Advances in Big Data Analytics: Algorithmic Stability and Data Cleansing

Ph.D. Dissertation Defense

Yuping (Allan) Lu

Dissertation Advisor: Michael A. Langston

Committee Members: Qing (Charles) Cao, Jitendra Kumar,

Audris Mockus

Overview

This dissertation studies the following:

- A pair of foundational issues in algorithmic stability (robustness and tuning), with application to clustering in high-throughput computational biology.
- An issue in data cleansing (outlier detection), with application to pre-processing in streaming meteorological measurement.

Collaborators

- Michael A. Langston (University of Tennessee)
- Jitendra Kumar (Oak Ridge National Laboratory)
- Charles A. Phillips (University of Tennessee)
- Elissa J. Chesler (The Jackson Laboratory)
- Nathan Collier (Oak Ridge National Laboratory)
- Bhargavi Krishna (Oak Ridge National Laboratory)

Outline

- Introduction
- Algorithmic Stability Part I: Robustness
- Algorithmic Stability Part II: Tuning
- Data Cleansing: Outlier Detection
- Conclusions

Big Data

- The concept of big data was probably first mentioned in 1997 by Michael Cox and David Ellsworth when they worked on the visualization of computational fluid dynamics.
- Three Vs: volume, velocity and variety.
- Big data comes from a wide variety of fields. Examples include meteorology, genomics, neuroscience, social networks, public health, sensors, retail, financial services, transportation, web search, telecommunications and many other domains.
- Most big data algorithms fall into one of several broad categories. These categories often overlap with no clear boundaries.

Experimental Data

- More than 500,000 microarray datasets are publicly available due to the cheap price of genome sequencing.

- We carefully selected gene co-expression datasets from Gene Expression Omnibus (GEO).
  - Baker's yeast (*S. cerevisiae*)
  - Fruit fly (*D. melanogaster*)
  - Bacteria (*E. coli*))
  - Mouse (*M. musculus*
  - Fungi (*P. chrysogenum*)

- Analysis was performed on the Compute and Data Environment for Science (CADES) clusters at ORNL.

Experimental Data

- Atmospheric Radiation Measurement (ARM) facility collects data from instruments deployed in ground stations across the globe.

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

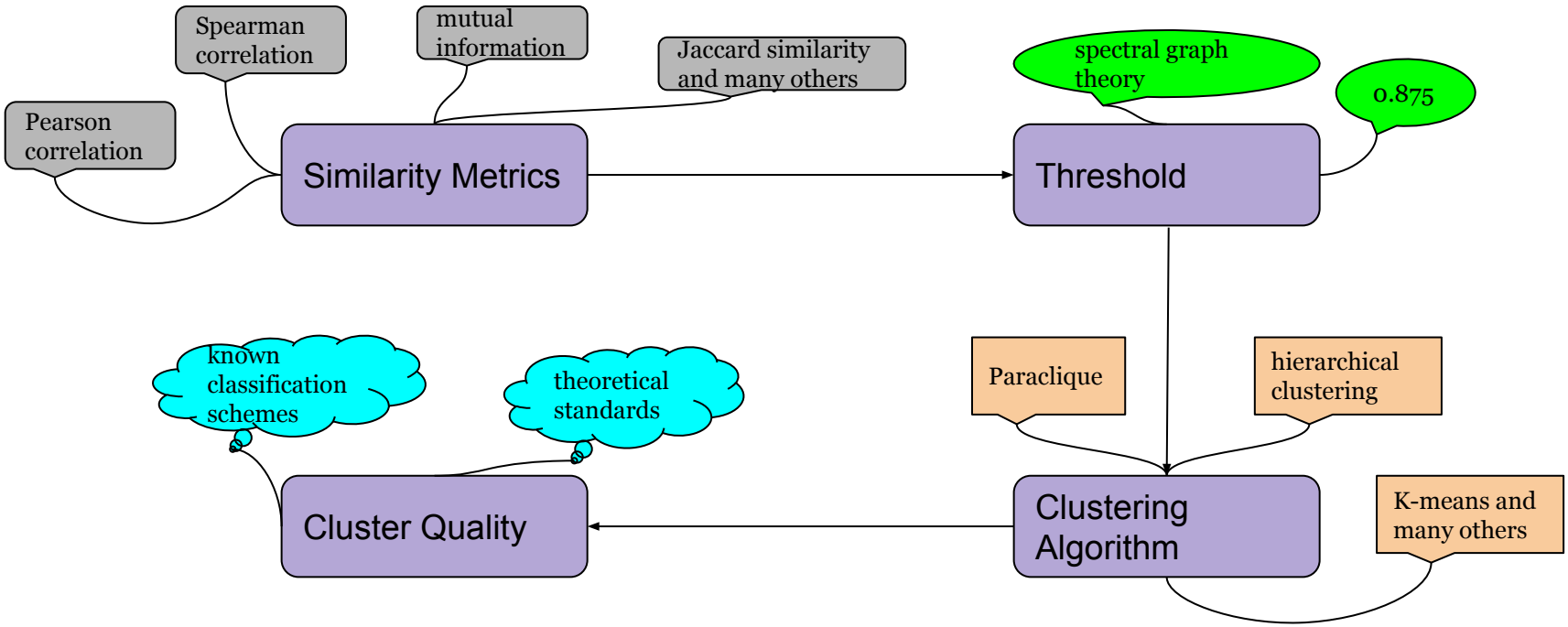
- We used data from Surface Meteorology Systems (MET) collected at the ARM Southern Great Plains (SGP) site in Oklahoma.

- Five core variables are used:
  - Air temperature
  - Vapor pressure
  - Atmospheric pressure
  - Relative humidity
  - Wind speed

Important Graph-Theoretic Concepts

- A graph **G=(V,E)** is formed by a set of vertices **V(G)** and a set of edges **E(G).**
- Graphs mentioned in the dissertation are simple, finite, undirected and unweighted, unless otherwise stated.
- A clique, or complete subgraph, is a subgraph in which each vertex is connected to every other vertex in that subgraph. A maximal clique is a clique to which no vertex can be added to form a larger clique. A maximum clique is a largest maximal clique.
- A paraclique is a near-clique, that is, one that is missing a handful of edges. It is designed to ameliorate the effects of noise, and is constructed by first finding a maximum clique, ***C***, and then adding vertices adjacent to most but not all of ***C*** in a tightly controlled fashion.

Graph Clustering Workflow



Outline

- Introduction
- Algorithmic Stability Part I: Robustness
- Algorithmic Stability Part II: Tuning
- Data Cleansing: Outlier Detection
- Conclusions

Clustering Basics

A clustering algorithm classifies a set of objects into subsets using some measure of similarity between each object pair.

Measuring cluster quality:

- Problem: ground truth is often unknown.
- known classification schemes (e.g. domain-specific knowledge such as ontological enrichment)
- theoretical standards (e.g. modularity, clustering coefficient, silhouette coefficient, etc.)

Robustness Motivation

- Clustering algorithms typically have one or more adjustable settings.
- Such a setting may represent, for example, a preset variable, a parameter of interest, or various sorts of initial assignments.
- A question of interest then is this: to what degree do the clusters produced vary as setting values change?

