International Journal of Natural and Engineering Sciences 2 (1): 21-27, 2008
ISSN: 1307-1149, www.nobel.gen.tr

Stochastic Modeling of Suspended Sediment from Yesilirmak Basin, Turkey

Fazli OZTURK[1*] Kadri YUREKLI[2] Halit APAYDIN[1]

[1] Ankara Univ., Faculty of Agriculture, Dept. of Farm Structures and Irrigation, 06110 Dıskapi – Ankara/TURKEY

[2] Gaziosmanpasa Univ., Faculty of Agriculture, Dept. of Farm Structures and Irrigation, 60250 Tasliciftlik – Tokat/TURKEY

[*] Corresponding Author
e-mail: fozturk@agri.ankara.edu.tr

Received: June 18, 2007
Accepted: September 30, 2007

Abstract

This paper presents a methodology on modeling of monthly-suspended sediment data from the Yesilirmak Basin, Turkey. For this purpose, linear stochastic models known as either Box-Jenkins or ARIMA (autoregressive-moving average) were used to simulate monthly sediment data. Diagnostic checks were done for all the models selected by considering the autocorrelation function (ACF) and partial autocorrelation function (PACF). The models that have the minimum Schwarz Bayesian Criterion (SBC) and fulfilled all the diagnostic checks were assumed to best represent the monthly sediment data for sediment-sampling stations in the Yesilirmak Basin. Modeled data were compared to observed data for a 10-year period, confirming that the generated data preserve the basic statistical properties of the original series.

Key words: Yesilirmak Basin, Turkey, suspended sediment, ARIMA model, Box-Jenkins.

INTRODUCTION

Approximately 154 billion cubic meters of soil is annually removed from Turkey by water erosion [1]. In this sense, soil and water loss is one of major problem in Turkey [2,1]. It has been recognized that soil loss from agriculture land is in itself a serious economic loss. Transported sediment by water erosion is subsequently deposited in stream channels, lakes, reservoir, and harbors, requiring costly remedial measures.

Sedimentation has a major impact on the useful life and economic viability of many reservoirs by causing a reduction in the storage capacity. Also with deposition at the head of reservoir, it can cause an increase in flood levels in the contributing rivers upstream. Therefore, it is important for the engineer to make an assessment of the likely quantity of sediment deposits throughout the design life of a reservoir. The useful life may be computed by determining the total time required to fill the critical storage volume. In addition, sediment transport to the lakes, reservoirs and rivers may adversely affect the aquatic life.

The annual sediment load of a stream is an important factor for determining dead storage volume of a dam. The annual sediment load of the stream is generally determined either from direct measurements of the sediment load throughout the year or from any of the many sediment transport equations that are available today. Direct measurement of the sediment load in a stream, which is the most reliable method, is very expensive and thus is not done for as many streams as the measurement of water discharge. On the other hand most of the sediment transport equations require detailed information on the flow and sediment characteristics and generally do not agree with each other making it difficult to choose the best equation for a given stream. Because of these problems researchers are always looking for simpler and easy to use relationships between sediment load and water discharge or drainage area of the stream [3].

Runoff and sediment load information can provide insights into catchment degradation especially in relation to identifying critical erosion source areas. As catchments remain in a dynamic state, the runoff and sediment load fluctuate over time. If sufficiently long time series data are available it may be possible, to some extent, to identify the reliable maximum and minimum values and the underlying distribution. However generally such long time series data, particularly for small catchments, are not available. It is, therefore, necessary to deploy some means to simulate sufficiently long time series data with similar statistical characteristics as that of the historical data.

Due to a lack of long-term sediment concentration data, stochastic time series models of sediment yield (t/yr) have not been attempted as widely as runoff records. Rice [4] and Hadley et al. [5] have strongly advocated the need for adopting stochastic approach for modeling sediment yield. Gurnell and Fend [6] turned autocorrelation to an advantage by estimating a Box-Jenkins transfer function between discharge and suspended sediment concentration series. Univariate autoregressive integrated moving average (ARIMA) stochastic time series models of discharge and suspended sediment concentration series were developed for an estimation period, and the simple transfer function used to bring them into phase and apply a scaling factor. The present study is an attempt to apply the ARIMA (Box-Jenkins) model in simulating sediment load data of the Yesilirmak Basin.

