Cluster: Silhouette and C-index

Haiwei Zhou

2023-03-22

Prepare Dataset

Monthly CNG Usage data is divided by the number of meters and the area of the building. The dataset we use covers data from 2017 till 2022. The data is log-transformed after adding a small positive constant: the 1% quantile of positive values in the dataset which is 3.5×10^{-5}.
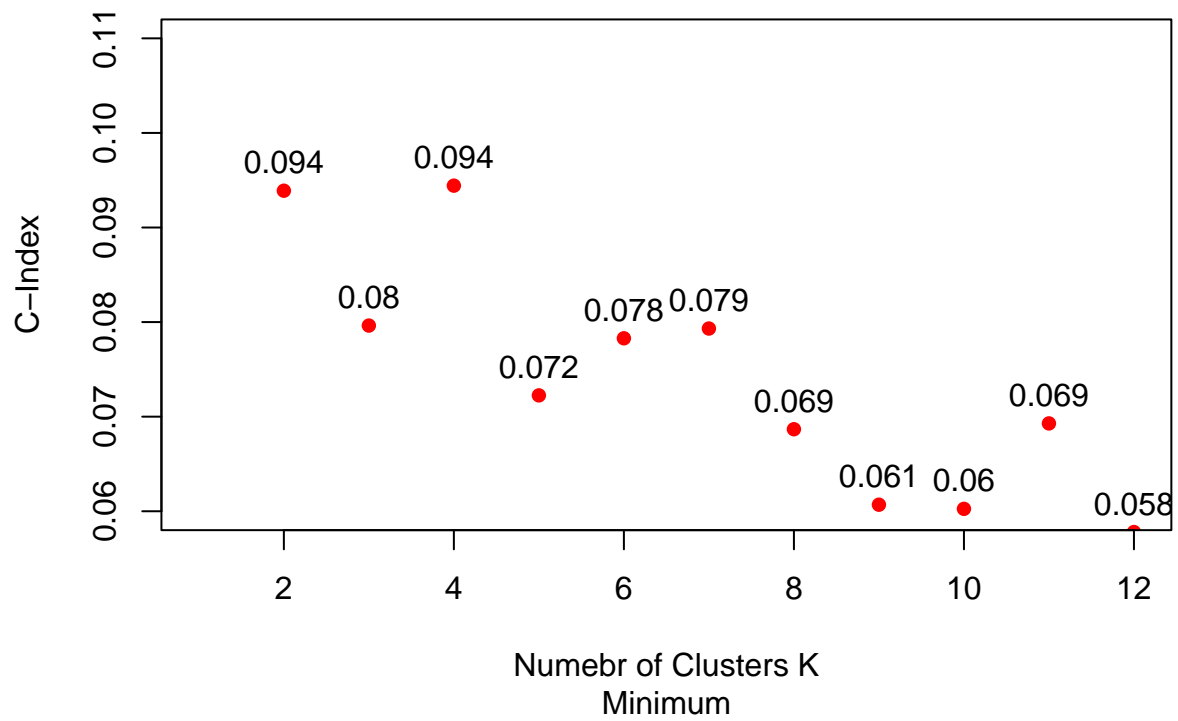
Clustering Using Kmeans Method, with C-index and Silhouette Index

The number of clusters are selected by silhouette index (from -1 to 1 and the closer to 1 the better) and C-index (The smaller the better). There are some missing values in the result of silhouette index because if some cluster including only ONE account exists, the calculation of silhouette index would return NaN. (Divided by 0 in the procedure)

