Outlier Detection for ARM Data

Yuping Lu[1], Jitendra Kumar[2], Nathan Collier[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 datasets 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

Pearson Correlation Coefficient was first introduced by Karl Pearson [3]. It is used to measure the linear correlation between two variables. Pearson correlation coefficient is calculated from the covariance of two variables divided by the multiplication of the standard deviation of those two variables. Thus the value falls in [-1, 1]. If the value is close to -1, it means those two variables are highly negatively related. On the other hand, then the two variables are strongly positively related. If the value is near 0, it means those two variables don't have linear relation.

B. Singular Spectrum Analysis

Singular Spectrum Analysis (SSA) is a popular method for time series data analysis [4], [5]. The general idea is to use a subset of the decomposition of trajectory matrix to approximate it. Many applications can be found in [4]. For example, SSA can be applied to monitor volcanic activity [6]. It can also be used to extract trend [7]. Different from the classic SSA method, we defined our own version of SSA to best work on ARM data. Figure 2 is a demonstration of the workflow of SSA. Below is the formal description of the algorithm.

Assume we have an ARM time series data Y of length T.

$$Y = (y_1, \ \ldots, \ y_T)$$

Here $T > 2$ and y_i is not empty. Let L $(1 < L \leq T/2)$ be the window size and $K = T - L + 1$. In general, the algorithm contains two main parts: decomposition and reconstruction.

1) The first step is to form trajectory matrix $\mathbf{X}$ from vector Y by embedding subsets of Y. These subsets of Y X_i are lagged vectors of length L.

$$X_i = (y_i, \ \ldots, \ y_{L+i-1})^T \quad (1 \leq i \leq K)$$
$$\mathbf{X} = [X_i, \ \ldots, \ X_K]$$

Thus the trajectory matrix is

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ y_3 & y_4 & y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_T \end{pmatrix} \quad (1)$$

where $x_{ij} = y_{i+j-1}$. We can see from equation 1 that matrix $\mathbf{X}$ has equal elements on anti-diagonals and therefore it is Hankel matrix.

