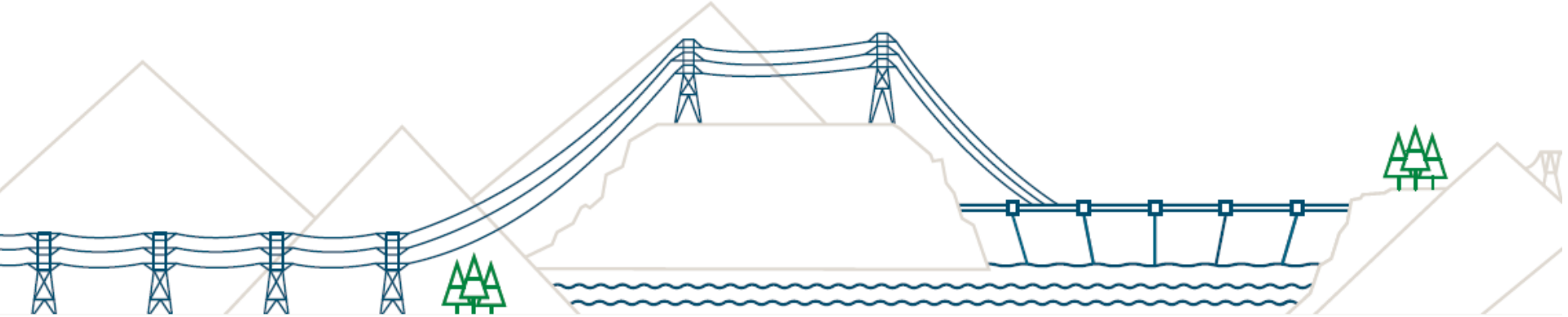
Applied Clustering Techniques

Jing Dong



Nov 31, 2016

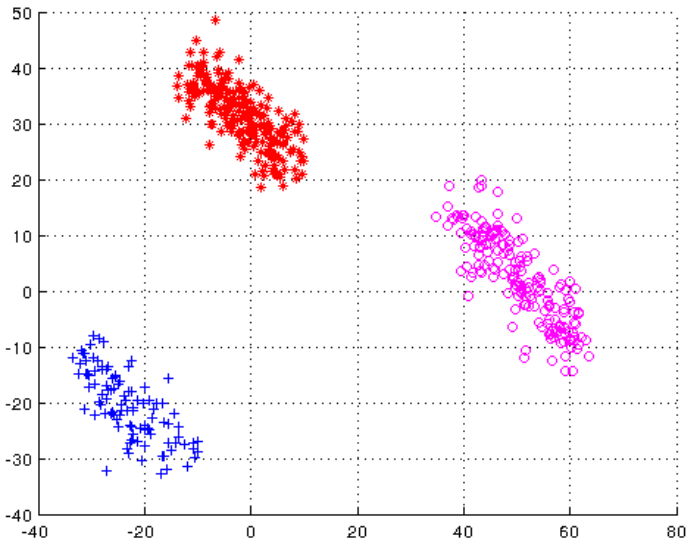
BC Hydro
Power smart

What is cluster analysis?

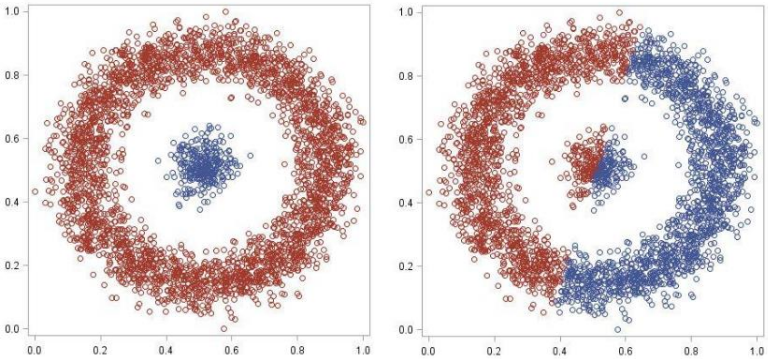
BC Hydro
Power smart

What is Cluster Analysis?

- Cluster:
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters



- Clustering is far from simple
  - Quantify similarity
  - Interpret results

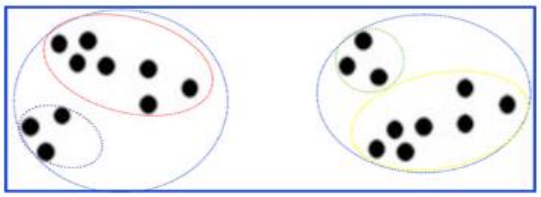


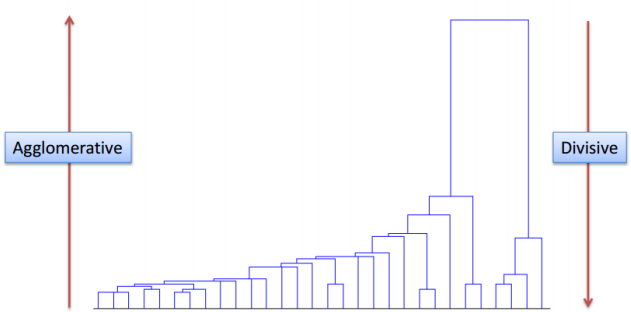
Which is more similar to a orange: a banana or an green apple?

