Lecture 11

Exploratory data analysis Part 2 and smoothing

Arnab Hazra

Illustration of log-transformation

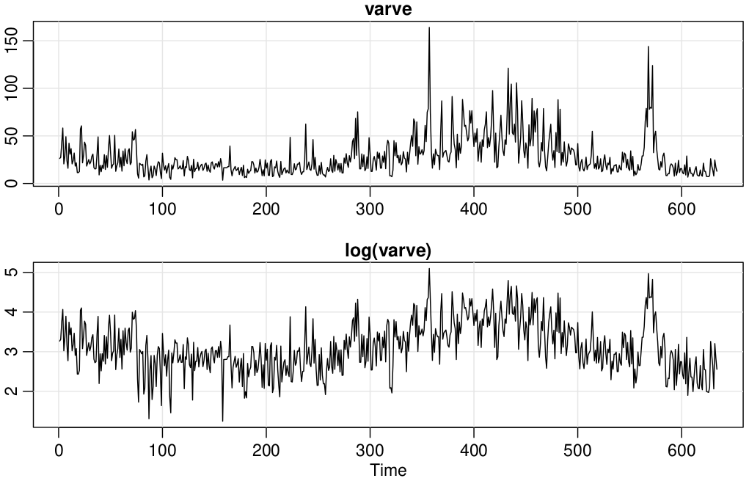


Fig. 2.7. *Glacial varve thicknesses (top) from Massachusetts for $n = 634$ years compared with* log *transformed thicknesses (bottom).*

Logarithmic and Box-Cox transformations

- Obvious aberrations can contribute nonstationary as well as nonlinear behavior in observed time series.

- In such cases, transformations may be useful to equalize the variability over the length of a single series.

- A particularly useful transformation is $Y_t = \log(X_t)$ which tends to suppress larger fluctuations that occur over portions of the series where the underlying values are larger.

- Other possibilities are power transformations in the Box-Cox family of the form $Y_t = (X_t^\lambda - 1)/\lambda$ if $\lambda \neq 0$ and $Y_t = \log(X_t)$ with $\lambda = 0$.

SOI and Fish Population (Recap)

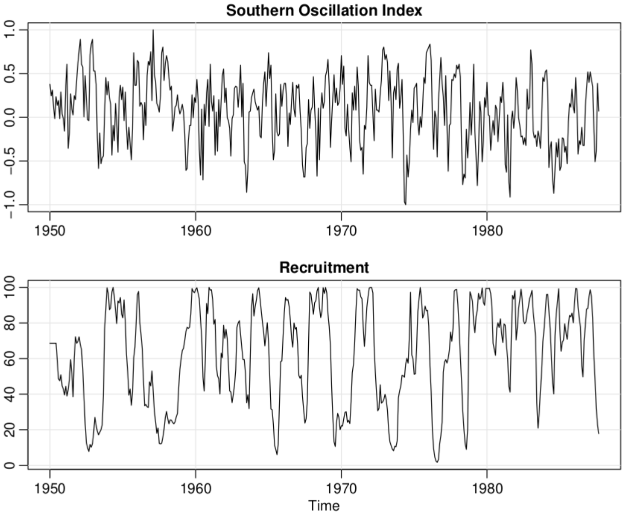
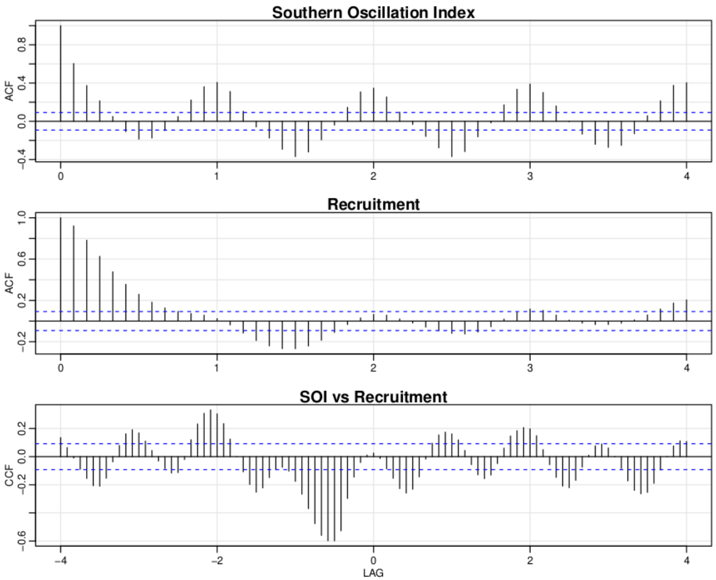
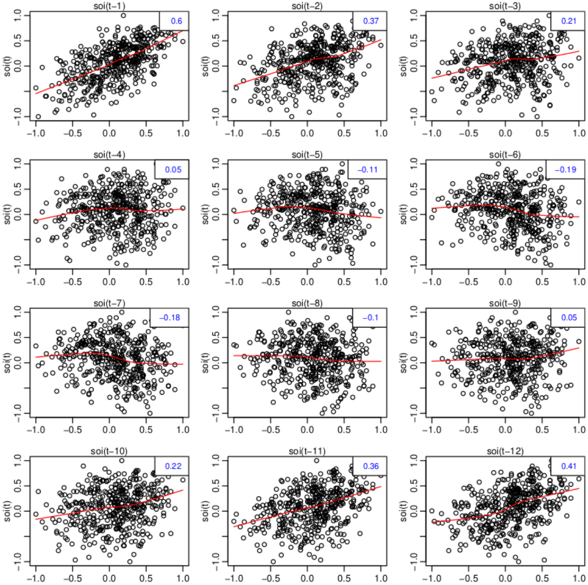