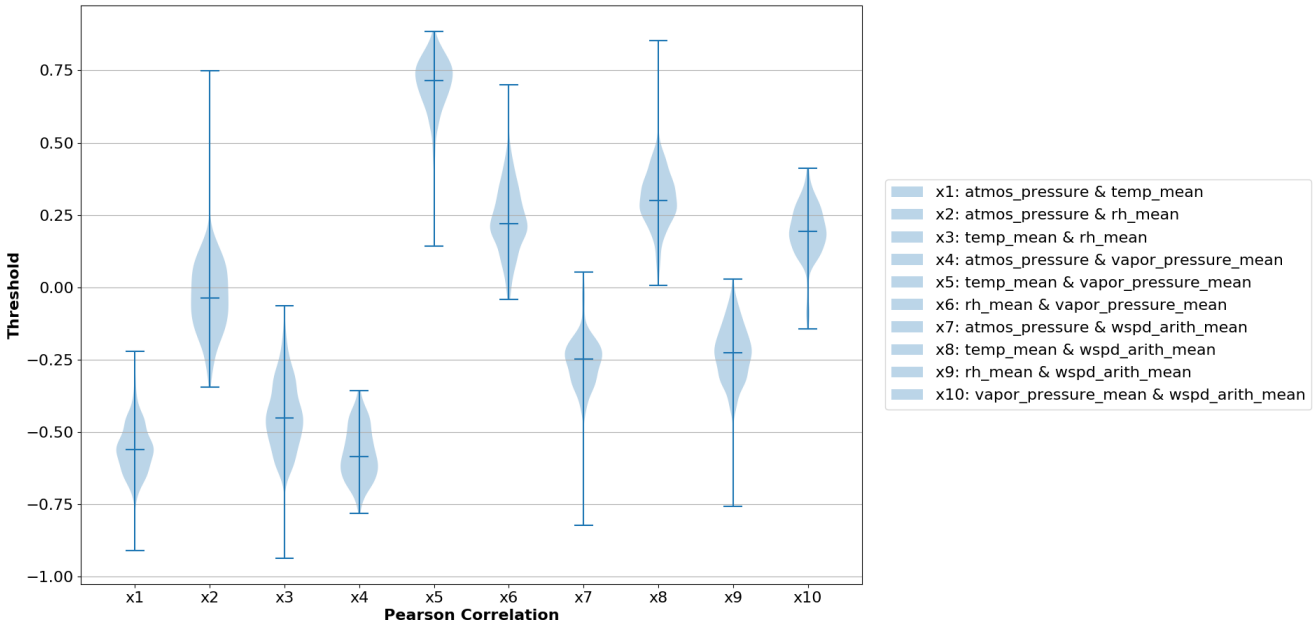

Fig. 1. Violin plot: Spring 5 variables from SGPMET

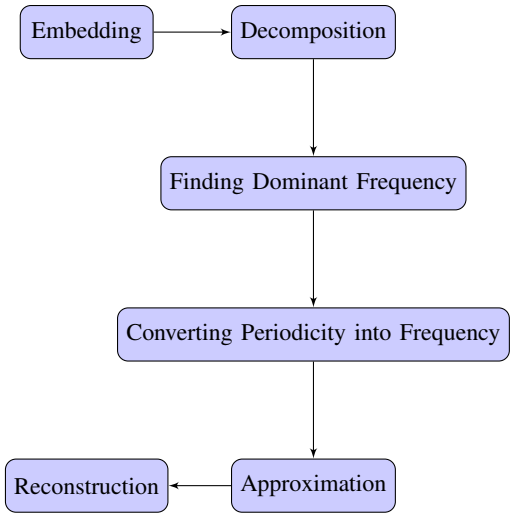

Fig. 2. Flowchart of SSA

2) Assume matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^T$, thus we perform the singular value decomposition (SVD) on $\mathbf{S}$. The eigenvalues of S are denoted by $\lambda_1, \ldots, \lambda_L$ in the decreasing order of magnitude $(\lambda_1 \geq \ldots \geq \lambda_L \geq 0)$ and corresponding eigenvectors are denoted by $P_1, \ldots, P_L$. Let $d = rank\ \mathbf{X}$ and $V_i = \mathbf{X}^T P_i / \sqrt{\lambda_i} (i = 1, \ldots, d)$. Thus, the trajectory matrix X can also be written as

$$\mathbf{X} = \mathbf{X_1} + \ldots + \mathbf{X_d} \tag{2}$$

where $\mathbf{X_i} = \sqrt{\lambda_i} P_i V_i^T$.

3) In this step, we use Fast Fourier transform (FFT) to find the dominant frequency of each eigenvector [8] . Algorithm 1 shows the process.

Algorithm 1: Dominant Frequency Finder

Input : λ of $\mathbf{S}$ and corresponding eigenvectors $\mathbf{P}$
Output: Dominant frequency of each eigenvector

```
fftfreq ← Discrete Fourier Transform sample frequencies
fft ← Discrete Fourier Transform
len ← size of λ
frequencies ← zero vector of size len
fs ← fftfreq(λ)
ix ← indices that sort fs
fs ← fs[ix]
for i in range(len) do
    p1 ← abs(fft(P[:,i]))
    ps ← p1**2
    ps ← ps[ix]
    frequencies[i] ← fs[index of the maximum value in ps]
end
return abs(frequencies)
```

4) Convert periodicity into frequency. Let P be the vector of user specified periodicity.

