12.714 Computational Data Analysis

Alan Chave (alan@whoi.edu)

Thomas Herring (tah@mit.edu),

http://geoweb.mit.edu/~tah/12.714

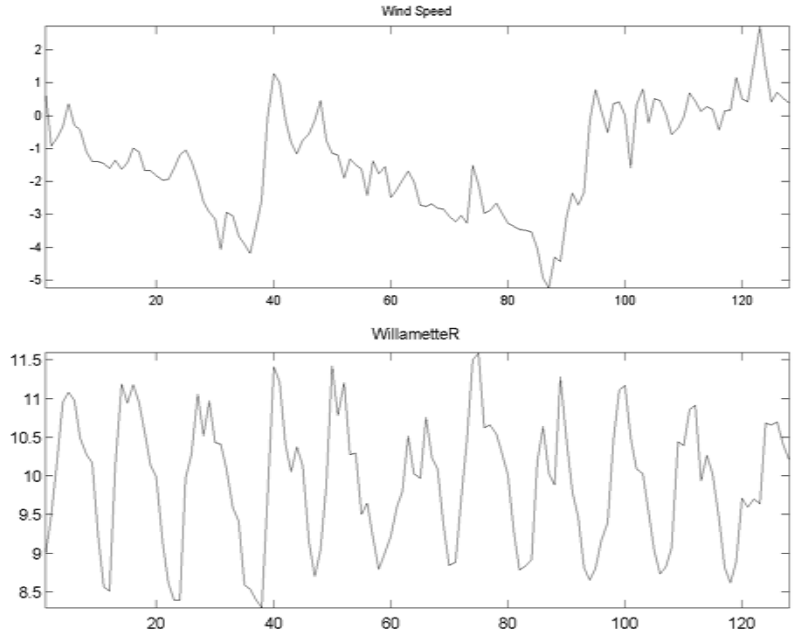
Introduction to Spectral Analysis

- Topics Today
  - Aspects of Time series analysis
  - Spectral Analysis for Simple Time Series Models
  - Non-Parametric Estimation of Spectra
  - Parametric Estimation of Spectra
  - Use of Spectral analysis
- Material based on Chapter 1 of Percival and Walden, Spectral Analysis for Physical Applications, Cambridge Univ. Press, pp. 583, 1993. In notes we use PW as abbreviation for this text.
- Web contents and Matlab scripts are available at: http://geoweb.mit.edu/~tah/12.714

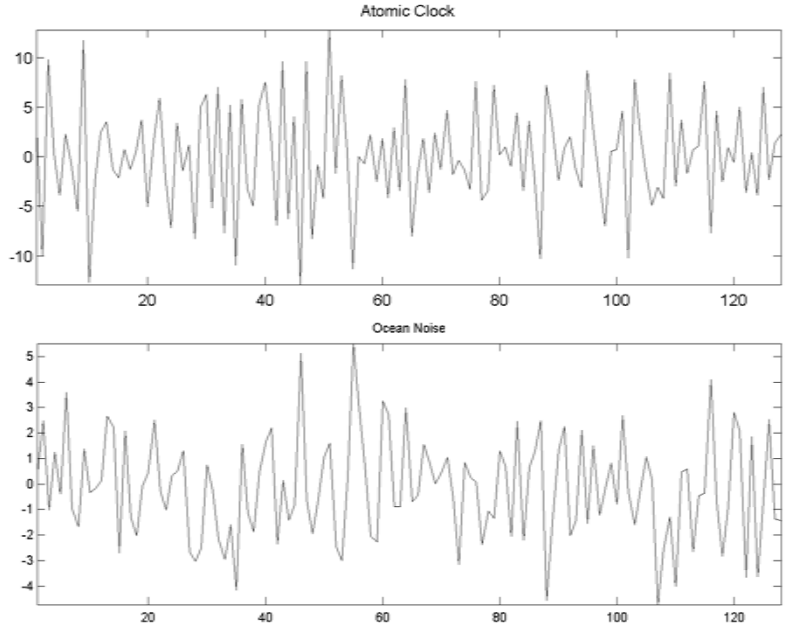
Aspects of Time Series Analysis

- Time Series: A set of observations made sequentially in time (or space)
- These series are often serially correlated (i.e., one value is not independent of an other) and the aim of time series analysis is to reveal the nature of correlations
- Spectral analysis is a subset of time series analysis
- Goal: To develop quantitative ways to characterize time series analysis and explain how they differ/relate.
- Two approaches: Analysis in time domain and frequency domain.

