INDIAN INSTITUTE OF SCIENCE

STOCHASTIC HYDROLOGY

Lecture -12

Course Instructor : Prof. P. P. MUJUMDAR
Department of Civil Engg., IISc.

Summary of the previous lecture

- Data Extension & Forecasting
  - Moving average
  - Double moving average
- Data Generation – Uncorrelated Data
- Data Generation – Serially Correlated Data
  - First order Markov Model

Data Generation – Serially Correlated Data

First order stationary Markov model
Or
Thomas Fiering model (Stationary)

$$X_{j+1} = \mu_x + \rho_1 \left(X_j - \mu_x \right) + t_{j+1} \sigma_x \sqrt{1 - \rho_1^2}$$

t_{j+1} → Standard normal deviate

- Stationary w.r.t mean, variance and lag-one correlation
- Known sample estimates of μ_x, σ_x, ρ_1
- Assume X_1 (= μ_x)
- Generate values X_2, X_3, X_4, X_5 ……

Data Generation – Serially Correlated Data

First order Markov model with non-stationarity:

- First order stationary Markov model assumes that the process is stationary in mean, variance and lag-one auto correlation.
- The model is generalized to account for non-stationarity (mainly due to seasonality/periodicity) in hydrologic data.
- A main application of this generalised model is in generating the monthly stream flows with pronounced seasonality.
- Periodicity may affect not only the mean, but all the moments of data including the serial correlations.

Data Generation – Serially Correlated Data

$$X_{j+1} = \mu_x + \rho_1 \left(X_j - \mu_x \right) + t_{j+1} \sigma_x \sqrt{1 - \rho_1^2}$$

Stationary First order Markov Model

First order Markov model with non-stationarity, for stream flow generation:

$$X_{i,j+1} = \mu_{j+1} + \rho_j \frac{\sigma_{j+1}}{\sigma_j} \left(X_{ij} - \mu_j \right) + t_{i,j+1} \sigma_{j+1} \sqrt{1 - \rho_j^2}$$

ρ_j is serial correlation between flows of j^{th} month and $j+1^{th}$ month.

$t_{i,\, j+1} \sim N(0, 1)$

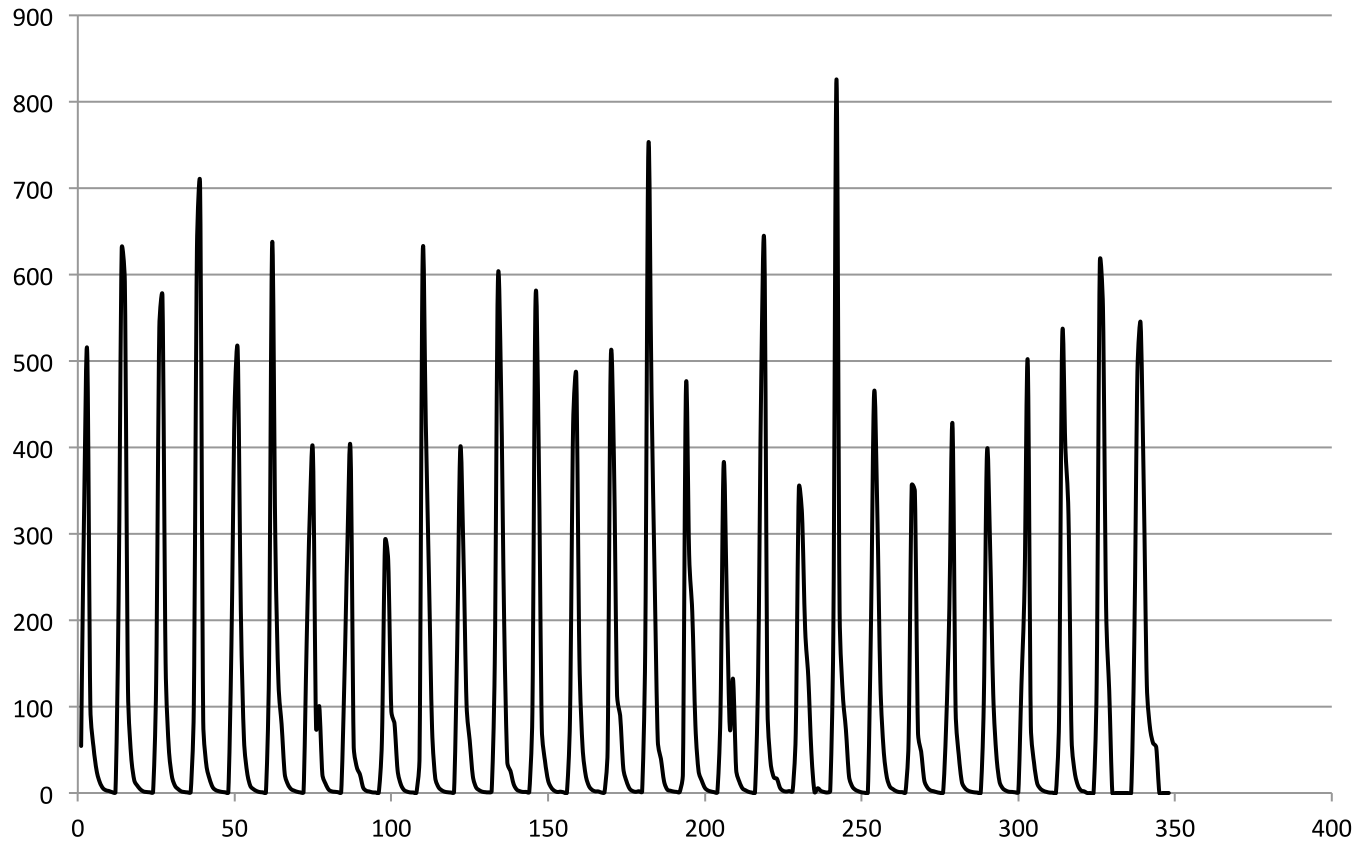
Example-1

The monthly stream flow (in cumec) for a river is available for 29 years (12 years data is given here)

SL. NO.	YEAR	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY
1	1979-80	54.60	325.40	509.50	99.40	53.50	25.80	12.50	5.60	3.10	2.20	0.90	0.81
2	1980-81	220.78	629.16	591.32	120.33	43.33	14.83	8.41	4.05	1.73	1.12	0.85	0.96
3	1981-82	131.30	538.89	574.21	151.06	53.03	19.49	8.38	4.51	1.89	1.11	0.74	1.06
4	1982-83	100.19	630.02	702.07	83.29	32.45	16.60	6.80	3.33	2.03	1.23	0.85	0.65
5	1983-84	171.30	444.30	512.30	211.00	62.40	24.00	8.40	4.50	2.30	1.10	0.80	0.60
6	1984-85	147.80	636.20	293.50	127.70	79.70	22.10	10.10	4.60	2.70	1.40	0.70	0.90
7	1985-86	174.50	323.30	393.20	75.40	100.60	21.80	10.90	4.00	1.90	1.40	1.00	0.70
8	1986-87	126.40	288.30	395.30	54.40	29.80	21.40	6.40	2.60	1.70	0.70	0.60	0.50
9	1987-88	60.50	291.00	269.60	95.09	80.84	26.39	10.37	3.68	1.65	0.71	0.62	0.38
10	1988-89	40.95	620.00	427.60	251.80	74.73	17.71	7.05	3.33	1.51	0.87	0.59	0.90
11	1989-90	167.10	398.80	277.80	102.70	61.10	19.54	6.79	3.33	1.52	0.96	0.77	1.93
12	1990-91	150.80	591.50	471.20	197.00	35.67	25.62	10.52	4.02	2.10	1.22	1.32	1.16

Example-1 (contd.)

Time series of monthly stream flow for 29 years



Example-1 (contd.)

S.No.	Month	Mean	Stdev.	Lag-1 correlation
1	JUN	117.49	52.24	0.348
2	JUL	474.50	150.18	0.154
3	AUG	421.39	126.53	0.169
4	SEP	145.94	77.65	0.365
5	OCT	66.61	30.67	0.490
6	NOV	22.99	13.26	0.798
7	DEC	10.30	9.82	0.955
8	JAN	5.55	9.16	-0.385
9	FEB	1.91	0.74	0.733
10	MAR	1.09	0.54	0.654
11	APR	0.76	0.51	0.676
12	MAY	0.80	0.60	-0.005

Example-1 (contd.)

Assume $X_{1,1} = \mu_1 = 117.49$;

$\sigma_1 = 52.24$, $\rho_1 = 0.348$

$\mu_2 = 474.5$, $\sigma_2 = 150.18$,

$$X_{1,2} = \mu_2 + \rho_1 \frac{\sigma_2}{\sigma_1}\left(X_{1,1} - \mu_1\right) + t_{1,2}\sigma_2\sqrt{1-\rho_1^2}$$

$$= 474.5 + 0.348\frac{150.18}{52.24}\left(117.49 - 117.49\right)$$

$$+ 0.335 * 150.18\sqrt{1-0.348^2}$$

$= 521.67$

Example-1 (contd.)

$X_{1,2}$ =521.67, μ_2 = 474.5; σ_2 = 150.18, ρ_2 = 0.154

μ_3 = 421.39, σ_3 = 126.53, ~~ρ_3 = 0.154~~

$$X_{1,3} = 421.39 + 0.154\frac{126.53}{150.18}(521.67 - 474.5)$$

$$+ 0.377 * 126.53\sqrt{1 - 0.154^2}$$

= 474.64

Example-1 (contd.)

$X_{1,3}$ =474.64, μ_3 = 421.39; σ_3 = 126.53, ρ_3 = 0.169

μ_4 = 145.94, σ_4 = 77.65,

$$X_{1,4} = 145.94 + 0.169\frac{77.65}{126.53}\left(474.64 - 421.39\right) + 0.379 * 77.65\sqrt{1 - 0.169^2}$$

= 180.45

.

