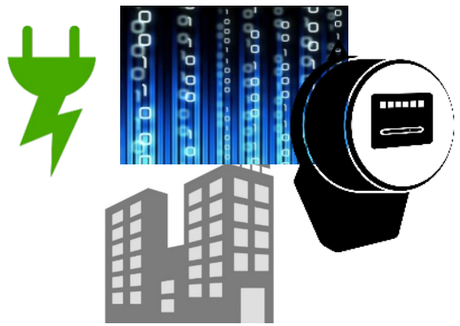
Overview

- **Introduction and problem statement**
- Case-study datasets
- Comparative assessment of clustering techniques
- Two-stage clustering
- Conclusion and future directions

High-resolution building energy data



Smart meter data

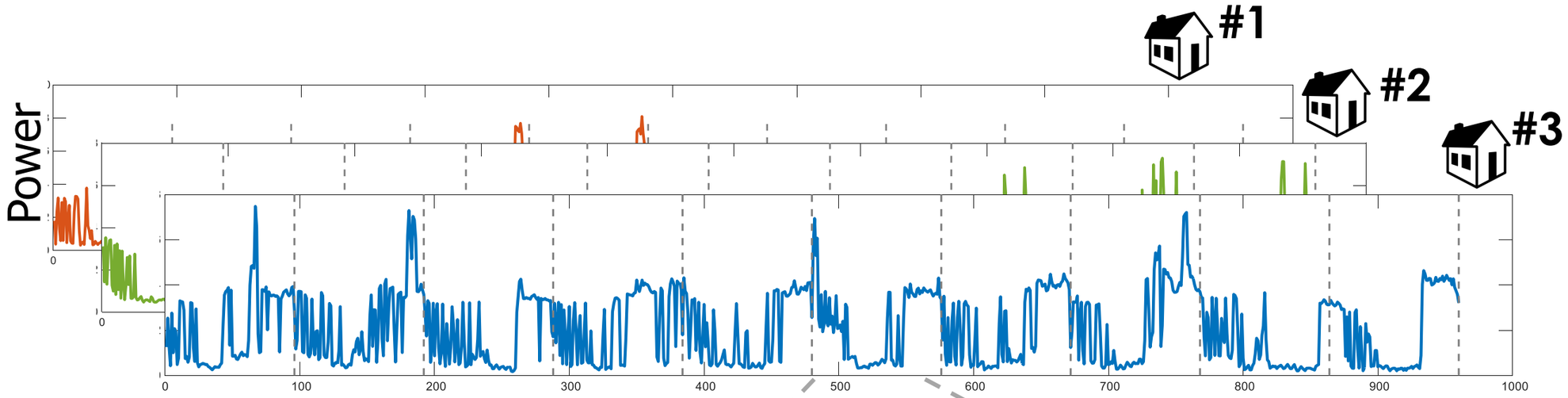
Many buildings

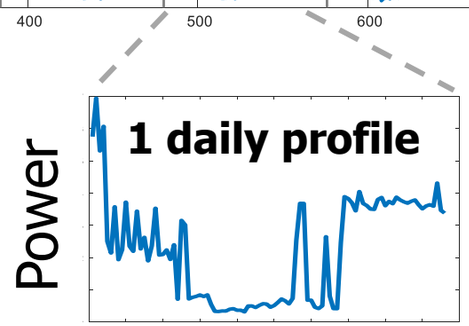
Insight for utilities

- Employing massive smart meter data promotes the efficiency of the power grid.
- Provides insight on consumption behaviors and lifestyles of the consumers at neighborhood scale.

Complex and varied energy patterns

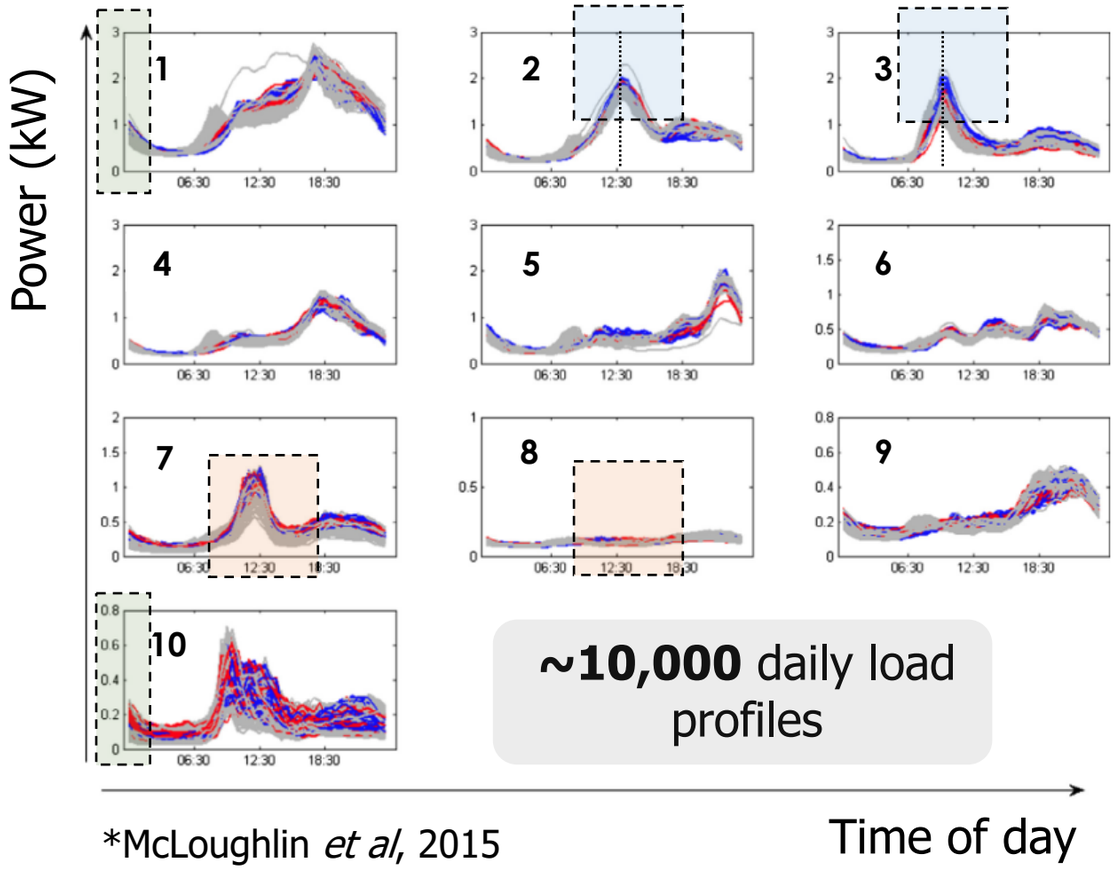
Time-series





- Decision-makers needs the distribution of **typical consumption patterns**. However, daily load profiles show **highly volatile behaviors**.

