SESSION 3:
DATA ANALYTICS

htp://uiuc-cse.github.io/matlab-sp17/

http://pad.software-carpentry.org/matlab-sp17

OUTLINE

- Intro to basic statistical functions
- Working on a real data set (data_ColetoCreek.csv)
    - Data description and data access
    - Data cleaning
    - Descriptive statistics
    - Data smoothing
    - Correlation

BASIC STATISTICAL FUNCTIONS

```
clear all;
clc;

y = rand(30,1)*100; %data
min(y)
max(y)
mode(y)
std(y)
avg = @(x) sum(x)/length(x)
avg(y)
mean(y)
z = y(y>50)
idx = find(y <50)
```

DATA DESCRIPTION AND DATA ACCESS

- Weather data of Coleto Creek Reservoir from 2003 to 2014
  **source:NOAA (http://www.noaa.gov/)
- 4197 daily observations
- The columns of the data: [Latitude, Longitude, Date, ET, Prcp, Tmax, Tmin]
- To access data – ‘textread’ function
  - Read list of numbers, one per line:
    You can use the asterisk (*) in a field to ignore that field.
    [c1 c2] = textread(‘file’, ‘%f %f %*f %*f %*f %*f %*f’, more options...)
  - Read a matrix of numbers:
    Matrix = textread(‘file’, ‘’, more options…)