$X_{1,12}$

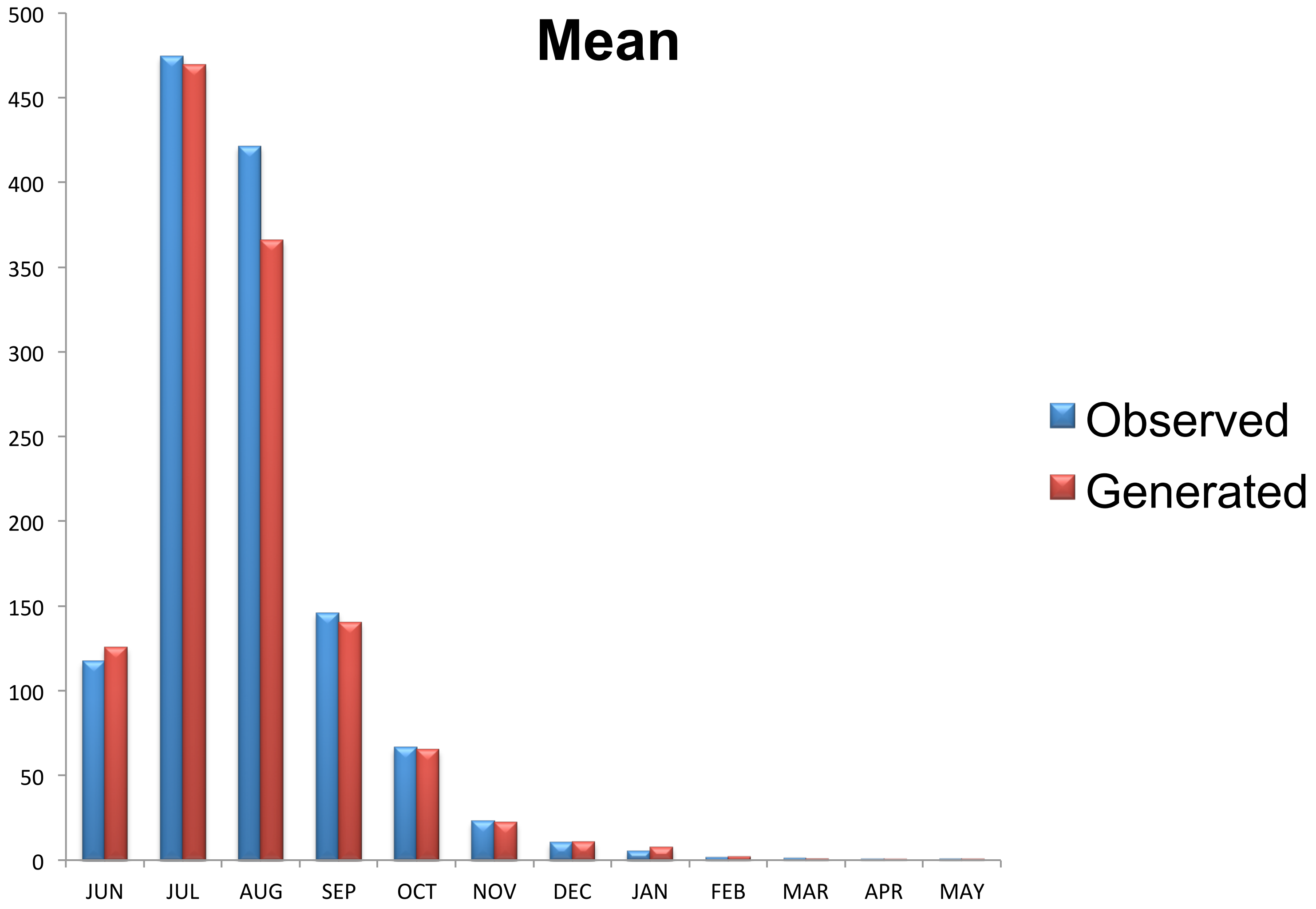
$$X_{2,1} = \mu_1 + \rho_{12}\frac{\sigma_1}{\sigma_{12}}\left(X_{1,12} - \mu_{12}\right) + t_{2,1}\sigma_1\sqrt{1 - \rho_{12}^2}$$

Example-1 (contd.)

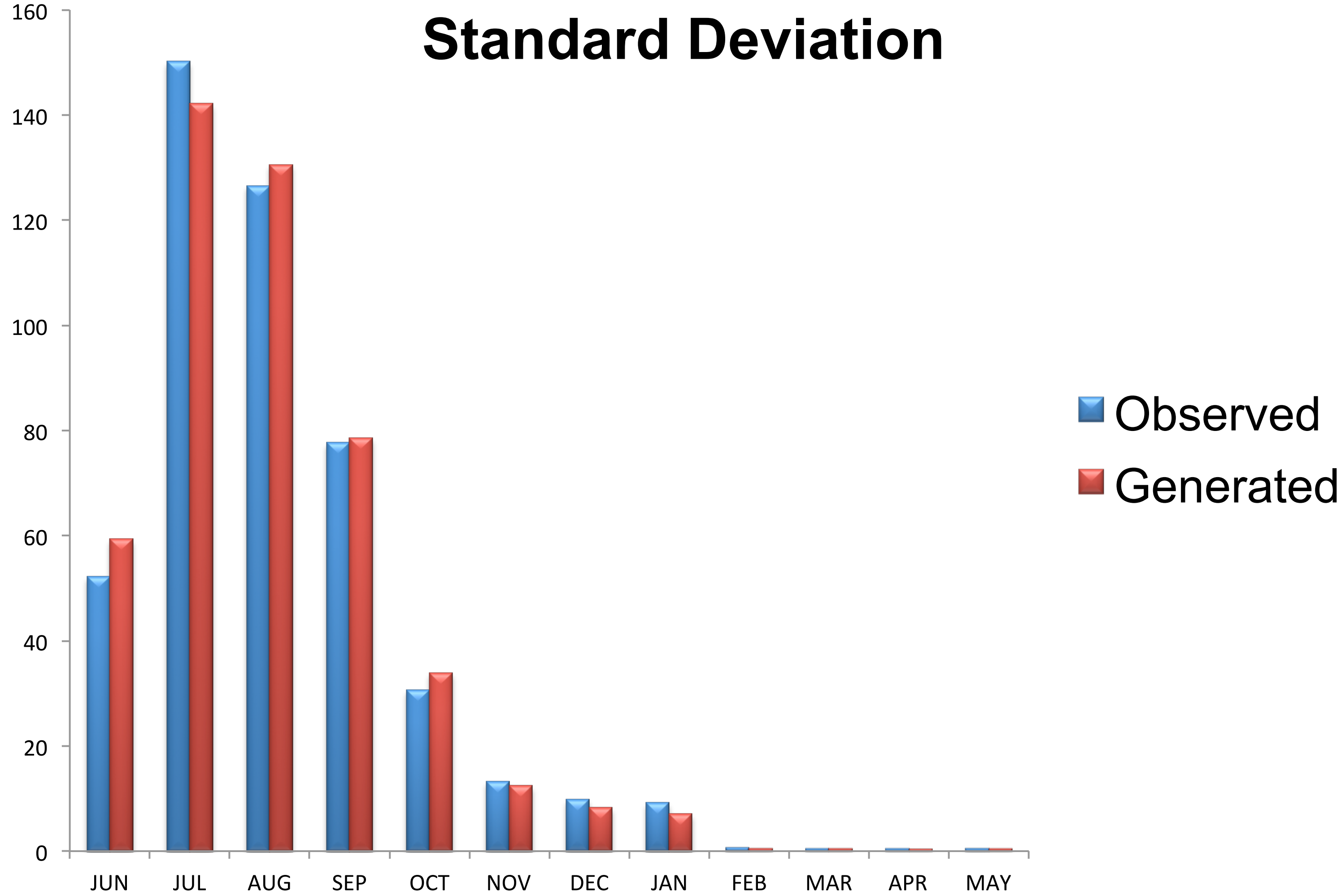
Generated:

S.No.	Month	Mean	Stdev.	Lag-1 correlation
1	JUN	125.69	59.30	0.516
2	JUL	469.36	142.10	-0.116
3	AUG	365.98	130.60	0.080
4	SEP	140.40	78.60	0.352
5	OCT	65.28	33.89	0.754
6	NOV	22.33	12.40	0.789
7	DEC	10.68	8.30	0.923
8	JAN	7.69	7.13	-0.154
9	FEB	1.95	0.59	0.728
10	MAR	0.95	0.52	0.791
11	APR	0.60	0.47	0.735
12	MAY	0.68	0.50	0.041

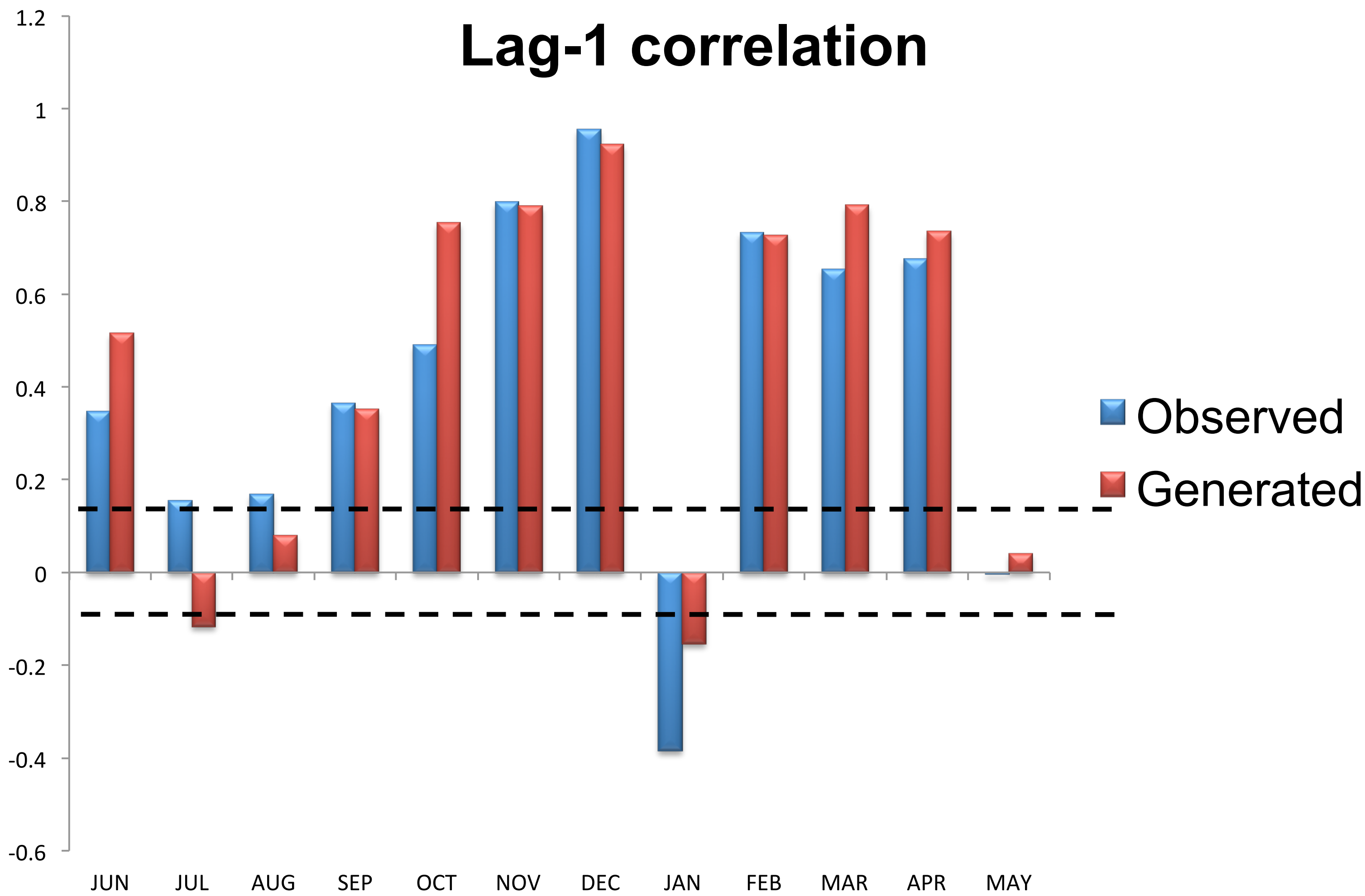
Example-1 (contd.)



