SOON: the Station Observation Outlier fiNder

S. Dal Gesso[1*], Elisa Arnone[1], Marco Venturini[1], Marco Cucchi[1], Marcello Petitta[1,2,3]
[1]Amigo s.r.l., Rome, ITALY
[2]ENEA, SSPT-MET-CLIM, Rome, Italy
[3]EURAC, Institute for applied remote sensing, Bolzano, Italy
* e-mail: sara.dalgesso@amigoclimate.com

Introduction

In the climate change era, it is fundamental to monitor the availability of water resources. One of the possible causes for a change in the water availability is related to variations in the meteorological conditions. To track this change, ground-based observations are one of the commonly used measurements (e.g. Klein Tank AM and Können GP, 2003). However, these datasets might include both extreme but realistic values and erroneous information. A necessary but not trivial preliminary process for exploiting the observations is to filter the former while retailing the latter (e.g. Jiménez 2010).

The Station Observation Outlier fiNder (SOON) is a highly innovative algorithm, that identifies errors in large dataflows. SOON can be used on historical datasets as well as in real-time dataflows. A first prototype has been tested on 8 years (2007-2014) of hourly data recorded by about 10000 stations around Europe, which includes 7 meteorological variables: temperature, dewpoint temperature, pressure, precipitation, wind speed, wind gusts, and cloudiness. The dataset belongs to the Ubimet archive and has been provided within the EDI incubator programme.

Materials and methods

The idea behind SOON is that data flows can be simultaneously screened from different perspectives. This approach maximizes the exploitation of the available information. At the same time, the process becomes resilient to the absence of information. More practically, SOON detects the errors through machine learning (ML) by checking the internal consistency of the dataflow from a temporal, spatial and parametric perspective. To implement the ML modelling, the dataset is split into the training and the test dataset, corresponding to the periods 2007-2013 and 2014, respectively.

SOON is constituted by three modules:

1. **Data Pre-processing**: preparation of dataset for the ML modelling, i.e. (i) filtering of unphysical measurements based on realistic ranges; (ii) assigning to each station the corresponding pre-defined cluster, calculated through a K-means technique with K=20 (MacQueen, 1967) on the basis of the geographic location, elevation, and monthly values of average temperature and cumulated precipitation; (iii) calculating the standardized variables based on the average and standard deviation computed over the training set.

2. **ML modelling:** application of Random Forest (RF; Breiman, 2000) regressor to predict hourly values at each station. Three types of RF models are built to leverage the three information domains, i.e. time, space and parameter consistency. Models are built cluster by cluster and differs mainly in the inputs. The time consistency is evaluated on the basis of the last 24 hours of observations, the space consistency through the measurements of the five nearest neighbor stations, and the parametric consistency based on the values of previously identified correlated variables.

3. **Errors Detection:** error identification on the basis of a *score.* For each information domain, the score is computed as the difference between the prediction and the observation divided by three times the root mean squared error (RMSE) calculated over the training set. The final score is the weighted average of the three previously defined scores. Finally, values overcoming 1 are detected as errors. The weights are estimated by exploring a synthetic dataset, which includes artificial random errors.

More specifically, the weight is defined as the product of Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve and the sensitivity at the threshold, i.e. 1.

Results and concluding remarks

SOON identifies as errors between 0.5 and 1% of the data of 2014. An example timeseries is displayed in Figure 1. The variation of temperature (T) observed at one station over 2014 is depicted as a grey line. The data that are identified as outliers by the initial filtering of the pre-processing module, and by the temporal, spatial and parametric components are highlighted with different colors. Finally, the observations that are classified as errors through the final score are shown as red circles. It is noticeable that the most evident errors are identified correctly. It is also worth stressing that not all the observations that are classified as outliers by one of the ML components present a final score that is beyond the threshold. In this sense, SOON reduces the effects of likely limitations in the reliability of the ML modelling components, and of using possible erroneous values as inputs of the RF regressors. In other cases, the classification of the different ML components are not all available. In this respect, SOON maximizes the available information, and enables the error identification even in fragmented datasets.