Examples



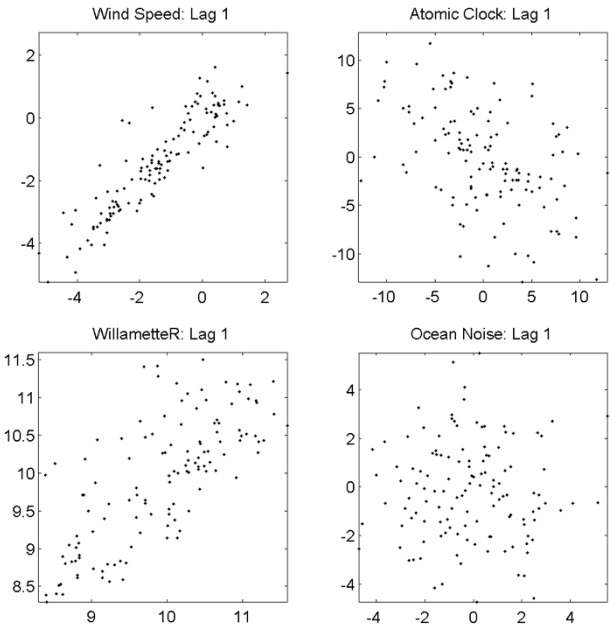
More Examples



Aim of time series analysis

- Aim: Develop quantitative means to characterize time series.
- Two broad classes:
  - Time domain techniques
  - Frequency domain methods: Spectral Analysis
- Time Domain: Lagged Scatter Plots
  - Plot the values in the time series against each other a fixed lag between them.
  - Lag 1 plot shown next

Lag 1 Plots



Notes:

Wind Speed falls on a positive slope

Willamette River is somewhat similar but more spread out

The Atomic points have negative slope

Ocean seem radomly scattered.

7

Autocorrelation sequence (acs)

- To measure the linear relationship between two ordered collections use the Pearson product moment correlation coefficient:

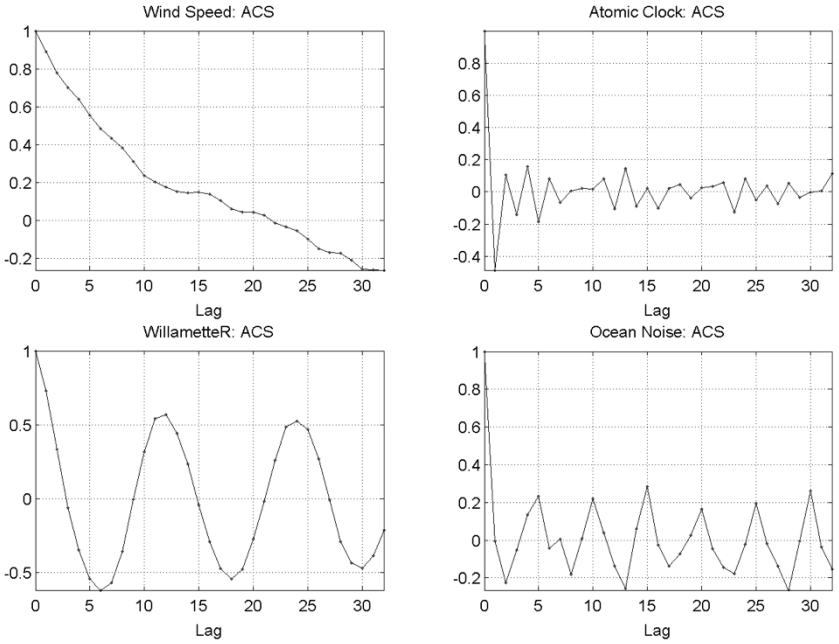
$$\hat{\rho} = \frac{\sum (y_t - \bar{y})(z_t - \bar{z})}{\sqrt{\sum (y_t - \bar{y})^2 \sum (z_t - \bar{z})^2}}$$

where $\bar{y}$ and $\bar{z}$ are sample means. Using $y_t = x_{t+k}$ and $z_t = x_t$

$$\hat{\rho}_k = \frac{\sum_{t=1}^{N-k} (x_{t+k} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{N} (x_t - \bar{x})^2}$$

Note ρ_o=1. The ρ_k for a sequence of k lags is called a sample autocorrelation sequence (acs).