Example-1 (contd.)



Example-1 (contd.)



Data Generation – Serially Correlated Data

$$Y_{i,j+1} = \mu_{y_{j+1}} + \rho_{y_j} \frac{\sigma_{y_{j+1}}}{\sigma_{y_j}} \left(Y_{ij} - \mu_{y_j} \right) + t_{i,j+1} \sigma_{y_{j+1}} \sqrt{1 - \rho_{y_j}^2}$$

Where $Y_{i,j+1} = \ln (X_{i,\ j+1})$

$\mu_{y_j}, \sigma_{y_j}, \rho_{y_j}$ refer to the mean, standard deviation and lag one correlation of logarithms of original data

Example-2

The logarithms of stream flow (in cumec) of example-1are constructed

SL. NO.	YEAR	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY
1	1979-80	4.00	5.79	6.23	4.60	3.98	3.25	2.53	1.72	1.13	0.79	-0.11	-0.21
2	1980-81	5.40	6.44	6.38	4.79	3.77	2.70	2.13	1.40	0.55	0.11	-0.16	-0.04
3	1981-82	4.88	6.29	6.35	5.02	3.97	2.97	2.13	1.51	0.64	0.10	-0.30	0.06
4	1982-83	4.61	6.45	6.55	4.42	3.48	2.81	1.92	1.20	0.71	0.21	-0.16	-0.43
5	1983-84	5.14	6.10	6.24	5.35	4.13	3.18	2.13	1.50	0.83	0.10	-0.22	-0.51
6	1984-85	5.00	6.46	5.68	4.85	4.38	3.10	2.31	1.53	0.99	0.34	-0.36	-0.11
7	1985-86	5.16	5.78	5.97	4.32	4.61	3.08	2.39	1.39	0.64	0.34	0.00	-0.36
8	1986-87	4.84	5.66	5.98	4.00	3.39	3.06	1.86	0.96	0.53	-0.36	-0.51	-0.69
9	1987-88	4.10	5.67	5.60	4.55	4.39	3.27	2.34	1.30	0.50	-0.34	-0.48	-0.97
10	1988-89	3.71	6.43	6.06	5.53	4.31	2.87	1.95	1.20	0.41	-0.14	-0.54	-0.11
11	1989-90	5.12	5.99	5.63	4.63	4.11	2.97	1.91	1.20	0.42	-0.04	-0.26	0.66
12	1990-91	5.02	6.38	6.16	5.28	3.57	3.24	2.35	1.39	0.74	0.20	0.28	0.15

Example-2 (contd.)

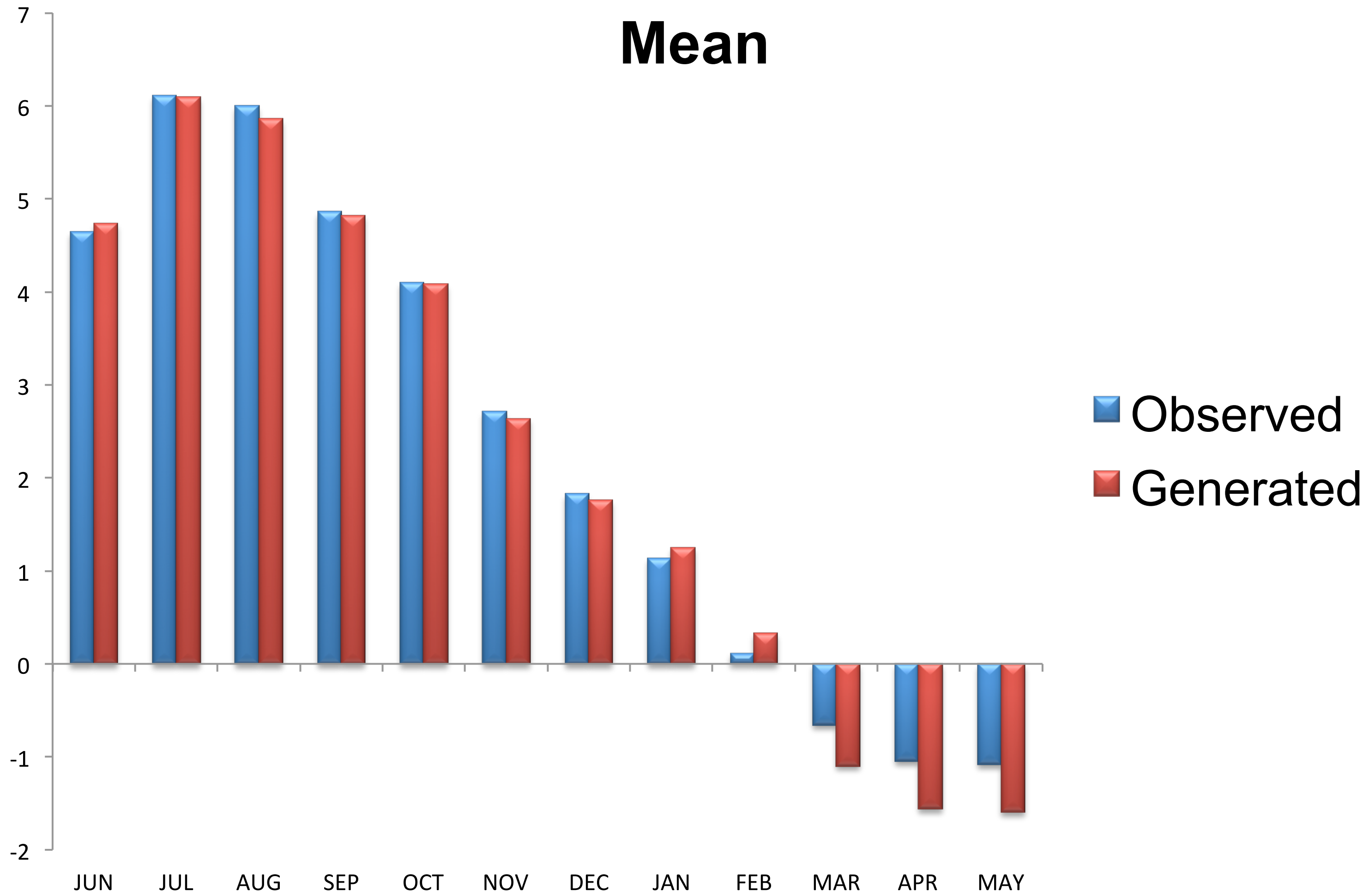
S.No.	Month	Mean	Stdev.	Lag-1 correlation
1	JUN	4.64	0.54	0.239
2	JUL	6.11	0.33	0.114
3	AUG	6.00	0.31	0.183
4	SEP	4.86	0.49	0.409
5	OCT	4.10	0.44	-0.163
6	NOV	2.71	2.02	0.955
7	DEC	1.83	1.91	0.967
8	JAN	1.13	1.77	0.474
9	FEB	0.12	2.15	0.800
10	MAR	-0.66	2.42	0.985
11	APR	-1.05	2.32	0.973
12	MAY	-1.08	2.35	-0.105

Example-2 (contd.)

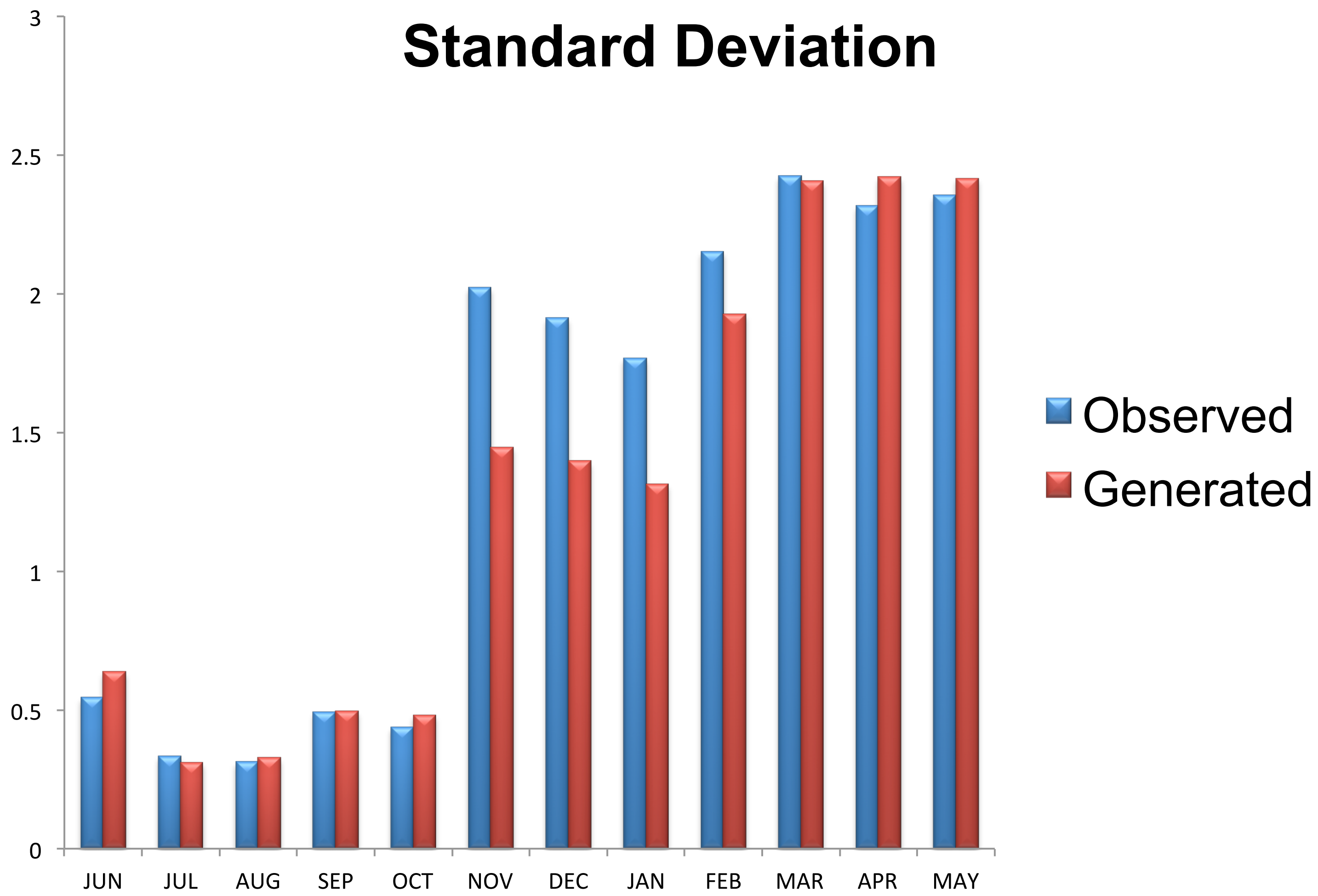
Generated:

S.No.	Month	Mean	Stdev.	Lag-1 correlation
1	JUN	4.74	0.64	0.442
2	JUL	6.09	0.31	-0.153
3	AUG	5.86	0.33	0.090
4	SEP	4.82	0.50	0.385
5	OCT	4.08	0.48	0.007
6	NOV	2.64	1.45	0.926
7	DEC	1.76	1.40	0.924
8	JAN	1.25	1.31	0.519
9	FEB	0.33	1.93	0.801
10	MAR	-1.10	2.41	0.991
11	APR	-1.56	2.42	0.978
12	MAY	-1.60	2.42	0.086

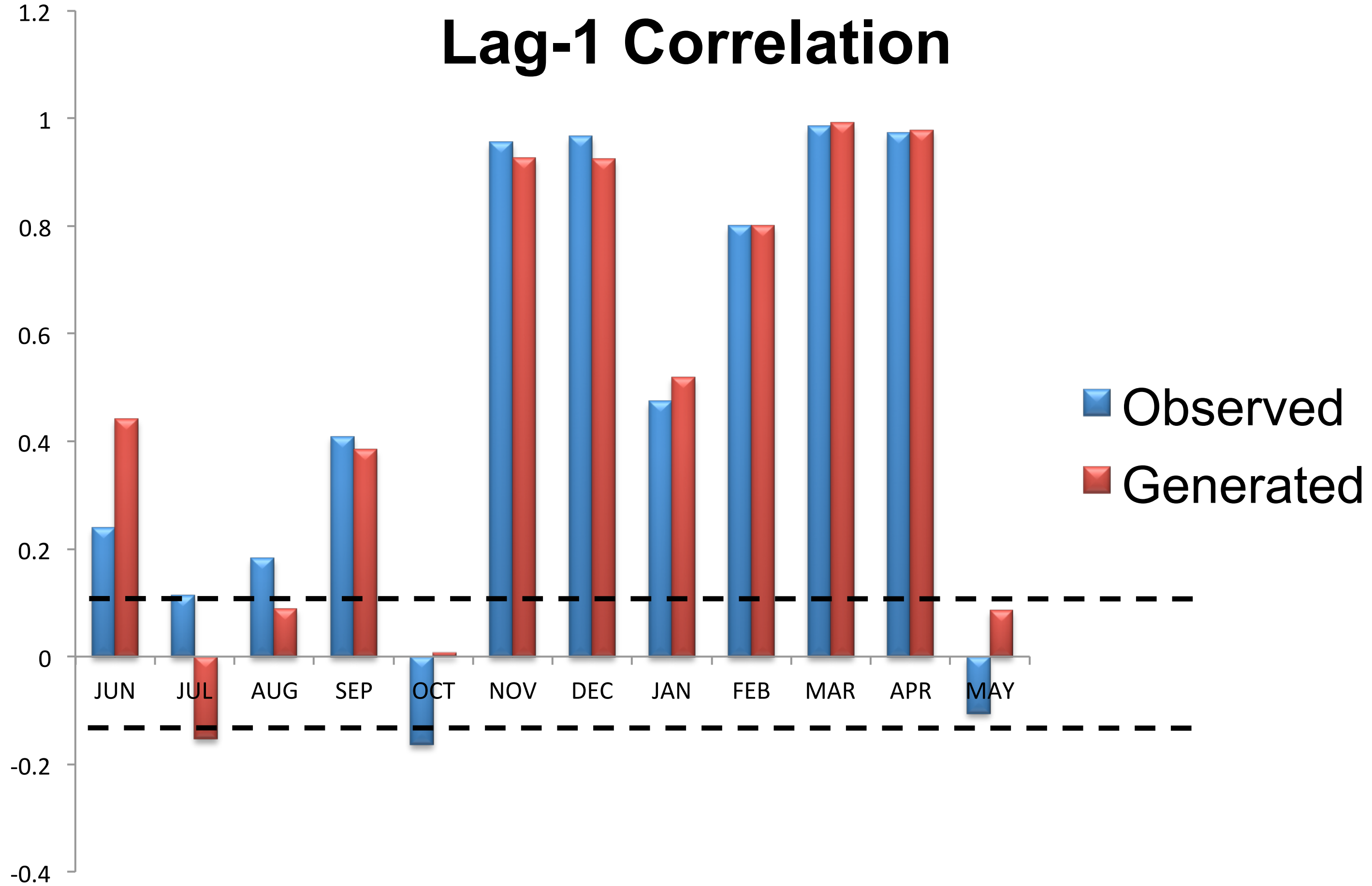
Example-2 (contd.)



Example-2 (contd.)

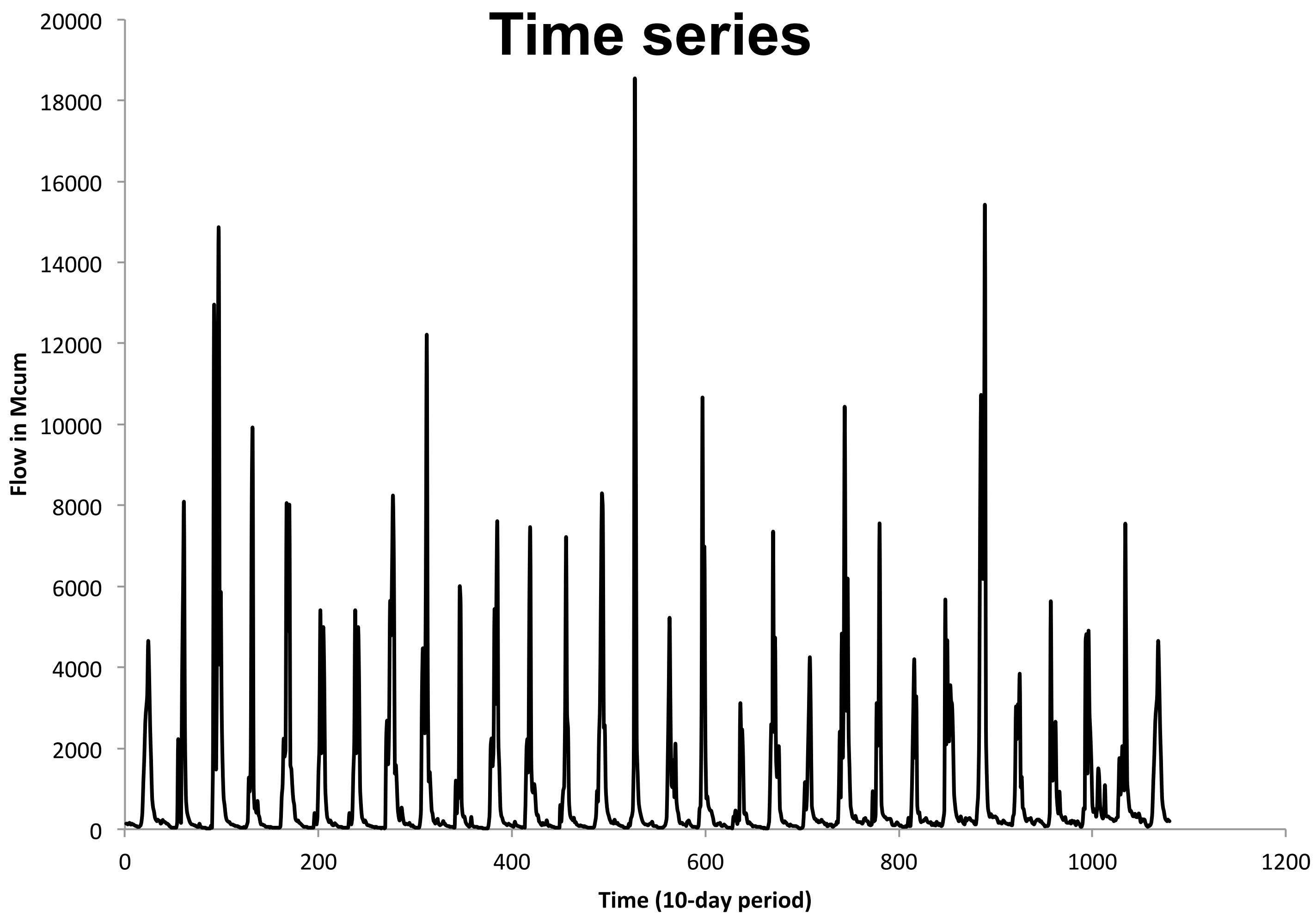


Example-2 (contd.)