Daily profile segmentation (*Clustering*)



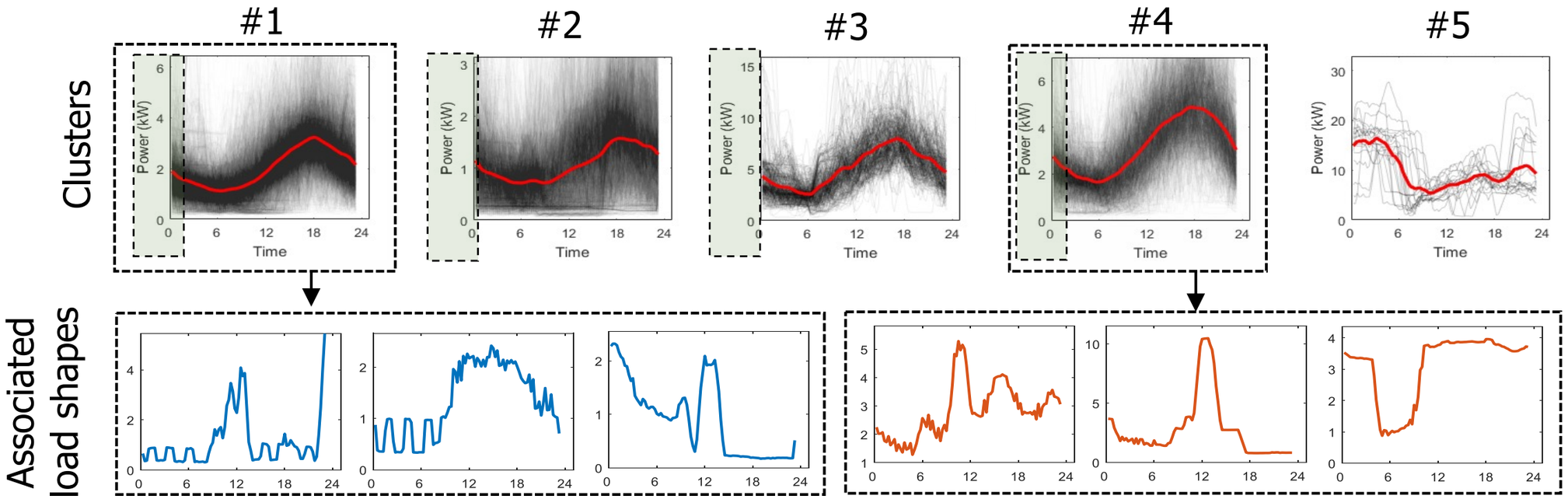
*McLoughlin *et al*, 2015

Important features in clustering:

1. Peak detection
2. Peak timing
3. Energy volume

Problem statement

- Clustering based on validation indices can result in simplified patterns that do not represent the shape or energy volume of their associated load profiles.

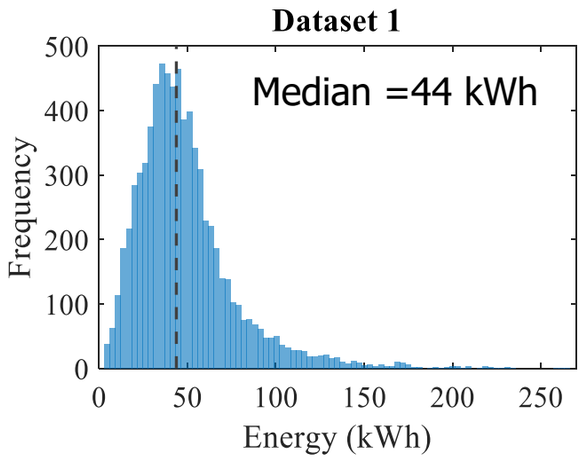


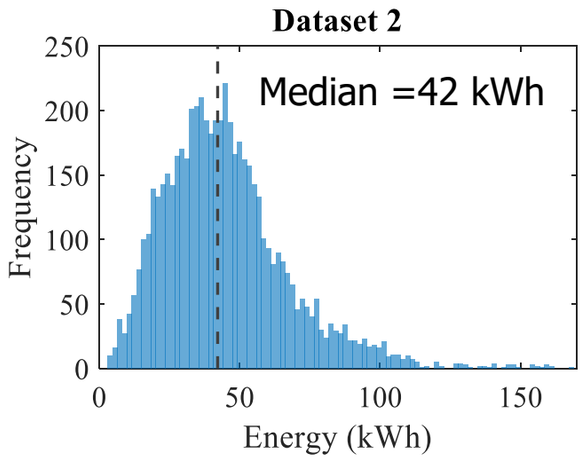
Overview

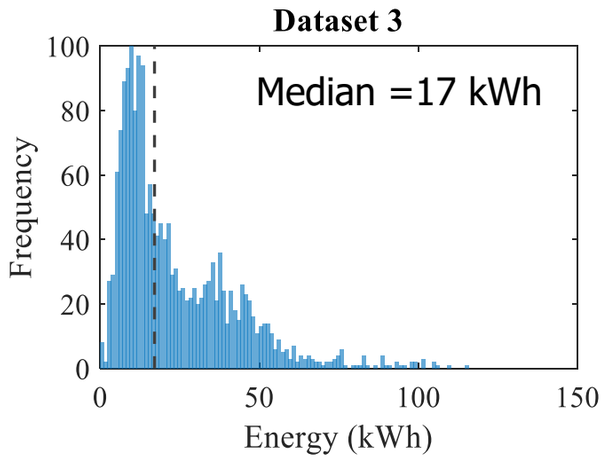
- Introduction and problem statement
- **Case-study datasets**
- Comparative assessment of clustering techniques
- Two-stage clustering
- Conclusion and future directions