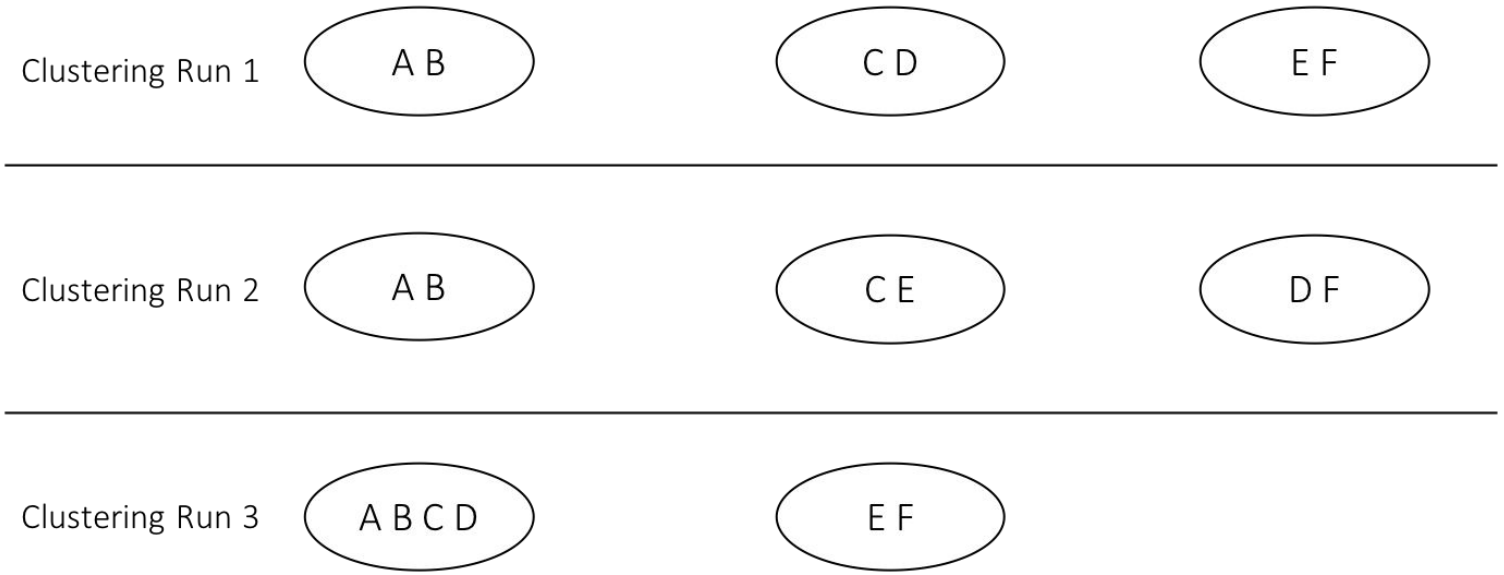
Robustness Definition

If a pair of items appear together once, will they appear together consistently?

$$R = t / (dr)$$

- ***t*** **-** the total number of (not necessarily distinct) pairs of objects that appear together in some cluster summed over all runs.
- ***d*** **-** the number of distinct pairs of objects that appear together in some cluster produced by some run.
- ***r*** **-** the number of times the clustering algorithm was run, each run using a different value for some setting of interest.
- Robustness lies in the interval **(0, 1]**.

An Example

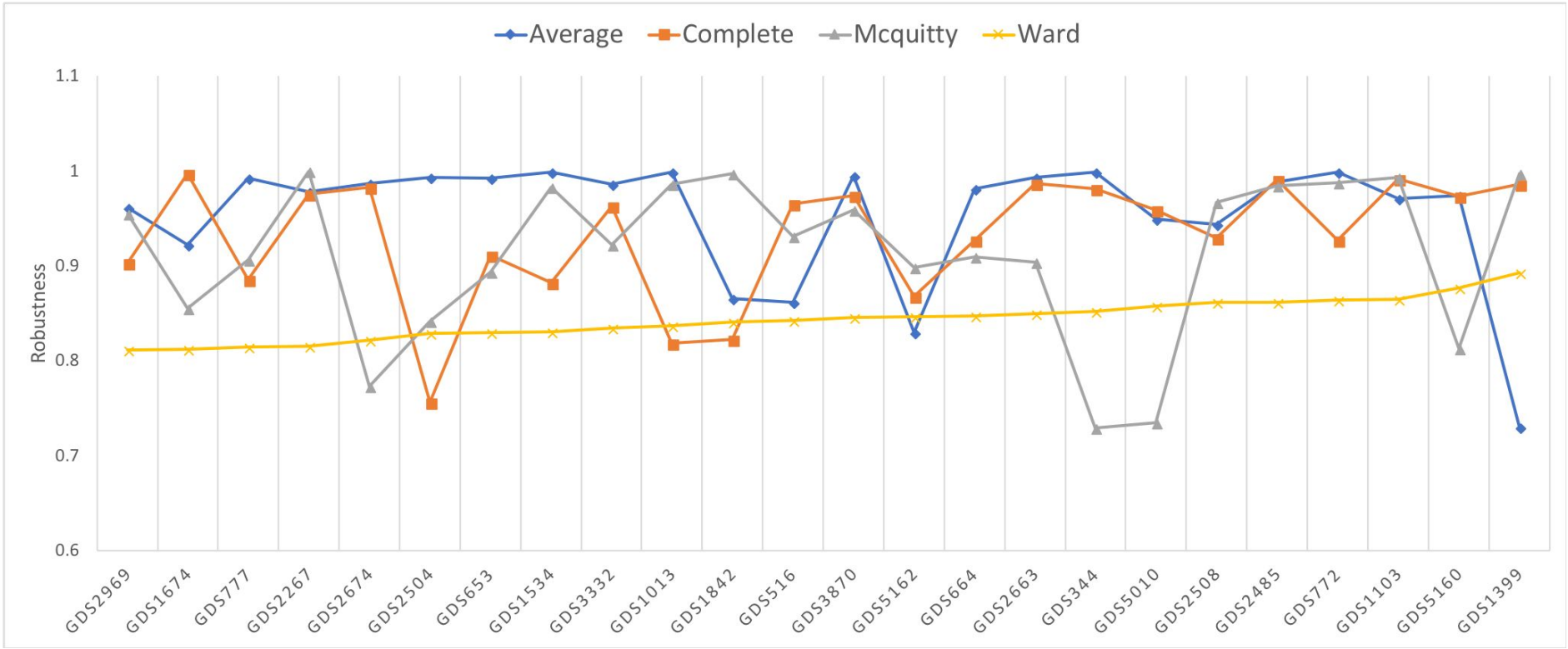


(A,B): 3/3; (A,C): 1/3; (A,D): 1/3; (B,C): 1/3; (B,D): 1/3; (C,D): 2/3; (C,E): 1/3; (D,F): 1/3; and (E,F): 2/3. Robustness = 0.481

