The Iraqi Journal For Mechanical And Material Engineering, Vol.16, No4, 2016

STOCHASTIC ANALYSIS OF DAILY INFLOW OF (HILLA RIVER) IN BABYLON PROVINCE

Nesrin J. AL-Mansori[1] **Saif S. ALquzweeni[2]** **Rasha S. AL-kizwini[3]**
nassrin20052001@yahoo.com **saif_alkizwini2012@yahoo.com** **rasha_alkizwini2015@yahoo.com**

[1,3]Environmental Engineering Dept, Collage of Engineering, University of Babylon.,
[2]Civil Engineering Dept, Collage of Engineering, University of Babylon.,

ABSTRACT :

The daily time series (stochastic) of the inflow discharge of (Al-Hilla River) at Babylon was analyzed as a time series. The data used for the analysis was the daily series during (2004-2014). The series was tested for non-homogeneity and found to be nonhomogeneous. A significant positive jump was observed at 2009. This non-homogeneity was removed, the homogeneous series was then normalized transformation. The periodic component of the series was fitted using harmonic analyses, and removed from the series to obtain the dependent stochastic component. This component was then modeled using first order autoregressive model (Markovian chain). The above analysis was conducted using the data for the period (2004-2012), the remaining two-years (2013-2014) of the observed data was left for the verification of the model. The observed model was used to generate future series. Those series were compared with the observed series using t-test. The comparison indicates the capability of the model to produce acceptable future data.

KEYWORDS ; daily discharge, nonhomogeneous, periodic component, and dependent stochastic.

التنبؤ بالبيانات اليومية لتصاريف شط الحلة المار في مدينة بابل

ا.م. د. نسرين جاسم حسين المنصوري **ا.م. سيف صلاح القزويني** **م. رشا صلاح القزويني**

الخلاصـــــة :

تم في هذا البحث تحليل البيانات اليومية لتصاريف شط الحلة في محافظة بابل (مدينة الحلة) ، حيث تم اخذ المعدل الشهري للبيانات المتوفرة من سنة (2004-2014).(دائرة الموارد المائية- بابل، بيانات). وتم ازالة عدم التجانس في البيانات وتم اختبار البيانات لمعرفة فيما اذا كانت متجانسة ام غير متجانسة ووجد انها غير متجانسة عند سنة 2000. ثم تم توزيع البيانات توزيعا طبيعيا. بعد ذلك تم اخذ القيم المعدلة لإزالة المركبة الدورية عنها وذلك للحصول على الدالة المستقلة التي تم نمذجتها بموديل من نوع سلسلة Markovian. ان التحليل اعلاه تم تطبيقه على البيانات من سنة 2004-2012 حيث تركت قيم التصاريف للسنتين المتبقيتين (2013-2014) لاستخدامها في التحقق من نتائج الموديل . ومن خلال النموذج الذي تم التوصل اليه تم التنبؤ بقيم لسنتين مستقبليه لمقارنتها مع القيم المقاسة وباستخدام اختبار t-test وقد بينت النتائج امكانية الاعتماد على النموذج لا عطاء نتائج مستقبلية مقبولة .

INTRODUCTION :

Most of activities associated with the operation, design, and planning of future water schemes require forecasts of future inflow. Efficient long term operation of a stream could be achieved using forecasted monthly inflows, while effective short term operation of a stream requires forecasted daily inflows. Short term operation of a stream is much important during rainy months rather than during drought months. This is due to the high variation of daily inflow during these rainy months. However, forecasting low daily inflow values is also important for proper schemes operation to meet the downstream demands. Hence, providing models for estimating daily inflows is essential for any operation. [Al-Suhaili,1985]

During the last 25 years, Time Series Analysis had become one of the most important and widely used branches of Mathematical Statistics. The technique of time-series analysis uses estimated statistical parameters to build a mathematical model. This model is capable of describing the evolution of possible sequences of events in time, at the site of observations, which have the same statistical properties as the historical sample. [Al-Suhaili,1985].

In this research, the data of daily discharge in (Al-Hilla River) from Euphrates River will be utilized to build the mathematical model to predict the future discharge. Numbers of projects were by now under construction in the southern of Iraq after 2003, such as water treatment plants and projects related to restoration of marshes and irrigation projects. The modeling of daily discharge could be used to predict future values, that are useful in the operation of such project. Moreover these data are useful also in planning, design of new projects. While [Al-Suhaili,1985] used daily stream flow data of Tigris river to build stochastic model. The data were collected from four measurement station (Mosul, Fatha region , Baghdad, Kut) located downstream of a river. Two different single site model were used to detect the changes in the stream flow data ,Autoregressive.(AR) and Autoregressive Integrated Moving Average.(ARIMA) models and a multisite model (MATLABS).