Case-study datasets

Dataset	Location	Number of buildings	Duration	# of daily profiles	# of samples per profile
Dataset 1	Austin, TX	129	60 days	7535	96
Dataset 2	Austin, TX	100	60 days	5676	96
Dataset 3	Boulder, CO	31	60 days	1790	96





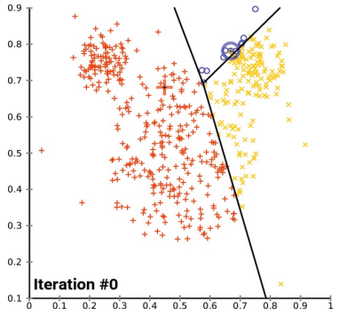


Overview

- Introduction and problem statement
- Case-study datasets
- **Comparative assessment of clustering techniques**
- Two-stage clustering
- Conclusion and future directions

Clustering techniques

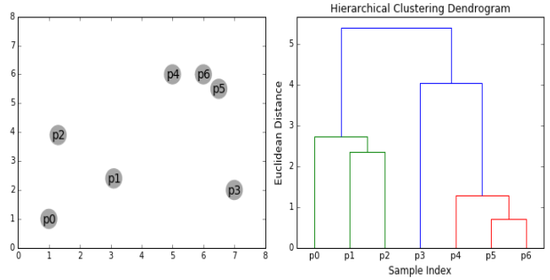
K-means



- Distance measure-based

$$SSE = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - C_k\|^2$$

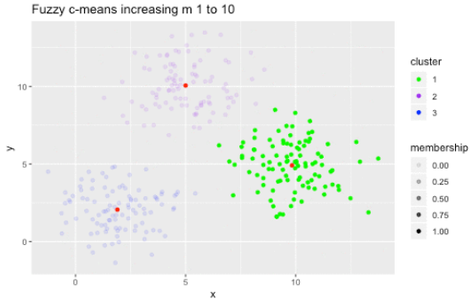
Hierarchical clustering



- Tree structure

$$d_{ij} = \|X_i - X_J\|^2$$

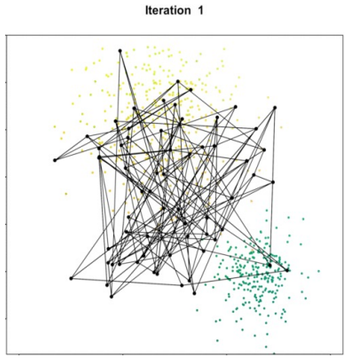
Fuzzy c-means



- Degree of fuzziness

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik}^{m} \|x_i - C_k\|^2$$

SOM



- Neural network-based

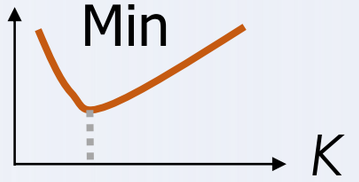
$$n_i(t+1) = n_i(t) + \alpha(t)\theta_{bi}(x(t) - n_i(t))$$

Ref for images: algobeans.com, wikipedia.com

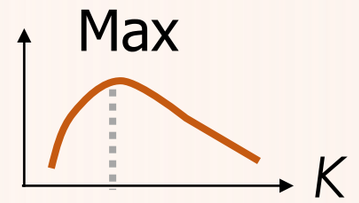
Cluster validation index (CVI)

Selection criteria

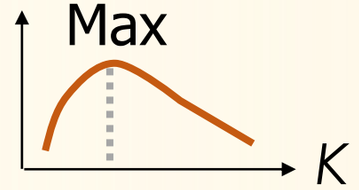
Davies-Bouldin: $$DBI = \frac{1}{K}\sum_i max_{j,j\neq i}\left[\frac{1}{\|C_i\|}\sum_{x\in C_i} d(x,\mu_i) + \frac{1}{\|C_j\|}\sum_{x\in C_j} d(x,\mu_j)\right]$$



Sillhouette: $$SIL = \frac{1}{K}\sum_i \left\{\frac{1}{\|C_i\|}\sum_{x\in C_i} \frac{b(i)-a(i)}{\max(b(i),a(i))}\right\}$$



Calinski-Harabasz: $$CHI = \frac{\sum_i \|C_i\| * d^2(\mu_i,\mu_j)/(K-1)}{\sum_i \sum_{x\in C_i} * d^2(x,\mu_j)/(N-K)}$$



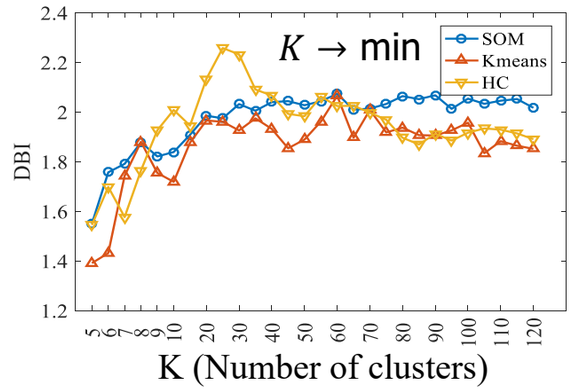
Within cluster sum of square error: $$WCSSE = \sum_i \sum_{x\in C_i} \|x-\mu_i\|^2$$

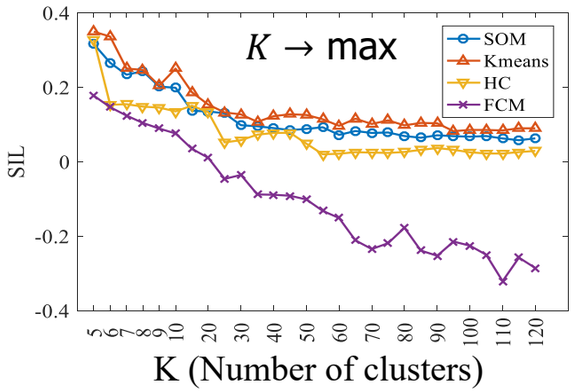
Comparative assessment

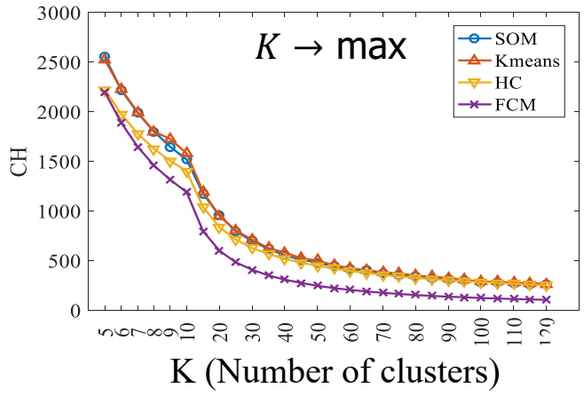
- Different combinations of Clustering techniques and Validation Index for evaluation