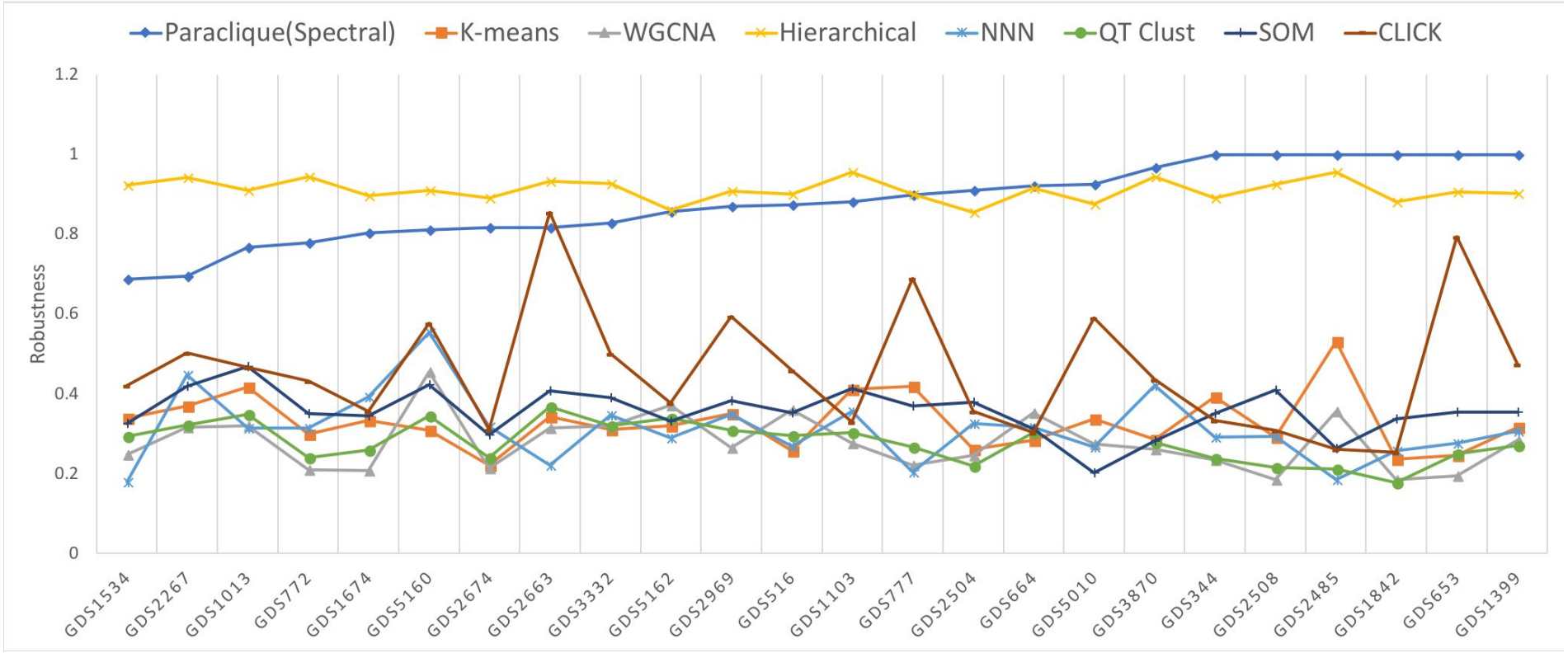
Clustering Methods Tested for Robustness

Hierarchical	**Setting**	**Graph-based**	**Setting**
Average	Number of clusters	CLICK	Cluster homogeneity
Complete	Number of clusters	NNN	Min neighborhood size
Mcquitty	Number of clusters	Paraclique	Starting clique
Ward	Number of clusters	WGCNA	Power
Partitioning	**Setting**	**Neural network**	**Setting**
K-means	Number of clusters	SOM	Grid type/size
QT Clust	Max cluster diameter		

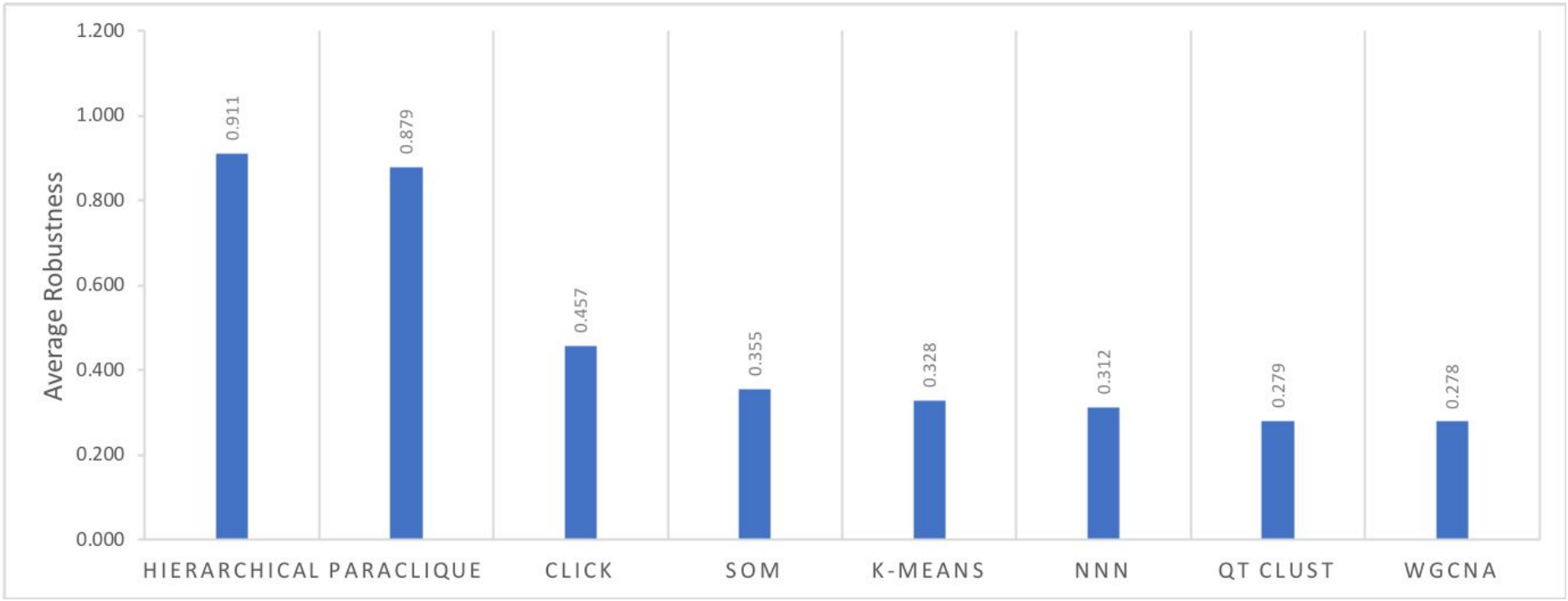
Robustness of Four Hierarchical Algorithms on 24 Transcriptomic Datasets