ACS for examples



Comments on ACS

- River data are negatively correlated at k=6 and positively at k=12 (lags are in months) consistent with visual inspection of the time series with a clear seasonal cycle
- The atomic clock and ocean noise data are weakly correlated for $k > 0$
- The wind speed data remain highly correlated for large lags
- In the figure 2 methods (that generate identical results were used: Straight evaluation of Pearson's formula and a Fourier method from Box, Jenkins, Reinsel, pages 30-34, 188.

Modeling time series

- The times series values can be regarded as realizations of corresponding random variables X_i, I=1,N. Time series modeling is determining the properties of the N random variables.
- The ACS values are estimates of population theoretical autocorrelation at each lag.

$$\rho_k = E\{(X_t - \mu)(X_{t+k} - \mu)\}/\sigma^2$$

where E{.} is expectation operation,

μ is expectation E{X} and σ^2 is variance E{(X - μ)^2 }

Note: If ρ_k, μ, σ do not depend on time: Stationary process

Time series modeling

- When X follows a multivariate Gaussian distribution, the model of the time series is completely specified by knowledge of ρ_k, μ, σ.
- However, these parameters have some deficiencies
  - Experience needed to know what a time series will look given ρ_k, μ, σ.
  - Estimates of ρ_k are not necessarily reliable for lags near the length of the time series. Noise in estimates increases at longer lags; adjacent estimates are often correlated. Lack of homogeneity can make ρ_k hard to interpret.
  - Statistical tests can be difficult
  - Even when ρ_k is well known, other ways of viewing data can provide insights.
- Spectral analysis provides another set of tools.

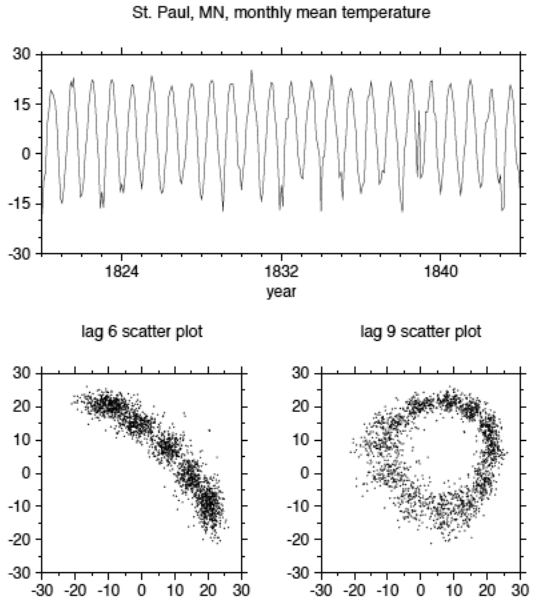
Comments on Lag Plots

- Question in ρ calculation should N/(N-k) scale result. In general answer is no. (Return to latter)

$$\hat{\rho}_k = \frac{\sum_{t=1}^{N-k}(x_{t+k}-\bar{x})(x_t-\bar{x})}{\sum_{t=1}^{N}(x_t-\bar{x})^2}$$

- There is a conflict with engineering definition of ACS (no division by variance).
- Certain types of functions can have interesting scatter plots (shown in next page)
- In Willamette River data, maybe the $E\{X_t\}$ depends on time due to seasonal nature. Such data can be made stationary with a mathematical trick.

St. Paul Temperature Record



24 year temperature record with monthly sampling.

Spectral Analysis for a simple time series model.

- Some problems with acs are lessened when a frequency domain characterization is used
- Key idea of spectrum is to express the time series as function of periodic functions

$$X_t = \mu + \sum_f [A(f)\cos(2\pi ft) + B(f)\sin(2\pi ft)]$$

- To represent a time series we will specific frequencies

$$X_t = \mu + \sum_{j=1}^{\lfloor N/2 \rfloor} [A_j \cos(2\pi f_j t) + B_j \sin(2\pi f_j t)] \quad t = 1,2,\cdots,N$$

$\lfloor z \rfloor$ is greatest integer less than or equal to z; $f_j = j/N \quad 1 \le j \le \lfloor N/2 \rfloor$

Spectral Model

- We assume: $E\{A_j\}=E\{B_j\}=0$, $E\{A_j^2\}=E\{B_j^2\}=\sigma_j^2$. Notice that each frequency has its own variance. Also assume that $E\{A_jA_k\} = E\{B_jB_k\} = E\{A_jB_k\} = 0$ for all j and k.
- With these assumptions: $E\{X_t\}=\mu$ and

$$\sigma^2 = E\{(X_t - \mu)^2\} = \sum_{j=1}^{\lfloor N/2 \rfloor} \sigma_j^2$$

$$\rho_k = \frac{\sum_{j=1}^{\lfloor N/2 \rfloor} \sigma_j^2 \cos(2\pi f_j k)}{\sum_{j=1}^{\lfloor N/2 \rfloor} \sigma_j^2}$$