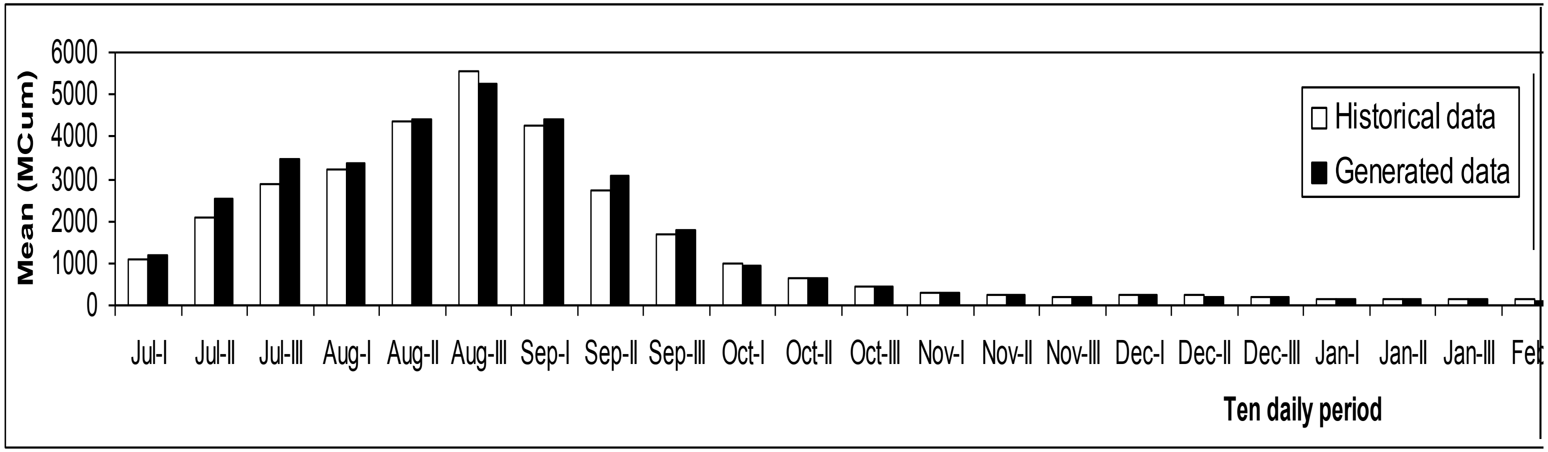


Example-3

(Generation of 10-day flows to Sardar Sarovar Reservoir)

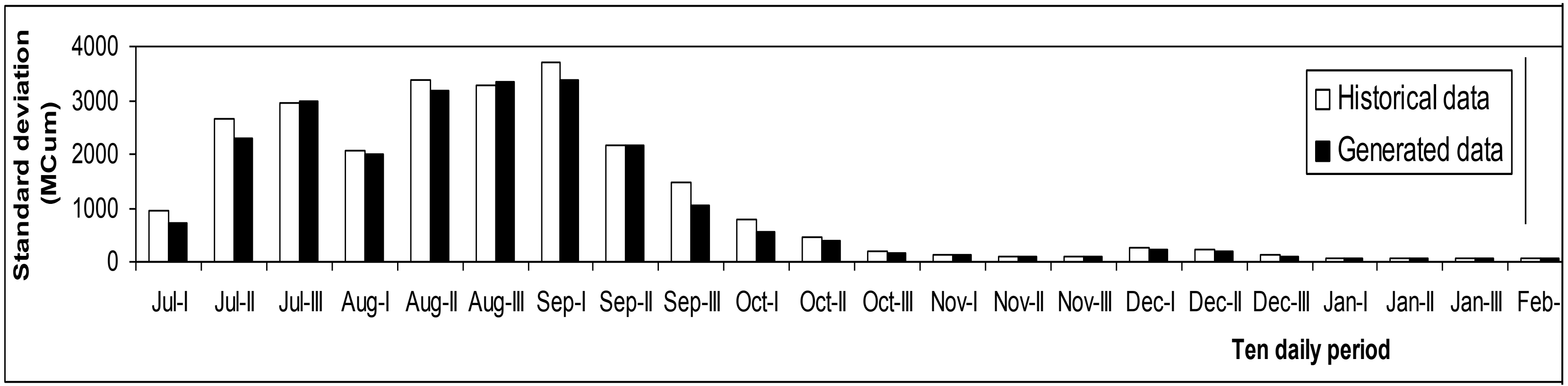


Example-3(contd.)



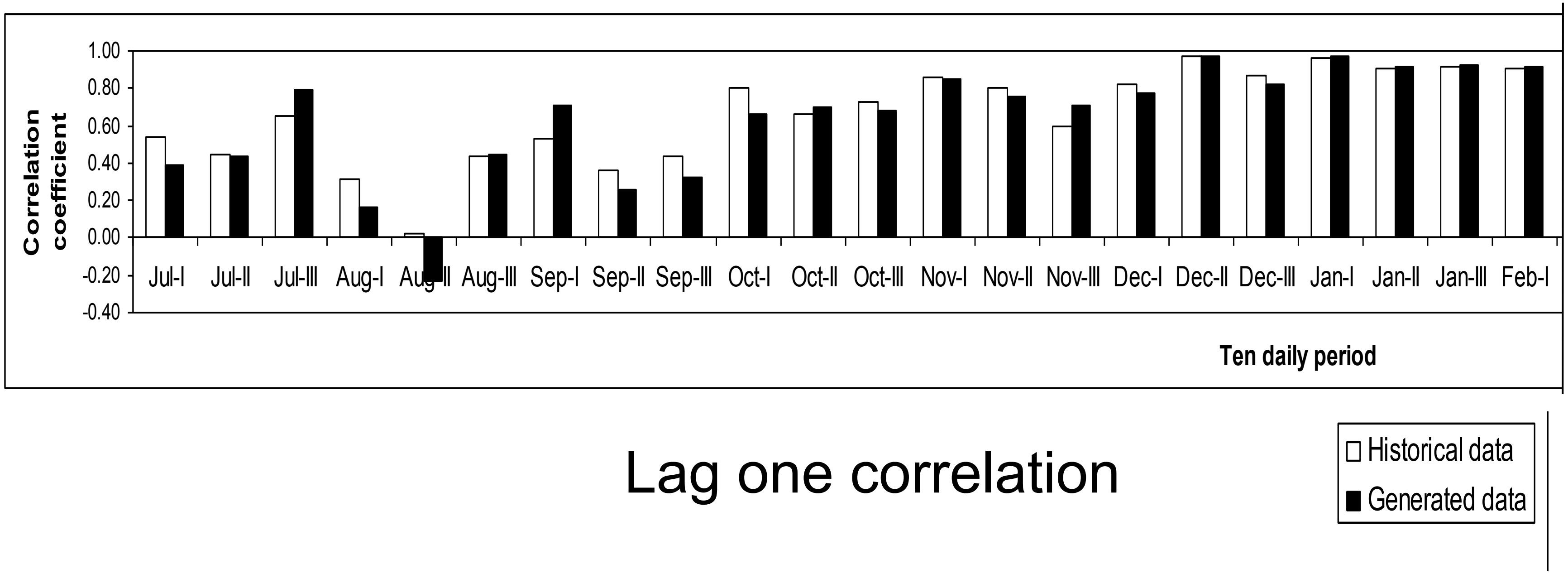
Mean Flows

Example-3(contd.)



Standard Deviation

Example-3(contd.)



Lag one correlation

Issues to be addressed in modeling flows

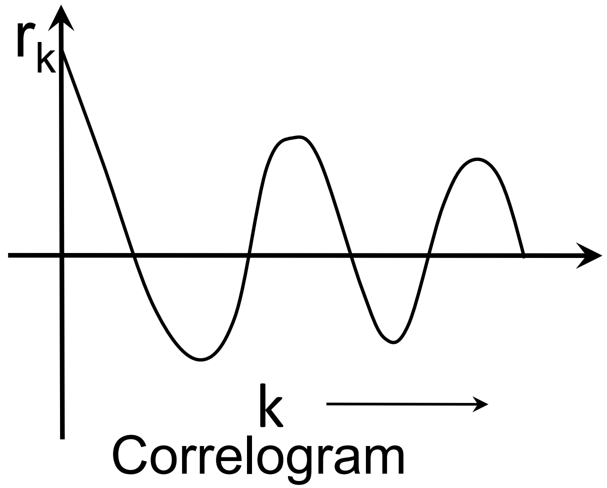
- Is it necessary to model peak flows?
- Time during the year when the peak occurs important?
- Volume of flow important?
- Duration of flow to be considered (Daily, weekly, monthly etc.)
- Dependence of the flow from one time period to another?
- Is time series of flows stationary?
- Is there evidence of trends or jumps?
- Quality and quantity of data available

Ref: C.T.Haan, 1995, Page no.291

FREQUENCY DOMAIN ANALYSIS

Frequency Domain Analysis

- Auto correlation function or correlogram is used for analyzing the time series.



- Time domain analysis

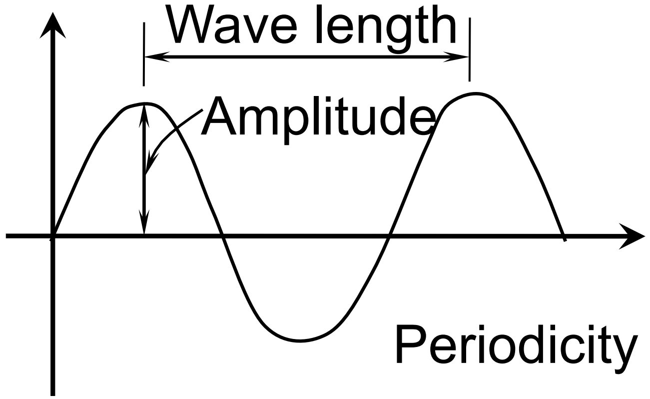
$$X_t = d_t + \varepsilon_t$$

- Periodicities in data can be determined by analyzing the time series in frequency domain.

Frequency Domain Analysis

- Spectral analysis or the frequency domain analysis: the time series is represented in the frequency domain instead of the time domain
- The observed time series is a random sample of a process over time and is made up of oscillations of all possible frequencies.
- Spectral analysis is used to identify the periodicities in the data.