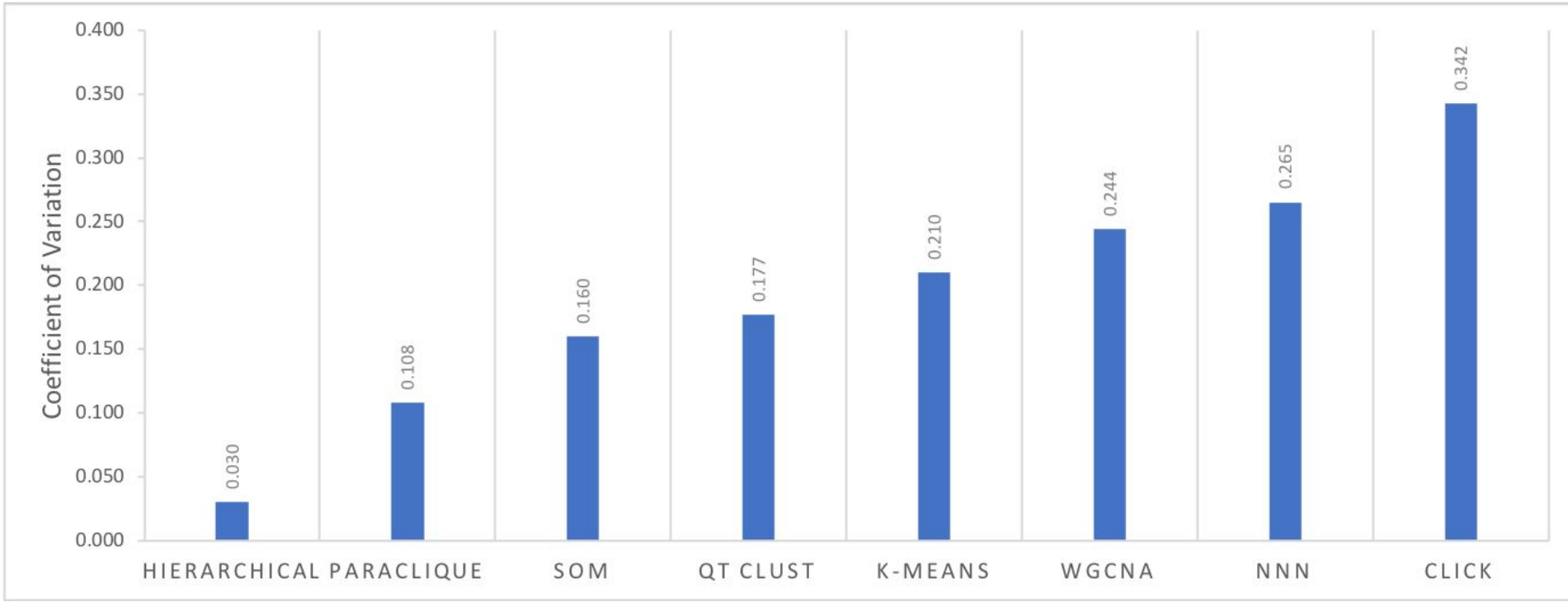
Robustness of All Algorithms Tested on 24 Transcriptomic Datasets



Average Robustness of Each Algorithm



Coefficient of Variation of Each Algorithm



Discussion

- Hierarchical methods display the highest overall robustness.
- WGCNA uses soft-power to construct its network, however, the topology of each weighted network changes with different powers, so that item pairs are not at all stable.
- For K-means, items often shift to different clusters as the number of clusters changes.
- It is a similar situation for SOM, QT Clust, CLICK and NNN.
- For paraclique, the high robustness with different starting cliques is likely due in part to the fact that many of these cliques have significant overlap.

Outline

- Introduction
- Algorithmic Stability Part I: Robustness
- Algorithmic Stability Part II: Tuning
- Data Cleansing: Outlier Detection
- Conclusions

Motivation

- When more than one maximum clique is present, deciding which to employ is usually left unspecified. In practice, graph clustering algorithms usually use the first maximum clique found.

- We empirically tested three different maximum clique selection strategies, comparing Gene Ontology (GO) enrichment p-values of paracliques produced by each.

28 yeast microarray expression datasets and experimental results obtained at a threshold of 0.80.

Dataset	Maximum Clique		Average Paraclique Edge Weights and Enrichment Scores					
	Size	Number	Highest	P-value	Lowest	P-value	Random	P-value
GDS344	87	6	0.9111	1.10E-49	0.9099	5.30E-50	0.9105	5.30E-50
GDS362	304	75184	0.9267	2.60E-09	0.9259	1.90E-10	0.9267	1.90E-10
GDS600	1736	40	0.9584	1.50E-06	0.9584	1.40E-06	0.9584	1.50E-06
GDS772	78	6	0.9134	2.70E-26	0.9118	2.70E-26	0.9118	2.70E-26
GDS777	87	15	0.9101	2.00E-08	0.9096	2.00E-08	0.9096	2.00E-08
GDS922	450	2160	0.9235	5.20E-11	0.9230	5.80E-11	0.9232	6.50E-11
GDS991	317	2468	0.9245	1.10E-95	0.9224	1.70E-85	0.9243	6.90E-97
GDS1013	269	19152	0.9127	3.30E-127	0.9112	6.90E-123	0.9123	3.30E-127
GDS1103	312	672	0.9293	8.10E-20	0.9283	8.10E-20	0.9290	9.40E-20
GDS1534	154	180	0.9140	3.40E-08	0.9133	1.20E-06	0.9137	3.40E-08
GDS1550	361	240	0.9469	2.60E-05	0.9459	2.50E-05	0.9464	2.60E-05
GDS1551	453	48	0.9408	5.30E-06	0.9405	4.80E-06	0.9405	4.80E-06
GDS1611	182	258	0.8847	3.90E-05	0.8839	3.70E-05	0.8845	3.70E-05
GDS1674	93	160	0.9102	8.00E-14	0.9078	1.40E-13	0.9090	1.40E-13
GDS2050	617	1152	0.9365	2.10E-32	0.9363	2.10E-32	0.9364	2.80E-32
GDS2079	1611	16	0.9563	8.30E-07	0.9563	8.30E-07	0.9563	4.50E-07
GDS2267	168	312	0.9058	7.50E-103	0.9035	3.00E-98	0.9058	2.60E-101
GDS2462	1351	13	0.9538	3.10E-46	0.9535	1.30E-43	0.9537	3.10E-46
GDS2508	49	11	0.9036	1.40E-03	0.8980	1.50E-03	0.9002	1.50E-03
GDS2522	428	13724	0.9321	1.40E-03	0.9313	1.50E-03	0.9318	2.10E-04
GDS2625	309	80	0.9191	3.00E-06	0.9187	2.80E-06	0.9189	2.80E-06
GDS2663	282	600	0.9283	5.80E-18	0.9269	4.40E-16	0.9273	4.40E-16
GDS2925	89	60	0.8940	1.10E-03	0.8930	3.80E-03	0.8934	4.40E-03
GDS2969	119	24	0.9161	1.80E-12	0.9143	1.80E-12	0.9148	1.80E-12
GDS3061	181	152	0.9218	2.80E-25	0.9198	2.80E-25	0.9208	2.80E-25
GDS3137	562	1088	0.9354	1.00E-04	0.9350	1.00E-04	0.9353	1.50E-04
GDS3198	383	2184	0.9333	3.50E-06	0.9327	1.70E-06	0.9331	2.80E-06
GDS3438	3424	2	0.9898	8.50E-11	0.9898	8.50E-11	0.9898	8.50E-11