In this model we define the spectrum by

$$S_j \equiv \sigma_j^2, 1 \le j \le \lfloor N/2 \rfloor$$

Spectral Model

- The equation for ρ_k says we can determine the acs and σ^2 if we know the spectrum and visa-versa.
- Non-parametric spectrum estimation estimates the coefficients A and B directly (first method developed).
- Estimation of the spectrum from data proceeds from the theoretical expressions using a non-parametric approach. For *N* data, we have 2[N/2] sinusoids plus the mean, or *M* = 2[N/2] + 1 quantities. ([.] is being used as the floor symbol).
- This is *N* for *N* odd, and *N* + 1 for *N* even, but $B_{N/2}$ is not used in the latter case because
  sin ($2\pi f_{N/2} t$) = sin (πt) = 0 for all integer *t*. Therefore, there are *M* = *N* parameters in the expression no matter if N is odd or even.

Fourier Orthogonality

- In working with Fourier series (and other types of basis functions), advantage can be taken of the orthogonal nature of the sine and cosine functions provided the frequencies are selected appropriately.

$$\sum_{t=1}^{N} \cos(2\pi f_k t)\cos(2\pi f_j t) = \begin{cases} 0 & \text{if } k \neq j \\ N/2 & \text{if } k = j \end{cases}$$

$$\sum_{t=1}^{N} \sin(2\pi f_k t)\sin(2\pi f_j t) = \begin{cases} 0 & \text{if } k \neq j \\ N/2 & \text{if } k = j \end{cases}$$

$$\sum_{t=1}^{N} \cos(2\pi f_k t)\sin(2\pi f_j t) = 0 \text{ for all } k \text{ and } j.$$

$$\text{for } f_j = j/N \text{ and } f_k = k/N$$

Non-Parametric Spectrum Estimation

- Due to the orthogonality of sines and cosines when the time series is multiplied and summed over integer periods, the equations for the A_j and B_j coefficients can be found by:

$$A_j = \frac{2}{N}\sum_{t=1}^{N} X_t \cos(2\pi f_j t)$$

if $1 \le j \le N/2$ and, if N is even

$$A_{N/2} = \frac{1}{N}\sum_{t=1}^{N} X_t \cos(2\pi f_{N/2} t) \quad \text{(Only cos term here)}$$

$$B_j = \frac{2}{N}\sum_{t=1}^{N} X_t \sin(2\pi f_j t); \overline{X} \equiv \frac{1}{N}\sum X_t = \mu$$

Spectrum Estimation

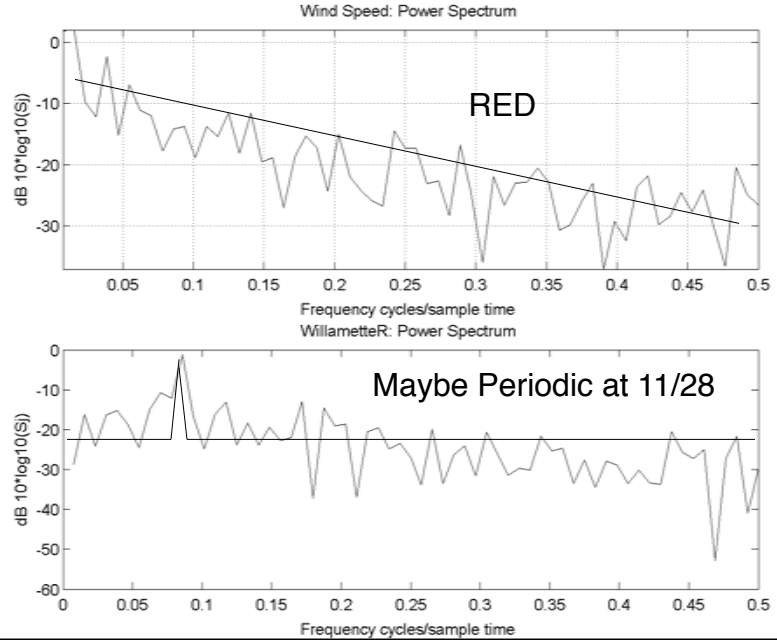
- The estimates of S_j are simply given by the previous equations evaluated with the realization of X_t denoted x_t

$$\hat{S}_j = \frac{A_j^2 + B_j^2}{2} \text{ with } A_j^2, B_j^2 \text{ evaluated with } x_t$$