5) Build an approximation of X by taking a subset of the decomposition. This approximation is formed by taking eigenvectors whose dominant frequency is close to the tar-

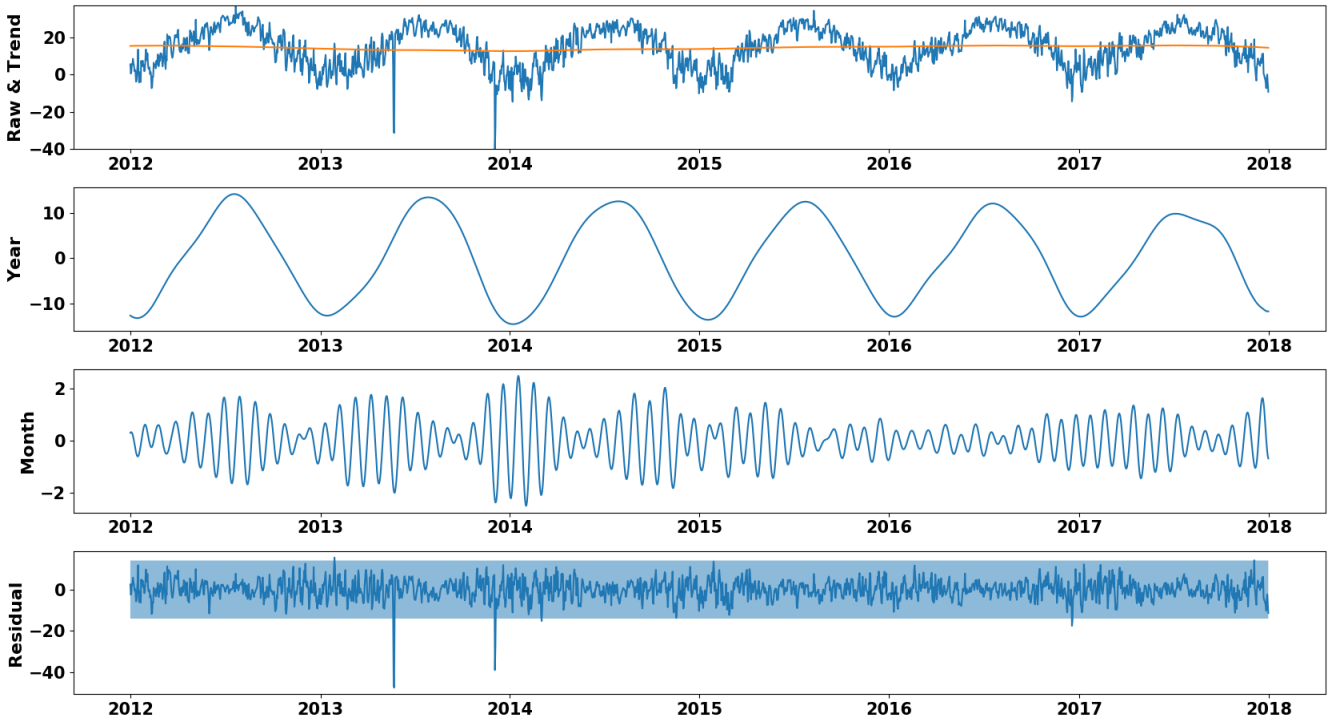

Fig. 3. Example of SSA application on ARM data. E33 temp_mean data full decomposition.

geted values.

6) Now we reconstruct the signal by taking a mean of all the approximations.

Mention the trend is flat. What parameters do we pick for SSA.

C. K-means

k-means is a partitioning clustering algorithm [9], [10]. It starts with the k centroids user specified, and assigns the points to the nearest centroid. Then it computes the new k centroids and assign other points to these centroids again. The process repeats until it converges.

IV. Results and Discussion

SSA is an univariate method. K-means is a multivariate method. Results and pics go here. Comparison metric: DQR database.

How do you pick extreme values as outliers? Some methods do not work. Here we use the three sigma rule to extract outliers [11].

Precision and recall was first defined in [12]. It is commonly used to measure the quality of classification tasks [13]. Precision is calculated from True Positives divided by the sum of True Positives and False Positives. On the other hand, recall is measured from True Positives divided by the sum of True Positives and False Negatives. In this paper, detected outliers in the DQR database are the ground truth. So we treated these as True Positives. Thus detected outliers not in the DQR database are False Positives. Undetected values which in the DQR database are False Negatives, and which not in the DQR database are True Negatives. Analysis of table 2 and 3 goes here.

