4. Data smoothing, Trend analysis and Harmonic analysis of data

Dr. Prasad Modak

Environmental Management Centre, Mumbai

Why do data smoothing?

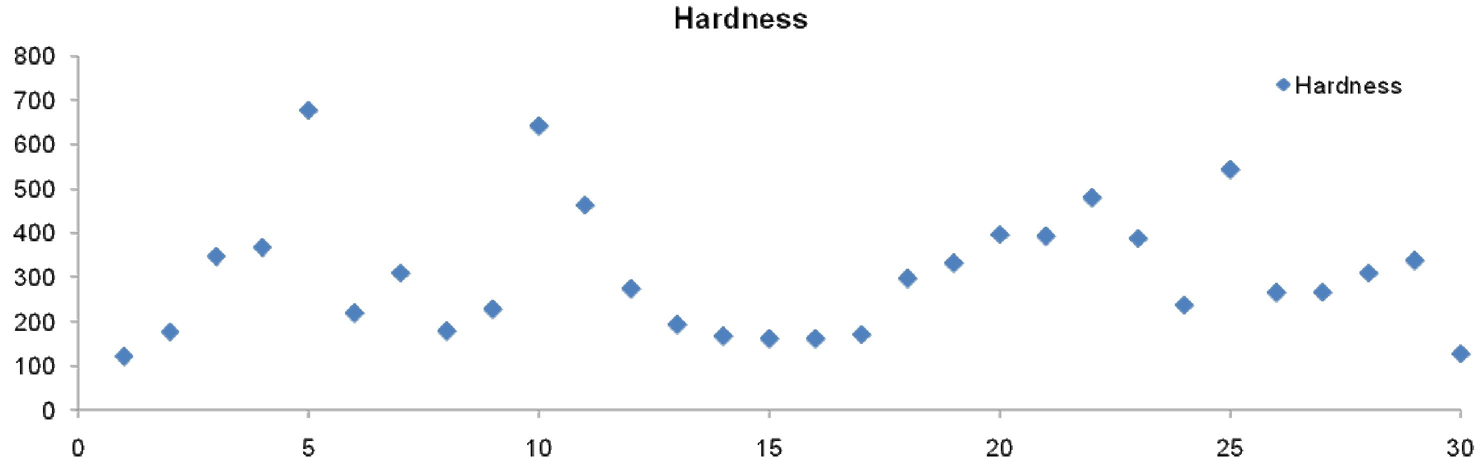
- Data smoothing is useful to reduce “noise” in the data.
- Minimizes the effects of cyclic trends (like seasonality) in data
- Reveal underlying patterns in the data
- When to data smoothing
  - Trend analysis of data
  - Long term reporting (annual report)

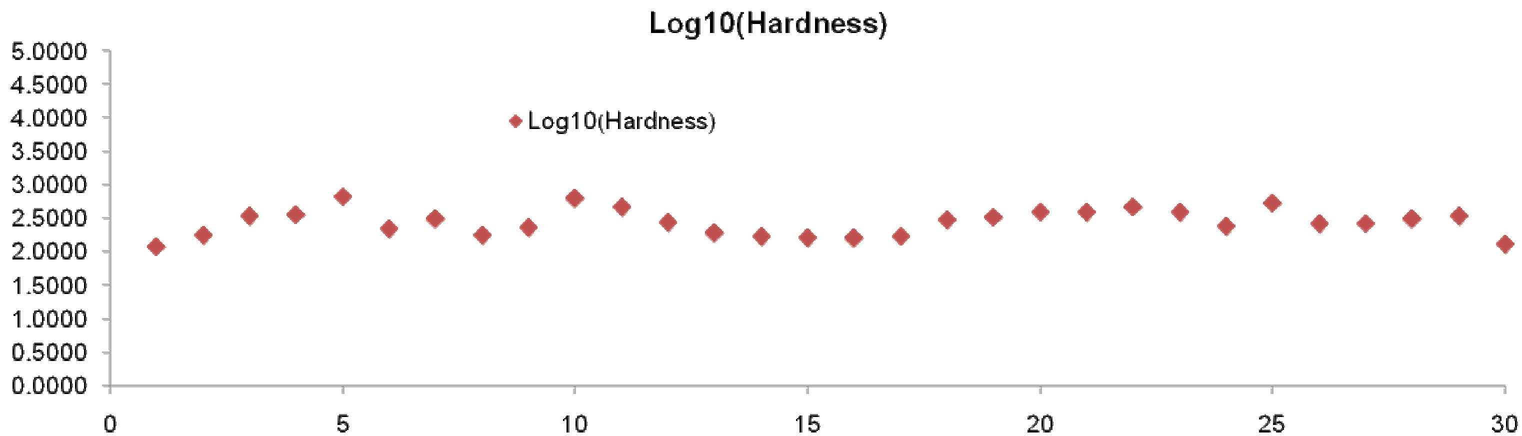
How to do Data Smoothening?