[Mustafa, et al. 2012] used modelling technique (Artificial neural network) represent the non-liner relationship among the input and output variables of a water resources system by using a grouped neurons or nodes in layers

[Zhiyong and Hiroshi ,2006] designed a structure of input/mixing/output model (called it the three-step model) and represented each process of material input, mixing and output by a stochastic numerical methods. Data of (Biological Oxidation Demand) BOD_5 and (Dissolved Oxygen) DO concentrations of river Cam used to build the model. The stochastic numerical methods applied in this paper are discrete time approximation methods. Additional equations representing mixing processes and biochemical reactions also used in model. The researcher concluded that the theory of stochastic differential equations is a beneficial tool for studying water pollution. [Kadriand and Ahmet 2006] had analyzed the daily discharge data of each month from three gauge stations on Cekerek Stream for forecasting using stochastic approaches. Initially non-parametric test (Mann-Kendall) was used to identify the trend during the study period. The two approaches of stochastic modeling, ARIMA and Thomas-Fiering models were used to simulate the monthly-minimum daily discharge data of each month. The error estimates (RMSE:Root Mean Square Errorand MAE:Mean Absolute Error) forecasts from both approaches were compared to identify the most suitable approach for reliable forecast. The two error estimates calculated for two approaches indicate that ARIMA model appear slightly better than Thomas-Fiering model. However, both approaches were

identified as an appropriate method for simulating the monthly-minimum daily discharge data of each month from three gauge stations on Cekerek Stream.
Generally, a hydrologic time series may consist of four components depending on the type of variable and the averaging time interval. Daily discharge series has four components may exist and can be considered to arise from a combination of those components, which are termed the jump component (Jt), trend component (Tt), periodic or cyclic component (Pt) and stochastic or random component (εt).These components may be formulated by:

$$Qt= Jt +Tt +Pt +\varepsilon t \quad (1)$$

The first three components represents the deterministic part of the process while the fourth component represents the non-deterministic part, therefore those three components should be detected and identified by suitable formulations and decomposed from the stochastic component.

Aim of Research:

The daily time series(stochastic) of the inflow discharge of (Al-Hilla River) at Babylon was analyzed as a time series. The data used for the analysis was the daily series during (2004-2014). The observed model was used to generate future series**.**
In this research, the data of daily discharge in (Al-Hilla River) from Euphrates River will be utilized to build the mathematical model to predict the future discharge

Materials and Methods

The procedure used for data analysis may be summarized by the following steps:

1- Test and Removal of Non-Homogeneity

The modeling process required a set data to be homogenous. Hence, the first step before starting the analysis is to test the homogeneity of the data series. If the test indicates non-homogeneity, then this non-homogeneity should be removed. This was achieved by plotting the average monthly data and computing the annual mean and standard deviation for each year then using the spilt-sample approach which divides the entire sample into two sub-samples. Then testing the differences between the means and standard deviations of these two sub-samples at the 95 percent probability level of significance using the t-test method [Al-Suhaili,1985]. The data were tested for non-homogeneity and found to be non-homogeneous at year 2009, see **Fig. 1**. The calculated t-value is greater than the tabulated t-value.

For non-homogeneity removal, [Yevjevich,1972] suggest fitting linear regression equations for both annual means and annual standard deviations, then applying the following equation:-

$$YJ,t = \frac{XJ,t - XJ}{SJ}\ Sd2 + Av2 \quad (2)$$

where:
j, t: the annual and seasonal positions of the observations, respectively.
Y: transformed series (homogeneous)
X : historical non-homogeneous series.
Av_2,Sd_2: the average and standard deviation of the second sub –sample respectively, and
Xj, Sj : linear regression equations for annual means and standard deviations against years.

Thus the data are divided into two sub-samples, the first (6) years (2004-2009) and the second (5) years (2010-2014), The first sub-sample will be transformed according to the above equation. The two sub-samples were then tested again using t-test to check the homogeneity. The result of applying the above equation are shown in **Fig. 1** and **Table 1** below, hence the data are homogeneous. Since the calculated t-values is less than tabulated t-value.

