.5.1]

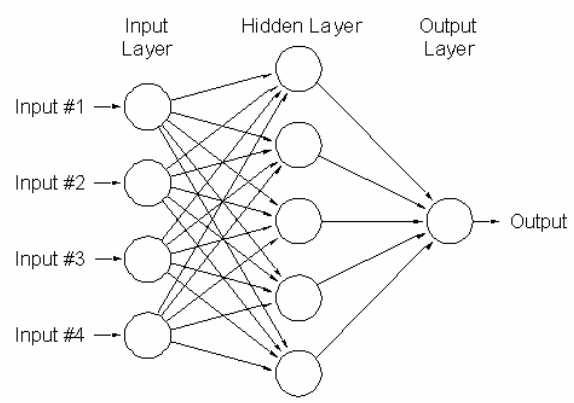

Figure 4. An artificial neural network with one hidden layer.

$$Q_X = \frac{W_A Q_A + W_B Q_B + ...}{W_A + W_B + ...} \quad (13$$

where $W = \frac{1}{X^2 + Y^2}$ and X, Y are coordinates of each around station.

Simple Regression Method (SLR)

In this method, the monthly discharge of station X is estimated using only station A as the following equation:

$$Q_X = a + bQ_A \quad (14$$

where a and b are regression coefficients.

Multivariables Regression Method (MLR)

In this method, the monthly discharge of station X is estimated using stations A, B… as following equation:

$$Q_X = a + bQ_A + cQ_B + ... \quad (15$$

where a , b and c are regression coefficients.

Autoregressive Method (AR)

In this method, the monthly discharge of station X at time t is estimated using the monthly discharge of the same station in time t-1,t-2 ,…..t-k as following equation:

$$Q_t = a + a_1 Q_{t-1} + a_2 Q_{t-2} + ... + a_k Q_{t-k} \quad (16$$

where $a, a_1, a_2, ..., a_k$ are coefficients of equation.

Artificial Neural Networks Method (ANN)

An ANN consists of input, hidden and output layers and each layer includes an array of processing elements (Figure 4). A typical neural network is fully connected, which means that there is a connection between each of the neurons in any given layer with each of the neurons in the next layer.

A processing element is a model whose components are analogous to the components of an actual neuron. The array of input parameters is stored in the input layer and each input variable is represented by a neuron. Each of these inputs is modified by a weight whose function is analogous to that of the synaptic junction in a biological neuron. The processing element consists of two parts. The first part simply aggregates the weighted inputs; the second parts is essentially a nonlinear filter, usually called the activation function.

[DOR: 20.1001.1.16807073.2010.12.1.5.1]

Table 3. Evaluation of monthly discharge reconstruction methods based on RMSE index in the sub-basins of great Karoon River basin.

Sub-basin code	Reconstruction priorities						
	1	2	3	4	5	6	7
3-3-1	ANN	MLR	SLR	FR	NR	AR	GR
3-3-2	MLR	SLR	NR	ANN	FR	AR	GR
3-3-4	ANN	SLR	MLR	NR	FR	AR	GR
3-4-1	ANN	MLR	NR	FR	SLR	AR	GR
3-4-4	ANN	SLR	MLR	NR	AR	GR	FR

Training of an artificial neural network involves two phases. In the first phase or forward pass, the input signals propagate from the network input to the output. In the second phase or reverse pass, the calculated error signals propagate backward through the network, where they are used to adjust the weights. The calculation of input is carried out, layer by layer, in the forward direction. The output of one layer is input to the next layer. This training method is known as the standard back propagation training method that was used in this research.

RESULTS AND DISCUSSION

According to the values of calculated RMSE, the accuracy order for each method of data reconstruction was determined in the hydrometric stations of different groups. Therefore, whichever method has the least RMSE in a given station was chosen as the best method. In each of sub-basins, the best method was determined by following the aforementioned procedures in stations of each sub-basin whose results are shown in Table 3. Three out of eight sub-basins did not have suitable conditions for reconstruction and have been omitted in this table. According to Table 3, the MLR method was chosen as the best method in sub-basin 3-3-2 and ANN was selected in the other sub-basin, whereas the FR method was in order of accuracy fourth to seventh in all of the sub-basins.

Considering the frequency of each method in Table 2, the accuracy order of the methods was determined for the great Karoon River basin. For example, the ANN method was repeated four times as the first order of accuracy. Thus, the first method in reconstruction of monthly discharge data of great Karoon River basin was the ANN method. The other methods are presented with regard to their accuracy order in Table 4. According to this table, the ANN, SLR and MLR methods were chosen as the first to third order of accuracy and FR was in fifth order of accuracy for reconstruction of the monthly discharge data of great Karoon River basin.

We can also determine the relative accuracy of monthly discharge reconstruction methods based on the nearness of the results to the line of $Y=X$. For example, Figures 5 to 10 show the results for sub-basin 3-3-1.

Table 4. Priority of monthly discharge reconstruction methods in great Karoon River basin.

Reconstruction method	ANN	SLR	MLR	NR	FR	AR	GR
Reconstruction priority	1	2	3	4	5	6	7

[DOR: 20.1001.1.16807073.2010.12.1.5.1]

CONCLUSION

In this research, seven methods were used for the reconstruction of monthly discharge data of great Karoon River basin. These methods, in order of accuracy, are the ANN, SLR, MLR, NR, FR, AR and GR methods. Thus, FR method is not a suitable method for reconstruction of monthly discharge data in the studied river basin. This method may be applied for reconstruction of annual discharge data which is another research subject conducted by the authors.