As a future development, SOON will be further extended by including a module that distinguishes between errors and extreme but realistic observations. To this end, the extreme value theory will be employed. With this final module, SOON will become a unique trans-sectoral tool for identifying errors in any type of instrumental network, including weather stations and smart water sensors.

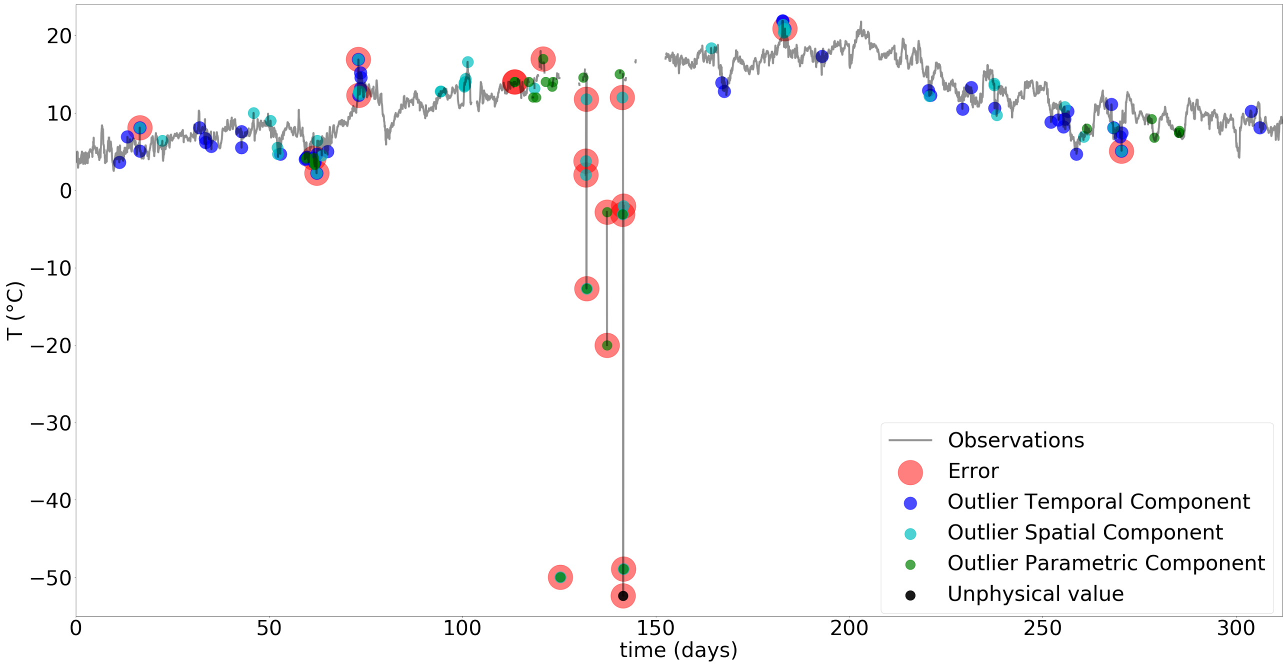

Figure 1. Example timeseries of Temperature for one station. The red dots indicate the errors identified by SOON.

References

Klein Tank AM and Können GP (2003). Trends in Indices of Daily Temperature and Precipitation Extremes in Europe, 1946–99. Journal of Climate, 16, 3665–3680, https://doi.org/10.1175/1520-0442(2003)016<3665:TIIODT>2.0.CO;2

Jiménez PA, González-Rouco JF, Navarro J, Montávez JP, and García-Bustamante E (2010). Quality Assurance of Surface Wind Observations from Automated Weather Stations. Journal of Atmospheric and Oceanic Technology, 27, 1101–1122, https://doi.org/10.1175/2010JTECHA1404.1

MacQueen, JB (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967. https://projecteuclid.org/euclid.bsmsp/1200512992

Breiman, L *(2000).* Some infinity theory for predictor ensembles*. Technical Report 579, Statistics Dept. UCB*