End-to-end Anomaly Detection, Correction and Prediction of Missing Values in Historical Daily Temperature Timeseries*

Alioscha-Perez M.[1], Oveneke M.C.[1], Diaz A.[1], Bertrand C.[2], and Sahli H.[1,3]

Vrije Universiteit Brussel, ETRO Department (ETRO-VUB), Brussels, Belgium
{maperezg,mcovenek,aberengu,hsahli}@etrovub.be
Royal Meteorological Institute of Belgium (RMI), Brussels, Belgium
Interuniversity Microelectronics Center (IMEC), Heverlee, Belgium

Abstract. In this paper, we present a deep learning solution to *detect and correct* anomalous values present in historical temperature timeseries, that are likely associated to human and weather instruments errors. Our solution consists in a joint peaks detection and end-to-end sequence prediction involving synchronous measurements of individual meteorological stations along with their neighboring peers. We designed our models in a way that the false positive rate (FPR) of the anomaly detection is minimized and the accuracy maximized, so that the historical records are corrected as less as possible. The method was applied to temperature records of 24 meteorological stations in Belgium, and allowed to automatically correct more than 80% of all errors in both max/min daily temperature records by modifying less than 15% of all the timeseries values, with an overall detection accuracy of 90%. The corrected temperature timeseries yielded a perfect match with respect to errors-free signals in several climate indicators. Our method can be potentially applied to other historical timeseries such as precipitation.

1 Introduction

Studies of extreme climate rely extensively on historical records documenting weather conditions from the past, which are usually available in the form of weather ledgers or logbooks with a registry of the state of the weather at different locations and time. However, it is known that human errors and weather instruments malfunctioning are a major source of errors already present in many weather registries. To address this problem, we present a deep learning solution for anomaly detection, automated correction and prediction of missing values in historical temperature records. Our solution outputs cleaned timeseries where most of anomalies are replaced by predicted values. In addition, it optionally allows to perform a manual revision (i.e. human in the loop) of some or all the corrections, if wanted or necessary.

* Supported by the Belgian Science Policy Office (Belspo) thru the Bel-Hornet project.

2 Our Approach

Our approach consists in a joint peaks detection and end-to-end sequence prediction involving synchronous measurements of individual meteorological stations along with their neighboring peers. Following a Neural Architecture Search (NAS) approach, our solution minimizes the false positive rate (FPR) of the anomaly detection and maximizes the overall accuracy, so that the historical records are corrected as less as possible. This preserves the historical timeseries as original as possible, while still allowing to correct most of erroneous values.

Results: For our study we included more than $170,000$ daily extreme temperature values involving 24 meteorological stations in Belgium, comprising two time-periods used for training and testing, respectively. We assessed the performance of our solution in many indices [3] widely adopted for climate studies [2]. Our results are detailed in Table (1), and confirm a perfect match between the errors-free signal and the signal obtained with our solution.

Table 1. Results in climate indicators frequently used for climate analysis for all the 24 meteorological stations under study; given values are per-year and per-station.

Indicator	Raw timeseries	Errors-free timeseries	Proposed solution
Hot days	9	5	5
Summer days	31	27	27
Tropical nights	79	76	76
Frost days	47	48	48
Icy days	7	7	7

Conclusions: We presented a solution for the automatic detection and correction of anomalies, likely associated to human and instrument errors, that are present in daily historical temperature records. In addition, our solution allows to predict some of the missing values in the timeseries under study. In future works, we will explore ways to optimize our NAS model via SVRG [1].

Bibliography

[1] Alioscha-Perez, M., Oveneke, M.C., Dongmei, J., Sahli, H.: Multiple kernel learning via multi-epochs svrg. In: 9th NIPS Workshop on Optimization for Machine Learning. vol. 12 (2016)

[2] Frich, P., Alexander, L.V., Della-Marta, P., Gleason, B., Haylock, M., Tank, A.K., Peterson, T.: Observed coherent changes in climatic extremes during the second half of the twentieth century. Climate research **19**(3), 193–212 (2002)

[3] Zhang, X., Alexander, L., Hegerl, G.C., Jones, P., Tank, A.K., Peterson, T.C., Trewin, B., Zwiers, F.W.: Indices for monitoring changes in extremes based on daily temperature and precipitation data. Wiley Interdisciplinary Reviews: Climate Change **2**(6), 851–870 (2011)