2- Transformation to Normally Distributed Data

It is of common practice in time- series analysis to transform the data to the normal distribution. This means, to remove the skewness in the data and try to make it nearly zero. For the normalization process several transformations may be used to normalize the data, but the most common one is the power transformation. The power transformation used in this research is the one suggested by [Box, and Jenkins,1976] (see equation (4) below. The application of this transformation begins with the estimation of the transformation coefficient value (λ). This coefficient has a value between (-2) and (2) and is strongly related to the skewness coefficient (Cs).**Table 2** shows values of Cs computed for each transformed series, using different λ -values. [Al-Suhaili,1985]

The values above were found to best fitted by a second polynomial equations.

$$\lambda_Q = 0.0032Cs^2 + 0.2901Cs + 0.053 \qquad (3)$$

In order to find the λ-value that will normalize the data, the skewness coefficient Cs in the above equations are substituted by zero, which gives a λ value as 0.053This value will be used to get the transformed series according to the Box and Cox transformation as follows:

$$Y = \frac{(X-1)^{\lambda}}{\lambda} \qquad (4)$$

where: y: The transformed Series, x : The original Series Data. **Table 3** shows the monthly means and standard deviat6ions for the original and transformed series.

3- Determination of the Independent Stochastic Component

The series obtained after the removal of non-homogeneity and non-stationary (periodic component of mean and standard deviation) is termed as the dependent stochastic component

of the process and denoted as (εp,t). The values of monthly mean and standard deviations were used to find the value of the independent stochastic component by the following equation:-

$$\varepsilon_{p,t} = \frac{Y_{p,t} - \mu_t}{\sigma_t} \quad (5)$$

where: εp,t = is the dependent stochastic component. μt =is the mean value of yp,t data at position p(month). σt =is the standard deviation value of yp,t data at position p(month). The values of εp,t may be fitted by a suitable model whose parameters will depend directly or indirectly on the amount of existing correlation represented by the lag rk serial correlation coefficient model (rk), **Fig.2** One of the most familiar models, are the autoregressive model. It is preferable to try the first degree model, and then check its adequacy to remove the dependency of the εp,t series in the first degree model fails to remove the dependency, the second degree model will be used, and so on. The first degree autoregressive model required the calculation of lag-one (r_1) serial correlation coefficient, which was found to be (r_1=0.721)

The model is represented as the relation between the dependent stochastic component (εp,t) and the independent stochastic component (ζp,t). The independent stochastic series (ζp,t) is a series of random numbers usually with zero mean and unit variance.

As mentioned above one of the most used models is the first order autoregressive model (Markov model). This model express the relationship between the εp,t and ζp,t as follows:

$$\varepsilon_{p,t} = a * \varepsilon_{p,t-1} + \sqrt{1-a^2}\zeta_{p,t} \quad (6)$$

where: a= r_1, Substituting the value of a =(0.721) in the above equation (6) ,the independent stochastic component ζp,t, for both parameters could be found using:

$$\zeta_{p,t} = \frac{\varepsilon_{p,t} - 0.45626 * \varepsilon_{p,t-1}}{0.8898} \quad (7)$$

In order to test the validity of the proposed first order autoregressive model, the correlogram of the ζp,t component should be found and tested. This correlogram is shown below which show that the first order autoregressive model, is suitable since the values of the serial correlation coefficient are fluctuated around the zero-values. Hence the proposed model was capable of removing the dependency between the values of εp,t. **Fig. 3**.

4- Model Verification

Upon the completion of the first four steps above, the model parameters were found. As mentioned before the observed data series were divided into two parts (2004-2011), was used for the analysis (i.e., models parameters estimation), the other part (2012-2013), will be used now for model verification since the significant positive jump was observed at 2009, based on Seasonal Time Series. (Brock Well and Davis (1991))

Usually in practice, the model is used to generate future values (series). The model validity will be decided upon the comparison between the statistical properties of the generated series with those of the observed one that was not used in the estimation of the parameters of the model. [Wegman, 2000]

The Microsoft Excel program was used for generating future series as shown in Table 4. The generation process begins by generating a standardized normally distributed random series (i.e., with zero mean and unit variance) then, using those as ζp,t values to generate the εp,t values using the first autoregressive model. The daily discharge values were found using a reverse process of the analysis conducted in steps (2-4).

Table 5. show the generated monthly discharge values respectively using the two generated randomized series, proposed to be for years 2012 and 2013. The observed monthly discharge for these 2-years are shown in Table 6.

