Outlier Detection for ARM Data

Yuping Lu[1], Jitendra Kumar[2] and Michael A. Langston[1]

***Abstract*— Outliers are common in ARM data. These outliers could be either an instrument failure or extreme weather event. Multiple methods are available to detect these outliers from the huge ARM datasets. We combined Pearson Correlation Coefficient, Singular Spectrum Analysis and K-means methods together as a whole framework to track down these outliers. Compared to the current outliers recorded in the DQR database, our results showed this framework is promising.**

I. Introduction

We will use this section to introduce the background of outlier detection for time series data. [1]

The Atmospheric Radiation Measurement (ARM) user facility was founded by the U.S. Department of Energy (DOE) in 1989 [2]. Since then, its aim is to be the platforms for the observation and study of Earth's climate. Huge ARM datasests are generated and stored in ARM data center daily. And outliers are pretty common in these datasets. Currently, these datasets are checked manually and outliers are stored in Data Quality Report (DQR) database to be fixed.

II. Datasets

ARM data center gathers data from multiple data sources. It ranges from *Atmospheric Profiling* to *Satellite Observations*. All these data are measured at different locations using different instruments. Each instrument may only work on a specified time range. For the raw netcdf dataset collected from each instrument, it contains multiple variables. In this paper, we only tested Surface Meteorology Systems (MET) data collected from the Southern Great Plains (SGP). There were total 24 instruments in SGP area and we chose 5 typical variables which are *temp_mean*, *vapor_pressure_mean*, *atmos_pressure*, *rh_mean* and *wspd_arith_mean* from multiple variables. Table 1 contains the detail of these datasets.

TABLE I

SGPMET datasets tested

Instrument	E1	E3	E4	E5	E6	E7
Begin Year	1996	1997	1996	1997	1997	1996
End Year	2008	2008	2010	2008	2010	2011
Instrument	E8	E9	E11	E13	E15	E20
Begin Year	1994	1994	1996	1994	1994	1994
End Year	2008	2017	2017	2017	2017	2010
Instrument	E21	E24	E25	E27	E31	E32
Begin Year	2000	1996	1997	2004	2012	2012
End Year	2017	2008	2001	2009	2017	2017
Instrument	E33	E34	E35	E36	E37	E38
Begin Year	2012	2012	2012	2012	2012	2012
End Year	2017	2017	2017	2017	2017	2017

[1]University of Tennessee, Knoxville, TN, USA
[2]Oak Ridge National Laboratory, Oak Ridge, TN, USA

III. Methodology

Mention methods we used in this paper and how do we preprocess the data.

A. Pearson Correlation Coefficient

PCC goes here [3].

B. Singular Spectrum Analysis

SSA goes here [4], [5].

C. K-means

k-means goes here [6].

IV. Results and Discussion

Results and pics go here. Comparison metric: DQR database. Add precision and recall result here [7].

V. Conclusions

We presented a combined model to detect outliers for ARM data. Future work: ML and tried methods working on multiple instruments multiple sites [8].

Acknowledgment

This research was supported by the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Office of Biological and Environmental Research.

References

[1] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.
[2] "Arm research facility," https://www.arm.gov/, accessed: 2018-06-22.
[3] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.
[4] N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for time series*. Springer Science & Business Media, 2013.
[5] T. Alexandrov, "A method of trend extraction using singular spectrum analysis," *arXiv preprint arXiv:0804.3367*, 2008.
[6] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.
[7] J. W. Perry, A. Kent, and M. M. Berry, "Machine literature searching x. machine language; factors underlying its design and development," *Journal of the Association for Information Science and Technology*, vol. 6, no. 4, pp. 242–254, 1955.
[8] J. D. Phillips, W. Schwanghart, and T. Heckmann, "Graph theory in the geosciences," *Earth-Science Reviews*, vol. 143, pp. 147–160, 2015.