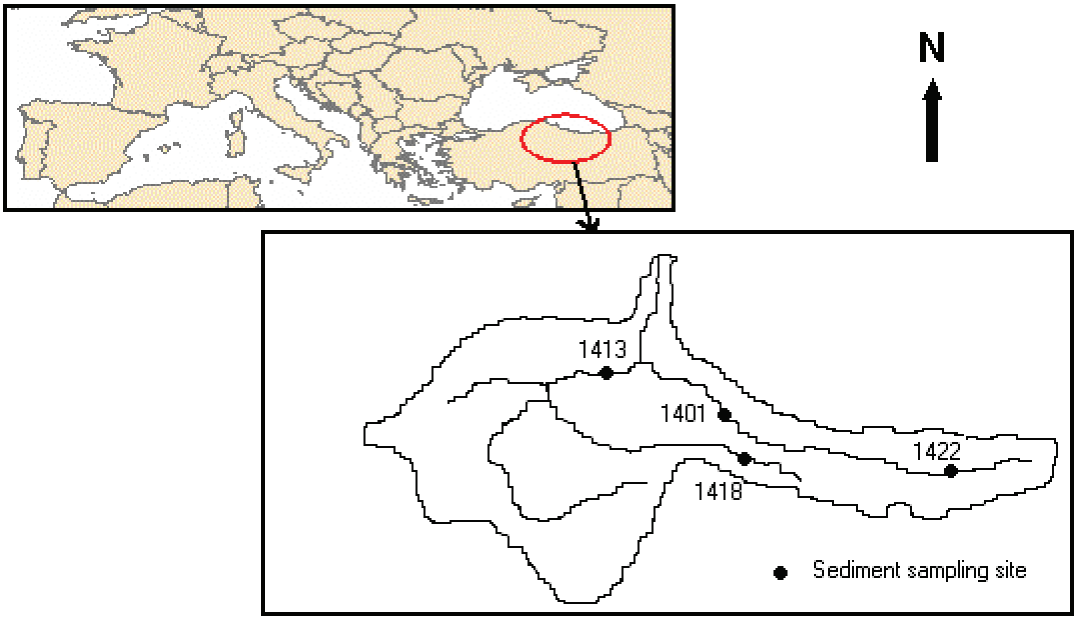

Figure 1. Sediment-sampling sites within Yesilirmak Basin

MATERIAL AND METHOD

Study area

Monthly sediment data from four sampling stations (numbered 1401, 1413, 1418 and 1422), which are managed by General Directorate of Electric Power Research Survey and Development Administration (EIE), in the Yesilirmak Basin, were used as the input data for this study. The approximate locations of the sediment-sampling stations are given in Figure 1 and a summary of identification number and names and drainage areas for the sediment-sampling stations are presented in Table 1.

Table 1. Sediment-sampling stations identification in Yesilirmak Basin

Station Number	Station Name	Drainage Area, km²	Number of Years of Data
1401	Kelkit-Fatli	10048.8	20
1413	Yesilırmak-Durucasu	21667.2	20
1418	Yesilırmak-Gomeleonu	1608.0	14
1422	Kelkit-Cicek Buku	1714.0	15

Yesilirmak Basin is bounded between a latitude 39º 30' and 41º 21' N and 34º 40' and 39º 48' E longitude. It covers approximately 3 873 280 ha, equivalent to about 5% of Turkey's total area. Land use categories equate to around 39 % forest, 39 % cropland and 19 % pasture in the basin. Yesilirmak Basin is located on the north Anatolia fault line that is one of the most effective faults in the world. This basin has been subjected to phases of major tectonic activity (Hercynian, and Alpine Orogeny) as well as epiorogenic movements. Therefore, bedrock is extensively faulted and folded. Major lithologies include sandstone, claystone, andesite, volcanic bressica and tuffs. The headwaters of the Yesilirmak River and most of its tributaries originate in the mountains that form the eastern and southern boundaries of the study area reaching elevation of 2800 m. The major tributaries to the Yesilirmak River are Kelkit, Cekerek, Corum Cat and Tersakan Streams. The length of the main river channel is about 519 km [7].