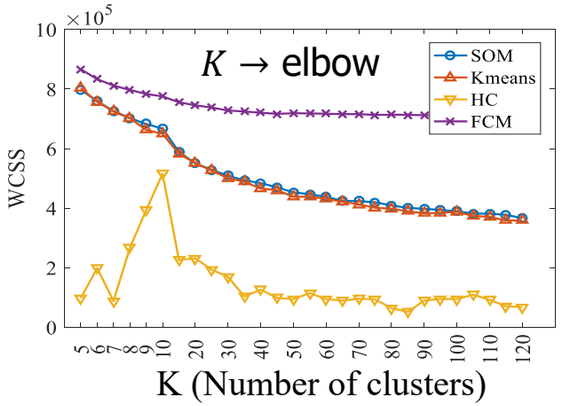
		Validation index			
		Davies-Bouldin	Sillhouette	Calinski-Harabasz	WCSS
Clustering technique	K-means	?	?	?	?
	Hierarchical	?	?	?	?
	Fuzzy c-means	?	?	?	?
	SOM	?	?	?	?

CVI results





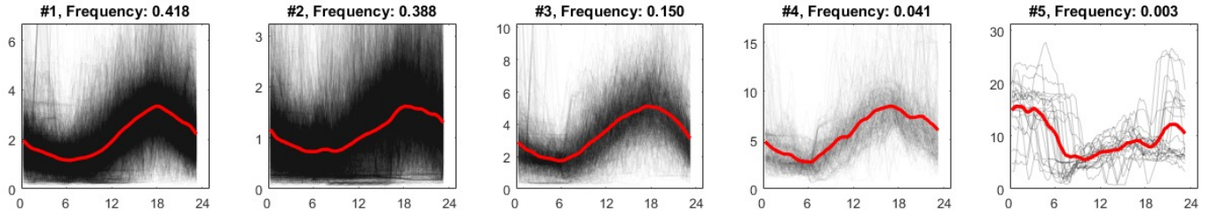




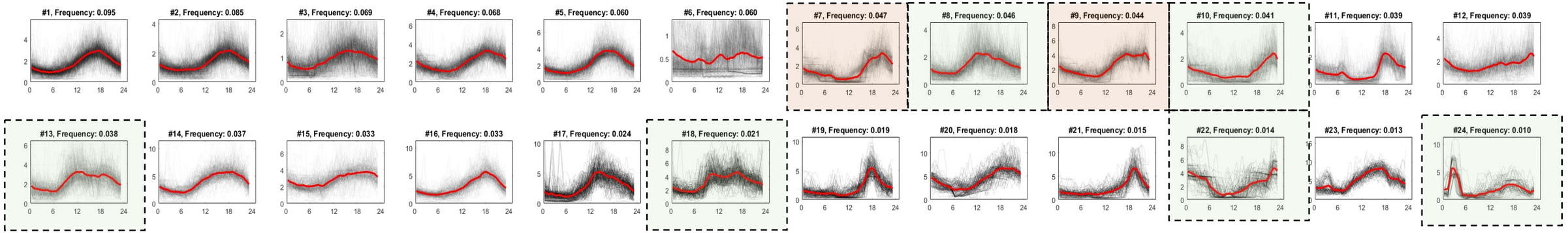
- Three out of four metrics estimate a low number of clusters (**K=5**).
- Only WCSS selects a higher number (~**K=30**).

Compared two scenarios
K=5 versus K=30

Empirical investigation



K=5



K=30

Going with higher number reveals distinct patterns.

However, it results in some similar (redundant) clusters.

Comparative assessment

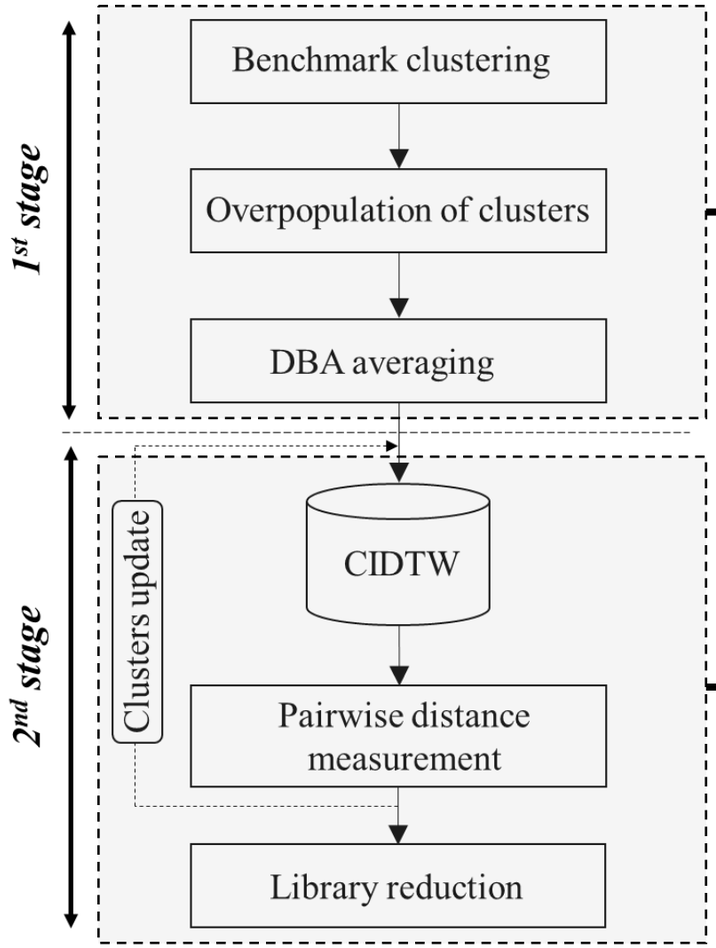
		Validation index			
		Davies-Bouldin	Sillhouette	Calinski-Harabasz	WCSS
Clustering technique	K-means	✓	✓	✓	✓
	Hierarchical	✓	✓	✓	✓
	Fuzzy c-means	✓	✓	✓	✓
	SOM	✓	✓	✓	✓

Limitation: Low number of clusters can't retrieve realistic energy profiles. High number of clusters could solve this problem but results in similar (redundant) clusters.