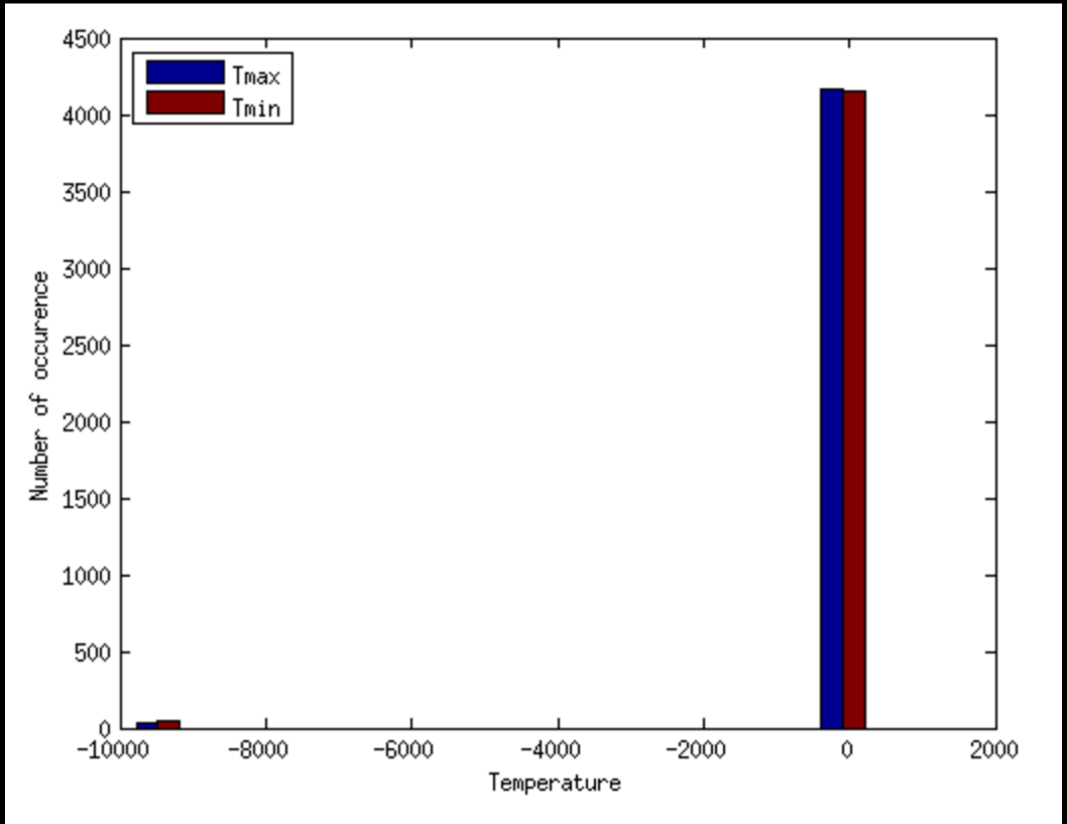
DATA CLEANING & DESCRIPTIVE STATISTICS

- Cleaning missing values and outliers
  - hist- outlier
  - setdiff
- Plot over time (time series) and boxplot
- Perform basic statistical analysis

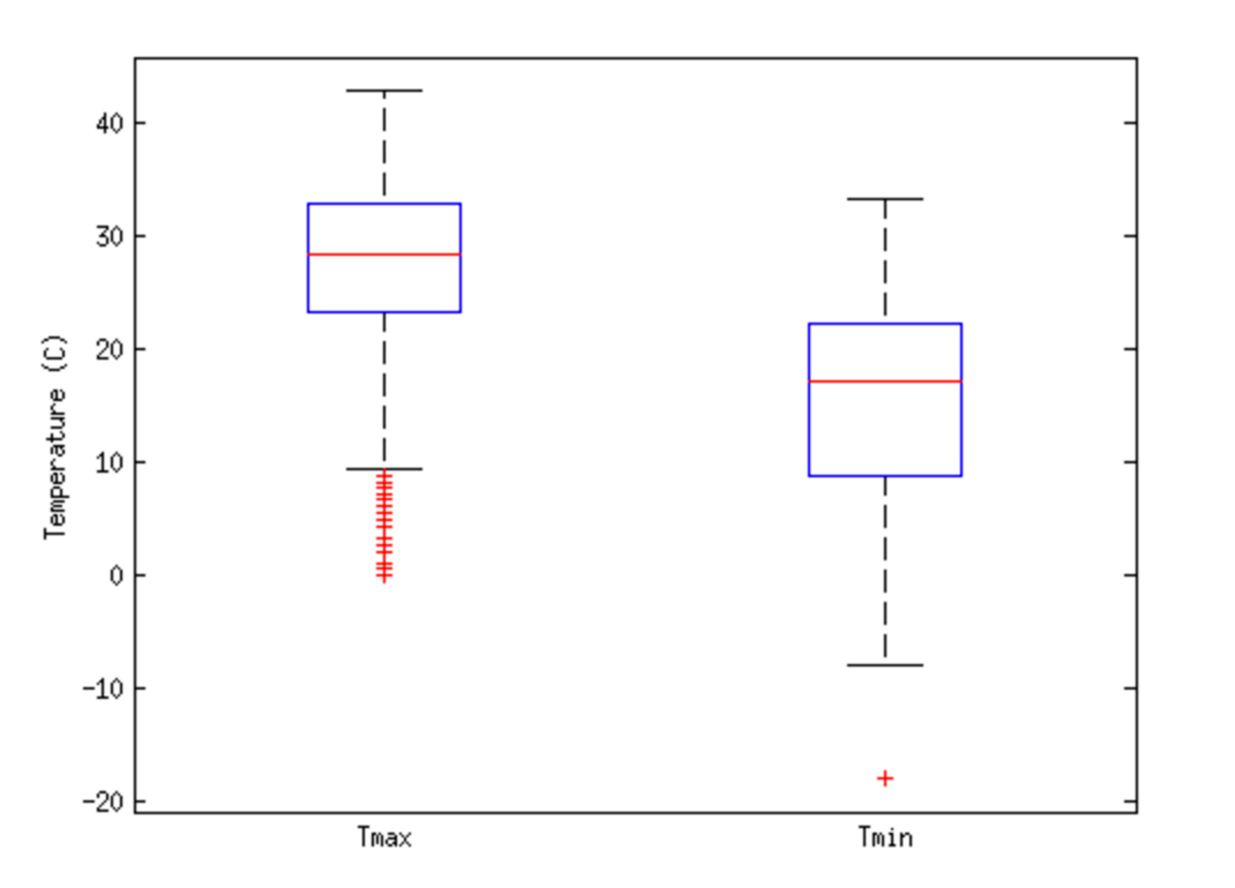
for subset of data:

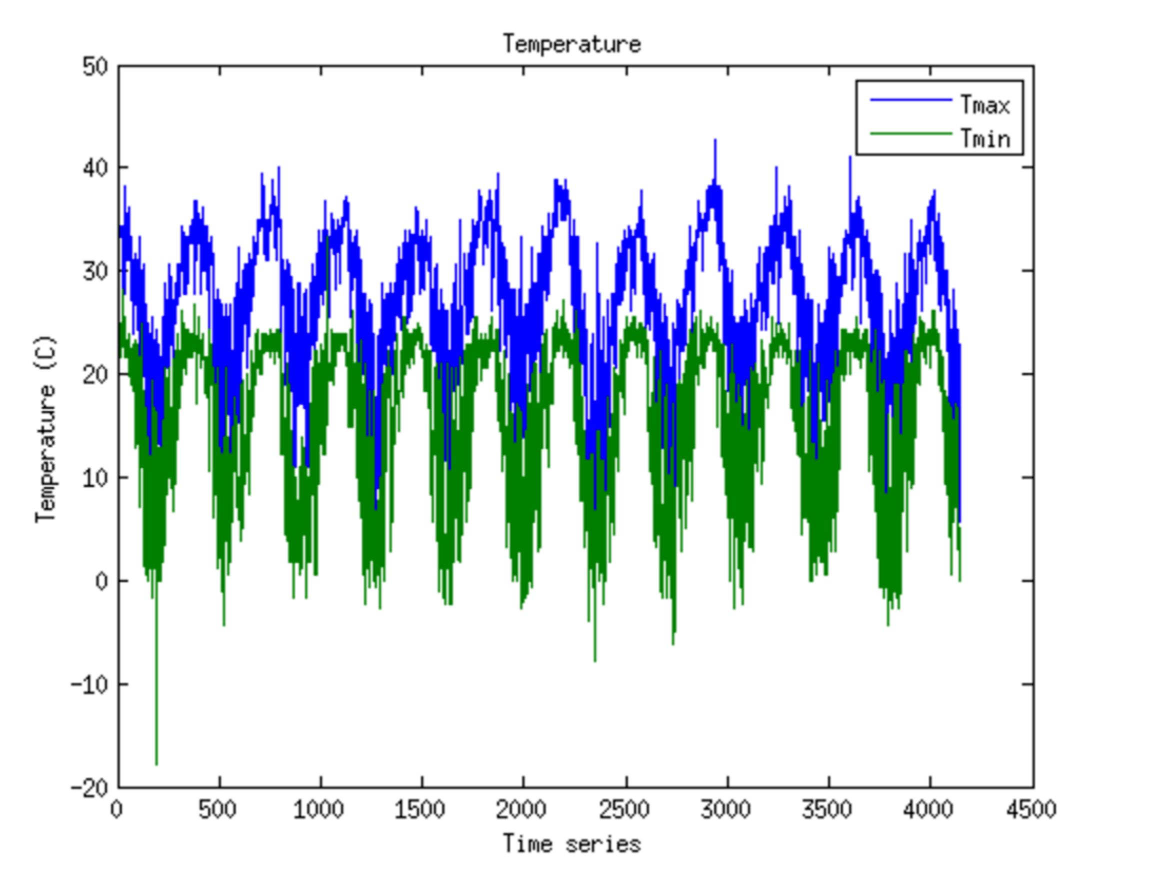
  - min, max, mean, median, mode, std
- Extract data that meets certain condition

  indices= find(data(:,k)>a)
  data(indices,:)