- How to smooth data ?
  - Apply transformation to data (e.g. Log)
  - Use Moving Average (MA)
  - Use Exponentially Weighed Moving Average (EWMA)

Using Log Scale

- Take $\log_{10}$(number) of numbers
- Very effective for large numbers



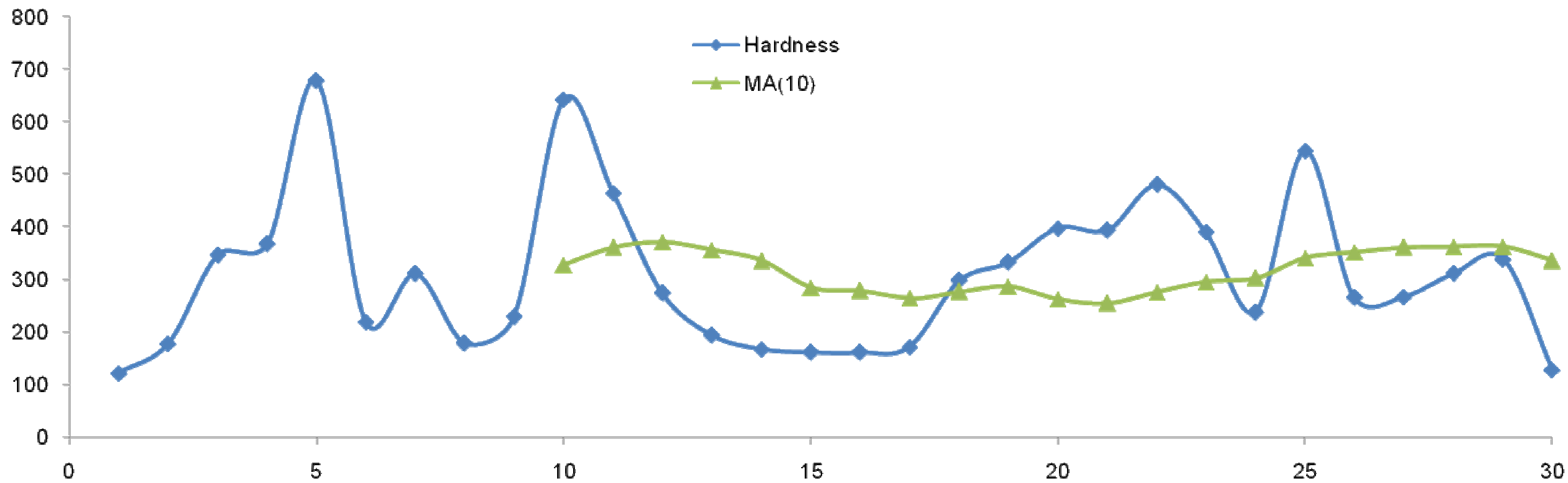


Using Moving Average

- MA is average of values for a defined period of time
- Simple average of the most recent '*k*' data points
- Ignores the k-1th day from calculations
- Excellent way to suppress spikes and establish trends

