Potsdam University, Department of Physics WS 2021/2022
Numerical methods for ClEWS A. Pikovsky, R. Cestnik

Time Series Analysis

Time series analysis is one of the universal tools in studies of natural and social processes. It is often based on the theory of random processes and it is aimed to extract properties of these processes from the observations. In a typical situation one observes a continuous variable $x \in \mathbb{R}$, e.g. temperature, at times $t_1, t_2, t_3, \ldots, t_N$, yielding a time series $x_1, x_2, \ldots, x_N$. Usually, observation times are equidistant (daily averages or noon temperature). There are also multivariate time series, where several variables, e.g. temperature, atmospheric pressure, precipitation, etc.) are measured.

In this project, several methods of time series analysis should be applied to a time series of daily averaged temperatures. These measurements from a weather station in Stockholm are in the file `TG_STAID000010.txt`. The top lines in this file describe the data set.

Literature:

1. H. Gilgen, Univariate time series analysis in geosciences. Spriner, 2006
2. Online manual: S. Prabhakaran, Time Series Analysis in Python – A Comprehensive Guide with Examples, https://www.machinelearningplus.com/time-series/time-series-analysis-python/
3. Chatfield, The analysis of time series: An introduction (CRC Press)
4. P. J. Brockwell and R. A. Davis, Introduction to Time Series and Forecasting, Springer 2016

Task 1: Load the data from the file and trim the ends such that it start on a January 1st and ends on December 31st. Draw the time series of daily averaged temperatures. Compare the time series over the first and the last ten years in a seperate diagram.

Task 2: Calculate the maximal and the minimal values of the temperature. At which days did these extreme events occur? Calculate the empirical mean value

$$\overline{x} = \frac{1}{N}\sum_{t=1}^{N} x_t$$

and the empirical variance

$$\mathrm{var}(x) = \frac{1}{N-1}\sum_{t=1}^{N}(x_t - \overline{x})^2$$

over the whole time series. Calculate and plot the mean and the variance of the temperature restricted to the months of a year, i.e. the average December temperature in Stockholm and its variability, in the month of the Nobel prize ceremonies.

The square root of the variance is the standard deviation. Variations or errors around a mean value can be indicated with errorbars twice the length of the standard deviation in plots.

Time series are often a superposition of responses from several influences and forces acting on different time scales. Some of these components are fluctuating from day to day or weekly, due to changes in the weather, some are periodic, e.g. the yearly seasonal temperature changes, and some changes occur slowly over decades due to changes in the climate.

Let us assume a representation of the time series as

$$x_t = a_t + p_t + n_t$$

corresponding to a slowly changing long time average a_t, a $T = 365.25d$ periodic component p_t and fast fluctuations n_t. The components p_t and n_t will cancel out in sliding window averages (moving average) over M years

$$a_t = \frac{1}{MT} \sum_{s=-MT/2}^{MT/2} x_{t+s}, \qquad M = \text{window size in years.}$$

Task 3: Calculate and draw the $M = 1, M = 10$ and $M = 20$ years moving averages a_t. For values a_t at the beginning and at the end of the time series the sliding window stretches into time intervals for which there is no data. Make a reasonable assumption and correction for these boundary effects.

The periodic component in the time series when a_t is removed

$$y_t = x_t - a_t = p_t + n_t$$

can be estimated either by another moving average over a shorter time interval

$$q_t = \frac{1}{m} \sum_{s=-m/2}^{m/2} y_{t+s},$$

e.g. $m = 14d$, $m = 31d$ or $m = 60d$, $q_t \approx p_t$ and $z_t = y_t - q_t \approx n_t$ or by Fourier analysis.

Let us assume the T periodic component p_t is exactly represented as

$$p_t = \sum_{k=1} \left[S_k \cdot \sin(2k\pi t/T) + C_k \cdot \cos(2k\pi t/T) \right].$$

With N an exact multiple of T we calculate

$$S_k = \frac{2}{N} \sum_{t=1}^{N} y_t \sin(2k\pi t/T)$$

$$C_k = \frac{2}{N} \sum_{t=1}^{N} y_t \cos(2k\pi t/T).$$

The two time series $z_t = y_t - q_t$ and $n_t = y_t - p_t$ represent fast fluctuations after removing the long time averages and the periodic component. These two time series should be highly correlated. The Pearson coefficient of correlation $-1 \leq \rho \leq 1$

$$\rho = \frac{\sum_t (n_t - \bar{n})(z_t - \bar{z})}{\sqrt{\text{var}(n)\text{var}(z)}}$$

tells you how well one time series can be linearly inferred from the other one.

Task 4: Calculate q_t $(m = 31)$ and $z_t = y_t - q_t$. Calculate S_1, C_1, S_2 and C_2 from y_t, from that p_t, and finally the fluctuations $n_t = y_t - p_t$. Plot p_t and q_t on top of y_t. Make a scatter plot of n_t vs. z_t and calculate the Pearson coefficient of correlation.

The last tasks are concerned with the statistics of the daily temperature fluctuations n_t. First, we want to approximate the short time temperature variability by a Gaussian distribution with mean μ and variance σ^2

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

and observe the difference to the Gaussian distribution.

Task 5: At which dates assumes n_t a maximum and a minimum, i.e. which where the seasonally most unusual temperatures and by how many degrees? Calculate the empirical mean μ and the variance of σ^2 of the fluctuations n_t. Plot a normalized histogram of the temperature fluctuations n_t and a Gaussian distribution of the same mean and variance on top. Draw the plot again in semi-log-y scale to observe the difference in the probabilities to the Gaussian prediction.

Bonus tasks

Bonus Task A : The long time averages a_t appear to be stationary for around $50,000 - 60,000$ days before they start to grow on average. Fit a linear function $a_t = At + B + u_t$ to the first half and to the second half of the $M = 20$ years time averaged time series a_t (by linear regression, i.e. minimizing the squared errors $\sum_t u_t^2$) and draw the estimated *trends* $At + B$ on top of the time series a_t.

The Pearson coefficient of correlation for n_t and the shifted time series $n_{t+\tau}$ is the autocorrelation function

$$c_\tau = \frac{\frac{1}{N-\tau} \sum_{t=1}^{N-\tau} (n_t - \bar{n})(n_{t+\tau} - \bar{n})}{\text{var}(n)}$$

The autocorrelation function tells you how fast correlations in the fluctuations decay, i.e. the time scale for the temperature predictability.

Bonus task B: Calculate and plot the autocorrelation function c_τ for the temperature fluctuations n_t (show only the first 100 days). Fit it with an exponential function to the first seven days and to days $7 \dots 40$. An exponential function $f(t) = Ae^{-\gamma\tau}$ appears linear in a semi-logarithmic scales $\log(f) = \log(A) - \gamma\tau$. You can estimate A and γ by linear regression to $\log |c_\tau|$.