# Time Series Analysis

Time series analysis is one of the universal tools in the 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 starts 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 top 5 largest and lowest values of the temperature. At which days did these extreme events occur? Calculate mean, variance and standard deviation over the whole time series. Calculate and plot the mean and the standard deviation of the temperature conditioned on the month of the year (temperature climograph) using either errorbars or boxplots. Be sure to explain the plot in the captions.

Time series are often a superposition of responses to several forces and processes 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 and note these assuptions in the report.

The one-year periodic component $p_t$ in the time series when $a_t$ is removed

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

can be estimated by Fourier analysis. Let us assume the $T$ periodic component $p_t$ is exactly represented as

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

We can calculate

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

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

as averages over sufficiently many years. Be sure to use the exact period $T = 365.25d$ to avoid going out of phase over such a long time series and to average over an integer multiple of $T$.

**Task 4:** Calculate $S_k$ and $C_k$ for $k = 1 \ldots 3$ from $y_t$, from that $p_t$ with Eq.1, and finally the fluctuations $n_t = y_t - p_t$. Plot $p_t$ on top of $y_t$ and plot $n_t$ separately.

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 $\mu$ and variance $\sigma^2$

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

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 temperature deviations? Report absolute temperature $x_t$ and deviation $n_t$ from seasonal norm on these dates. Calculate the empirical mean $\mu$ and the variance $\sigma^2$ of the fluctuations $n_t$. Plot a normalized histogram of the temperature fluctuations $n_t$ and a Gaussian distribution (Eq.2) of the same mean and variance on top. Draw the plot again in semilogy scale to observe the difference in the probabilities to the Gaussian prediction in the bulk and in the tails.

## Bonus tasks

**Bonus Task A :** The long time averages $a_t$ appear to be stationary for around 50000 - 60000 days before they start to grow on average. Fit a linear function $a_t \approx At + B$ to the first half and to the second half of the $M = 20$ years time averaged time series $a_t$ (by linear regression) 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

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

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 $c_\tau$ for the temperature fluctuations $n_t$ (only for $\tau = 0 \ldots 100$). Fit it with an exponential function to the first seven days and to days $7 \ldots 40$. An exponential function $f(\tau) = Ae^{-\gamma\tau}$ appears linear in a semi-logarithmic scales $\log(f) = \log(A) - \gamma\tau$. You can estimate $A$ and $\gamma$ by linear regression to $\log|c_\tau|$. $c_{\tau=0}$ is necessarily equal to one.