Overview

- Introduction and problem statement
- Case-study datasets
- Comparative assessment of clustering techniques
- **Two-stage clustering**
- Conclusion and future directions

Proposed two-stage clustering



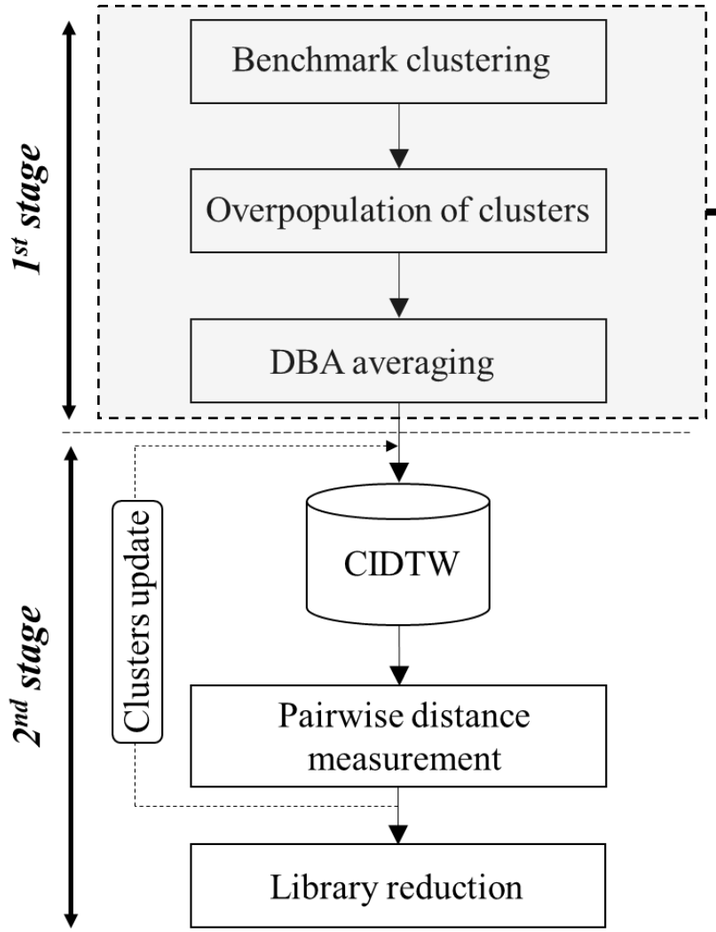
- **1st Stage**

  High number of clusters (Accuracy)

- **2nd Stage**

  Find relatively similar clusters and merge them

Proposed two-stage clustering

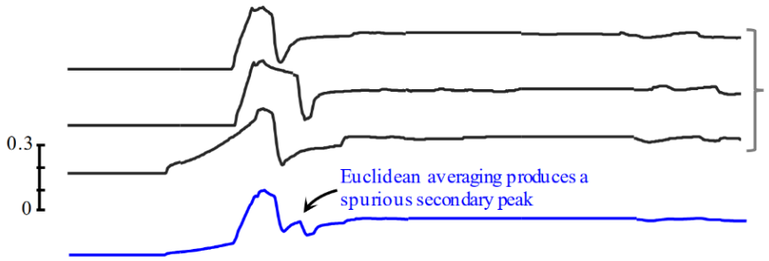


1st Stage

- Initial high number of clusters to reveal distinct patterns
- Transferring the clusters into a representative format

DBA averaging

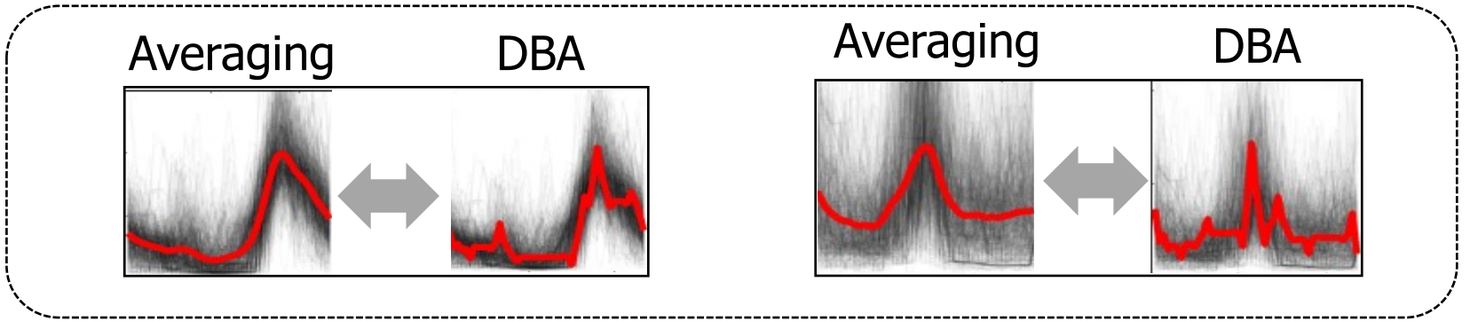
- Conventional averaging problem: Centroid is different from the profiles.