Comparisons Between Highest and Lowest Weight Maximum Cliques

p1 = 0.0000163

p2 = 0.00047

Threshold	Highest Better	No Difference	Lowest Better	Highest Better / Lowest Better	Binomial P-value
0.70	16	6	4	4	2.14E-03
0.71	10	7	6	1.667	9.96E-02
0.72	10	4	8	1.25	8.44E-02
0.73	11	6	9	1.222	9.96E-02
0.74	13	8	6	2.167	4.31E-02
0.75	14	5	8	1.75	2.15E-02
0.76	11	8	8	1.375	1.12E-01
0.77	13	8	7	1.857	5.36E-02
0.78	15	7	6	2.5	1.34E-02
0.79	9	10	8	1.125	1.61E-01
0.80	11	9	8	1.375	1.23E-01
0.81	12	7	8	1.5	7.47E-02
0.82	12	8	8	1.5	8.72E-02
0.83	9	12	7	1.286	1.58E-01
0.84	10	9	9	1.111	1.50E-01
0.85	11	8	9	1.222	1.23E-01
0.86	11	8	9	1.222	1.23E-01
0.87	8	13	7	1.143	1.42E-01
0.88	10	10	8	1.25	1.50E-01
0.89	10	10	8	1.25	1.50E-01
0.90	8	14	6	1.333	1.42E-01
Total	234	177	157	1.490	1.63E-05

Comparisons Between Highest and Random Weight Maximum Cliques

p1 = 0.00219

p2 = 0.0000278

Threshold	Highest Better	No Difference	Random Better	Highest Better / Random Better	Binomial P-value
0.70	17	3	6	2.833	6.29E-04
0.71	9	12	2	4.5	1.42E-01
0.72	8	6	8	1	1.67E-01
0.73	10	8	8	1.25	1.37E-01
0.74	11	12	4	2.75	1.12E-01
0.75	11	6	10	1.1	1.12E-01
0.76	13	9	5	2.6	4.31E-02
0.77	15	11	2	7.5	1.34E-02
0.78	11	10	7	1.571	1.23E-01
0.79	11	11	5	2.2	1.12E-01
0.80	9	10	9	1	1.58E-01
0.81	9	11	7	1.286	1.61E-01
0.82	10	11	7	1.429	1.50E-01
0.83	8	14	6	1.333	1.42E-01
0.84	12	12	4	3	8.72E-02
0.85	9	12	7	1.286	1.58E-01
0.86	7	14	7	1	1.09E-01
0.87	11	12	5	2.2	1.23E-01
0.88	9	13	6	1.5	1.58E-01
0.89	9	13	6	1.5	1.58E-01
0.90	7	15	6	1.167	1.09E-01
Total	216	225	127	1.701	2.19E-03

Comparisons Between Random and Lowest Weight Maximum Cliques

A random choice was better in 191 graphs, there was no difference in 215 graphs, and a lowest choice was better in 162 graphs.

Although the ratio was still above one (at 1.179), neither binomial test reached the level of significance, with $p = 0.035$ and $p = 0.0876$, respectively.

Discussion

- Sometimes there is little difference in enrichment *p*-values due to significant overlap between maximum cliques. In GDS344, for example, 84 of 87 vertices appear in all maximum cliques at a threshold of 0.8.

- The number of maximum cliques can vary greatly between datasets, and even between graphs constructed at different thresholds from the same dataset.

Outline

- Introduction
- Algorithmic Stability Part I: Robustness
- Algorithmic Stability Part II: Tuning
- Data Cleansing: Outlier Detection
- Conclusions

Motivation

- Collected datasets at ARM require high accuracy to enable rigorous study of atmospheric processes. But outliers are common due to instrument failure or extreme weather events.
- Currently, the datasets are checked manually by Data Quality Office (DQO).
- Thus an effective and efficient outlier and noise detection method is crucial for ARM to provide scientific users with high quality data for research.

Background

- Outlier detection is a common task in many application domains.
- Common techniques for outlier detection include signal processing, classification, clustering, nearest neighbor, density, statistical, information theory, spectral decomposition, etc.
- We compared three methods from pairwise, univariate, and multivariate perspective respectively.

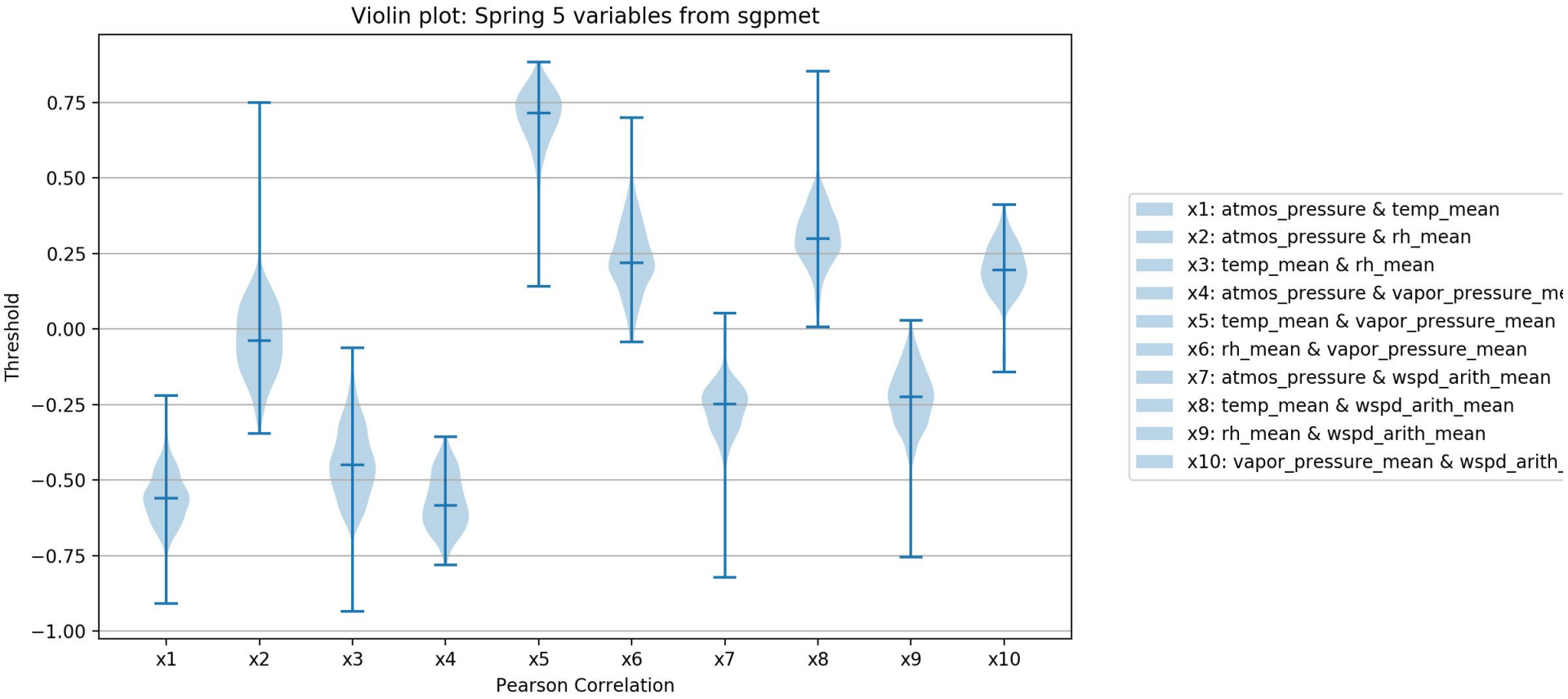
Pearson Correlation Coefficient

- Co-located meteorological variables measure different aspects of atmospheric conditions at any location, and driven by atmospheric physics are inherently correlated with each others. Any atmospheric phenomena at the location would affect all variables in an expected and correlated fashion.

- We performed pairwise comparisons of the five variables using Pearson correlation on data from 24 extended facilities.