Types of Clustering

Define the two major classes of clustering method

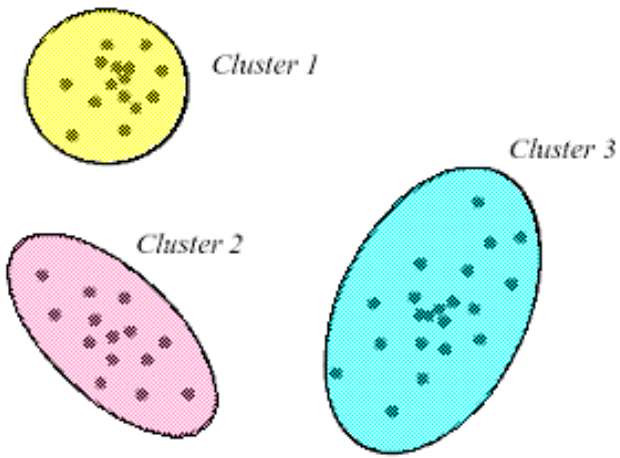
Hierarchical Clustering - Nested





Dendrogram

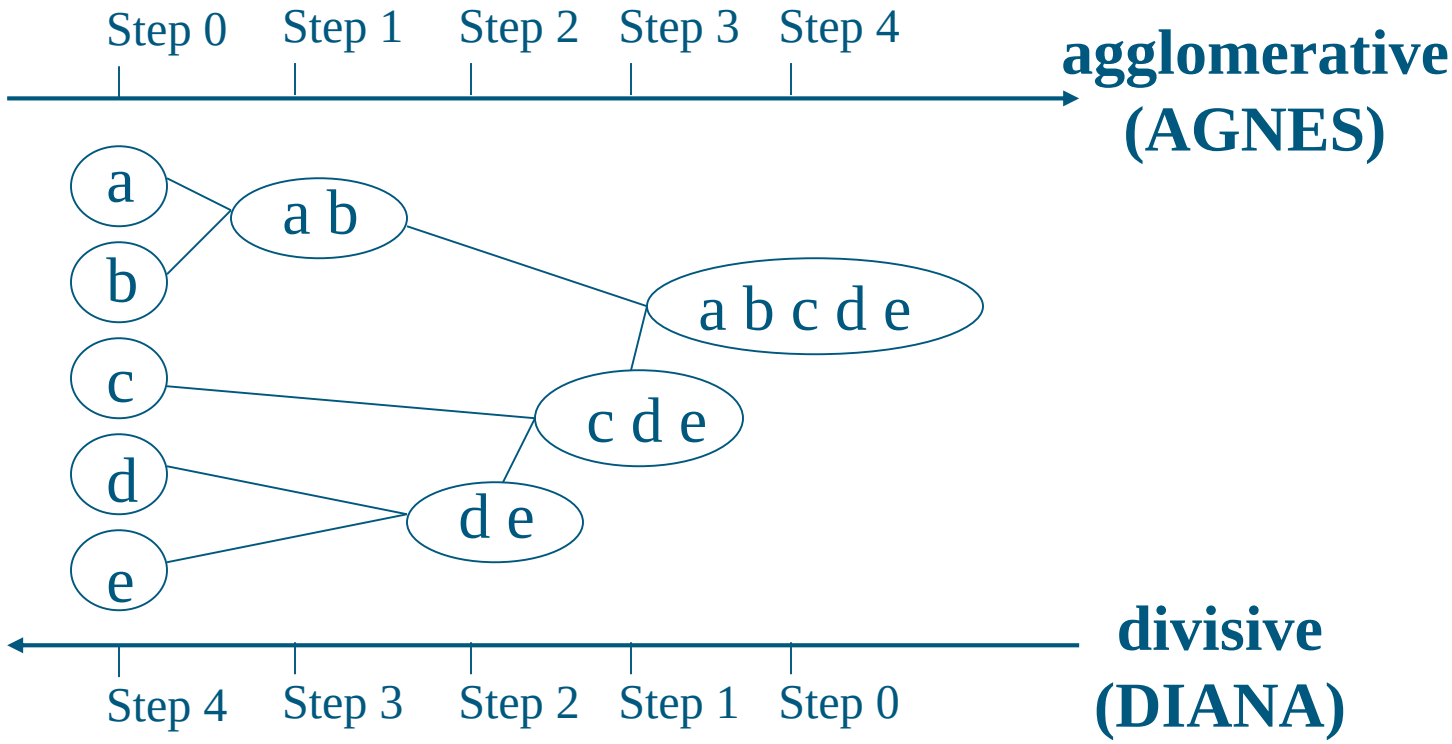
Partitive Clustering



BC Hydro
Power smart

Hierarchical Clustering

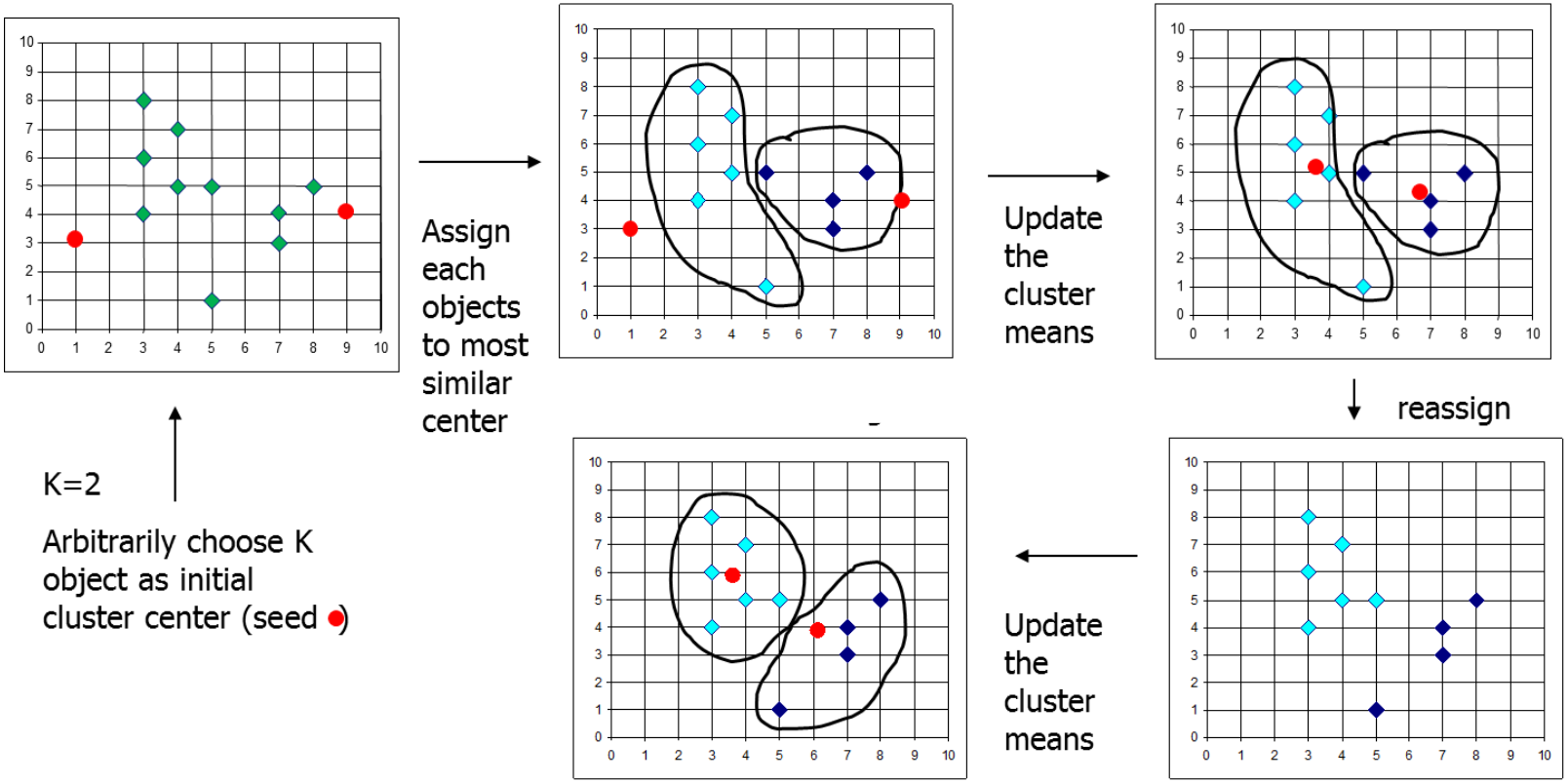
- Agglomerative: **Merge** the two clusters that are ***most similar***
- Divisive: **Partition** the observations that are ***least similar*** into two clusters



Partitive Clustering

Divide a data set into k clusters by trying to minimize some specified error functions.

k-means algorithm



https://www.youtube.com/watch?v=BVFG7fd1H30

Hierarchical vs Partitive

Hierarchical Clustering	Partitive Clustering
• Hierarchical methods do not scale up well. • Previous merges or divisions are irrevocable. • There are many hierarchical clustering methods, each defining cluster similarity in different ways and no one method is the “best”!	• Partitive methods scale up linearly with the number of observations. • **For a large dataset,** partitive methods might be the only practical choice. • Make you guess the number of clusters present • Be influenced by seed locations, outliers, and the order of the observations are read in

Application:

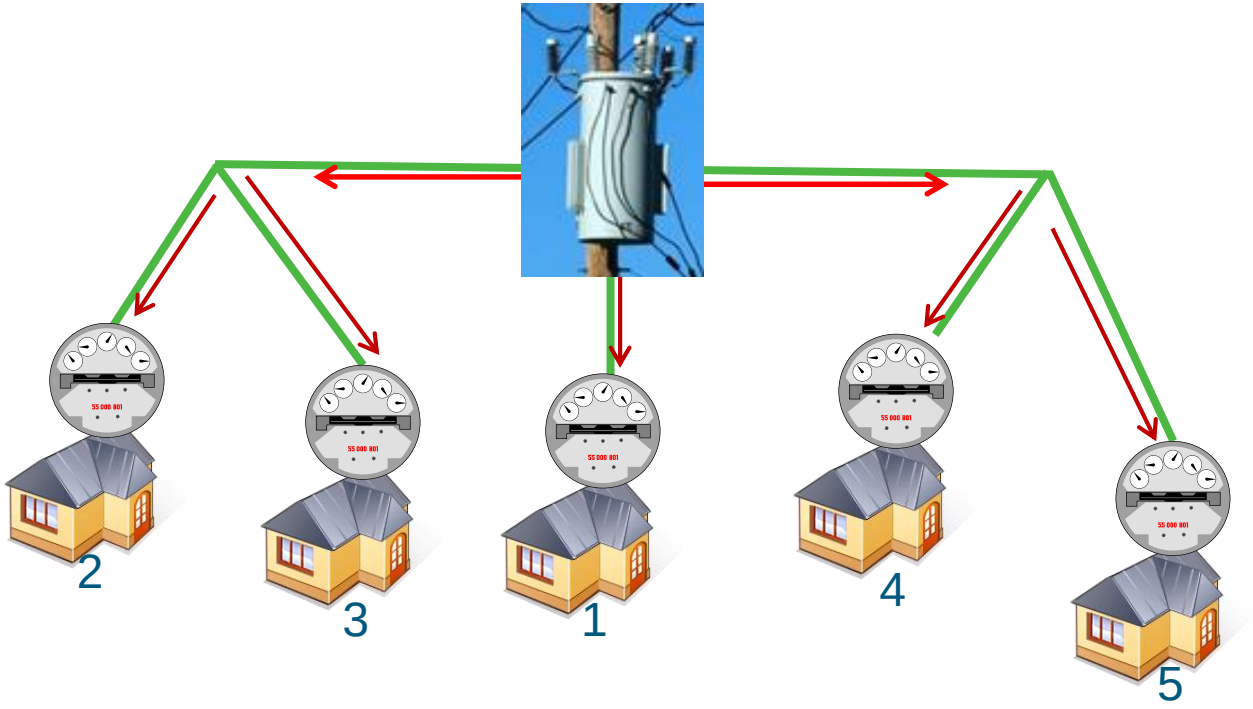
Transformer Topology Error Detection

(Hierarchical Clustering)

BC Hydro
Power smart

Topology

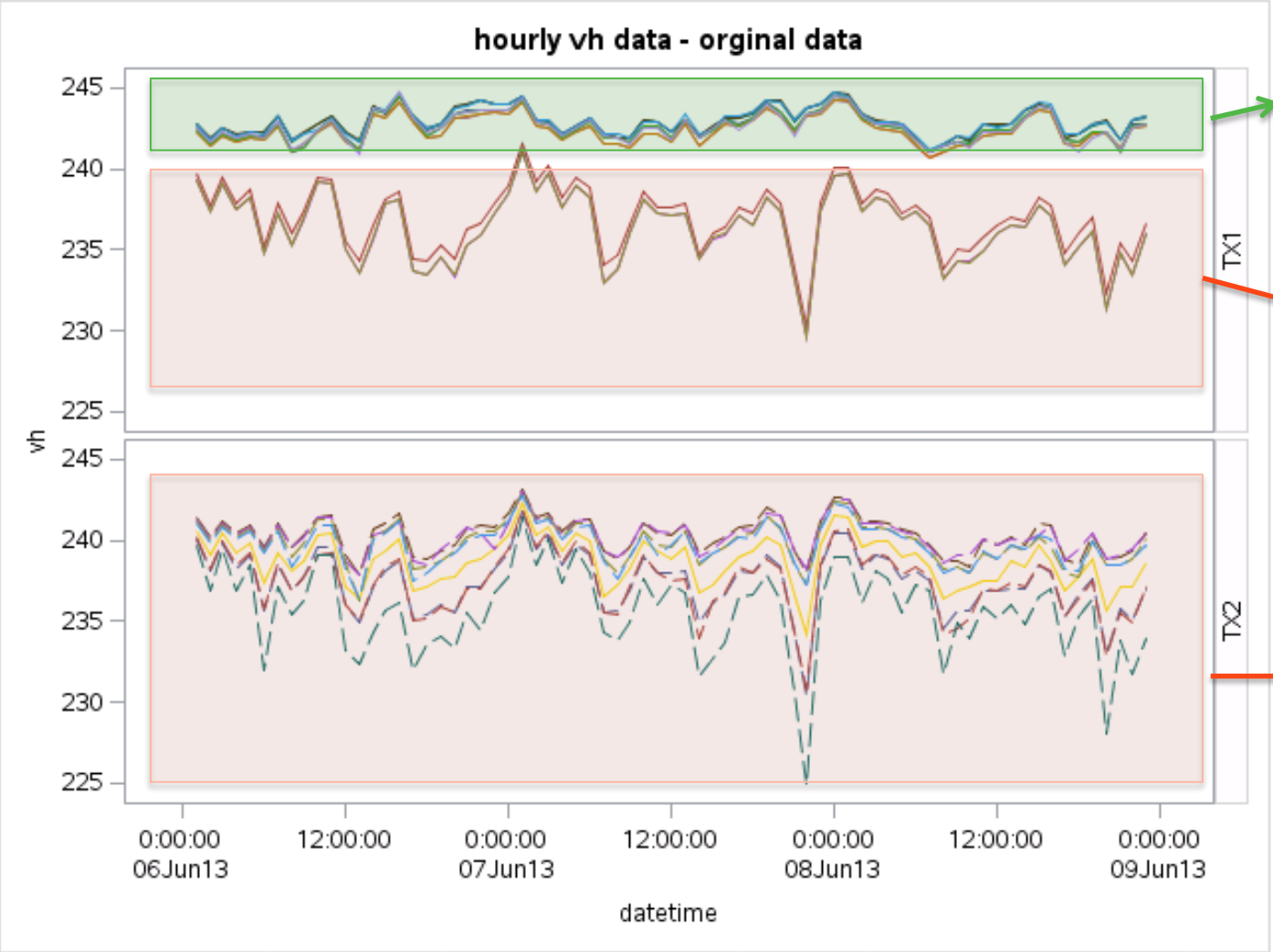
Transformer (12kV, 25kV)



Customer (110V/220V)

BC Hydro
Power smart

Customer 2, 5, 6 should belong to tx2



TX1:
1, 3, 4, 7,
8,9,10,11

TX1: 2, 5, 6

TX2: 12-19

How to group 19 time-series into two clusters?

BC Hydro
Power smart

Dataset

19 customers from 2 adjacent transformers.