Frequency Domain Analysis



n odd — n even

$$X_t = \alpha_0 + \sum_{k=1}^{\frac{n-1}{2},\frac{n}{2}} \left[\alpha_k \cos\left(2\pi f_k t\right) + \beta_k \sin\left(2\pi f_k t\right)\right] + \varepsilon_t$$

t = 1, 2, …. N

Frequency Domain Analysis

$$f_k = \frac{k}{N} \quad ;$$

k^{th} harmonic of the fundamental frequency (1/N)

N is the no. of observations

Periodicity (P):

$$P = \frac{1}{f_k}$$

Frequency Domain Analysis

$$\alpha_0 = \overline{x}$$

$$\alpha_k = \frac{2}{N}\sum_{t=1}^{n} x_t \cos\left(2\pi f_k t\right) \qquad k = 1, 2, \ldots. M$$

$$\beta_k = \frac{2}{N}\sum_{t=1}^{n} x_t \sin\left(2\pi f_k t\right) \qquad k = 1, 2, \ldots. M$$

M is maximum lag (typically considered up to 0.25N)

The above equations for α_k and β_k are valid up to k=N/2

Frequency Domain Analysis

When ‘N’ is odd, the expressions are true until

$$k = \frac{N}{2} - 1$$

$$\alpha_{N/2} = \frac{1}{N} \sum_{t=1}^{n} \left(-1\right)^{t} x_{t}$$

$$\beta_{N/2} = 0$$

Frequency Domain Analysis

- A variance spectrum divides the variance into no. of intervals or bands of frequency.
- Spectral density (I_k) is the amount of variance per interval of frequency.

$$I(k) = \frac{N}{2}\left[\alpha_k^2 + \beta_k^2\right] \qquad k = 1, 2, \ldots. M$$

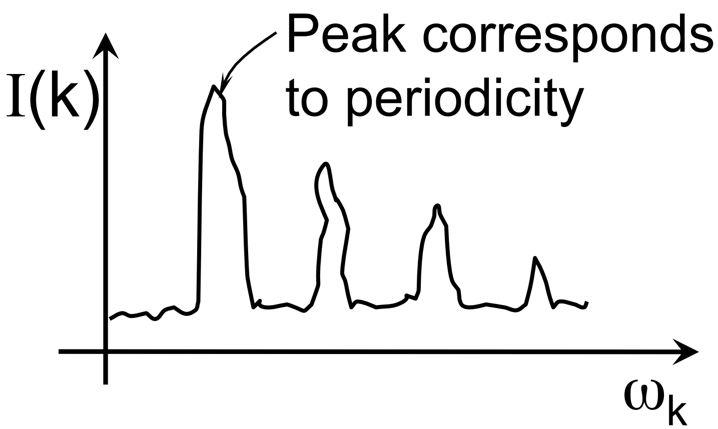
- Angular frequency

$$\omega_k = \frac{2\pi k}{N} \qquad k = 1, 2, \ldots. M$$

Frequency Domain Analysis

$$\omega_k = \frac{2\pi}{P}$$

A plot of ω_k vs $I(k)$ is called spectrum



Prominent spikes indicate periodicity

Example-2

Obtain ω_k and $\mathrm{I}(k)$ for k=1

t	X_t	$\cos\left(2\pi f_k t\right)$	$\sin\left(2\pi f_k t\right)$	$X_t \cos\left(2\pi f_k t\right)$	$X_t \sin\left(2\pi f_k t\right)$
1	105	0.809	0.5878	84.945	61.719
2	115	0.309	0.9511	35.535	109.3765
3	103	-0.309	0.9511	-31.827	97.9633
4	94	-0.809	0.5878	-76.046	55.2532
5	95	-1	0	-95	0
6	104	-0.809	-0.5878	-84.136	-61.1312
7	120	-0.309	-0.9511	-37.08	-114.132
8	121	0.309	-0.9511	37.389	-115.083
9	127	0.809	-0.5878	102.743	-74.6506
10	79	1	0	79	0
Σ				15.523	-40.6849

Example-2 (contd.)

$$f_k = \frac{k}{N}$$

$$= \frac{1}{10} = 0.1$$

$$\alpha_k = \frac{2}{N}\sum_{t=1}^{n} x_t \cos\left(2\pi f_k t\right)$$

$$= \frac{2}{10} \times \left(15.523\right)$$

$$= 3.1046$$

$$\beta_k = \frac{2}{N}\sum_{t=1}^{n} x_t \sin\left(2\pi f_k t\right)$$

$$= \frac{2}{10} \times \left(-40.6849\right)$$

$$= -8.13698$$

Example-2 (contd.)

$$I(k) = \frac{N}{2}\left[\alpha_k^2 + \beta_k^2\right]$$

$$= \frac{10}{2}\left[(3.1046)^2 + (-8.13698)^2\right]$$

$$= 379.245$$

$$\omega_k = \frac{2\pi k}{N}$$

$$= \frac{2\times\pi\times1}{10}$$

$$= 0.62832$$

Frequency Domain Analysis

- Spectral density is also called as line spectrum
- The line spectrum thus transforms the information from time domain to the frequency domain
- While the correlogram indicate merely the presence of periodicities in the data
- The spectral analysis helps indentify the significant periodicities themselves

Frequency Domain Analysis

- Spectral density is an inconsistent estimate
- The plot is not a smooth function
- The smoothened spectrum is called as power spectrum
- Power spectrum is a consistent estimate of spectral density

Frequency Domain Analysis

- Power spectrum – Fourier cosine transform of auto covariance function.

$$I(k) = 2\left[c_0 + 2\sum_{j=1}^{\frac{n-1}{2}} \lambda_j c_j \cos\left(2\pi f_k j\right)\right]$$

c_j = Auto covariance function

λ_j = lag window (or smoothing window)

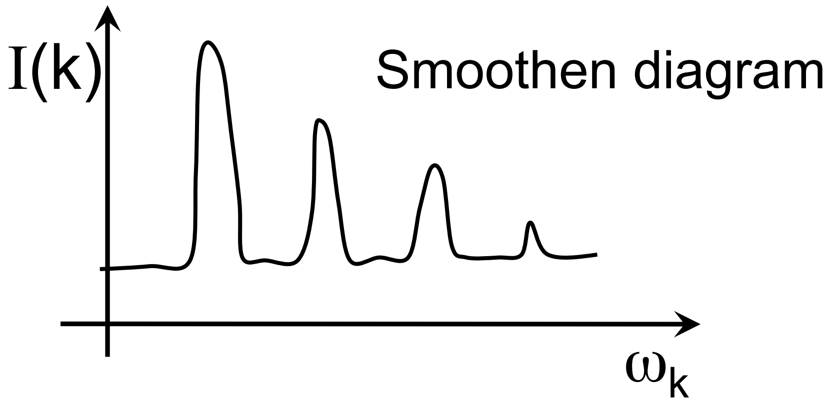
Different ways of estimating λ_j

Frequency Domain Analysis

Tukey window

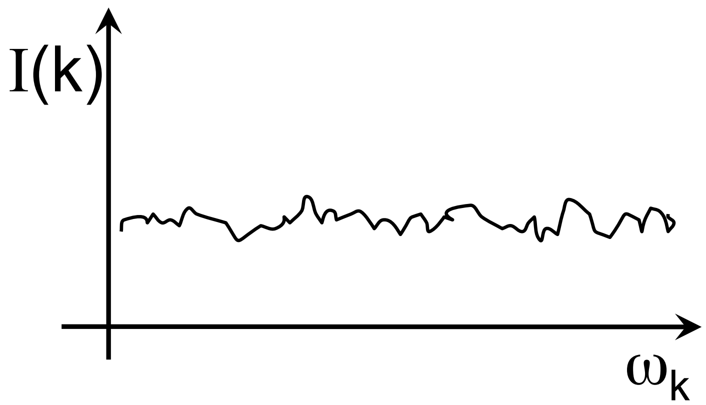
$$\lambda_j = \frac{1}{2}\left[1 + 2\cos\left(\frac{2\pi}{M^{'}}\right)\right]$$

M’ = Maximum lag (~ 0.25N)



Frequency Domain Analysis

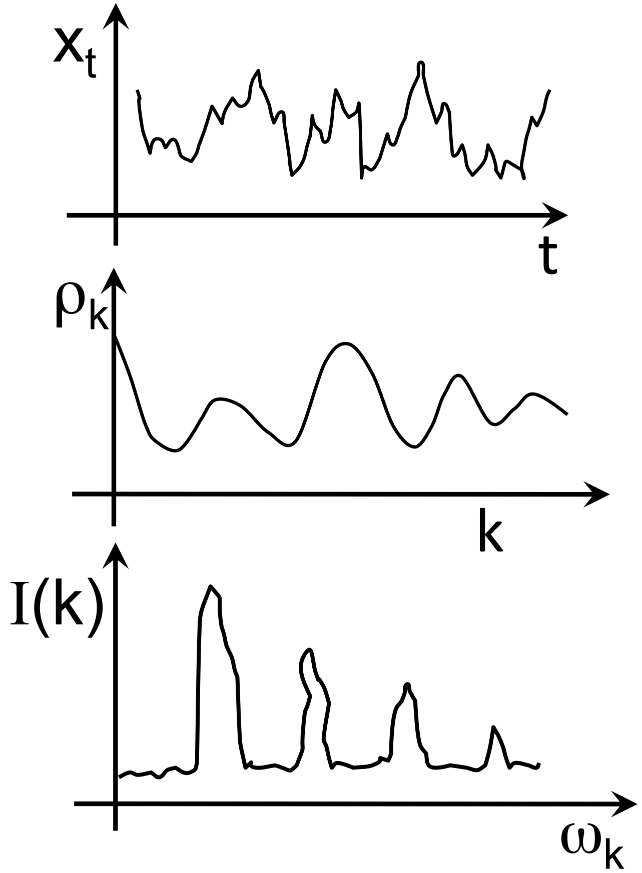
- Information content is extracted from spectrum.
- For a completely random series (e.g., uniformly distributed random numbers), the spectral density function is constant – termed as white noise
- White noise indicates that no frequency interval contains any more variance than any other frequency interval. (auto correlation function $\rho_k = 0$, for $k \neq 0$)



Frequency Domain Analysis

The steps for analyzing the data are as follows

- Plot the time series
- Plot the correlogram
- Plot the spectrum



Frequency Domain Analysis

- The spectrum shows prominent spikes (which represent the periodicities inherent in the data)
- The period corresponding to any value of ω_k may be computed by $2\pi/\omega_k$.
- A rough approximation can be made in neglecting the no. of spikes from the spectrum.
- To test the significance, the periodicities (which are approximated to be significant) are removed from the original series to get a new series $\{Z_t\}$, where

$$Z_t = X_t - Y_t ,$$

Frequency Domain Analysis

$$Y_t = \mu + \hat{\alpha}_1 \cos\left(\omega_1 t\right) + \hat{\beta}_1 \sin\left(\omega_1 t\right) + \hat{\alpha}_2 \cos\left(\omega_2 t\right) + \hat{\beta}_2 \sin\left(\omega_2 t\right) +$$
$$\text{..........} + \hat{\alpha}_d \cos\left(\omega_d t\right) + \hat{\beta}_d \sin\left(\omega_d t\right)$$

where d is no. of periodicities removed (which are assumed to be significant)

- The spectrum of new series Z_t is plotted and the spikes are observed.
- A wrong conclusion may be made that these spikes are significant. However they need to be analyzed for their statistical significance

Frequency Domain Analysis

Statistical significance of the periodicities:

The periodicities are tested for significance by defining a statistic '∩' as follows (Kashyap and Rao 1976)

$$\text{I} = \frac{\gamma^2 (N-2)}{4\hat{\rho}_1}$$

Where $\gamma^2 = \alpha^2 + \beta^2$ and

$$\hat{\rho}_1 = \frac{1}{N}\left[\sum_{t=1}^{N}\left\{x_t - \hat{\alpha}\cos(\omega_k t) - \hat{\beta}\sin(\omega_k t)\right\}\right]$$

Ref: Kashyap R L and Ramachandra Rao A 'Dynamic stochastic models from empirical data', Academic press, New York, 1976

Frequency Domain Analysis

The periodicity corresponding to ω_k is significant at level α only if

$$I \geq F(2, N-2)$$

Where F denotes F distribution

- This test examines the periodicity at a time and should be carried out on a series from which all periodicities (previously found significant) are removed.