$$\overline{y}_i(k) = \frac{1}{k}\sum_{j=i-k+1}^{i} y_j$$

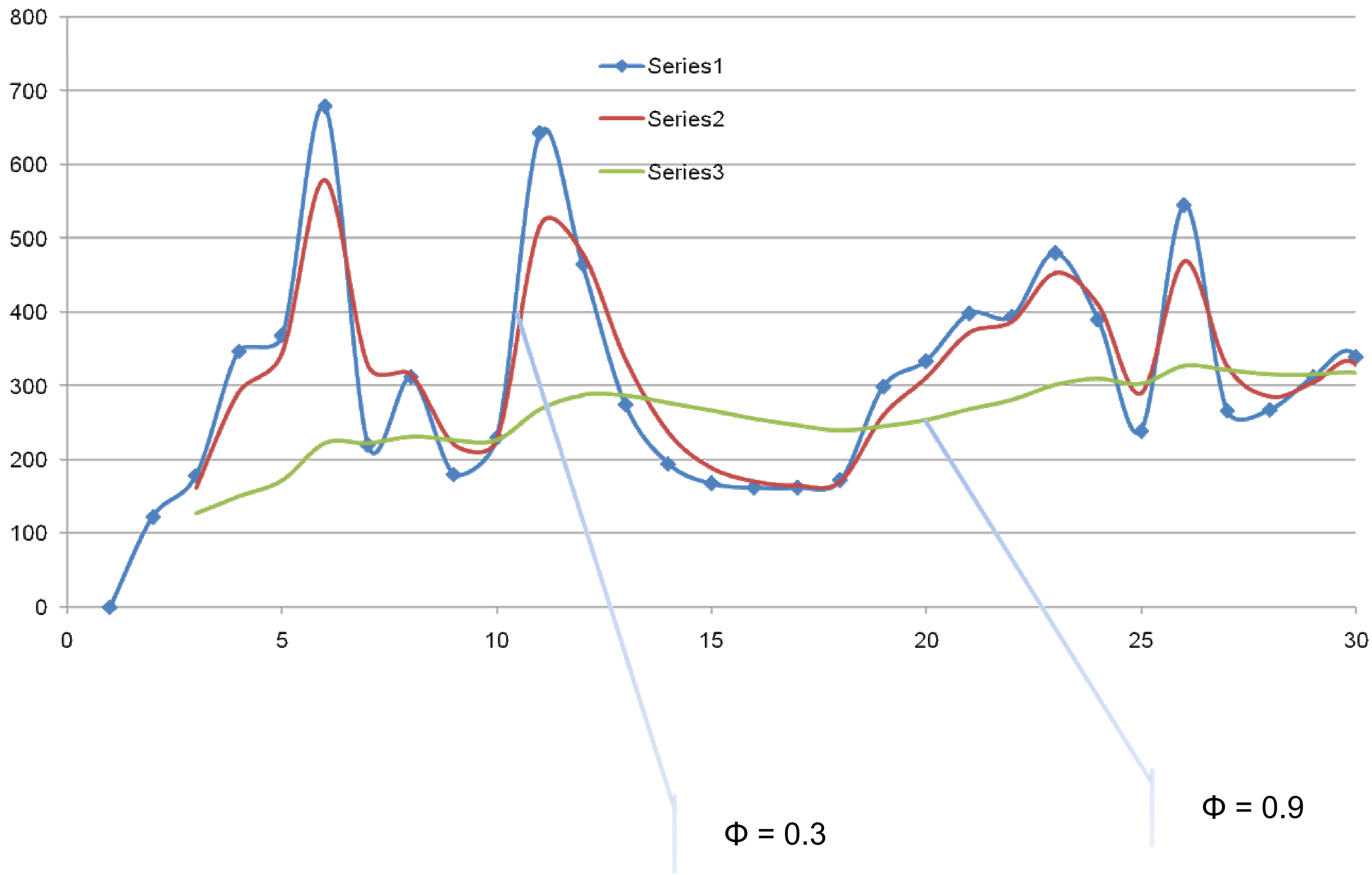
Where, i = k, k+1, ...n



Exponentially Weighted Moving Average

- EWMA gives more weightage to recent data than older data
- Influence of older data are essentially reduced in exponential manner during smoothening
- It has a feature Φ , which denotes the "memory" of the system
- Lower Φ denotes lower memory
- Equation is given by $\overline{Z}_i = (1-\Phi)\sum_{j=o}^{\infty}\Phi^j y_{i-j}$
- EWMA could be updated using: $\overline{Z}_i = \Phi\overline{Z}_{i-1} + (1-\Phi)y_i$