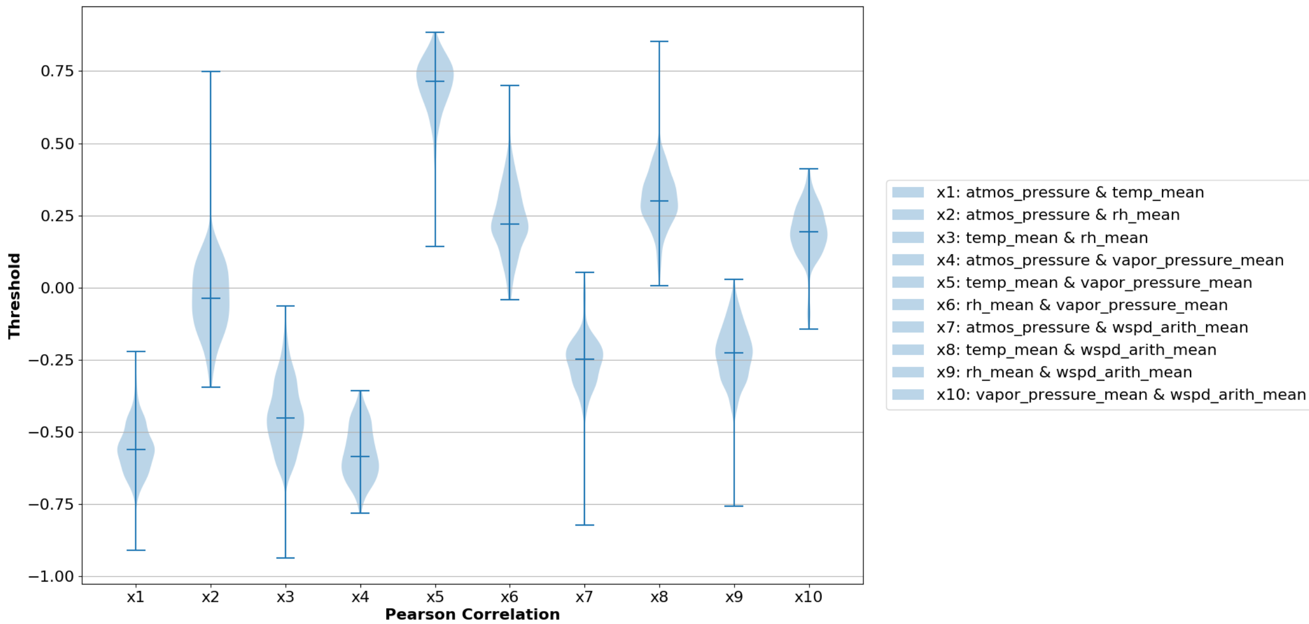

Fig. 1. Violin plot: Spring 5 variables from SGPMET