Time Series Analysis for Monthly Sediment Data

In order to analyze the monthly time series from the four sediment-sampling stations, a linear stochastic model, known as either Box-Jenkins or ARIMA, was used in this study. The seasonal ARIMA model [8] can be written as:

$$Ø(B)\Phi(B^s)(w_i - \mu) = \theta(B)\Theta(B^s)a_i \qquad (1)$$

$$w_i = (1-B)^d\,(1-B^s)^D\,x_i \qquad (2)$$

In equation 1, w_i should be taken as z_i if the series is stationary.

Box and Jenkins [9] recommended that model development consists of three stages (identification, estimation and diagnostic check) when an ARIMA model is applied to a particular problem. The identification stage is purposed to determine the differencing required to produce stationarity and also the order of both the seasonal and nonseasonal autoregressive (AR) and moving avarage (MA) operators for a given series. By plotting original series (monthly series), seasonality, trends in the mean and variance may be revealed. The non-parametric Spearman's Rho test can be applied to decide whether trend exists in the monthly data [10]:

$$R_{sp} = 1 - \frac{6 * \sum D_i^2}{n * (n^2 - 1)} \qquad (3)$$

$$D_i = K_{xi} - K_{yi} \qquad (4)$$

$$t_{cal} = R_{sp} * \left[\frac{(n-2)}{(1-R_{sp}^{2}} \right]^{1/2} \qquad (5)$$

To determine whether there is a trend, the t_{cal} value in equation 5 should be compared to the t-table critical value. If the t_{cal} value lies within the 5% significance interval, there is no trend for the data set.

Autocorrelation function (ACF) and partial autocorrelation function (PACF) should be used to gather information about the seasonal and nonseasonal AR and MA operators for the monthly series [11]. ACF measures the amount of linear dependence between observations in a time series. In general, for a MA (0,d,q) process, autocorrelation coefficient (r_k) with the order of k cuts off and is not significantly different from zero after lag q. If r_k tails off and does not truncate, this suggests that an AR term is needed to model the time series. When the process is a MA (0,d,q)x(0,D,Q), r_k truncates and is not significantly different from zero after lag q+sQ. If r_k attenuates at lags that are multiples of s, this implies the presence of a seasonal AR component. For an AR (p, d, 0) process, partial autocorrelation coefficient ($ø_{kk}$) with the order of k truncates and is not significantly different from zero after lag p. If $ø_{kk}$ tails off and this implies that a MA term is required. When the process is an AR (0, d, q) x (0, D, Q), $ø_{kk}$ cut off and is not significantly different from zero after lag p+sP. If $ø_{kk}$ damps out at lags that are multiples of s, this suggests the incorporation of a seasonal MA component into the model. The ACF for seasonal series should not exceed a maximum lag of approximately 5s (5s < n/4). PACF are usually calculated for 20 to about 40 lags (40 < n/4). For seasonal models, higher lags of the PACF may be required for identification.

Estimation stage consists of using the data to estimate and to make inferences about values of the parameters conditional on the tentatively identified model. In an ARIMA model, the residuals (a_i) are assumed to be independent, homoscedastic, and usually normally distributed. However, if the constant variance and normality assumptions are not true, they are often reasonably well satisfied when the observations are transformed by a Box-Cox transformation. The transformations can be expressed as either of the following equations [12]:

$$z_{i=1}^{n} = \lambda^{-1}\left[\left(x_{i=1}^{n} + c\right)^{\lambda} - 1\right] \qquad \lambda \neq 0 \qquad (6)$$

$$z_{i=1}^{n} = \ln\left(x_{i=1}^{n} + c\right) \qquad \lambda = 0 \qquad (7)$$

Box and Jenkins [9] cited that the model should be parsimonious. Therefore, they recommended the need to use as few model parameters as possible so that the model fulfils all the diagnostic checks. Akaike [13] suggests a mathematical formulation of the parsimony criterion of model building as AIC (Akaike Information Criterion) for the purpose of selecting an optimal model fit to a given data. Mathematical formulation of AIC is defined as:

$$\text{AIC (M)} = n \ln\sigma_a^2 + 2M \qquad (8)$$

Where M is the number of AR and MA parameters to estimate. The model that gives the minimum AIC is selected as a parsimonious model.