Fig. 1.5. Monthly SOI and Recruitment (estimated new fish), 1950-1987.

Sample ACF and CCF (Recap)



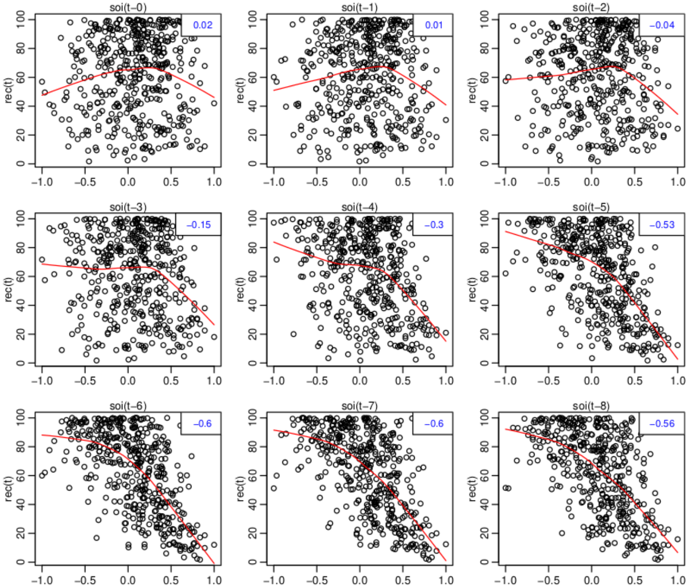
Scatterplot matrix of the lagged same series



Exploration of nonlinear relationships

- To check for nonlinear relations of this form, it is convenient to display a lagged scatterplot matrix.
- We notice that the lowess fits are approximately linear, so that the sample autocorrelations are meaningful.
- Also, we see strong positive linear relations at lags $h = 1, 2, 11, 12$, that is, between S_t and $S_{t-1}, S_{t-2}, S_{t-11}, S_{t-12}$, and a negative linear relation at lags $h = 6, 7$.
- Similarly, we might want to look at values of one series, say Recruitment, denoted R_t plotted against another series at various lags, say the SOI, S_{t-h}, to look for possible nonlinear relations between the two series.

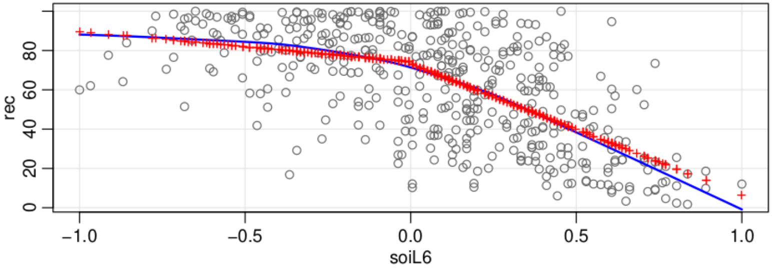
Scatterplot matrix of the lagged different series