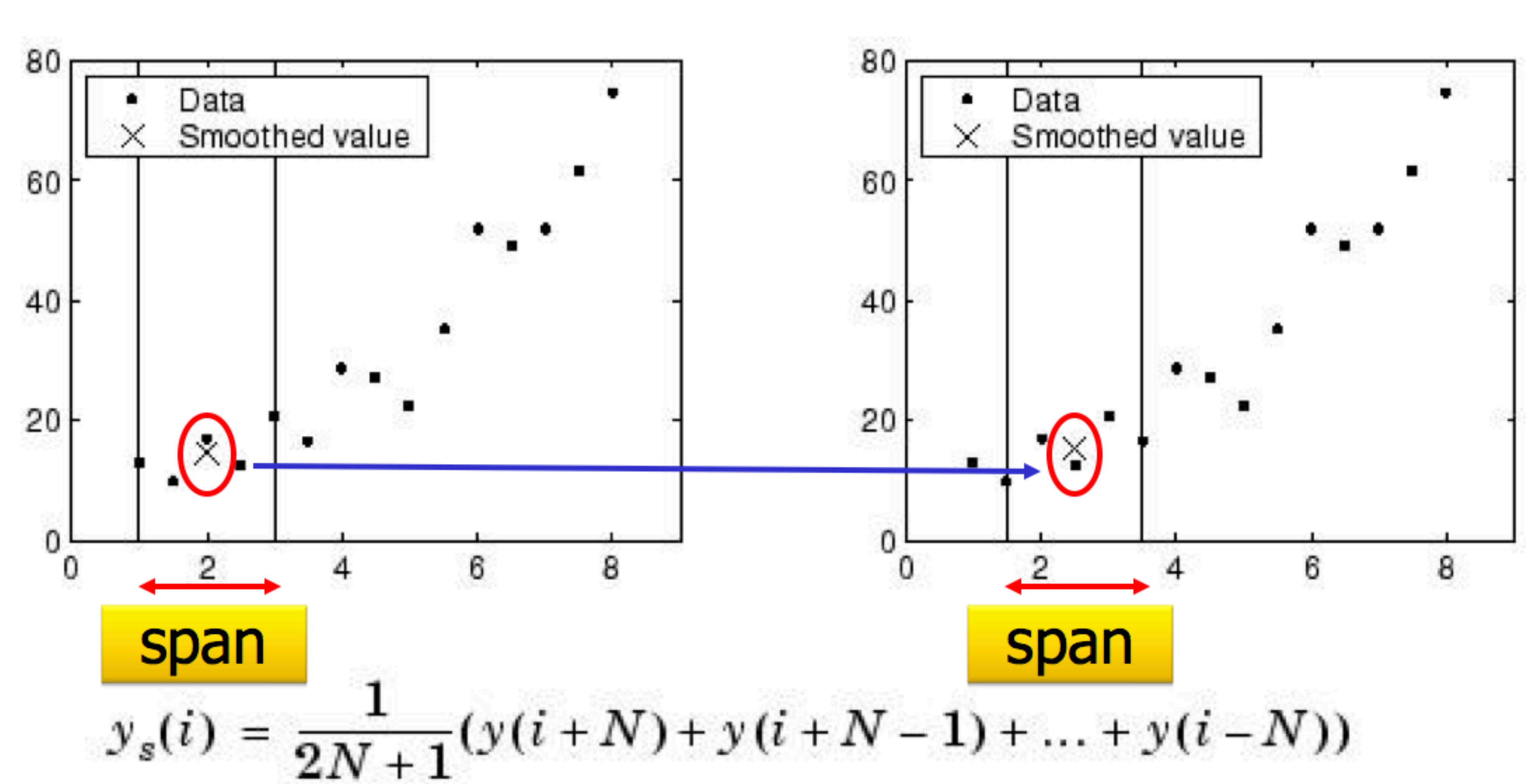


BOX PLOT, TIME SERIES PLOT





DATA SMOOTHING - MOVING AVERAGE



$$y_s(i) = \frac{1}{2N+1}(y(i+N)+y(i+N-1)+\ldots+y(i-N))$$

SMOOTH FUNCTION

- `x = linspace(0, 4 * pi, 1000);`
- `y = sin(x) + (rand(1,1000)-0.5)*0.2;`

Data:

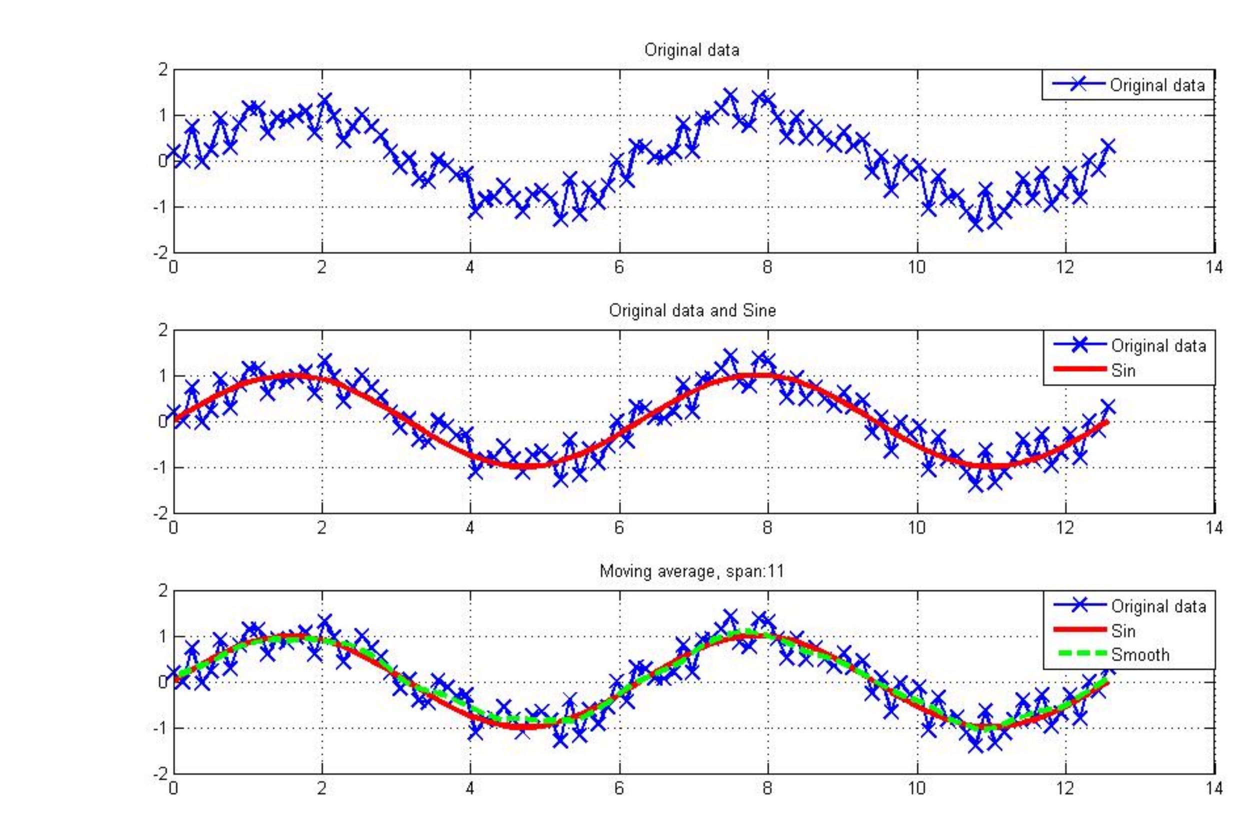
`y`

Generating Function:

`sin(x)`

Smoothed data:

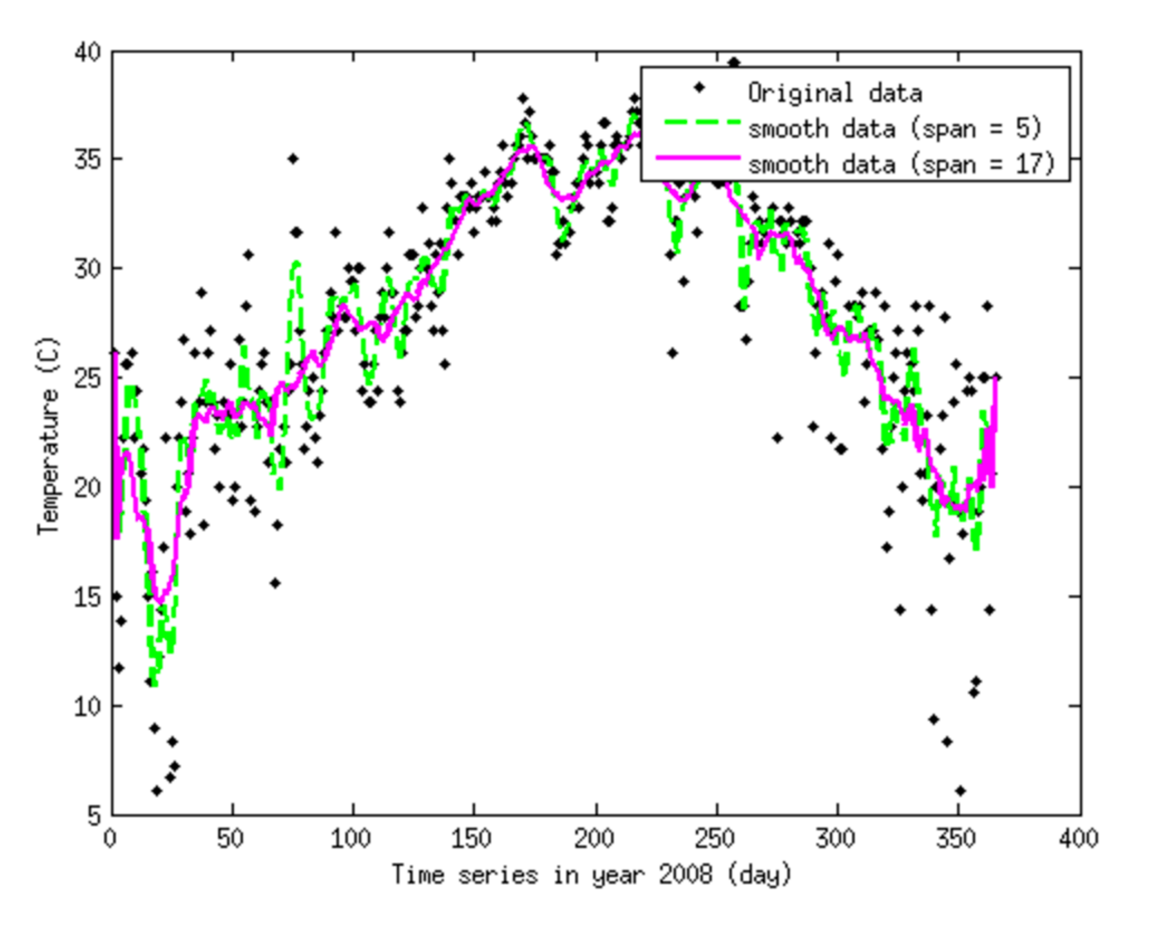
`smooth(y)`

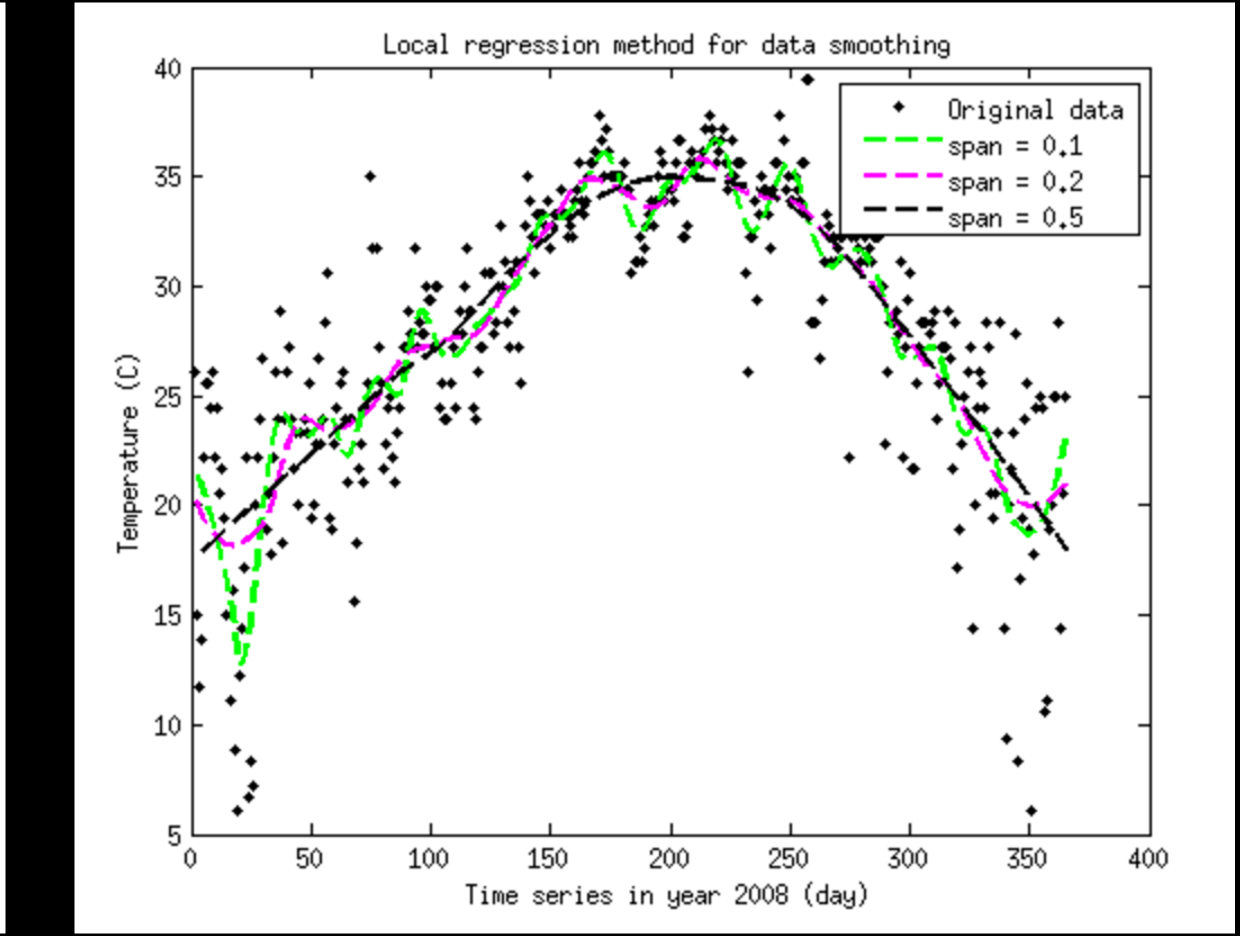

Original data