Shibata [14] has shown that the AIC criterion tends to overestimate the order of the autoregression. Akaike [15,16] developed a Bayesian extension of minimum AIC procedure, termed BIC. Similar to Akaike's BIC, Schwarz [17] suggested the Schwarz Bayesian Criterion (SBC):

$$\text{SBC (M)} = n \ln\sigma_a^2 + M \ln n \qquad (9)$$

Diagnostic checks determine whether residuals are independent, homoscedastic and normally distributed. The residual autocorrelation function (RACF) should be obtained to determine whether residual are white noise. There are two useful applications related to RACF for examining residuals. The first one is the periodogram drawn by plotting r_k (a) against lag k. If some of the RACF are significantly different from zero, this may mean that the present model is inadequate. The second one is Q (r) statistic suggested by Ljung-Box [18]. A test of this hypothesis can be done for the model adequacy by choosing a level of significance and then comparing the value of calculated χ^2 to χ^2-table of critical value. If the calculated χ^2 value is smaller than the χ^2-table critical value, the present model is adequate on the basis of available data. The Q (r) statistic is calculated by using:

$$Q(r) = n(n+2)\sum_{k=1}^{m}(n-k)^{-1} r_k(a)^2 \qquad (10)$$

$$r_k(a) = \sum_{i=k+1}^{n} a_i a_{i-k} \Big/ \sum_{i=1}^{n} a_i^2 \qquad (11)$$

The following test described by Breusch and Pagan [19] is very useful to determine whether a transformation of the data is needed. If there is a change in variance (heteroscedasticity) of residuals, a transformation is necessary for the data. For the test, the residuals from the model fit to the data are divided into two groups. Then, residual sum of squares (ESS_F, ESS_S) for these group are obtained. Breusch-Pagan test statistic (F_{cal}) is obtained from the following equation. If F_{cal} is smaller than F-table critical value, the residuals are assumed to be homoscedastic.

$$F_{cal} = \frac{ESS_S/(n_S - k_p)}{ESS_F/(n_F - k_p)} \approx F_{table}\left[(n_S - k_p),(n_F - k_p)\right] \qquad (12)$$

There are many standard tests available to check whether the residuals are normally distributed. Chow et al. [20] cited that if historical data are normally distributed, the graph of the cumulative distribution for these data should appear as a straight line when plotted on normal probability paper. Haan [21] expressed that the other way to check normality of residuals is the Kolmogorov-Smirnov method.

RESULTS AND DISCUSSION

To determine whether there is a trend in monthly data sequences from four sediment-sampling stations, the non-parametric test (Spearman's Rho test) at 5% significance level was applied to monthly series. The Spearman's Rho test results are given in Table 2. The t_{cal} value (-1.52) of data from sampling station 1422 is between t-table critical values (±1.96) at 5% significance level. This suggests that there is not a linear trend in this data sequence. However, the t_{cal} values of other stations

Table 2. The ARIMA models selected for the transformed data from the sediment sampling stations

Sampling Station	ARIMA Model	Model Statistics						
		Trend	AIC	SBC	Q (r)/p	K-S	B-P(F_{cal})	σ_a^2
1401	(0,1,1)(1,1,1)	-3.59	752.5	762.8	0.852	0.486	0.865	1.43
1413	(1,1,1)(0,1,1)	-3.85	721.7	732.0	0.414	0.662	0.797	1.22
1418	(0,1,1)(1,1,1)	-5.07	542.5	551.7	0.234	0.556	0.859	1.66
1422	(0,0,0)(0,1,1)	-1.52	527.7	530.8	0.721	0.379	1.130	1.22

AIC, Akaike Information Criterion
SBC, Schwarz Bayesian Criterion
Q (r)/p, Probability of Ljung-Box Q Statistic
K-S, Kolmogorov-Smirnov Statistic
B-P(F_{cal}), Breusch-Pagan Test statistic

(1401, 1413 and 1418) were not between t-table critical values (±1.96) at a 5% significance level, suggest these stations have linear trends (Table 2).