ACKNOWLEDGEMENTS

The authors appreciate Organization of Water Resources Management of Iran for their financial support for this research.

REFERENCES

1. Sadatinejad, S. J. 1997. Statistical Comparison of the Reconstruction Methods of Precipitation Gata in Isfahan Province. MSc. Thesis, Tarbiat Modares University, Iran.
2. Jamab Consulting Engineers Company. 1999. *Water Comprehensive Plan of Iran*.
3. Alizadeh, M. and Selvitabar, A. 1990. *Reconstruction of Discharge River*. Mahab-Ghods Consulting Engineers Company, Iran.
4. Lookzadeh, S. 2004. Evaluation of Several Different Reconstruction Methods for Precipitation Data in Central Alborz Region. MSc. Thesis, Tehran University, Iran.
5. Abeb, A. J., Solomatine, D. P. and Vennerker, R. G. W. 2000. Application of Adaptive Fuzzy Rule-based Methods for Reconstruction of Missing Precipitation Events. *Hydrol. Sci. J.,* **45(3):**425-436.
6. Bardossy, A. 1990. Fuzzy Regression in Hydrology. *Water Resour. Res.,* **26**: 1497-1508.
7. Bardossy, A., Bogardi, I. and Duckstein, L. 1990. Note on Fuzzy Regression. *Fuzzy Sets Syst.,* **37**: 65-75.
8. Chang, Y-H. O. and Ayyub, B. M. 2001. Fuzzy Regression Methods: A Comparative Assessment. *Fuzzy Sets Syst.*, **119(2):**187-203.
9. Chin, D. A. 1995. A Scale Model of Multivariate Rainfall Time Series. *J. Hydrol.*, **168(1):** 4-15.
10. Ishibuchi, H. 1992. Fuzzy Regression Analysis. *Fuzzy Theory and Systems*, **4**: 137-148.
11. Lohani, A.K., Goel, N. K. and Bhatia, K. K. S. 2006. Takagi-Sugeno Fuzzy Inference System for Modeling Stage-discharge Relationship. *J. Hydrol.,* **333**: 146-160.
12. MacCulloch, J. A.W. and Booth, M. 1970. Estimation of Basin Precipitation by Regression Equation.*Water Resour. Res.,* **16(6):** 1753-1758.
13. Sanchez, J. de A. and Gomez, A. T. 2003. Applications of Fuzzy Regression in Actuarial Analysis. *Journal of Risk and Insurance*, **70(4):** 797-802.
14. Savic, D. and Pedrycz, W. 1991. Evaluation of Fuzzy Regression Models. *Fuzzy Sets and Systems*, **39:** 51-63.
15. Selmins, A. 1987. Least Squares Model Fitting to Fuzzy Vector Data. *Fuzzy Sets Syst.*, **39:** 51-63.
16. Selvam, A. M., Pethkar, J. S. and Kulkarni, M. K. 1992. Signatures of Universal Spectrum for Atmospheric Interannual Variability in Rainfall Time Series the Indian Region. *J. Climatography*, **12(2):** 137-152.
17. Tanaka, H., Uejima, S. and Asai, K. 1982. Linear Regression Analysis with Fuzzy Model. *IEEE Trans. Syst. Man Cybern.,* **12(6):** 903-907.

[DOR: 20.1001.1.16807073.2010.12.1.5.1]

بررسی کارائی روش رگرسیون فازی در بازسازی داده های دبی ماهیانه ایستگاهای هیدرومتری حوزه آبخیز رودخانه کارون بزرگ

س. ج. ساداتی نژاد، م. شایان نژاد و ا. هنر بخش

چکیده

روش‌های متعددی برای باز سازی داده های هیدرولوژی وجود دارد که بسته به شرایط هر ایستگاه ممکن است یک روش خاص بهترین نتیجه را در پی داشته باشد. معمولاً برای برآورد داده‌های گمشده در یک ایستگاه, از ایستگاه‌های مجاور آن که داری تشابه هیدرولوژیکی, کلیماتولوژی و یا فیزیوگرافی هستند, استفاده می‌شود. اخیرا روش رگرسیون فازی برای بازسازی داده های هیدرومتری بکار رفته است. در این مقاله کارایی این روش در مقایسه با سایر روشهای مختلف بازسازی داده های دبی ماهانه ایستگاههای هیدرومتری حوزه آبریز کارون بزرگ مورد ارزیابی قرار گرفته است. در این تحقیق از روش حذف اعتباری استفاده گردید و پس از حذف عمدی داده‌های مشاهده‌ای, مقادیرآنها از طریق روش‌های مختلف برآورد گردید سپس با استفاده از آماره جذر میانگین مجذور مربعات خطا(RMSE) مناسبترین روش بازسازی تعیین گردید. نتایج تحقیق نشان داد که برای بازسازی دبی ماهانه ایستگاههای هیدرومتری حوضه کارون بزرگ، بترتیب اولویت عبارتند از: روش شبکه عصبی مصنوعی، روش رگرسیون خطی ساده، روش رگرسیون چند متغیره، روش نسبت نرمال، روش رگرسیون فازی روش اتو رگرسیو و روش گرفیکی.

[DOR: 20.1001.1.16807073.2010.12.1.5.1]