Each customer has 72 hourly voltage data.

datetime	kwh	vh	transformer_ID	ID	customer_ID
06JUN13:01:00:00	1.071	242.525	TX1	01	TX1-01
06JUN13:02:00:00	0.997	241.575	TX1	01	TX1-01
06JUN13:03:00:00	1.011	242.2	TX1	01	TX1-01
06JUN13:04:00:00	1.029	241.85	TX1	01	TX1-01
06JUN13:05:00:00	0.98	242.05	TX1	01	TX1-01
06JUN13:06:00:00	0.893	242	TX1	01	TX1-01
06JUN13:07:00:00	1.433	242.85	TX1	01	TX1-01
06JUN13:08:00:00	1.018	241.05	TX1	01	TX1-01
06JUN13:09:00:00	0.663	241.3	TX1	01	TX1-01
06JUN13:10:00:00	0.629	242.425	TX1	01	TX1-01
06JUN13:11:00:00	0.683	243.05	TX1	01	TX1-01
06JUN13:12:00:00	0.712	241.9	TX1	01	TX1-01
06JUN13:13:00:00	0.819	241.15	TX1	01	TX1-01
06JUN13:14:00:00	0.797	243.675	TX1	01	TX1-01
06JUN13:15:00:00	0.623	243.475	TX1	01	TX1-01
06JUN13:16:00:00	0.797	244.5	TX1	01	TX1-01
06JUN13:17:00:00	0.74	243.125	TX1	01	TX1-01
06JUN13:18:00:00	0.981	242.025	TX1	01	TX1-01

SAS Code – Step1 – Prepare Data

```
* step1: transpose long table to wide table;
* each customer has one row of hourly vh data;
proc transpose
    data    =   tx_hourly
    out     =   tx_hourly_trans (drop = _:);
    by      customer_id;
    var     vh;
run;
```

Drop any variable start with "_":

NAME
LABEL

	customer_ID	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9
1	TX1-01	242.525	241.575	242.2	241.85	242.05	242	242.85	241.05	241.3
2	TX1-02	239.775	237.75	239.5	237.9	238.725	235.225	237.9	236	237.35
3	TX1-03	242.75	241.825	242.5	242.1	242.275	242.25	243.25	241.7	242.225
4	TX1-04	242.8	241.875	242.525	242.175	242.3	242.3	243.3	241.75	242.25
5	TX1-05	239.4	237.35	239.125	237.475	238.275	234.75	237.225	235.275	236.95
6	TX1-06	239.4	237.325	239.125	237.475	238.275	234.75	237.225	235.25	236.975

SAS Code – Step2 – Calculate Distance

```
*Step2: calculate the range standardized Euclidean distance;
proc distance
    data   =   tx_hourly_trans
    method =   euclid
    out    =   vh_distance;
    var    interval(col:/std=range);
     id        customer_ID ;
run;
```

$$D = \sqrt{\sum_{h=1}^{72} (x_h - y_h)^2}$$

	customer_ID	TX1-01	TX1-02	TX1-03	TX1-04	TX1-05	TX1-06	TX1-07
1	TX1-01	0	.	.	.	.	.	.
2	TX1-02	6.5120524385	0	.	.	.	.	.
3	TX1-03	0.3969276116	6.8394064407	0	.	.	.	.
4	TX1-04	0.4473524941	6.8941839822	0.0632807125	0	.	.	.
5	TX1-05	7.2191736607	0.7299227273	7.5475645043	7.6025151064	0	.	.
6	TX1-06	7.2118795914	0.723737384	7.5403196224	7.5952682842	0.0260746612	0	.
7	TX1-07	0.4207415292	6.8814442586	0.1537986522	0.1572260622	7.5902301443	7.5830390543	0

SAS Code – Step 3 – Hierarchical Clustering

```
*Step3: generate hierarchical clustering solution
proc cluster
    data        =   vh_distance
    outtree     =   treedata
    method      =   median;
    id      customer_ID;
run;
```

method =
Specify the clustering method
(How to define **similarities**?)

Hierarchical Method	SAS options	Distance Data Ok?	Notes
Average Linkage	average	Yes	Winner, allow to use raw(coordinate) data
Two –Stage Linkage	twostage	Some Options	Can handle irregular shape directly
Ward's Method	ward	Yes	Winner, allow to use raw(coordinate) data
Centroid Linkage	centroid	Yes	Winner, allow to use raw(coordinate) data
Complete Linkage	complete	Yes	Loser
Density Linkage	density	Some Options	Can handle irregular shape directly
EML	eml	No	Loser, allow to use raw(coordinate) data
Flexible-Beta Method	flexible	Yes	
McQuitty's Similiarity	mcauitty	Yes	
Median Linkage	median	Yes	
Single Linkage	single	Yes	Loser, Can handle irregular shape directly

Result

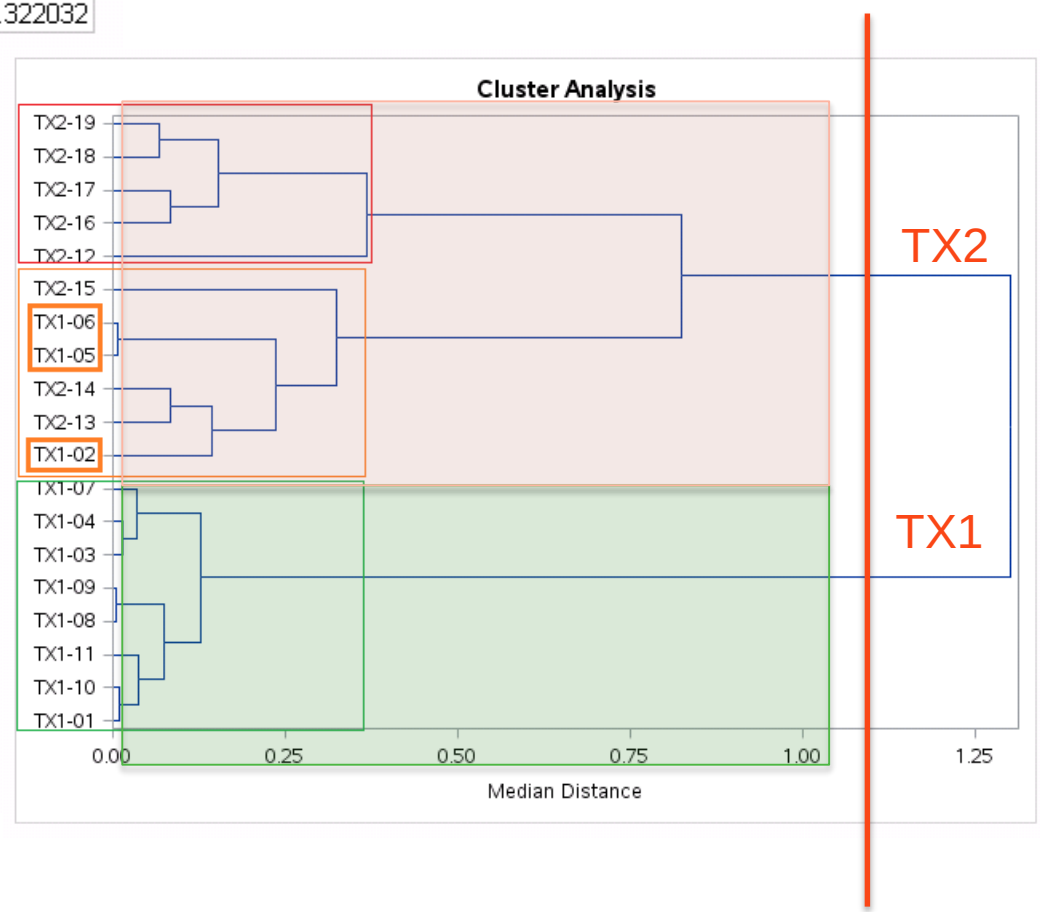
The number of lines that the vertical line crosses gives the number of clusters