The plots of the monthly data sequences reveal important information about the sediment data. The periodic peaks in the data reflect the seasonality of the observations. These plots for the sediment data of the sampling stations 1401, 1413 and 1418 also show that there is a linear component present. The plots of the ACFs drawn for the data sequences are examined in order to identify the form of the ARIMA model to estimate. The ACFs for monthly data follow an attenuating sine wave pattern that reflects the random periodicity of the data and possibly indicates the need for non-seasonal and/or seasonal AR terms in the model. For these type of series, the cyclic seasonal component was removed by taking the seasonal differencing operator as one (1). Similarly, for the data from sediment-sampling stations 1401, 1413 and 1418, the non-seasonal differencing operator was taken as one (1) to remove trend from the data set.

ACFs and PACFs were estimated for the monthly data. All the ACFs were significantly different from zero. Additional to this, Ljung-Box Q statistics were estimated. They emphasis that the ACFs obtained from the monthly data sequences are significantly different from zero; in other words, there was a linear dependence between monthly observations. However, the ACFs did not cut off but rather damp out. This may suggest the presence of autoregressive (AR) terms. The PACFs possess significant values at some lags but rather tail off. This may imply the presence of moving average (MA) terms. The ACFs have significant values at lags that are multiples of 12. This may stress that seasonal AR terms are required but these values attenuate. There are peaks on graphs of the PACFs at lags that are multiples of 12 may suggest seasonal MA terms, but these peaks damp out.

Based on these conclusions, alternative ARIMA models were estimated by considering the ACFs and PACFs graphs from the monthly data. The SBC was taken into account for obtaining a parsimonious model among these alternatives.

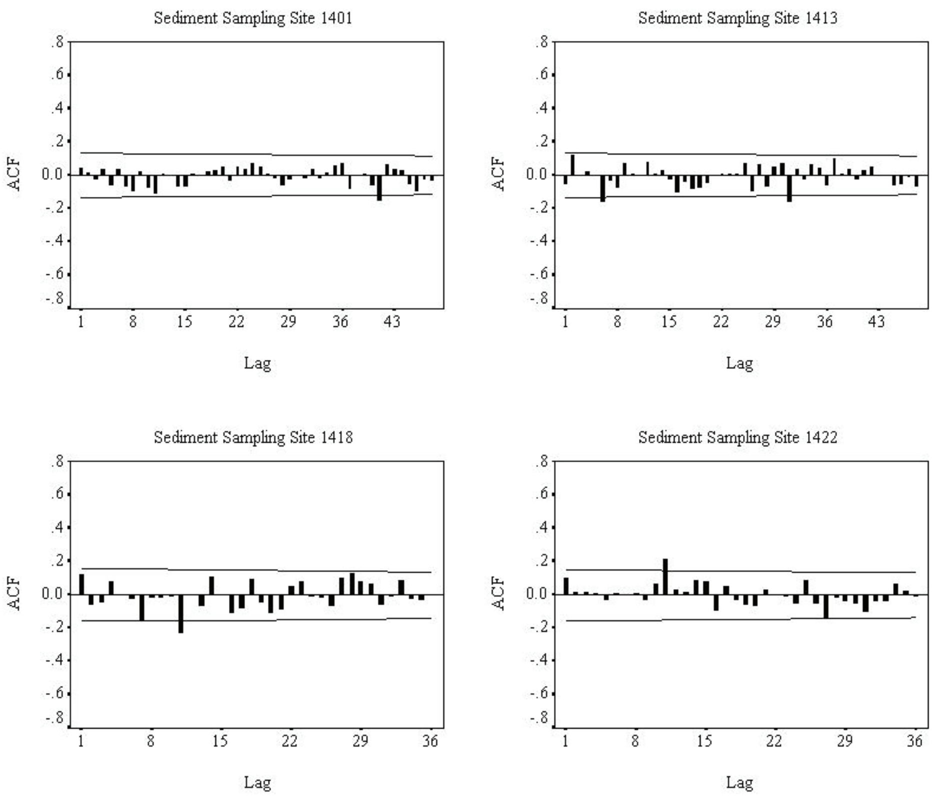

Figure 2. Residual ACF-suspended sediment data

The model that has the minimum SBC was assumed to be parsimonious. In addition to this, model parameters were analyzed at a 5% significance level by using t-tests to select the best model fit to the data. If there is any parameter greater than 5%, they were eliminated.