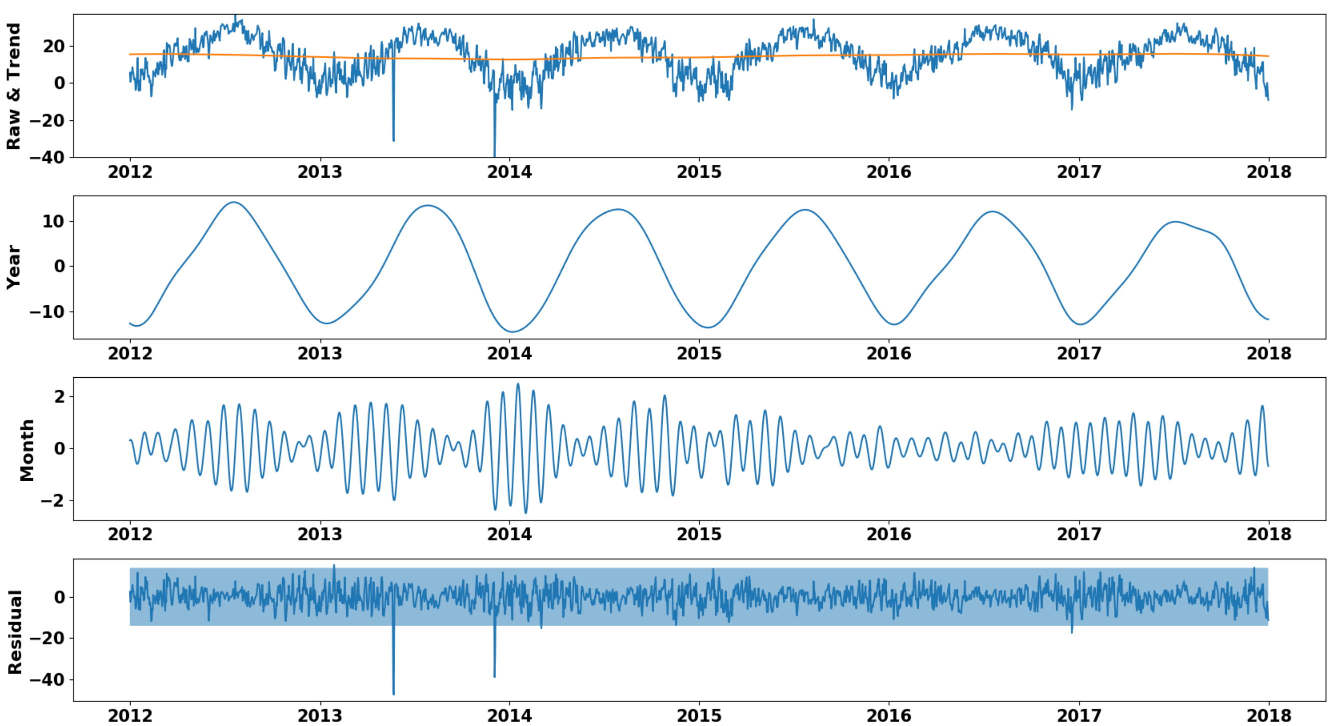

Fig. 2. Example of SSA application on ARM data. E33 temp_mean data full decomposition.