The CLUSTER Procedure
Median Hierarchical Cluster Analysis

Root-Mean-Square Distance Between Observations	4.322032

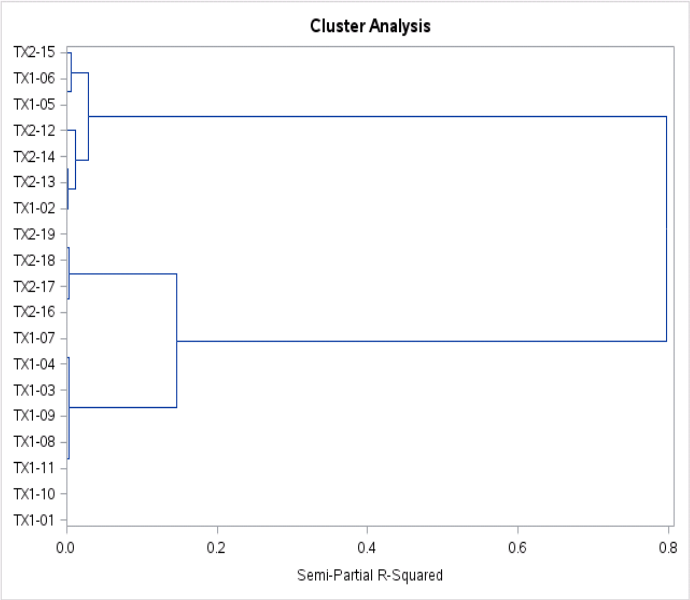
Cluster History

Number of Clusters	Clusters Joined		Freq	Norm Median Distance	Tie
18	TX1-08	TX1-09	2	0.0058	
17	TX1-05	TX1-06	2	0.006	
16	TX1-01	TX1-10	2	0.0087	
15	TX1-03	TX1-04	2	0.0146	
14	CL15	TX1-07	3	0.0352	
13	CL16	TX1-11	3	0.0377	
12	TX2-18	TX2-19	2	0.0667	
11	CL13	CL18	5	0.0744	
10	TX2-16	TX2-17	2	0.083	
9	TX2-13	TX2-14	2	0.0837	
8	CL11	CL14	8	0.1273	
7	TX1-02	CL9	3	0.1446	
6	CL10	CL12	4	0.1522	
5	CL7	CL17	5	0.2371	
4	CL5	TX2-15	6	0.3241	
3	TX2-12	CL6	5	0.3687	
2	CL4	CL3	11	0.8234	
1	CL8	CL2	19	1.2999	

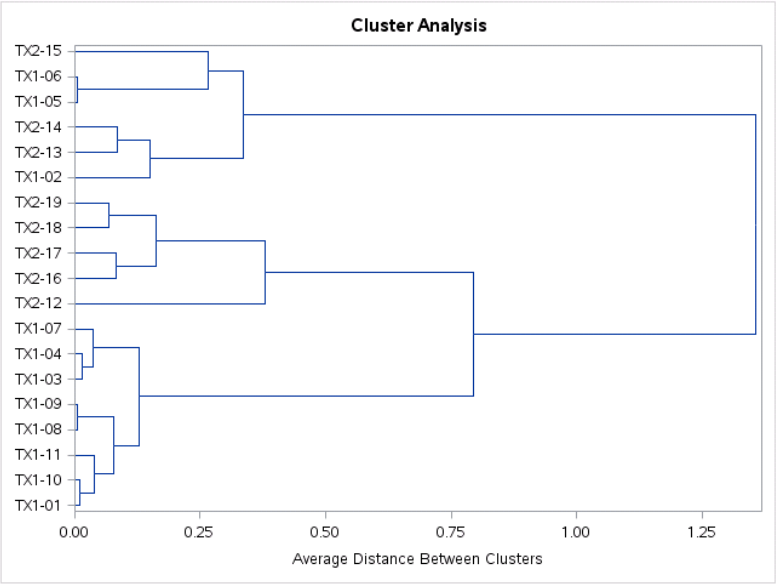


BC Hydro
Power smart

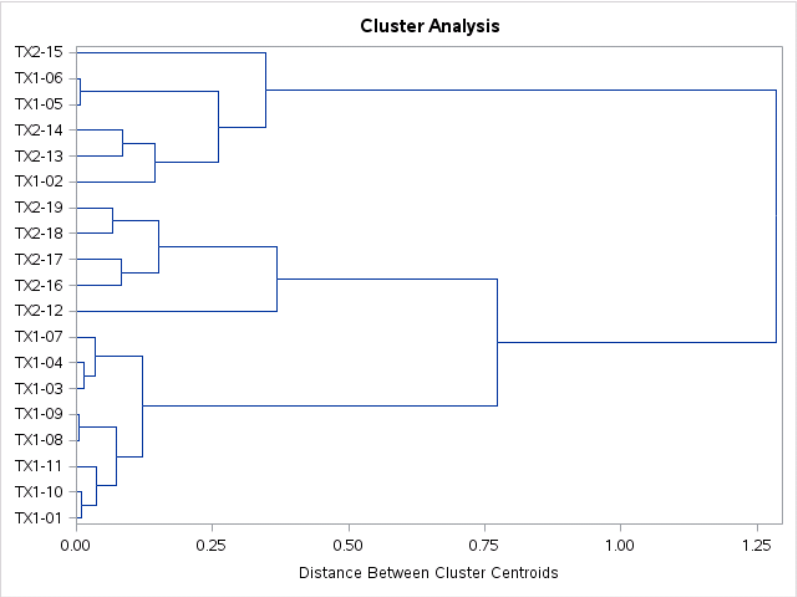
Wald's Method (wald)



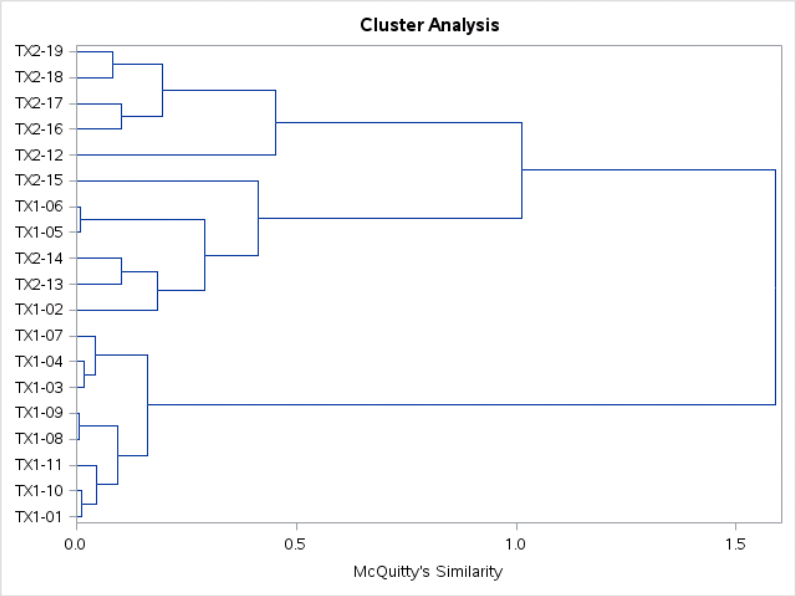
Average Linkage (average)