Diagnostic checks were applied in order to determine whether the residuals of the selected models from the ACF and PACF graphs were independent, homoscedastic and normally distributed. All the identified possible models using original monthly data for the sediment-sampling stations failed the diagnostic checks. Therefore, a Box-Cox transformation was required for original monthly sediment data of each station. By substituting λ and c = 0.0, as zero (0.0) for monthly data from all sediment-sampling stations in equations (6) and (7), a Box-Cox causes the residuals to be homoscedastic and approximately normally distributed.

The models that have the minimum SBC among the selected models fulfilled all the diagnostic checks were selected as the best model for monthly data from the sediment-sampling stations. The selected best models are presented in Table 2. The critical assumption of independence for the RACFs of the residuals was done by using the χ^2 distributed Ljung-Box Q statistic. The probabilities of Q statistics calculated for the best models were given in Table 2. Calculating the Ljung-Box Q statistic on the RACF, indicated that the residuals were not significantly different from a white noise series at 5% significance level. Inspection of the RACF and the residual integrated periodogram (Figure 2) confirmed a strong model fit.

In Table 2, test results from Kolmogorov-Simirnov method for the normality and test results from Breusch-Pagan for homoscedascity of the residuals are also presented. Table 2 shows that all the diagnostic checks for the residuals from each data set are completely fulfilled.

The value (V) of the parameters associated the standard errors (SEV), t-ratios and probabilities (<5%) for the standard errors were listed in Table 3. The standard errors calculated for the model parameters were rather small compared to the parameter values. Therefore, all of the parameters are significant and these parameters should be included in the models (Table 3).

Table 3. Statistical analysis for the model parameters

Sampling Station	Model Parameters	Variables in the Model			
		V	SEV	t-ratio	Probability
1401	θ_1	0.913	0.029	30.97	0.000
	Φ_1	-0.176	0.075	-2.34	0.020
	Θ_1	0.877	0.064	13.71	0.000
1413	$\varnothing_1$	0.259	0.068	3.80	0.000
	θ_1	0.950	0.031	30.56	0.000
	Θ_1	0.935	0.099	9.48	0.000
1418	θ_1	0.812	0.048	16.81	0.000
	Φ_1	-0.220	0.090	-2.43	0.016
	Θ_1	0.835	0.088	9.54	0.000
1422	Θ_1	0.857	0.075	11.50	0.000

Figure 3 shows the relationship between ten-years of monthly data observed at each sediment-sampling station and predicted data for the same years by using the selected best model for each sediment-sampling station. As shown in Figure 3, the predicted data follows the observed data very closely.

