TP1

Benjamin Pillot

December 10, 2020

Quality control procedure using the sigma test (Guttman et al., 1988; Shulski et al., 2014):

$$u_{min} - f \cdot s_{min} < x < u_{max} + f \cdot s_{max} \quad [0.1]$$

where :

x	value from time series
u_{min}	mean of the daily minimum for the given month
u_{max}	mean of the daily maximum for the given month
s_{min}	standard deviation of the daily minimum for the given month
s_{max}	standard deviation of the daily maximum for the given month
f	number of standard deviations by which data sample can differ from mean (3 corresponds to 99.73% confidence)

We want to flag the outliers in our time series based on the sigma test. Eventually, we shall plot nominal values in blue and the outliers in red.

1. Import the file ./FormationPython#examples/timeseries.csv using pandas.
2. Compute daily minimum and maximum values
3. Compute mean of daily minimum and maximum for each month
4. Compute standard deviation of daily maximum and minimum for each month
5. Apply equation [0.1] for each value in the time series. This is the tricky part. You have to use logical comparison between series of different size. There are multiple ways to do this. One very slow is to loop over the data in the original time series, get the month, and compare it with the right monthly values. A faster way is to loop over the values in the monthly series you have just computed and compare every value against the corresponding monthly sub-series in the original time series (you can make logical comparison between series and scalar). The fastest way (one-liner) is to rely on pandas groupby and apply built-in functions.
6. You should have a new series of Boolean flags (made of *True* and *False*). You can use it to extract two new series: one containing the nominal values and another one with the outliers.
7. Plot nominal values and outliers on the same graph.