Centroid Linkage (centroid)



McQuitty's Similarity (mcquitty)



SAS code – Step 4 – Assign Cluster ID

```
* Step4: obtain the cluster ID for each customer;
proc tree
    data     =   treedata
    out      =   customer_cluster
    nclusters =   2;
    id       customer_ID;
quit;
```

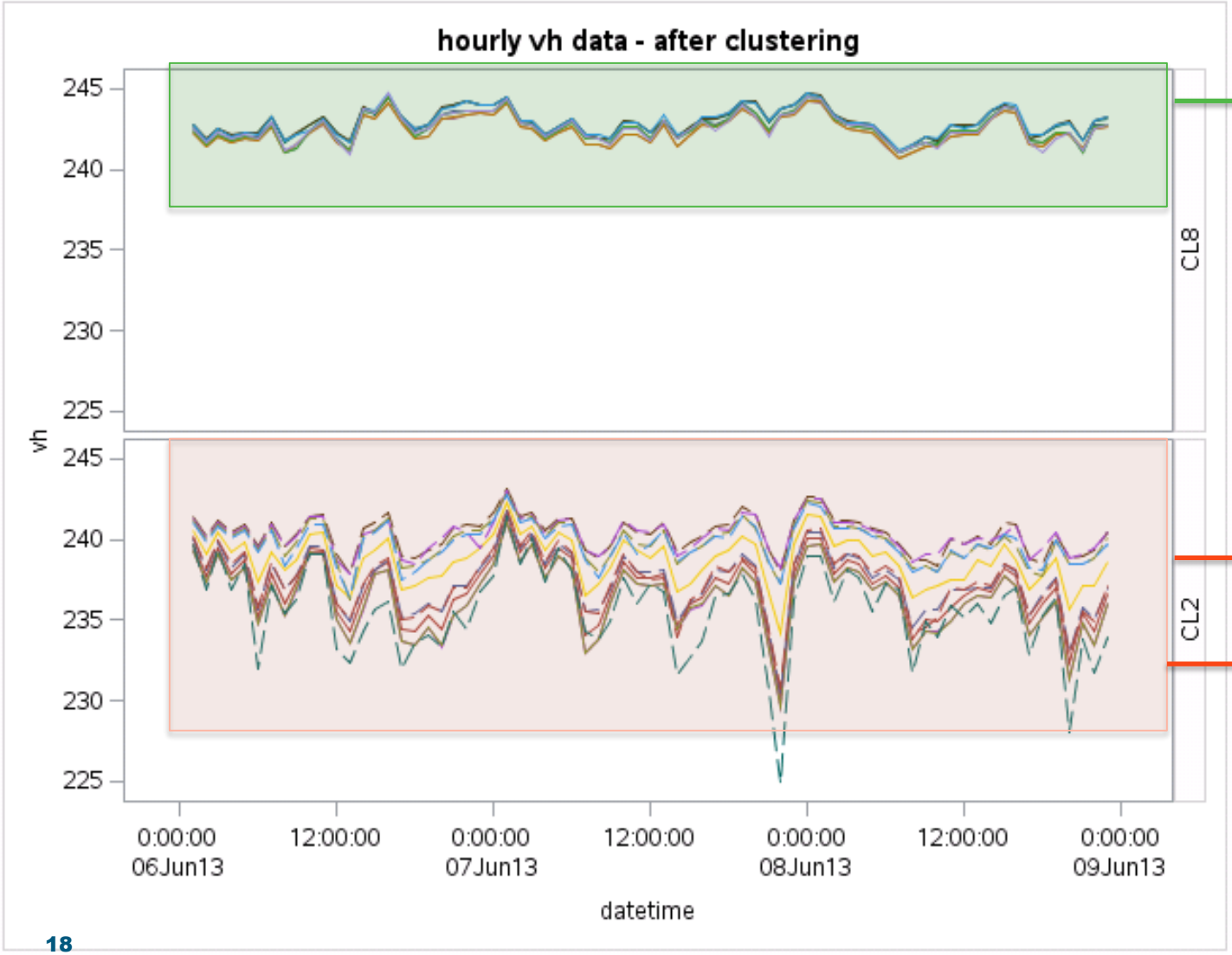
data =
Tree structure dataset was generate from previous **proc cluster**

nclusters =
Specifies the number of clusters desired in the out= dataset

	customer_ID	CLUSTER	CLUSNAME
1	TX1-01	1	CL8
2	TX1-02	2	CL2
3	TX1-03	1	CL8
4	TX1-04	1	CL8
5	TX1-05	2	CL2
6	TX1-06	2	CL2
7	TX1-07	1	CL8
8	TX1-08	1	CL8
9	TX1-09	1	CL8
10	TX1-10	1	CL8
11	TX1-11	1	CL8
12	TX2-12	2	CL2
13	TX2-13	2	CL2
14	TX2-14	2	CL2
15	TX2-15	2	CL2
16	TX2-16	2	CL2
17	TX2-17	2	CL2
18	TX2-18	2	CL2
19	TX2-19	2	CL2

BC Hydro
Power smart

Customer 2, 5, 6 are correctly assigned to tx2



TX1:
1, 3, 4, 7,
8,9,10,11

TX1: 2, 5, 6

TX2: 12-19

BC Hydro
Power smart

Application:

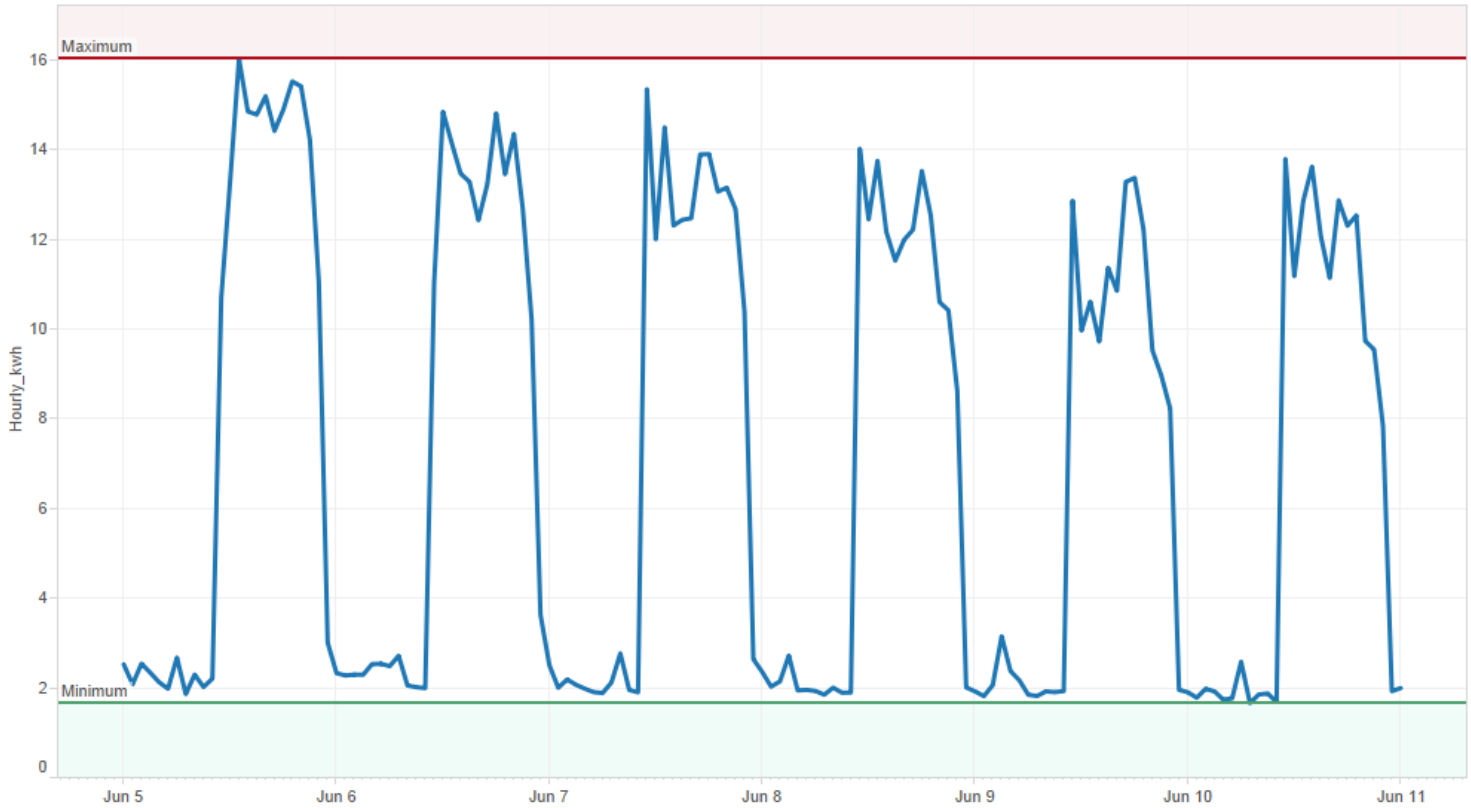
Special Load Shape Detection

(Partitive Clustering)