- Calculated Pearson correlation coefficients are stored as the expected values between two variables. If this pairwise Pearson correlation of two variables deviates far away from our expected historical correlation, we treat it as an outlier.

Pearson Correlation patterns for ten meteorological variable pairs during **spring** season across all the years.

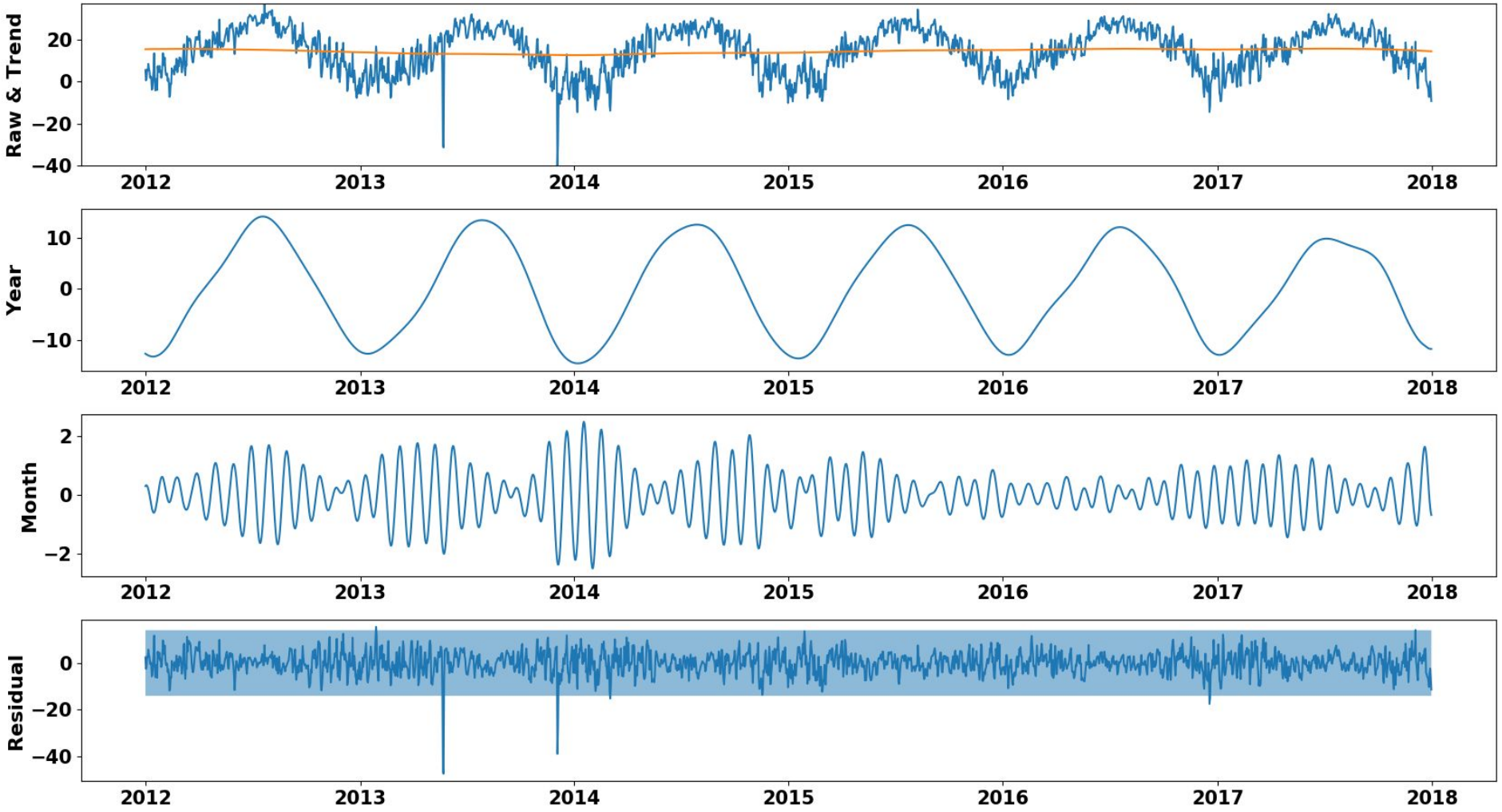


Singular Spectrum Analysis

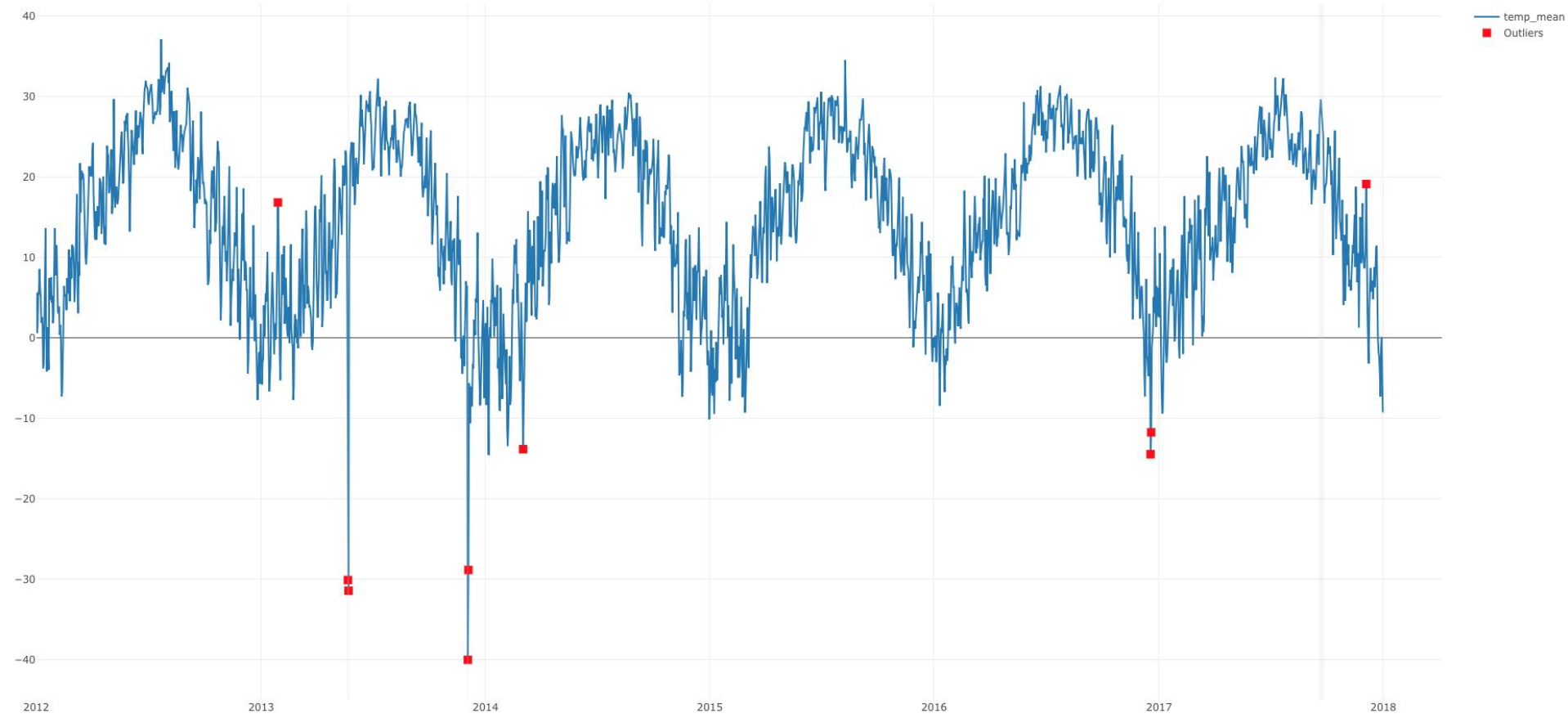
- Singular Spectrum Analysis (SSA) is a univariate time series analysis method which was applied to detect outliers by removing the anticipated annual and seasonal cycles from the signal to accentuate anomalies.