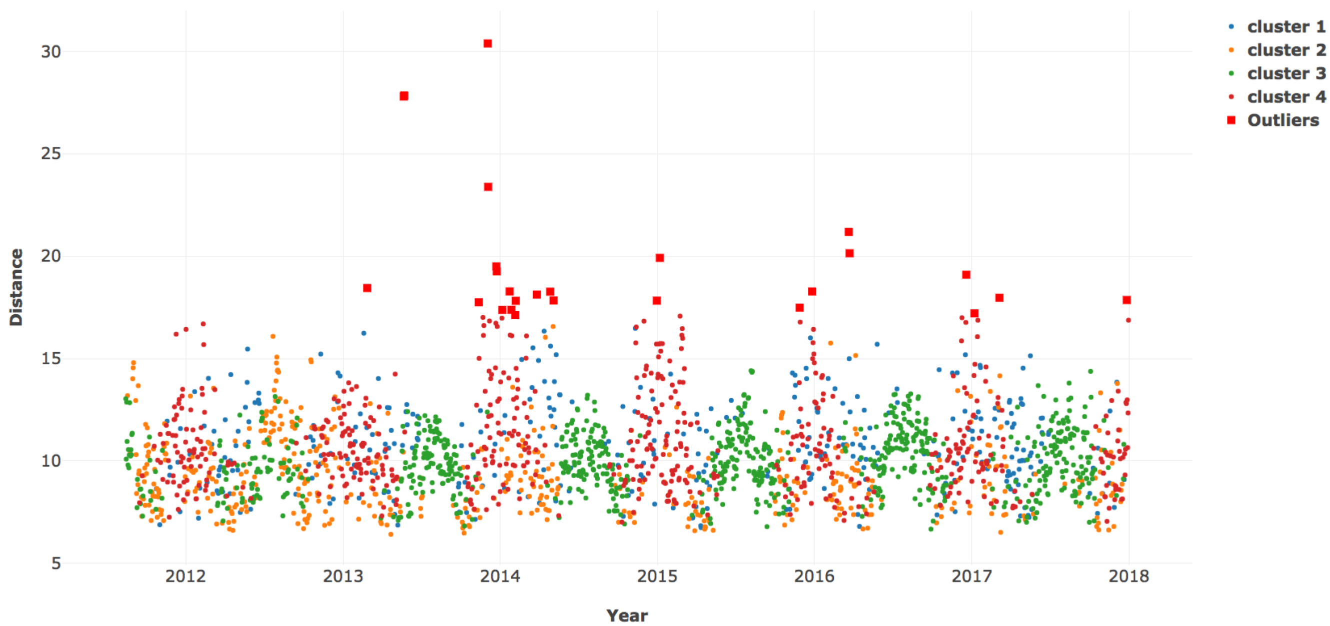

Fig. 3. E33 K-means

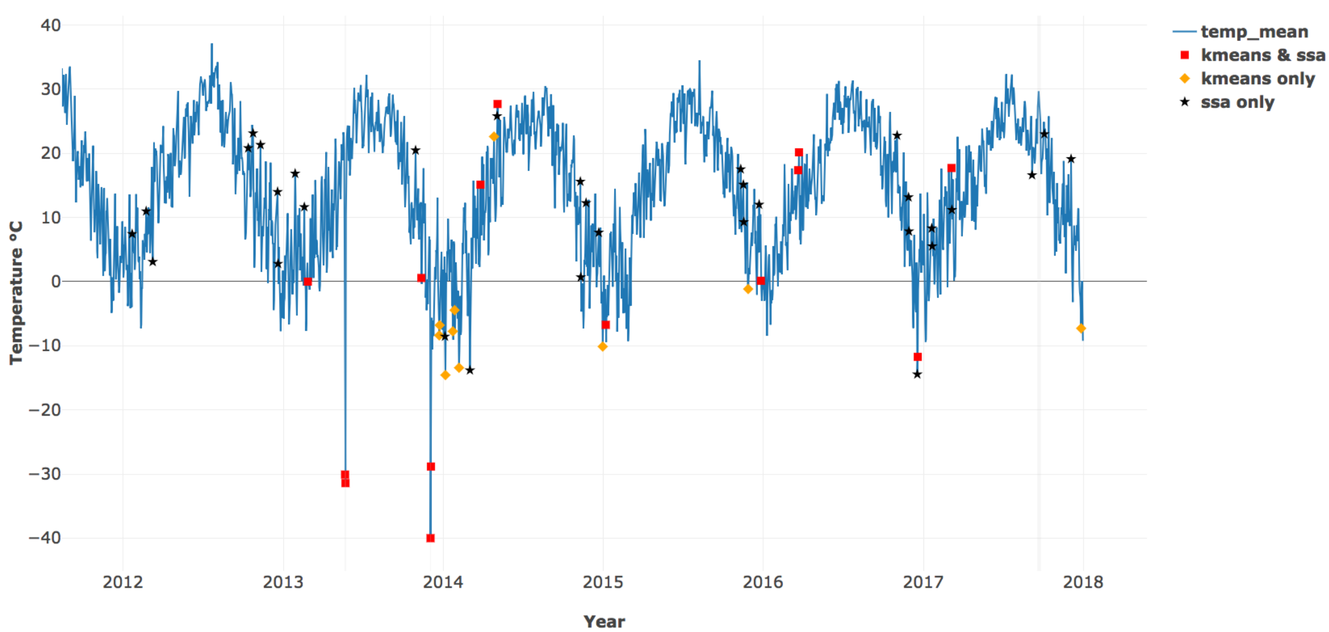

Fig. 4. E33 temp_mean combined