Frequency Domain Analysis

- A necessary condition in stochastic models is that the series being modeled must be free from any significant periodicities.
- One way of removing the periodicities from the time series is to simply transform the series into a standardized one.
- One method of standardizing the series $\{X_t\}$ is by expressing $\{X_t\}$ as the new series $\{Z_t\}$ where,

$$Z_t = \frac{\left(X_t - \bar{X}_i\right)}{S_i}$$

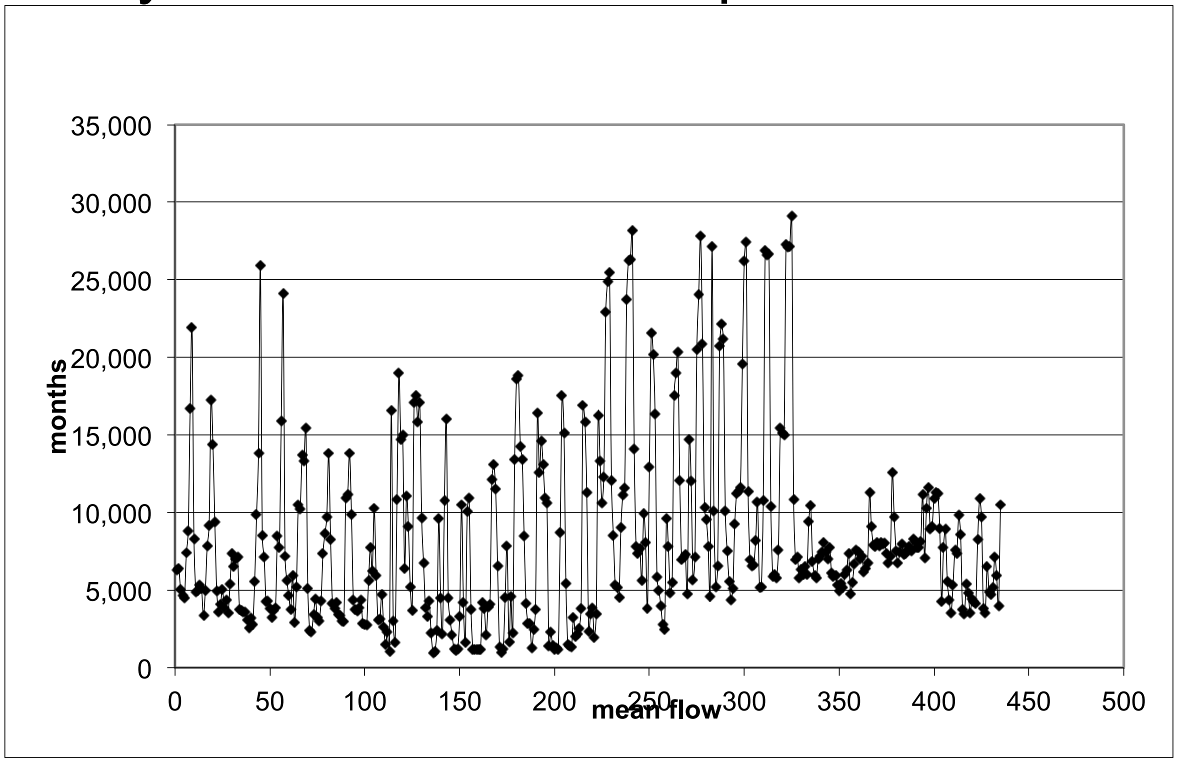
Frequency Domain Analysis

- X_i is the estimate of mean
- S_i is the estimate of the standard deviation
- The series has zero mean and unit variance.

- The series without periodicities is then obtained for which a stochastic (e.g., ARMA – Auto Regressive Moving Average) model is created.

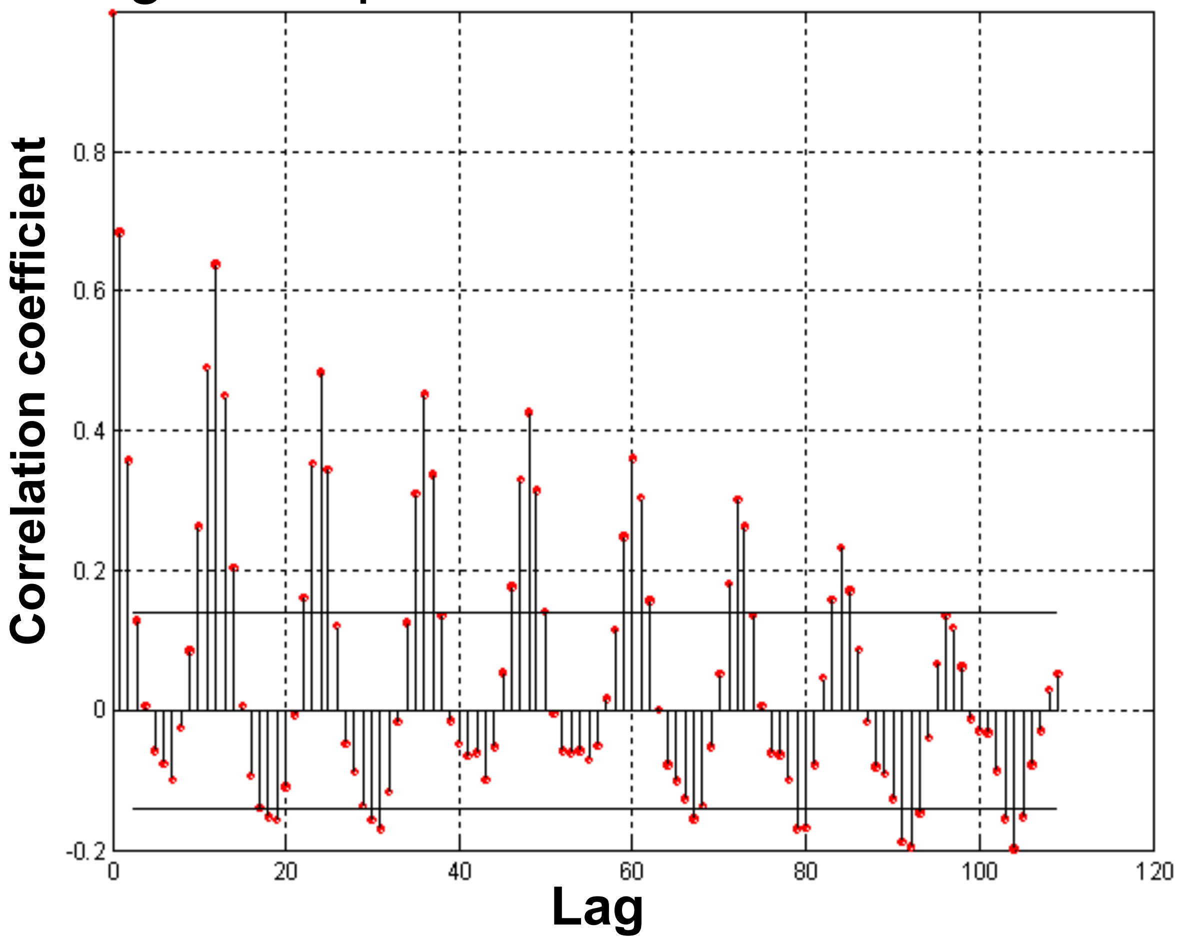
Example – 2

Monthly Stream flow (ft^2/sec) Statistics(1928-1964) for Missouri River near Wolf Point MT in Montana is selected for the study. The time series is plotted as follows