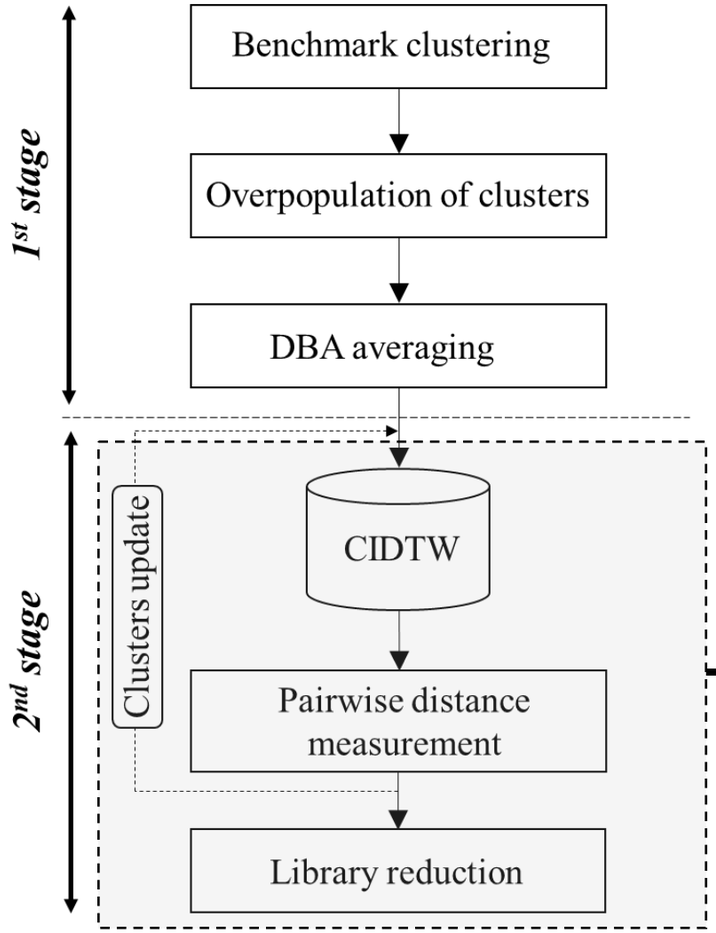
- *Petitjean et al., 2014*

DBA: Using an expectation maximization approach to refine the medoid of each group through finding the best set of alignments through iterations.

$$C_i'(t) = barycenter(assoc(C_i(t))) \qquad barycenter(X_1, X_2, \dots, X_m) = \frac{X_1 + X_2 + \cdots + X_m}{m}$$



Proposed two-stage clustering

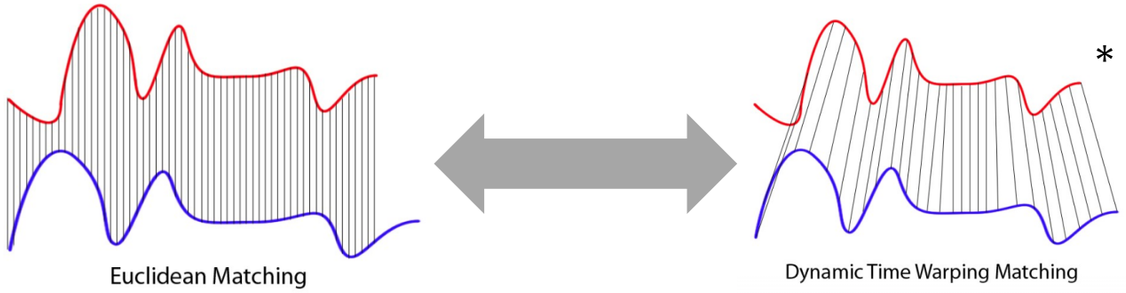


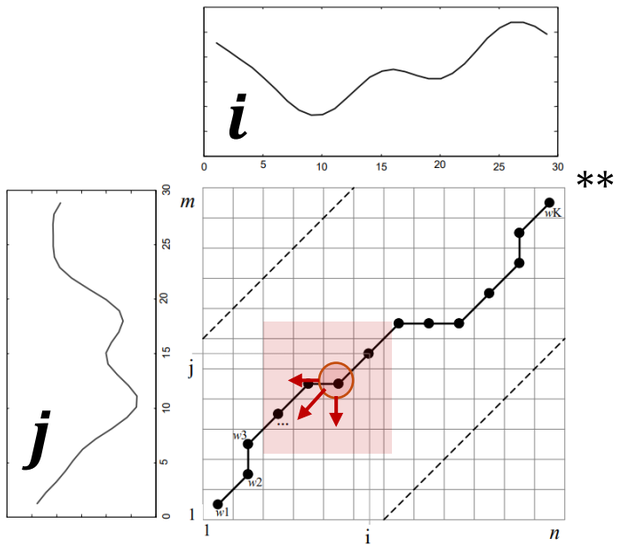
2nd Stage

- Selection of proper similarity measure for merging
- Updating the clusters by merging similar ones

CI-DTW *(complexity invariant dynamic time warping)*

Using a dynamic programming-based distance measure (CI-DTW) to find the similarity between pairs of clusters.





$$D(P_i, Q_j) = \delta(p_i, q_i) + \min \begin{cases} D(P_{i-1}, Q_{j-1}) \\ D(P_i, Q_{j-1}) \\ D(P_{i-1}, Q_j) \end{cases}$$

* https://towardsdatascience.com/

** Keogh and Pazzani, 2001

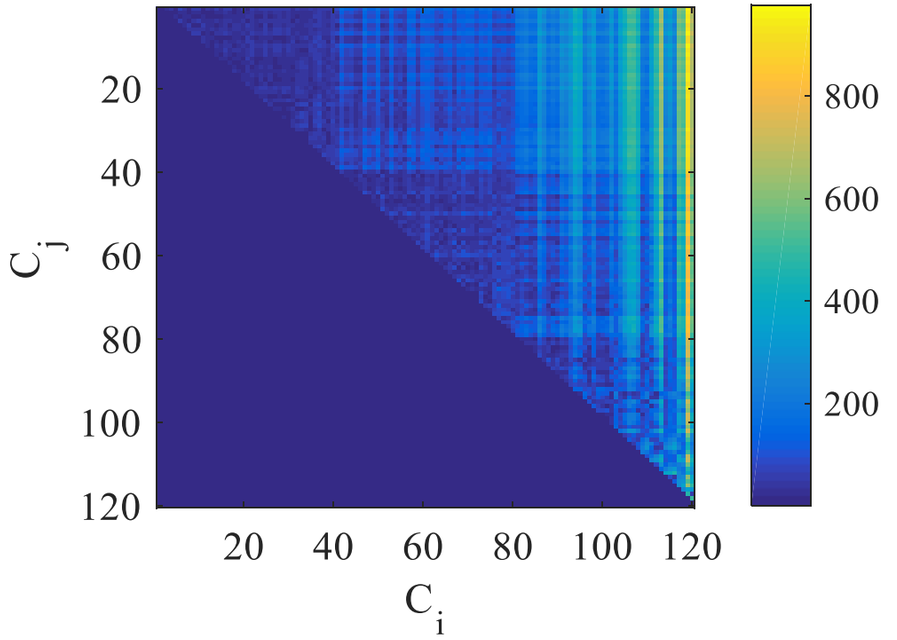
Cluster merging

K' initial clusters at the first stage is transferred into K final clusters. ($K < K'$).