Example of EWMA



Trend Analysis (Spearman's Rho)

- Trend in environmental data
  - Seasonal
  - Non-seasonal
- Spearman's Rank Correlation Coefficient is often denoted by ρ.
- Spearman's Rho can be calculated and then compared to standard values to test the significance of the trend.
- Positive or negative ρ value indicate positive or negative trend
- Trend could be compared with standard values at a definite level of significance (90% or 95%) and N-2 degrees of freedom

Spearman's Rho

- It is given by,

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

- D is difference between the serial and rank of a data point
- N is number of data point
- If ties exists between serial and rank

$$\rho = \frac{M - (D + T_x + T_y)}{\sqrt{(M - 2T_x)(M - 2T_y)}}$$

- Where

$$M = \frac{N^3 - N}{6} \qquad T_x = \frac{t_x^{\ 3} - t_x}{12} \qquad T_y = \frac{t_y^{\ 3} - t_y}{12}$$

- t_x = Number of ties in ranks given to data
- t_y = Number of ties in the time series

Example Spearman's Rho

DO of Bhima River at Takli

Month	2009	2010
Jan	6.22	5.95
Feb	6.12	6.56
Mar	6.01	4.61
Apr	4.2	5.94
May	5.1	5.25
Jun	5.2	6.9
Jul	5.49	6.25

source : MPCB website

Month	2009	Serial no.	Rank	Ties
Jan	4.2	4	1	
Feb	5.1	5	2	
Mar	5.2	6	3	
Apr	5.49	7	4	
May	6.01	3	5	
Jun	6.12	2	6	
Jul	6.22	1	7	
		D=	**92**	

Month	2010	Serial no.	Rank	Ties
Jan	4.61	3	1	
Feb	5.25	5	2	
Mar	5.94	4	3	
Apr	5.95	1	4	
May	6.25	7	5	
Jun	6.56	2	6	
Jul	6.9	6	7	
		D =	**44**	

Example ….contd..

N =	7

N-2 =	5

2009

ρ	-1.63988
	Negative

Value 0.8 @95% confidence level

Significant

N =	7

N-2 =	5

2010

ρ	-0.78274
	Negative

Value 0.8 @95% confidence level

Insignificant

- It can be inferred with 95% confidence that there was a significant downward trend in the DO concentration in Bhima River in the period 2009. This trend is however no longer significant in 2010.

- Discussion question: Why? What could be the reason?

Harmonic Analysis

- Examines the periodicities or cyclic changes in the data
- Environmental data is prone to seasonal/annual cyclicity
- Harmonic Analysis is used to "filter" variation in data due to seasonal effects. Data is modeled by

$$f(x) = \overline{X} + \sum_{n=1}^{\infty} \left(A_n \cos\frac{2\pi nt}{N} + B_n \sin\frac{2\pi nt}{N}\right)$$

- Where

$$A_n = \frac{2}{N} \sum_{i=1}^{N} X_i \cos\frac{2\pi nt}{N}$$

$$B_n = \frac{2}{N} \sum_{i=1}^{N} X_i \sin\frac{2\pi nt}{N}$$

$\overline{X}$ = Mean of the data; n = n^{th} harmonic; t = Time from start, like the 2^{nd} month will be 2 in monthly data; N = Total number of observations; T = The periodicity of the data, = n/N ; and; X_i = Data corresponding to time t.