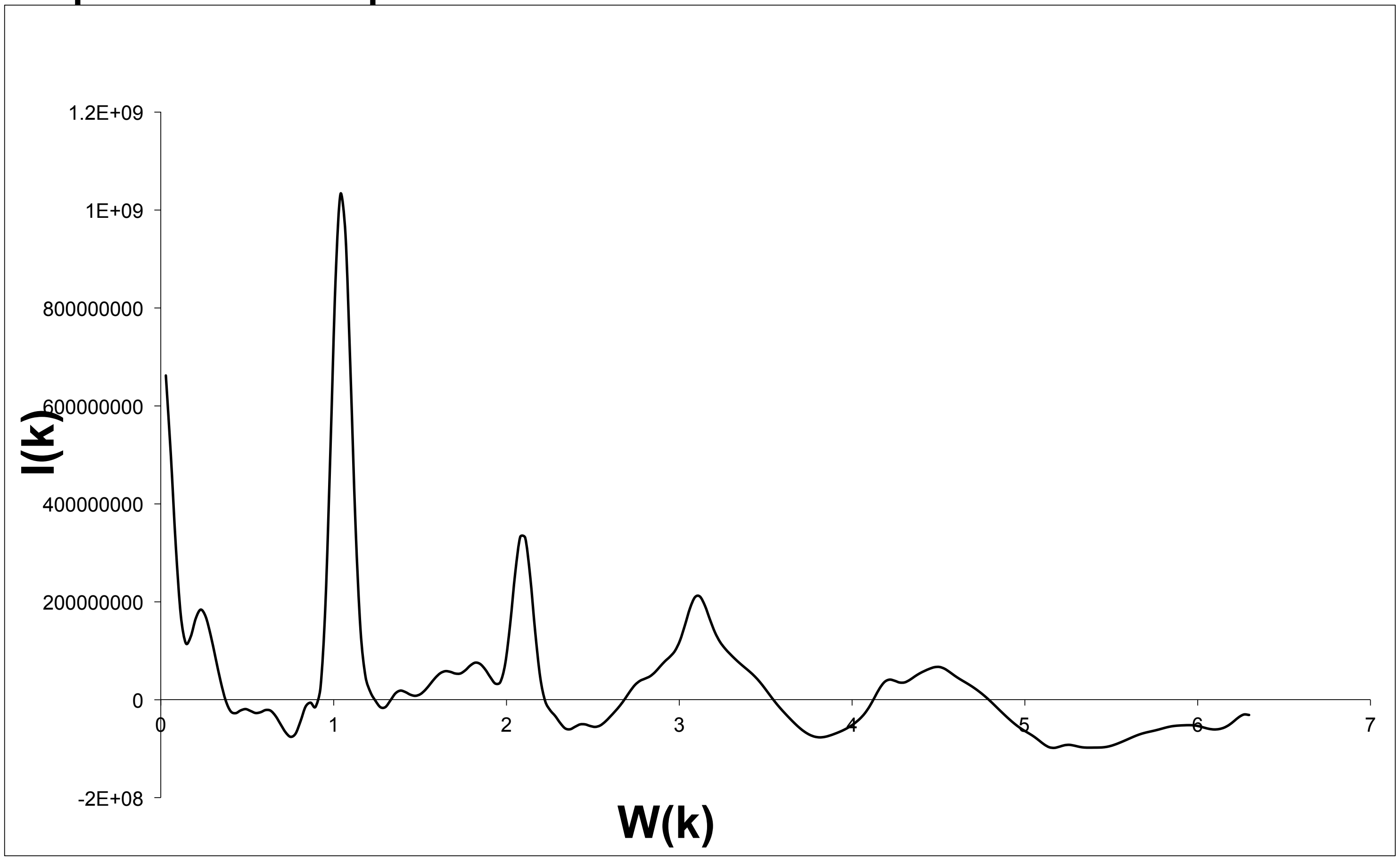
Example – 2 (contd.)

- Correlogram is plotted



Example – 2 (contd.)

- Spectrum is plotted



Frequency Domain Analysis

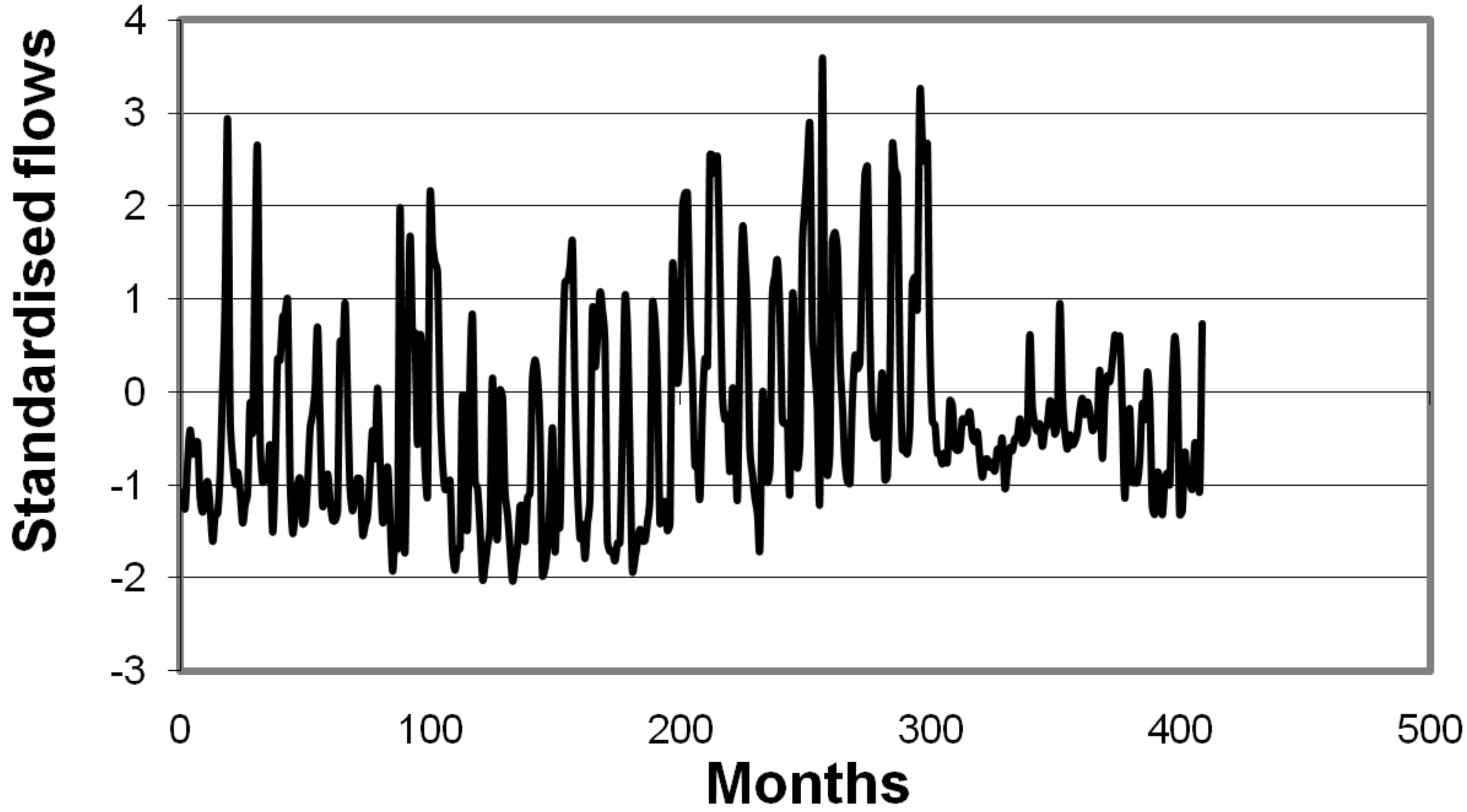
- Peaks represent the periodicities inherent in the data.
- The peaks correspond to w(k) = 0.0288 with a periodicity of 218 months,
- w(k) = 1.03 with a periodicity of 6 months,
- w(k) = 2.1088 with a periodicity of 3 months,
- w(k) = 3.1199 with a periodicity of 2 months and
- w(k) = 4.188 with a periodicity of 1.5 months and so on.
- The periodicities are tested for significance to avoid wrong conclusions.

Frequency Domain Analysis

- 218 months and 6 months periodicities are significant and 3 months, 2 months and 1.5 months periodicities are insignificant

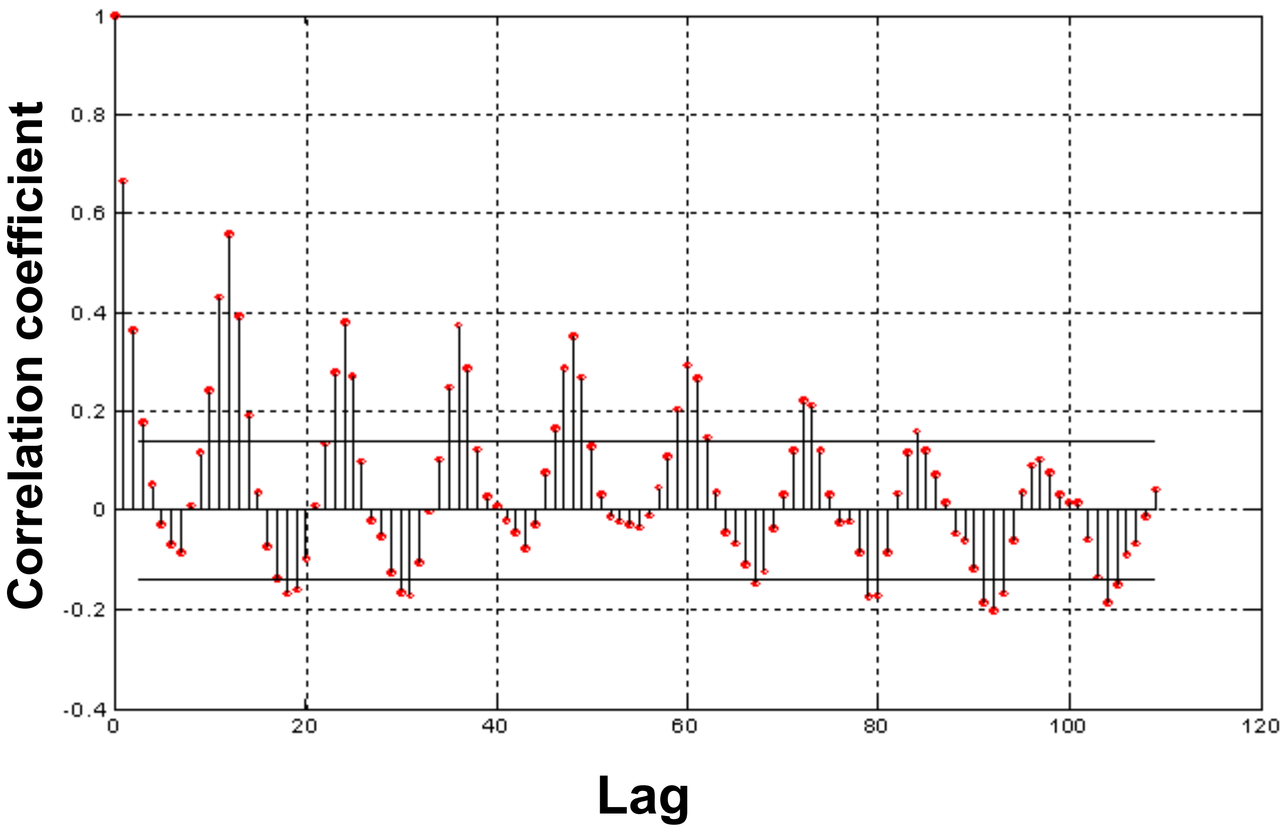
Frequency Domain Analysis

Time series of standardized data.



Frequency Domain Analysis

Correlogram of standardized data.



Frequency Domain Analysis

Spectrum of standardized data.