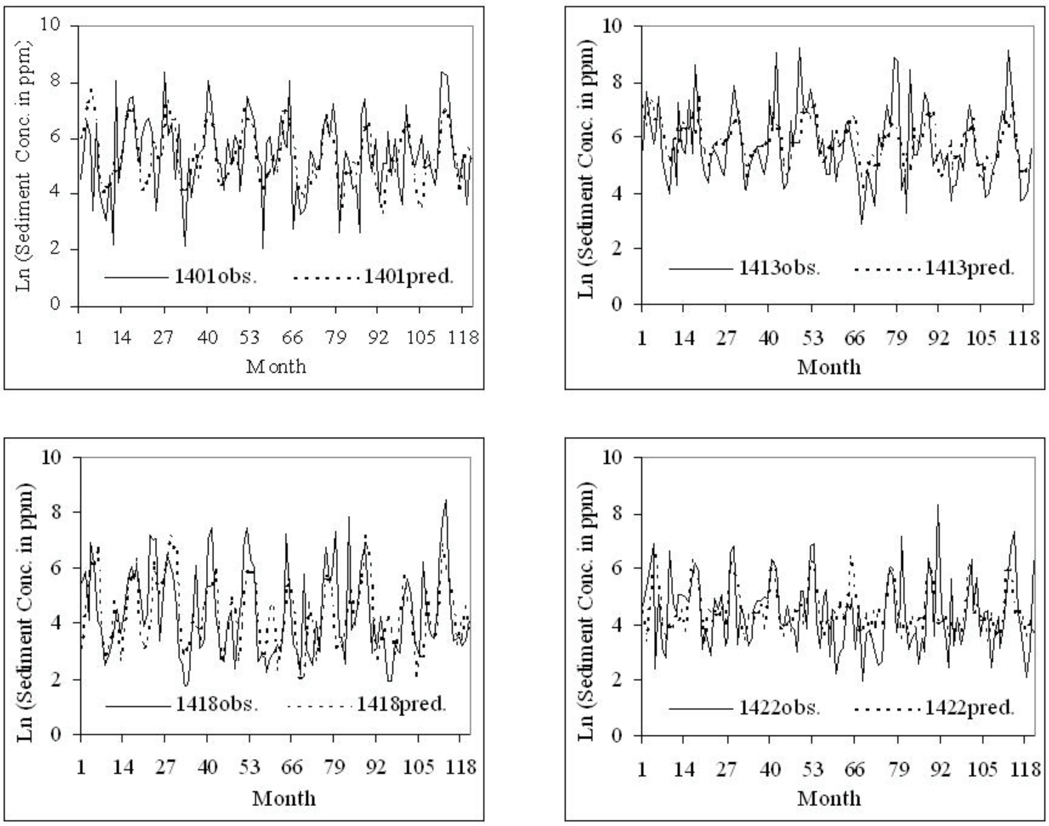

Figure 3. Comparison of observed and predicted data.

Table 4 shows some basic statistical properties of the observed and predicted data. Using t-test for the means and F-test for the variances to determine whether there is significant difference, the mean and variance values of the observed and predicted data for each sediment-sampling station were tested. The hypothesis that the mean and variance values of the generated data are not significantly different from those of the observed data can be accepted at 5 % significance level. Thus, the results show that generated data preserve the basic statistical properties of the observed series.

Table 4. Comparison of original data to synthetic data

Sampling Station	Mean	$t_{cal}<t_{table}$	Standard Deviation	$F_{cal}<F_{table}$
$1401_{Obs.}$	5.702	1.280<1.645	1.457	0.73<1.37
$1401_{Pred.}$	5.549		1.254	
$1413_{Obs.}$	6.074	-1.405<1.645	1.270	0.71<1.37
$1413_{Pred.}$	6.219		0.900	
$1418_{Obs.}$	4.478	-0.138<1.645	1.584	0.83<1.52
$1418_{Pred.}$	4.501		1.310	
$1422_{Obs.}$	4.447	-0.903<1.645	1.261	0.72<1.46
$1422_{Pred.}$	4.555		0.911	

CONCLUSIONS

In this study we presented a stochastic modeling method to simulate monthly-suspended sediment data. Monthly sediment data from four sampling stations in the Yesilirmak Basin were used for simulation purposes. The results from this study showed that the predicted series obtained from the model (Box-Jenkins ARIMA) preserve the basic statistical properties of the observed series. Therefore, this model proved to be a valuable tool to forecast the monthly-suspended sediment data from the Yesilirmak Basin, Turkey.