Original data and Sine

Moving average, span:11

SMOOTHING OVER OUR DATA SET

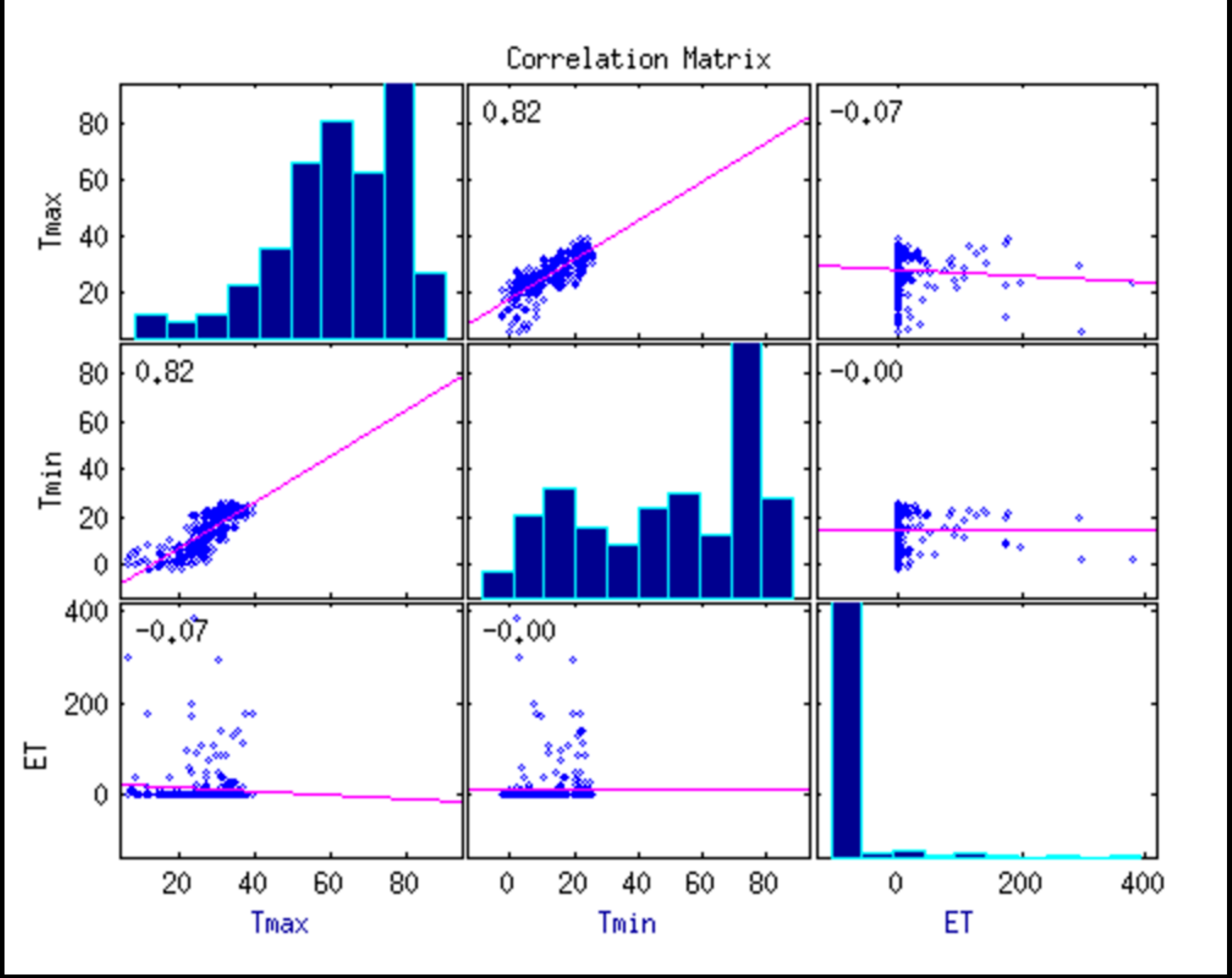




CORRELATION

- corrplot
- corrcoef

$$\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$


Correlation Matrix
Tmax
Tmin
ET
Tmax
Tmin
ET