Table 7 shows the values of the mean, standard deviation, skewness coefficient and kurtoses coefficient of observed data and generated one with t-test for monthly discharge means, t tabulated was 3.02.

CONCLUSIONS :-

1) The series of daily discharge of AL-Hilla River from Euphrates River at Babylon Province is non-homogeneous.
2) The suitable value of the power transformation parameter λ that can be used to transform data to the normal distribution was found to be 0.053.
3) The correlogram of the observed independent stochastic component indicate the capability of the first order autoregressive model to model to time-dependency of the dependent stochastic component.
4) The T-test result shows that the obtained model can presence future forecasted values for the daily discharge of AL-Hilla River.

Table 1: Mean and standard deviation of each sub-samples before and after applying the procedure of removal of non-homogeneity.

Daily discharge. Data,Q	Before Removal		After Removal	
	Mean	Standard deviation	Mean	Standard deviation
Set 1	234.388	75.876	179.123	72.735
Set 2	179.986	72.172	178.110	72.679

Table 2: Variation of Skewness Coefficient with Box and Cox transformations Coefficient for Daily Discharge Q.

λand Cs for Q								
λ	-0.6	-0.4	-0.2	0.2	0.6	0.8	1	1.2
Cs	-1.73	-1.4	-1.43	-0.73	-0.38	-0.23	-0.09	0.03

Table 3: Monthly means and standard deviations for the original homogeneous series and normalized series for Q in a period from 2004 to 2012.

Mon.	Original homogeneous Series(x)		Normalized series(y)	
	Mean	St.d	Mean	St.d
Jan.	149.491	42.454	27.6	0.19
Feb.	147.403	41.240	27.6	0.34
Mar.	170.973	78.138	27.6	0.12
Ap.	148.981	59.352	27.3	0.32
May	160.039	62.245	27.3	0.28
Jun.	294.517	78.831	27.6	0.18
Jul.	345.298	95.977	27.5	0.25
Aug	274.423	75.404	27.5	0.29
Sep.	261.368	64.218	27.5	0.16
Oct.	234.865	72.156	27.4	0.34
Nov	167.613	47.689	27.4	0.23
Dec.	160.948	48.002	27.4	0.20

Table 4: Daily Discharge Q Values of generated randomized numbers ($\zeta_{p,t}$) for 2013,2014 years.

months	$Rand_1$		$Rand_2$	
Jun.	0.287	0.265	0.125	0.289
Feb.	0.193	0.374	0.337	0.584
Mar.	0.038	0.638	0.604	0.787
Apr.	0.733	0.882	0.129	0.497
May.	0.811	0.670	0.677	0.178
Jun.	0.234	0.634	0.095	0.270
July.	0.890	0.823	0.343	0.613
Aug.	0.963	0.221	0.839	0.056
Sep.	0.234	0.563	0.704	0.564
Oct.	0.028	0.833	0.898	0.562
Nov.	0.901	0.365	0.571	0.762
Dec.	0.230	0.878	0.515	0.875

Table 5: generated monthly discharge (m^3/s) for 2013,2014 years

months	Q. $Rand_1$		Q. $Rand_2$	
	2012	2013	2012	2013
Jun.	794.9	897.8	794.9	887.8
Feb.	832.8	933.5	808.2	893.4
Mar.	818.0	889.0	815.3	886.0
Apr.	856.7	949.3	760.6	967.9
May.	1031.	1103.	992.0	1061.
Jun.	998.4	1019.	880.1	971.8
July.	933.1	1000.	923.0	1011.
Aug.	1017	1049.	1007.	1078.
Sep.	940.3	991.6	1028.	1008.
Oct.	931.9	1041.	1034.	1042.
Nov.	1044.2	1009.7	1049.	1005.
Dec.	1055.5	934.3	1052.	946.5

Table 6: observed discharge (m^3/s) for 2013,2014 years.

Year	JUN.	FEB.	MAR.	AP.	MAY.	JUN.	JUL.	AUG.	SEP	OCT	NOV.	DEC.
2013	115.6	137.5	242.9	268.1	278.3	355.9	303	131.2	127.6	150	117.2	131
2014	171	124.2	157.1	248.8	284.8	357.6	342.6	240.4	103	171	124.2	157.1

Table 7: statistical properties of observed data and generated series for Q

Mean	Observed data	$Rand_1$	$Rand_2$
	130.16	166.3603	162.1797
Standard deviation	13.26213	7.95224767	7.55870447
skeweness	0.483443	-0.0530578	-0.2132529
kurtoses	-0.27098	-0.7890046	-1.4118451

Fig.1 Split sample test of the original historical Data for Q.