where

σ^2

Harmonics …contd…

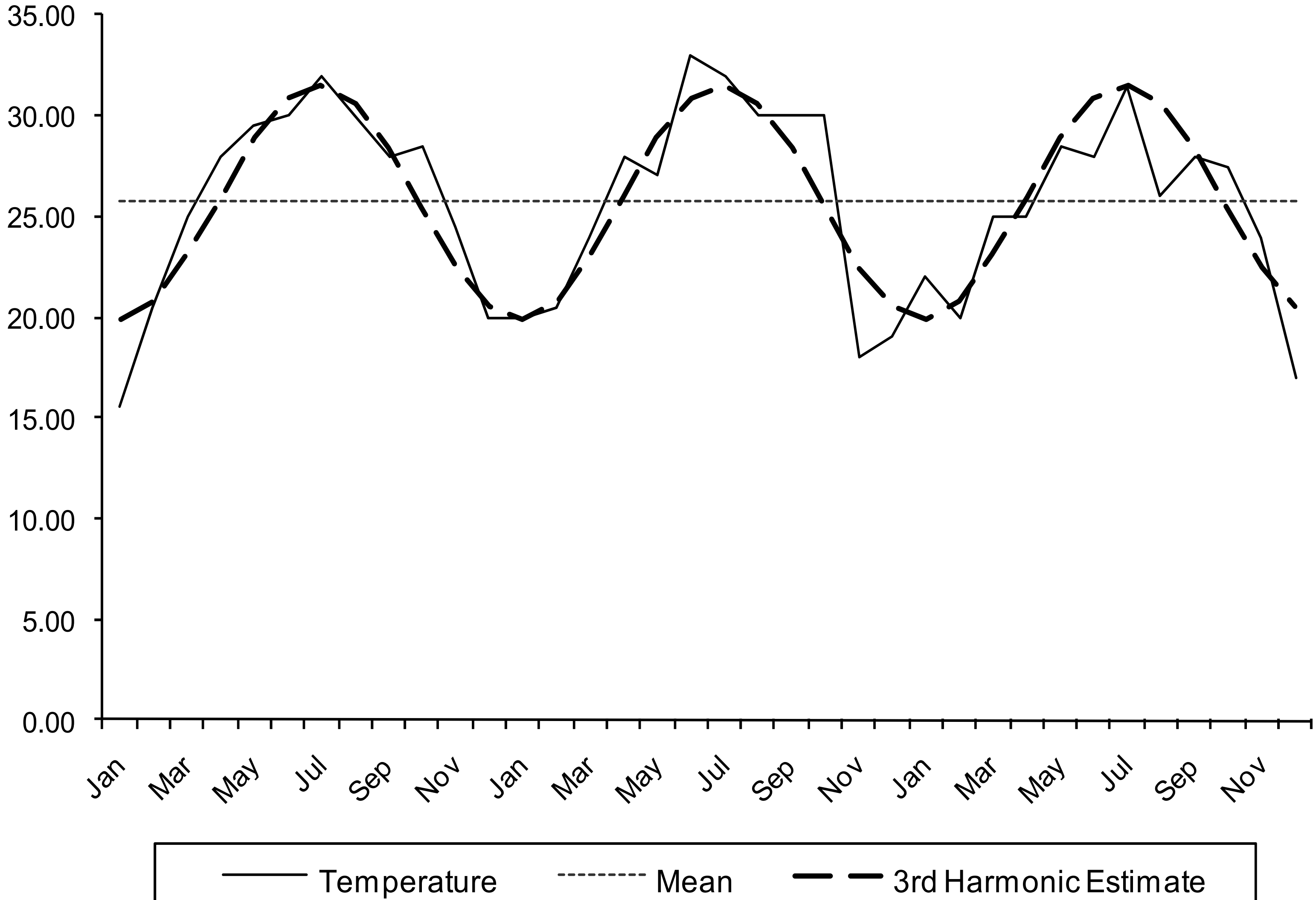
- In the harmonic series, if the calculation has been done for m harmonics, then sum up to n = 1 to m instead of ∞.
- Variance in the observed record accounted for by each harmonic is given by the following formula

$$\sigma^2 = \frac{1}{2}\sum_{t=1}^{N}(A_n{}^2 + B_n{}^2)$$

$$p = \frac{\sigma_n{}^2}{\sigma^2}$$

- where σ^2 is the total variance in the observed data.
- p indicates which period is dominant in the data.
- The largest value of p will show the most dominant period.

Example of Harmonic Analysis



Additional Points

- Trend Mapping – How could this mapping system be used for "source diagnosis"?
- Can Model for Harmonic Analyses be used for predictions?
- What if harmonics for flow and concentration show different lags or contribution to Variance?