Algorithm 1. Merging cluster centroids

Input: Overpopulated clustering results, cluster centroids with DBA, initial cluster number (K')

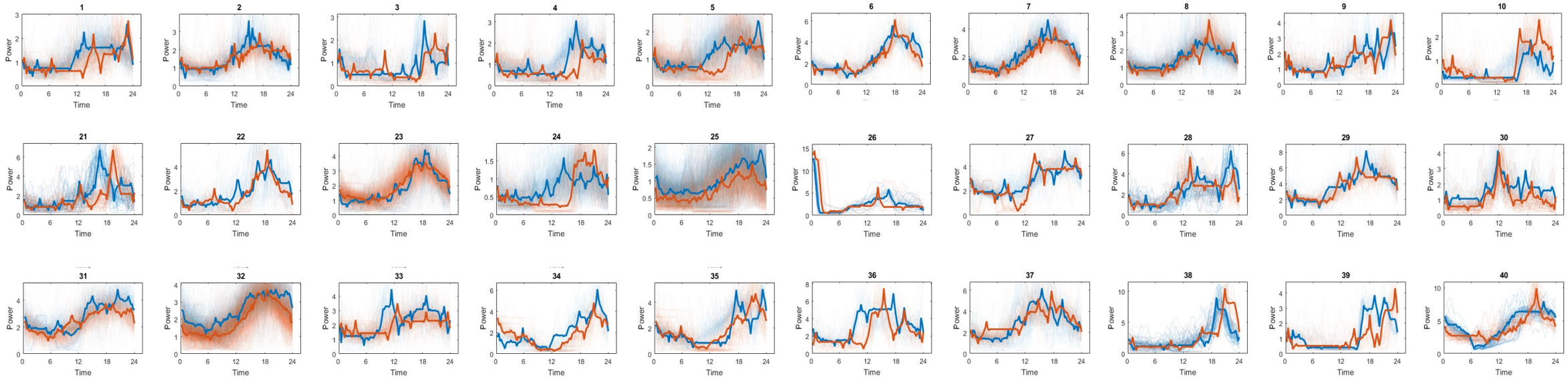
1: Set the target cluster number K.
2: While $K'>K$:
3: Find the closest cluster centroids C_i and C_j based on CI-DTW metric.
4: While $\|C_i\| + \|C_j\| > \tau * n$:
5: Find the next set of closest C_i and C_j.
6: Set $C_i = (n_i C_i + n_j C_j)/(n_i + n_j)$.
7: Delete C_j.
8: Update cluster index from cluster $j + 1$ to the last one.
9: $K' = K'$-1



Pairwise distance between all pairs of clusters (i and j) at one iteration.

Cluster merging results

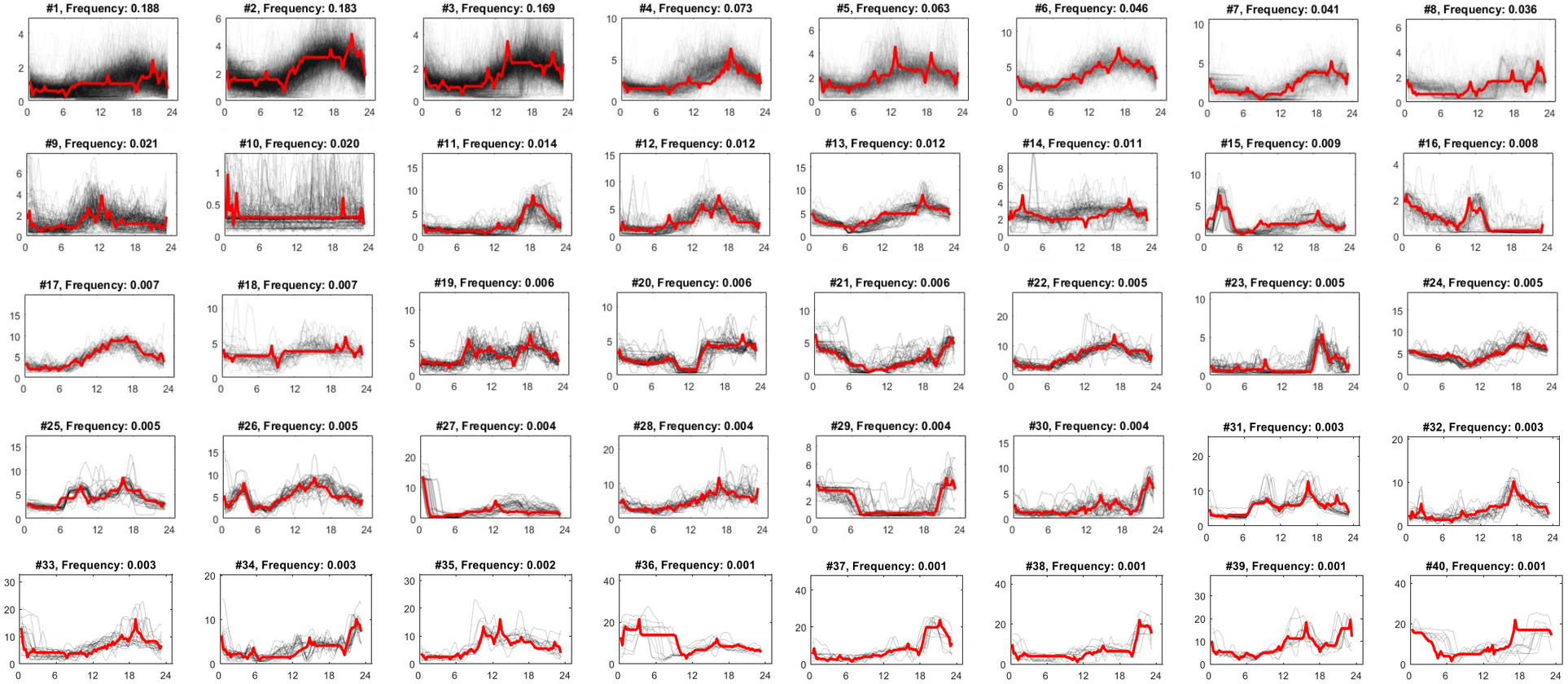
Higher iteration



Dataset1: K’=90 and K=40. First 40 iteration is shown in the plot.