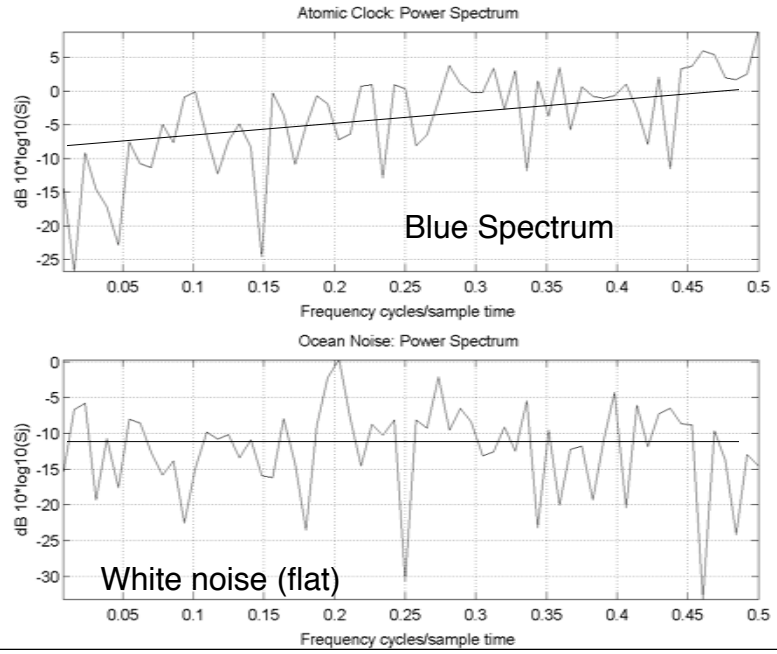
for $1 \le j \le N/2$ and if N is even

$$\hat{S}_{N/2} = \frac{1}{N^2}\left(\sum_{t=1}^{N} x_t \cos(2\pi f_{N/2} t)\right)^2$$

Spectra of Examples



Spectra Examples



Points about non-parametric estimates

- Since the A_j and B_j coefficients are uncorrelated, the S_j are approximately uncorrelated (true if A_j an B_j are Gaussian). This contrasts to the ρ_k estimated that are sequentially correlated.
- Statistical tests on S_j are easier to formulate because uncorrelated.
- Since S_j is only two-degree of freedom estimates of σ_j^2, there is variability in the spectrum. If the underlying process is smooth, then the spectrum can be smoothed.
- Later we see the variance of log(Sj) is similar for all j, thus noise easier to interpret.

Parametric Estimation of Spectra

- In some cases, the form of spectra may be known and the parameters of the form estimated from the data. The spectrum is generated by substituting the parameters estimates back in the form.
- Example:

This form derives from a "first-order autoregressive" process

Wind Speed and Atomic clock of this form

$$S_j(\alpha,\beta) = \frac{\beta}{1+\alpha^2 - 2\alpha\cos(2\pi f_j)}$$

$$\hat{\alpha} \approx \rho_1 \text{ and since } \sum \hat{S}_j(\alpha,\beta) \equiv \hat{\sigma}^2$$

$$\hat{\beta} = \hat{\sigma}^2 \left(\sum \frac{1}{1+\alpha^2 - 2\alpha\cos(2\pi f_j)} \right)^{-1}$$

Parametric estimates

- Two main difficulties with parametric estimates:
  - Difficult to characterize the statistical properties of the result spectrum. Without making additional assumptions, 95% confidence interval can not be calculated
  - Class of functional form may not be obvious: Too many parameters may not be reliably estimated; too few parameters might not represent the spectrum well
- Can be useful in treating a smaller data set when the form of the parametric model can be obtained from other larger data sets.

Uses of Spectral Analysis

- Testing theories: Physical models may predict a particular spectra shape.
- Investigating data: Spectral analysis allows general nature of time series to be inferred.
- Discriminating data: Allows differences between data to be assessed.
- Performing diagnostic tests: Analysis of spectrum of residuals after parameter fit. One caution some of the noise in the data is removed in the parameter estimation.
- Assessing predictability of time series.

Summary

- Today's class
  - Aspects of Time series analysis
  - Spectral Analysis for Simple Time Series Models
  - Non-Parametric Estimation of Spectra
  - Parametric Estimation of Spectra
  - Use of Spectral analysis
- Next Class
  - Stationary Stochastic processes.