- The general idea is to use a subset of the decomposition of trajectory matrix to approximate the original data.

Decomposition of air temperature data from MET instrument at facility E33 using SSA method to isolate various frequencies.



Mean temperature outliers detected by SSA on instrument E33 from 2012 to 2017.



http://yupinglu.me/arm-ssa/plotly/temp_mean/E33-2012-2017.html

K-means

- The southern plains, where SGP site is located, are known to experience frequent extreme storms occurring most often during spring and early summer.

- However, meteorological variables during such events won't be captured by Pearson Correlation as they may still follow known correlation structure at seasonal scales or by SSA method since any individual variable may not show large deviation.

- Multivariate approaches such as K-means clustering have been widely used to identify weather and climate regimes.

K-means

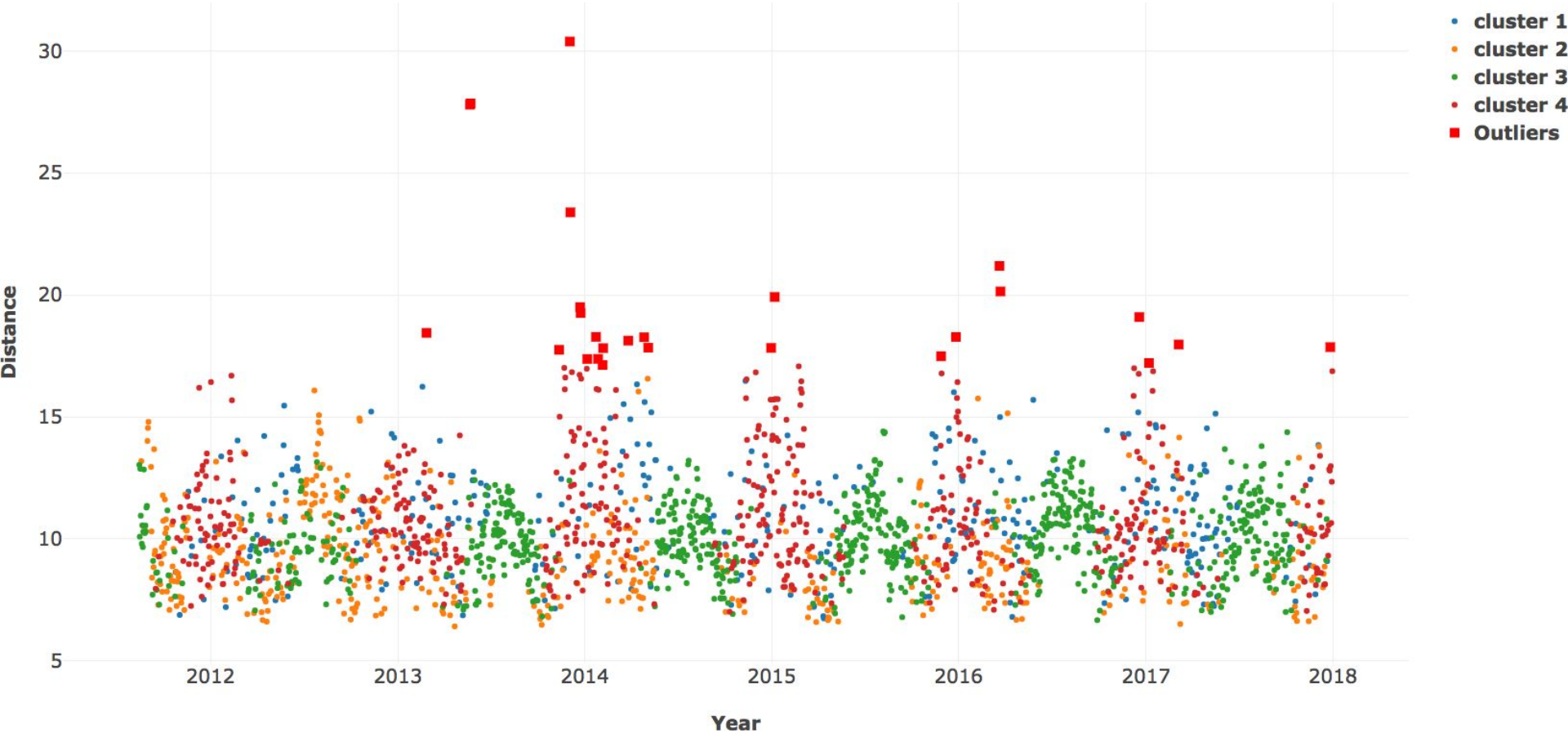
- We applied *k*-means clustering to ARM meteorological data.
- We then calculated the distance of each point within a cluster to its corresponding cluster centroid.
- Given known seasonal patterns at the site we set *k* to four to determine weather regimes for four seasons.

Algorithm 1: K-means Outlier Detection

Input : ARM time series data
Output: Outliers

```
outliers ← ∅
df ← ARM time series data
data ← df['atmos_pressure','temp_mean',
  'rh_mean','vapor_pressure_mean','wspd_arith_mean']
number_of_clusters ← 4
clusters ← K-means(data, number_of_clusters)
distances ← Distance between each point and its centroid
mean ← arithmetic mean of distances
sigma ← standard deviation of distances
threshold ← mean + 3 * sigma
for i in range(size of distances) do
    if distances[i] > threshold then
        outliers ← outliers ∪ distances[i]
    end
end
return outliers
```

Outliers detected by K-means on instrument E33 from 2012 to 2017.



http://yupinglu.me/arm-ssa/plotly/kmeans/E33.0.html

Evaluation of Detected Outliers

- We treated outliers detected in DQR database as True Positives. The equations below show the calculation of Precision and Recall.

$$\textbf{Precision} = \frac{\textbf{True Positives } (\textit{Outliers detected in DQR database})}{\textbf{True Positives} + \textbf{False Positives } (\textit{Outliers detected not in DQR database})}$$

$$\textbf{Recall} = \frac{\textbf{True Positives } (\textit{Outliers detected in DQR database})}{\textbf{True Positives} + \textbf{False Negatives } (\textit{Undetected records in DQR database})}$$