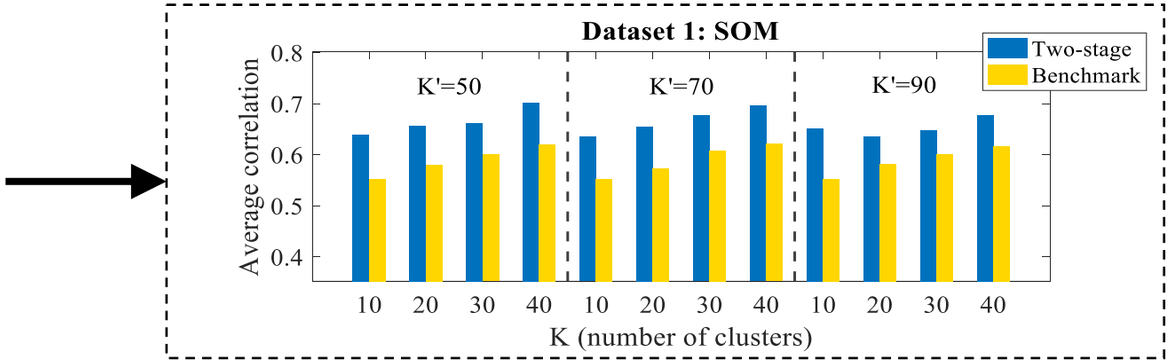
Merged clusters are subjectively close both in temporal patterns and peak magnitudes.

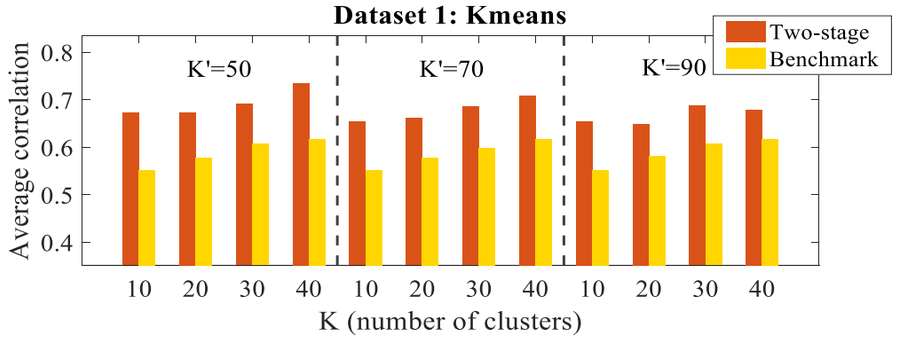
Two-stage clustering results

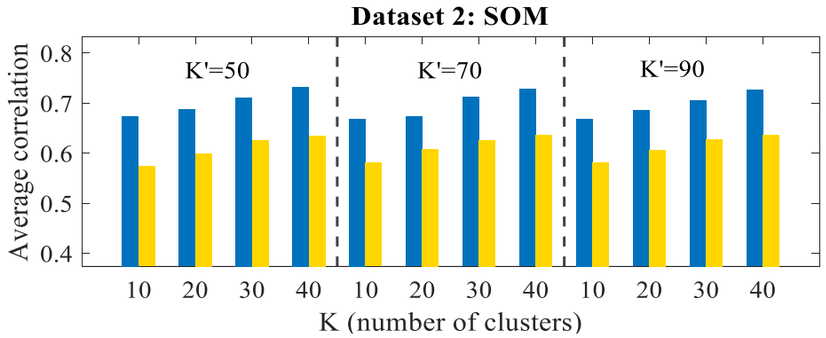


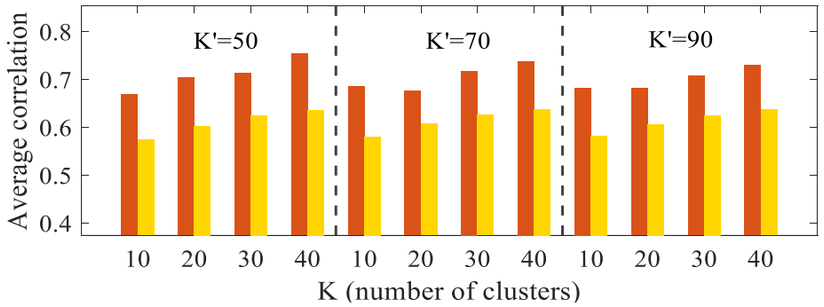
- Well-separated clusters
- Important features (peak detection, peak timing, and energy volume) are accentuated

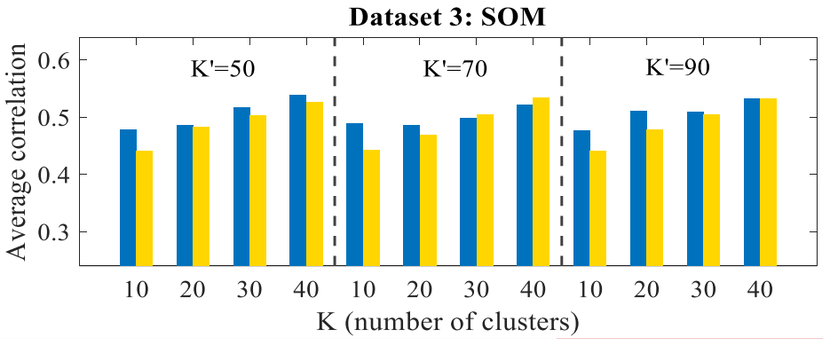
Quantified investigation *(Correlation)*: higher is better

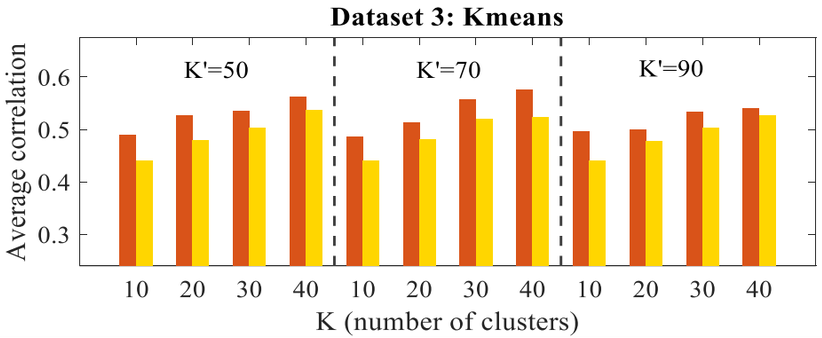












Quantified investigation *(WCSS)*: lower is better