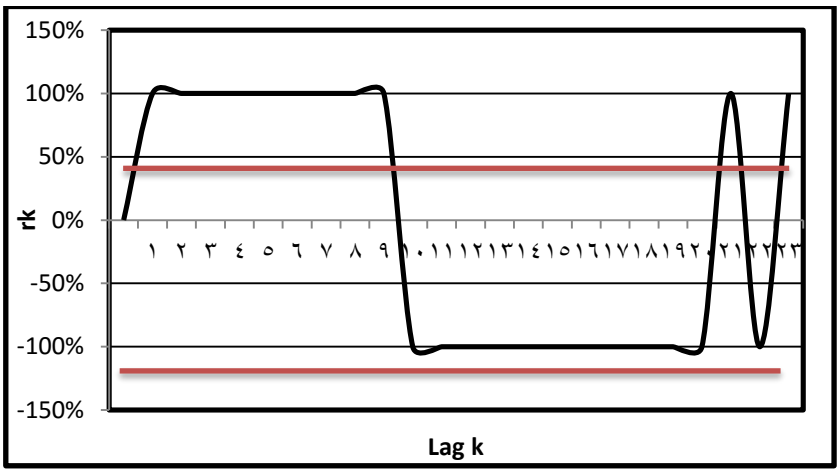

Fig.2.corrologram of the dependent stochastic component ($\varepsilon_{p,t}$), for Q data.

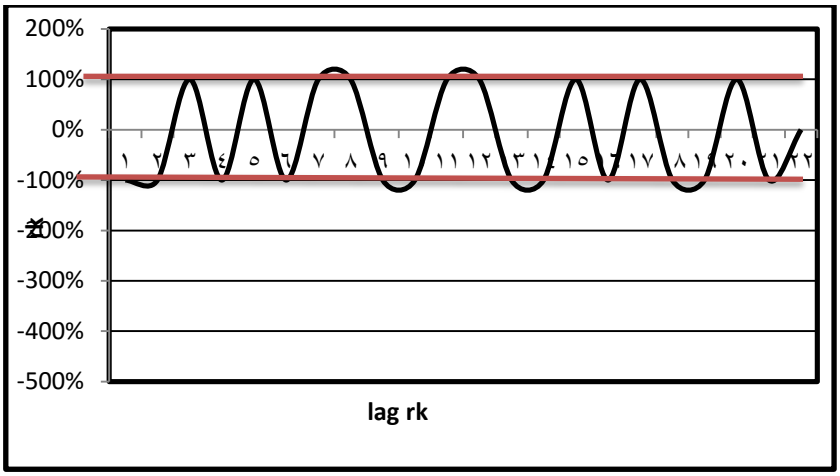

Fig.3.corrologram of the independent stochastic component ($\zeta_{p,t}$), for Q data.

REFERENCES :-

Box, G.E.P. and Jenkins,G. M. Time Series Analysis, Forecasting and Control. Holdened Day, San Fransisco. P.575.1976.

Brock Well , P.J. and Davis , R.A. (1991) , " Time Series Theory and Methods " , 2nd ed , Springer Verlag New York Inc , New York .

H. Zhiyong, and Hiroshi, M. Water Pollution Models based on Stochastic Differential Equations. Department of Earth and Environmental Sciences Graduate School of Environmental Studies, Nagoya University, PDF, Internet. 2006.

[Mustafa,M. R.; Isa, M. H. and Rezaur, R. B. Artificial neural networks modeling in water resources engineering: infrastructure and applications. World Academy of Science, Engineering and Technology, Vol. 6, No. 2,pp. 317-325. 2012.

Al-Suhaili, R. H. Stochastic Analysis of Daily Stream flow of Tigris River. M.Sc. Thesis, College of Engineering, University of Baghdad. 1985.

Yevjevich,W. M. Structural Analysis of Hydrology time Series, Hydrology Paper No.56,Fort Collins, Colorado, Nov. 1972.

Kadriand, Y. and Ahmet, K. Performances of Stochastic Approaches in Generating Low Stream flow Data for Drought Analysis". Journal of Spatial Hydrology Vol. 5, No. 1. Internet. 2006.

Wegman , E.J. (2000) , " Time Series Analysis – Theory , Data Analysis and Computation " , Addison-Wesley Publishing Company .