Dummy covariates

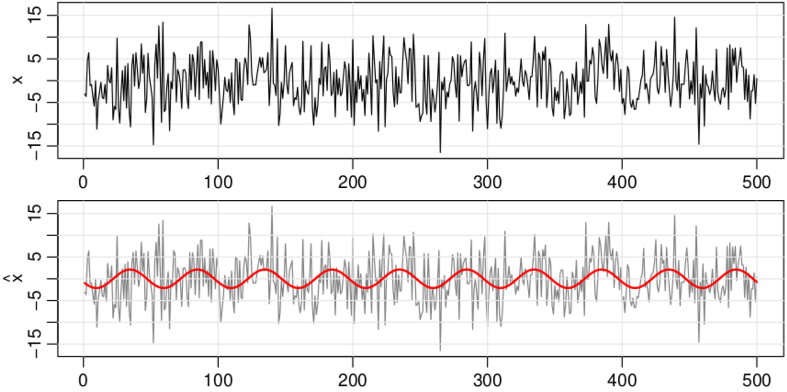
- The relationship between R_t and S_{t-6} is nonlinear and different when SOI is positive or negative.
- In this case, we may consider adding a dummy variable to account for this change. Define D_t a dummy variable that is 0 if $S_t < 0$ and 1 otherwise.
- In particular, we fit the model

$$R_t = \beta_0 + \beta_1 S_{t-6} + \beta_2 D_{t-6} + \beta_3 D_{t-6} S_{t-6} + W_t.$$



Discovering a signal in noise

- The data are simulated from $X_t = A\cos(2\pi\omega t + \phi) + W_t$.
- We have $A\cos(2\pi\omega t + \phi) = \beta_1 \cos(2\pi\omega t) + \beta_2 \sin(2\pi\omega t)$ where $\beta_1 = A\cos(\phi)$ and $\beta_2 = -A\sin(\phi)$.
- Assuming the frequency of oscillation $\omega = 1/50$ is known, we can fit a regression model $X_t = \beta_1 \cos(2\pi t/50) + \beta_2 \sin(2\pi t/50) + W_t$.



Moving average

- This method is useful in discovering certain traits in a time series, such as long-term trend and seasonal components.

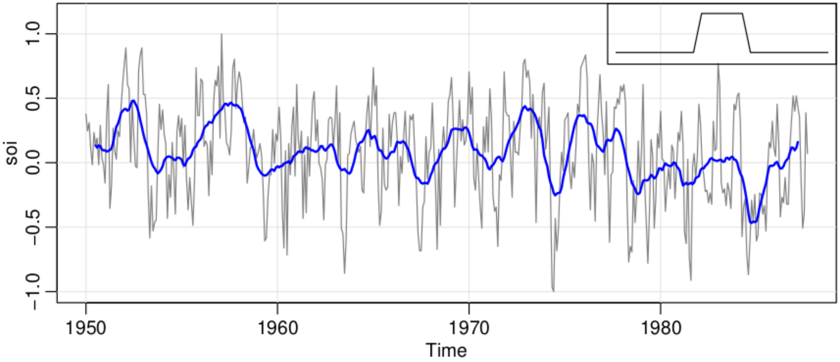
- In particular, if X_t represents the observations, then

$$m_t = \sum_{j=-k}^{k} a_j X_{t-j}$$

where $a_j = a_{-j} \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$ is a symmetric moving average of the data.

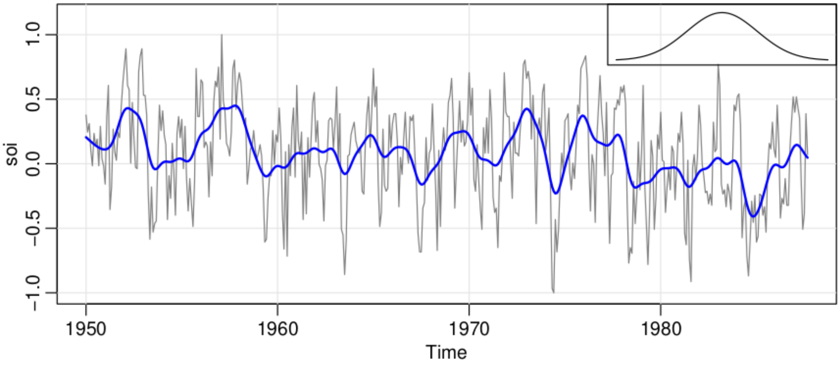
Moving average smoother: Illustration for SOI data

- Suppose we choose weights $a_0 = a_{\pm 1} = \ldots = a_{\pm 5} = 1/12$, and $a_{\pm 6} = 1/24$ and $k = 6$.
- This particular method filters out the obvious annual temperature cycle and helps emphasize the El Nino cycle.