Nomenclature

a_i : white noise time series value at time i
B : backward shift operator
c : constant for Box-Cox transformation
d : order of the nonseasonal differencing operator
D : order of the seasonal differencing operator
ESS_F : the residual sum of square for first group
ESS_S : the residual sum of square for second group
K_{xi} : rank of i^{th} observation in the historical data
K_{yi} : rank in the historical data of i^{th} observation in the ascended data
LBQ/P : probability for Q(r)
M : the number of AR and MA parameters
n : the number of observation
n_F : the number of residuals in the first group
n_S : the number of residuals in the second group
s : seasonal length
Q(r) : Ljung-Box statistic at lag m
$r_k(a)$: ACF of a_i at lag k
R_{sp} : rank order correlation coefficient
x_i : discrete time series value at time i
w_i : stationary series formed by differencing the x_i
z_i : transformation of x_i series

Greek Symbols

λ : exponent for Box-Cox transformation
μ : mean level of the w_i series (if D+d>0 often $\mu \approx 0$)
$Ø_i$: i^{th} nonseasonal AR parameter
Φ_i : i^{th} seasonal AR parameter
θ_i : i^{th} nonseasonal MA parameter
Θ_i : i^{th} seasonal MA parameter

REFERENCES

[1] Ozturk F. 1995. Turkiye'nin Toprak ve Su Koruma Sorunlari. In: Kulturteknige Giris (Ed. A. Balaban). Ankara Universitesi Ziraat Fakultesi Yayinlari, Ankara. (in Turkish).

[2] Cevik B. 1988. Toprak ve Su Koruma Muhendisligi. Cukurova Universitesi Ziraat Fakultesi Yayinlari, Adana (in Turkish).

[3] Ozturk, F. 1996. Suspended sediment yield of rivers in Turkey. Erosion and Sediment Yield: Global and Regional Perspectives. In: International Association Hydrological Sciences (Eds. D.E. Walling and B.W. Webb). Great Yarmouth, UK.

[4] Rice RM. 1982. Sedimentation in The Chaparall: How Do You Handle Unusual Events? In: Sediment Budgets and Routing in Forested Drainage Basins (Eds. F.J. Swanson, R.J. Janda, T. Dunne and D.N. Swanston) PNW-141. U.S. Forest Service Report.

[5] Hadley RF, Lal R, Onstad CA, Walling DE, Yaira A. 1985. Recent Developments in Erosion and Sediment Yield Studies. IHP, Unesco Project, A.1.3.1.

[6] Gurnell AM, Fenn CR. 1984. Box-Jenkins transfer function models apply to suspended sediment concentration-discharge relationships in a proglacial stream. Arctic and Alpine Research.16: 93-106.

[7] Anonymous. 1970. Yesilirmak Havzasi Topraklari. Topraksu Genel Mudurlugu Yayinlari, Ankara (in Turkish).

[8] Janacek G, Swift L. 1993. Time Series Forecasting, Simulation, Application. Ellis Horwood, New York.

[9] Box GEP, Jenkins GM. 1976. Time Series Analysis Forecasting and Control. Holden-Day, San Francisco.

[10] Gibbons JD. 1997. Nonparametric Methods for Quantitative Analysis. American Sciences Press, Inc., Ohio.

[11] Hipel KW, McLeod AI, Lennox WC. 1977. Advances in Box-Jenkins modeling: I. model construction. Water Resources Research. 13(3): 567-575.

[12] Wei WWS. 1990. Time Series Analysis. Addison-Wesley Publishing Company Inc., New York.

[13] Akaike H. 1974. A look at the statistical model identification. IEEE Transactions on Automatic Control AC.19(6): 716-723.

[14] Shibata R. 1976. Selection of the order of an autoregressive model by Akaike's information criterion. Biometrika. 63: 117-126.

[15] Akaike H. 1978. A Bayesian analysis of the minimum AIC procedure. Ann. Inst. Statist. Math. 30A: 9-14.

[16] Akaike H. 1979. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. Biometrika. 66: 237-242.

[17] Schwartz G. 1978. Estimating the dimension of a model. Ann. Statist. 6: 461-464.

[18] Ljung GM, Box GEP. 1978. On a measure of lack of fit in time series models. Biometrika. 65(2): 297-303.

[19] Breusch T, Pagan A. 1979. A simple test of heteroscedasticity and random coefficient variation. Econometrica. 47: 1287-1294.

[20] Chow VT, Maidment DR, Mays LW. 1988. Applied Hydrology. McGraw-Hill Book Company, New York.

[21]. Haan CT. 1977. Statistics Methods in Hydrology. Iowa State Press, Iowa.