BC Hydro
Power smart

Special Load Shape

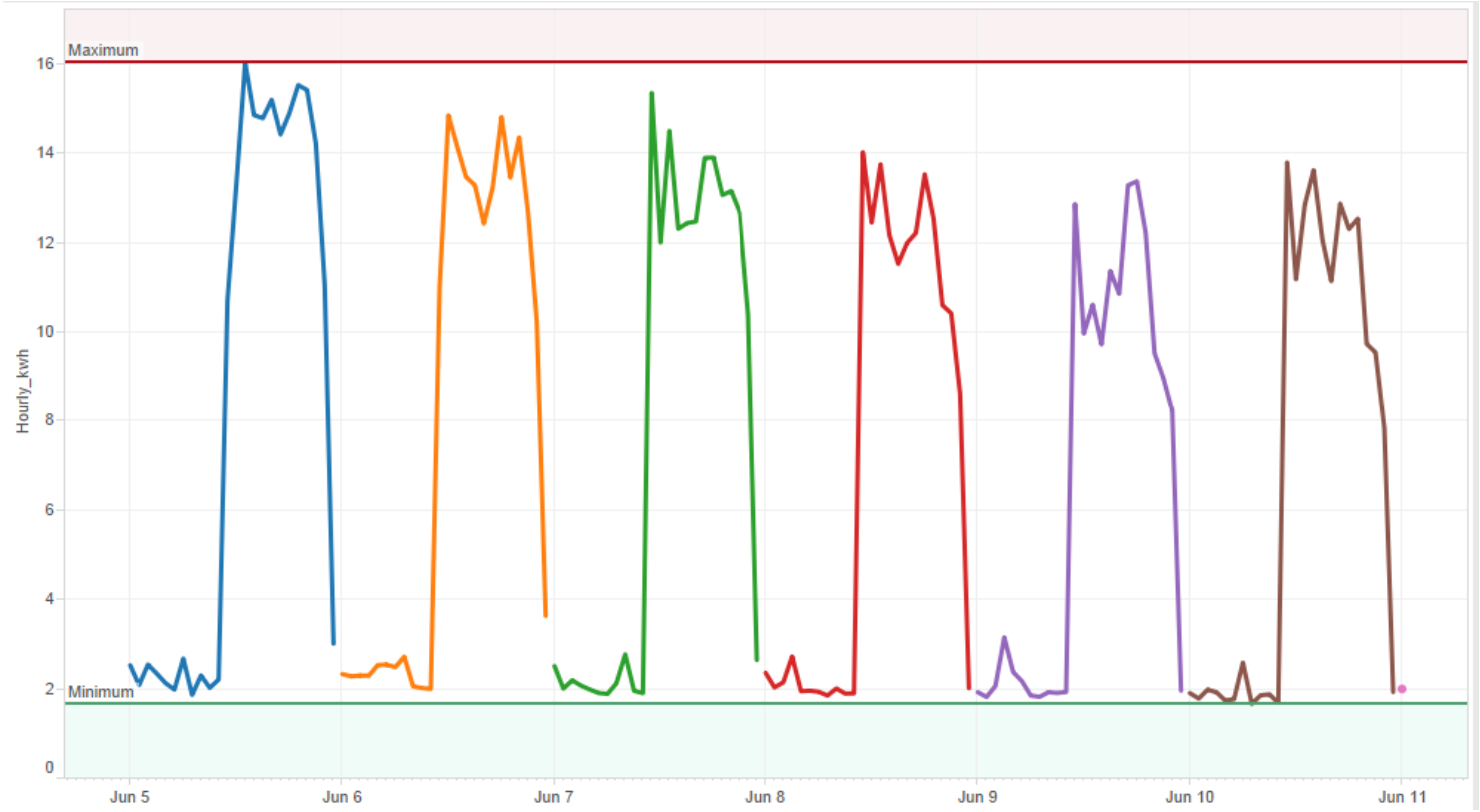
Restaurant hourly kwh



BC Hydro
Power smart

Special Load Shape

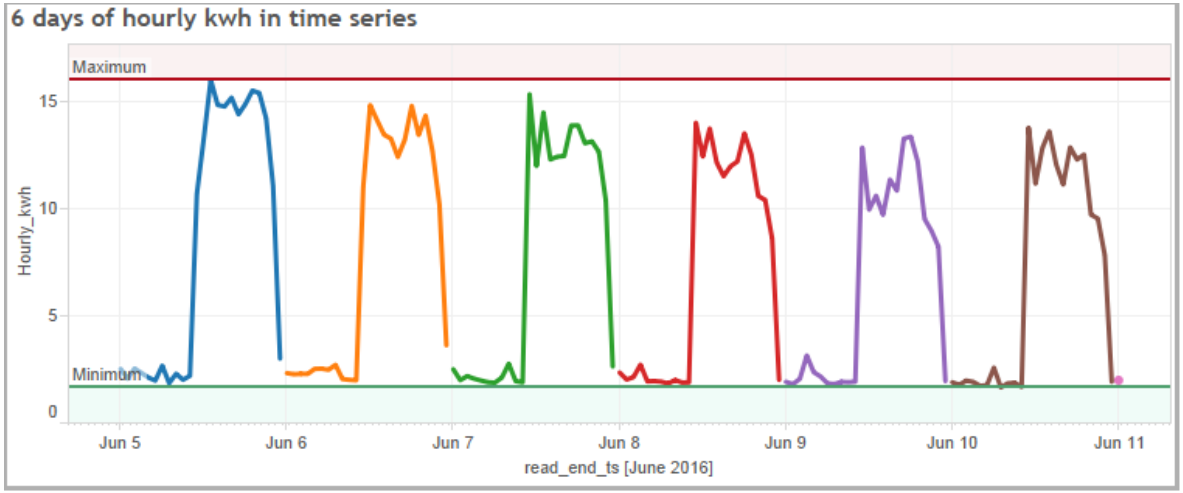
Restaurant hourly kwh – colored by day

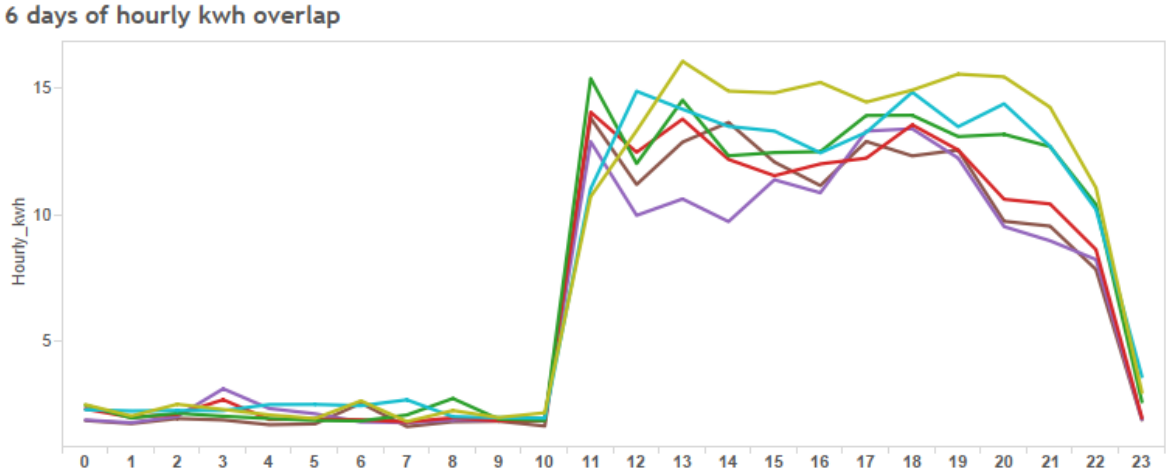


BC Hydro
Power smart

Special Load Shape

Restaurant hourly kwh – plot on 24 hours on x-axis

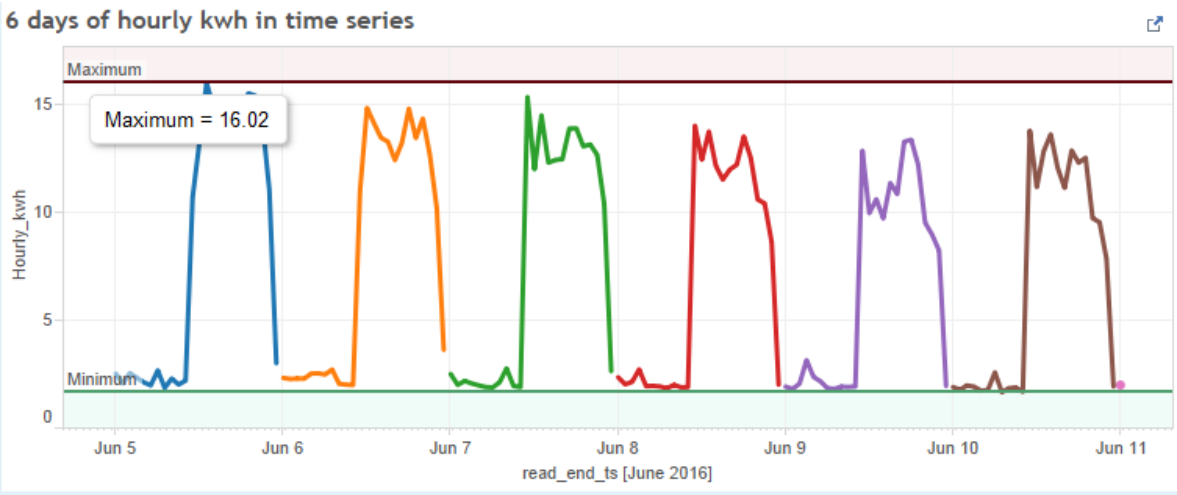




BC Hydro
Power smart

Special Load Shape

Restaurant hourly kwh– plot on 24 hours on x-axis



BC Hydro
Power smart

SAS Code – K-mean clustering

PROC FASTCLUS

```
** k-mean cluster analysis;
PROC FASTCLUS
    DATA        =   hourly_kwh
    MAXC        =   2
    MAXITER     =   10
    REPLACE     =   FULL
    out         =   cluster_matrix
    ;
    VAR         kwh_net;
    by          customer_ID;
RUN;
```

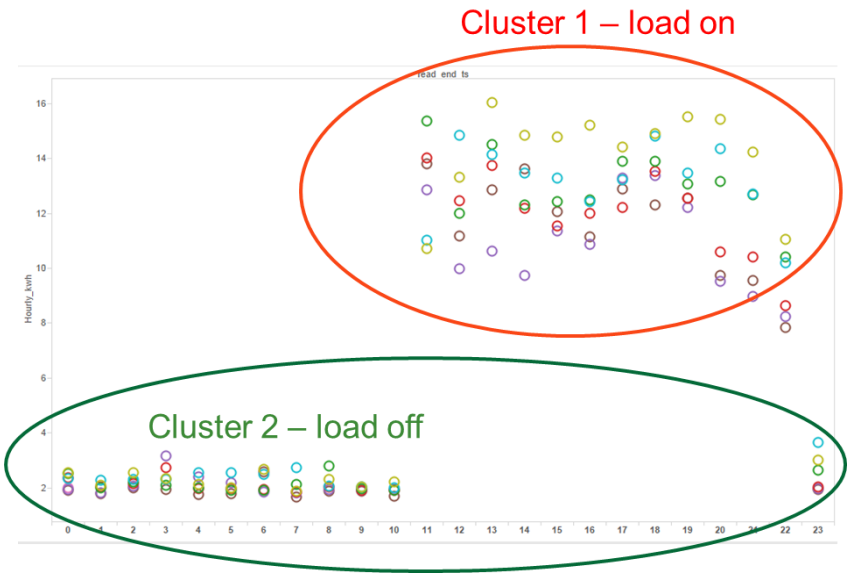
MAXC =
specifies maximum number of clusters

MAXITER =
specifies maximum number of iterations

REPLACE =
specifies seed replacement method

Out =
specifies output SAS data set containing original data and cluster assignments

K-mean results



The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=2 Maxiter=10 Converge=0.02

Initial Seeds	
Cluster	kwh_net
1	16.02100000
2	1.66100000

Minimum Distance Between Initial Seeds =	14.36

Iteration History			
		Relative Change in Cluster Seeds	
Iteration	Criterion	1	2
1	2.7908	0.2337	0.0508
2	1.3610	0.0129	0.0167

Convergence criterion is satisfied.

Criterion Based on Final Seeds =	1.3441

Cluster Means	
Cluster	kwh_net
1	12.48015278
2	2.14998630

Cluster Standard Deviations	
Cluster	kwh_net
1	1.885760366
2	0.362462086

BC Hydro
Power smart

Other Applications

- **Health:** Identifying groups of patients with similar behavioral patterns and health-related outcomes
- **Marking:** customer segmentation to develop targeted marketing
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **Utility:** Customer Behavior Analysis
- **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- **IoT:** Text Mining, Image analysis, Web cluster engines