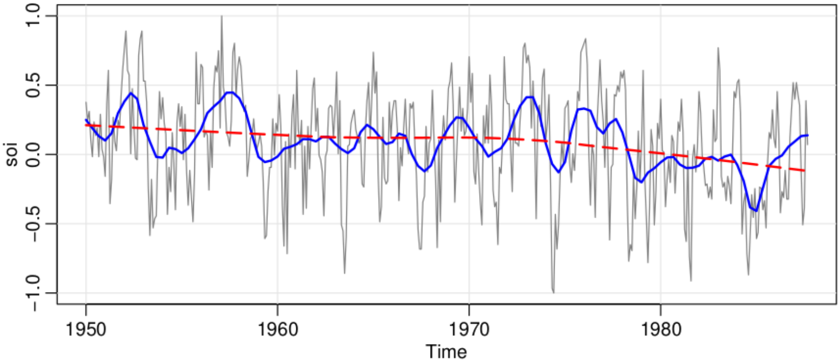
Kernel smoother: Illustration for SOI data

- kernel smoothing of the SOI series, where mt is $m_t = \sum_{i=1}^{T} w_i(t) X_i$, where $w_i(t) = K\left(\frac{t-i}{b}\right) / \sum_{j=1}^{T} K\left(\frac{t-j}{b}\right)$.
- Here the typical choice is $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$.



Lowess smoother: Illustration for SOI data

- The technique is based on k-nearest neighbors regression, wherein one uses only the data $\{X_{t-k/2}, \ldots, X_t, \ldots, X_{t+k/2}\}$ to predict X_t via regression, and then sets $m_t = \hat{X}_t$.
- Here one (blue) smoother uses 5% of the data and another (red) uses 2/3 of the data to obtain an estimate of the El Nino cycle of the data.



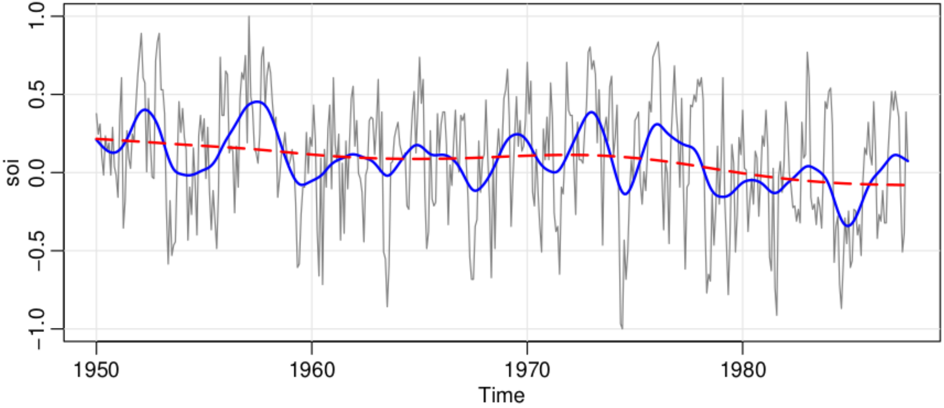
Spline smoother

- An obvious way to smooth data would be to fit a polynomial regression in terms of time.
- For example, a cubic polynomial would have $X_t = m_t + W_t$ where $m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$.
- An extension of polynomial regression is to first divide time $t = 1, \ldots, T$, into k intervals, $[t_0 = 1, t_1], [t_1 + 1, t_2], \ldots, [t_{k-1} + 1, t_k = T]$ and then, in each interval, one fits a polynomial regression; the values $t_0, t_1, \ldots, t_k$ are called knots.
- A related method is smoothing splines, which minimizes a compromise between the fit and the degree of smoothness given by

$$\sum_{t=1}^{T} (X_t - m_t)^2 + \lambda \int (m_t^{''})^2 dt$$

- The degree of smoothness is controlled by $\lambda > 0$.

Spline smoother: Illustration for SOI data



Thank you!