Results

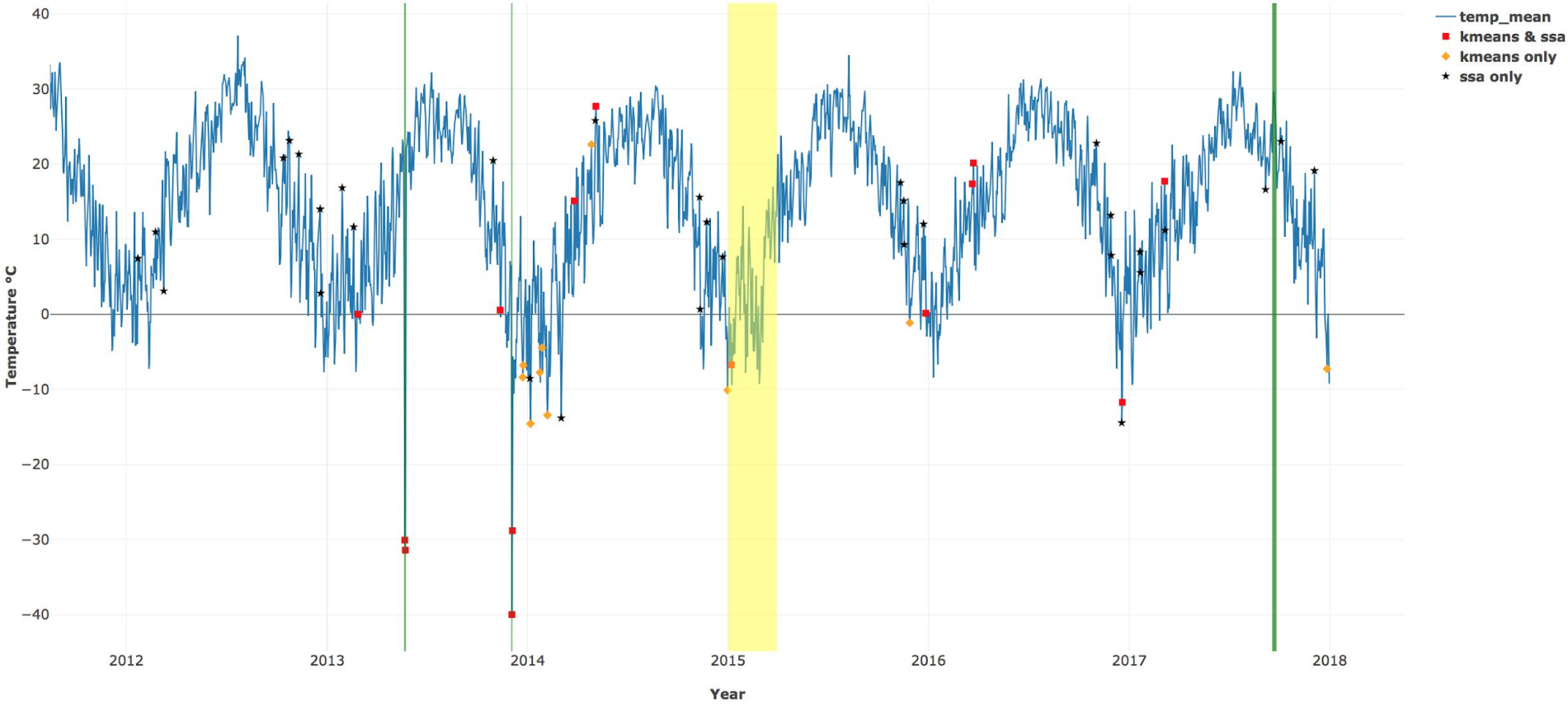
	Outlier Set Size
SSA	922
K-means	508
Intersection	378
Symmetric Difference	674

SSA and K-means outlier set size

Method	Variable	Precision	Recall
SSA	Air Temperature	16.00%	1.20%
SSA	Vapor Pressure	20.70%	1.40%
SSA	Atmospheric Pressure	0.00%	0.00%
SSA	Relative Humidity	14.80%	0.50%
SSA	Wind Speed	0.60%	1.50%
Kmeans	All Variables	13.00%	1.90%
Combined	All Variables	11.10%	4.10%

Precision and recall of SSA and K-means

Outliers detected at facility E33 for air temperature by Pearson correlation, SSA and k-means algorithms. Outliers detected by Pearson correlation are in the shaded yellow region. Outliers detected by both SSA and k-means algorithms are shown by red squares, while those identified by SSA and k-means only are indicated by black stars and orange diamonds respectively. DQR records are denoted by the vertical green shaded areas.



http://yupinglu.me/arm-ssa/plotly/combined/E33.0.html

Discussion

- Pearson correlation coefficient is a pairwise comparison method, however, if the two variables deviate in the same direction, their correlation may not change significantly and thus may go undetected. Due to seasonal nature of the analysis, it was not able to identify outliers that persisted at hours to days only.

- Univariate SSA method was very effective at identifying outliers with extreme high and low values in the time series but required the input data to be consistent with no missing values.

- K-means could be used to detect extreme storms and weather events but it was hard to tell which variable mainly caused the abnormality.

- However, these drawbacks could be easily overcome by combining methods together to detect outliers from three different angles.

Outline

- Introduction
- Algorithmic Stability Part I: Robustness
- Algorithmic Stability Part II: Tuning
- Data Cleansing: Outlier Detection
- Conclusions

Our Work in the Field

- "**A Robustness Metric for Biological Data Clustering Algorithms**"

  [14th International Symposium on Bioinformatics Research and Applications (ISBRA 2018), Beijing, China]

- "**Clique Selection and its Effect on Paraclique Enrichment: An Experimental Study**"

  [Under review]

- "**Detecting Outliers in Streaming Time Series Data from ARM Distributed Sensors**"

  [IEEE International Conference on Data Mining Workshop (ICDMW 2018), Singapore]

Summary of Contributions

- We demonstrated how the robustness of clustering algorithms can be measured and compared. Our tests on transcriptomic data show that hierarchical methods and the paraclique algorithm have higher robustness scores than other clustering algorithms.

- We performed empirical testing on three different maximum clique selection strategies and found that selecting a maximum clique with highest average edge weight tends to produce superior results on transcriptomic data.

- We described a novel automated framework for outlier detection in meteorological data. Experimental results show that 88.9% of outliers detected by the framework are not found in the database.

Future Research Directions: Robustness

- Extend the metric to overlapping clustering algorithms.

- Test the metric on other types of biological data and data from other domains such as communications, transportation and social networks.

Future Research Directions: Tuning

- Test whether the selection strategies have the same effect on the second or deeper level paracliques.
- Test other selection strategies. For instance, instead of restricting the choice to maximum cliques, one might choose the (not necessarily maximum) clique with the highest total edge weight.
- Further analyses on larger and more diverse datasets may also reveal greater improvements.

Future Research Directions: Outlier Detection

- Test the framework on meteorological data collected from other sites, and even other types of ARM data, for example, weather radar data and satellite observation data.

- Multivariate SSA methods and machine learning could also be explored in an effort to detect outliers more effectively.

Future Research Directions

- This dissertation only focused on three tasks in big data analytics for biological data and meteorological data. Future tasks include classification, regression, visualization and many others.

- Other technologies could also be explored, for example, Spark, Cassandra, Neo4j, etc.

Acknowledgements

- I would like to thank all my collaborators, without whom this dissertation would not have been possible.

- Michael A. Langston (University of Tennessee)
- Jitendra Kumar (Oak Ridge National Laboratory)
- Charles A. Phillips (University of Tennessee)
- Elissa J. Chesler (The Jackson Laboratory)
- Nathan Collier (Oak Ridge National Laboratory)
- Bhargavi Krishna (Oak Ridge National Laboratory)

THANK YOU
ANY QUESTIONS?