Algorithm 2: K-means Outlier Detection

Input : ARM time series data
Output: Outliers

```
fftfreq ← Discrete Fourier Transform sample frequencies
fft ← Discrete Fourier Transform
len ← size of λ
frequencies ← zero vector of size len
fs ← fftfreq(λ)
ix ← indices that sort fs
fs ← fs[ix]
for i in range(len) do
    p1 ← abs(fft(P[:,i]))
    ps ← p1**2
    ps ← ps[ix]
    frequencies[i] ← fs[index of the maximum value in ps]
end
return abs(frequencies)
```

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 [14].

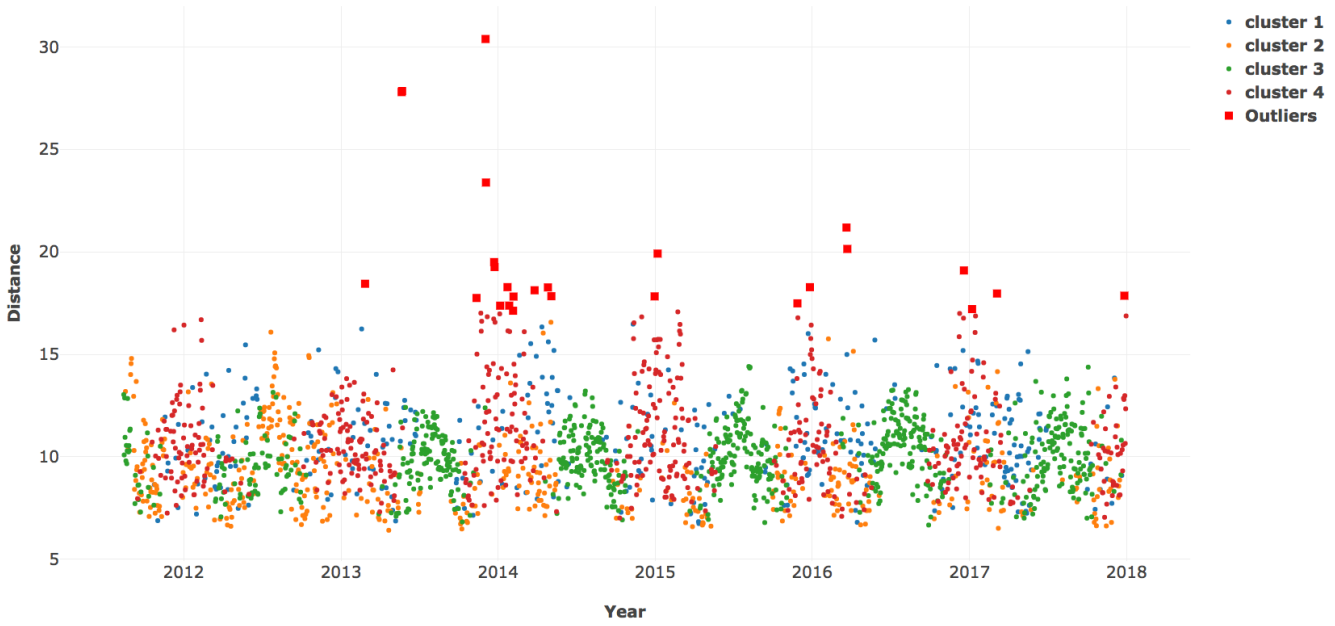

Fig. 4. E33 K-means

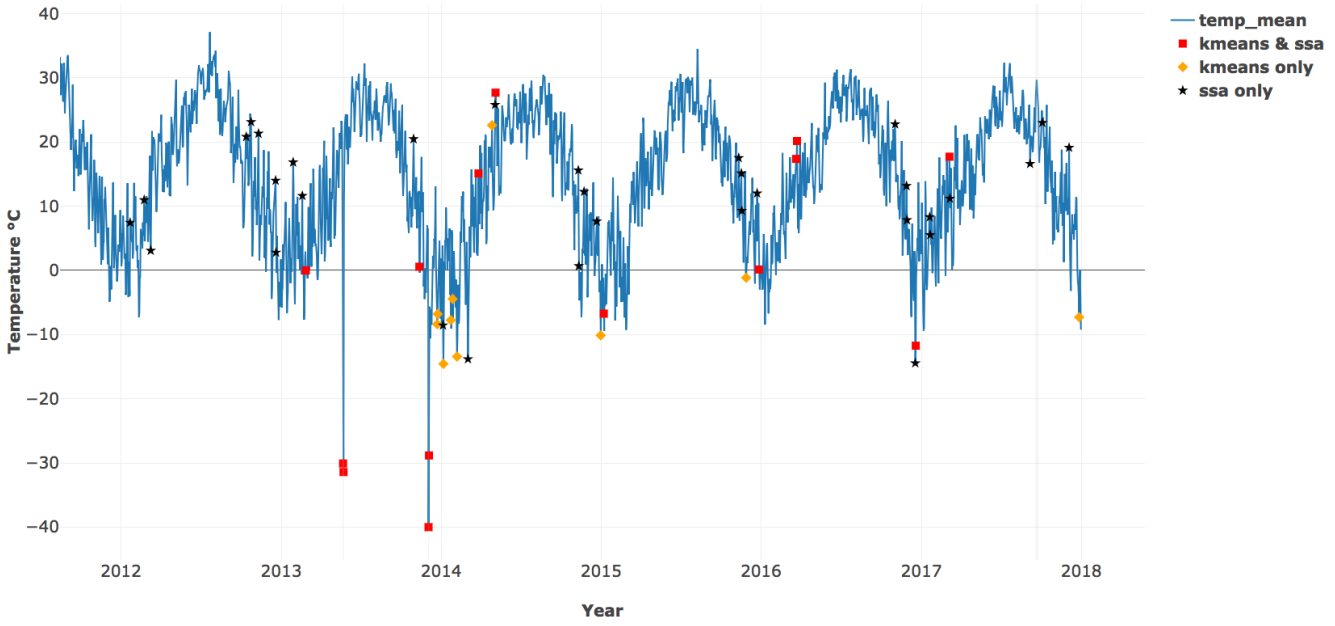

Fig. 5. E33 temp_mean combined

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.

TABLE II

PRECISION AND RECALL OF SSA AND K-MEANS

Method	Variable	Precision	Recall
SSA	temp_mean	16.00%	1.20%
SSA	vapor_pressure_mean	20.70%	1.40%
SSA	atmos_pressure	0.00%	0.00%
SSA	rh_mean	14.80%	0.50%
SSA	wspd_arith_mean	0.60%	1.50%
Kmeans	5 together	12.90%	1.90%
Combined	5 together	11.10%	4.10%

TABLE III

COMPARISON OF SSA AND K-MEANS OUTLIER SET SIZE

	Outlier Set Size
SSA	922
K-means	508
Intersection	378
Symmetric Difference	674

[4] N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for time series*. Springer Science & Business Media, 2013.

[5] N. Golyandina and A. Korobeynikov, "Basic singular spectrum analysis and forecasting with r," *Computational Statistics & Data Analysis*, vol. 71, pp. 934–954, 2014.

[6] E. Bozzo, R. Carniel, and D. Fasino, "Relationship between singular spectrum analysis and fourier analysis: Theory and application to the monitoring of volcanic activity," *Computers & Mathematics with Applications*, vol. 60, no. 3, pp. 812–820, 2010.

[7] T. Alexandrov, "A method of trend extraction using singular spectrum analysis," *arXiv preprint arXiv:0804.3367*, 2008.

[8] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.

[9] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[10] 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.

[11] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.

[12] 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.

[13] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[14] J. D. Phillips, W. Schwanghart, and T. Heckmann, "Graph theory in the geosciences," *Earth-Science Reviews*, vol. 143, pp. 147–160, 2015.