Silouette Index of each cluster number

0.445

0.304

0.258 0.253 0.246

0.214 0.213

Silouette Index

Numebr of Clusters K
Maximum
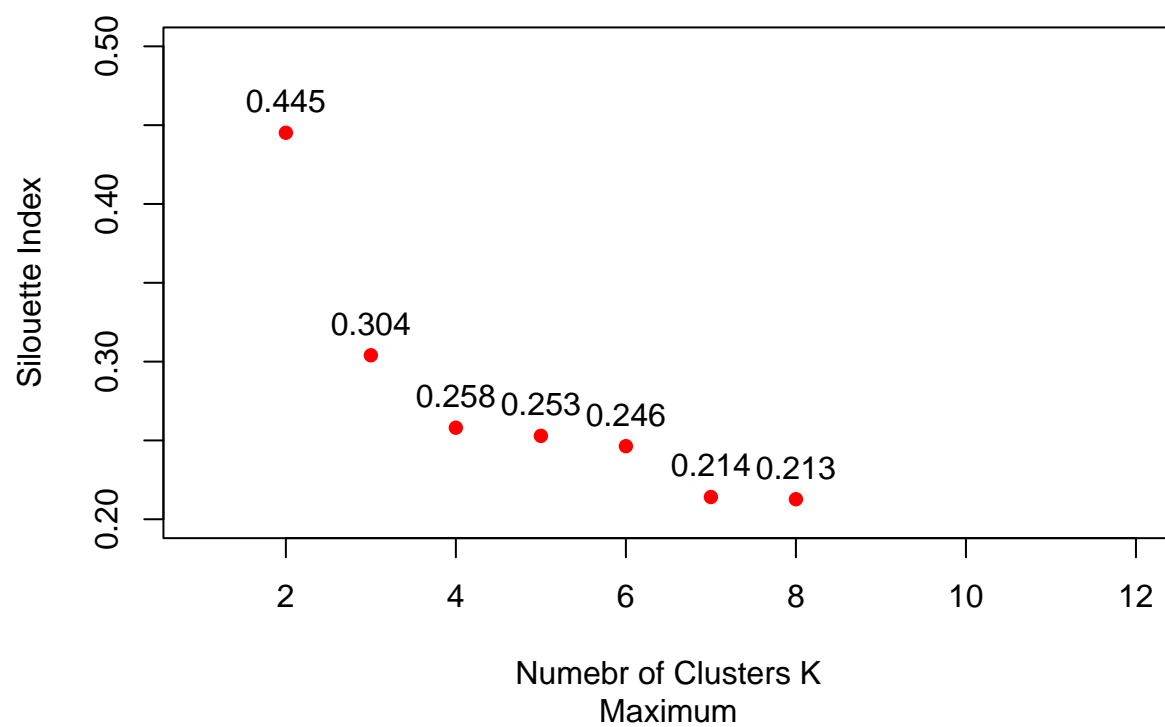
C−Index of each cluster number

Dunn Index of each cluster number

Numebr of Clusters K
Maximum
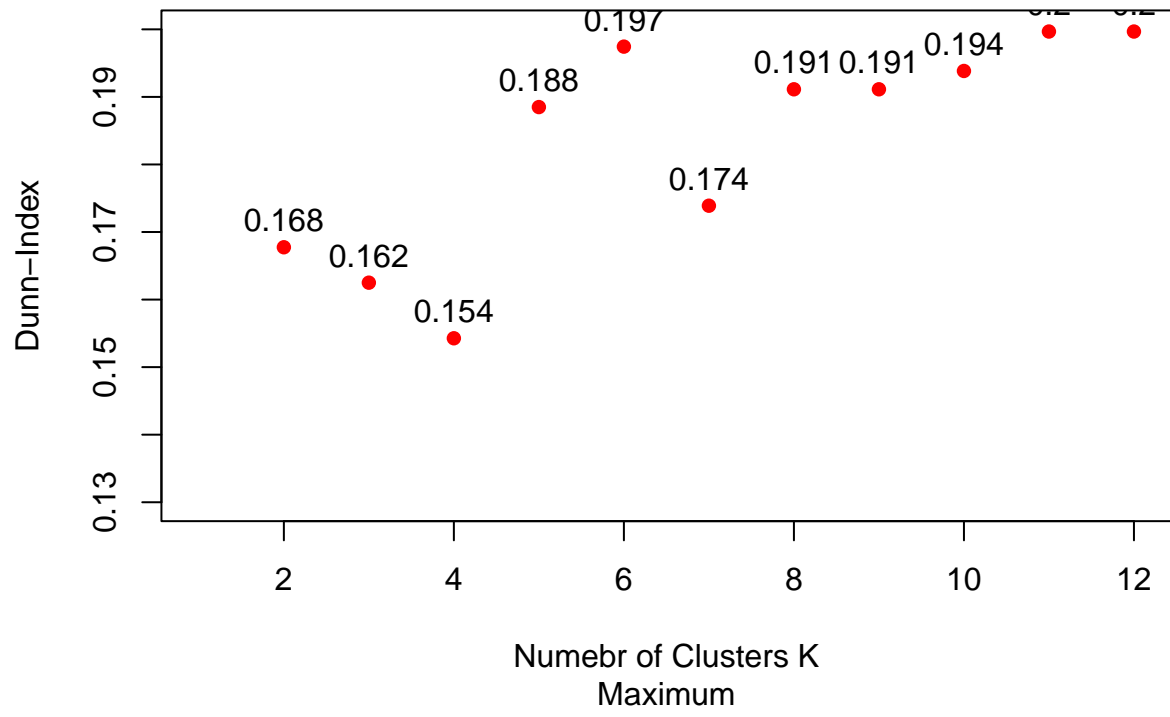
McClain Rao Index of each cluster number

0.571

0.555

0.554

0.56   0.56

0.533

0.52

0.513

0.506

0.525

0.482

McClain Rao Index

Numebr of Clusters K
Minimum
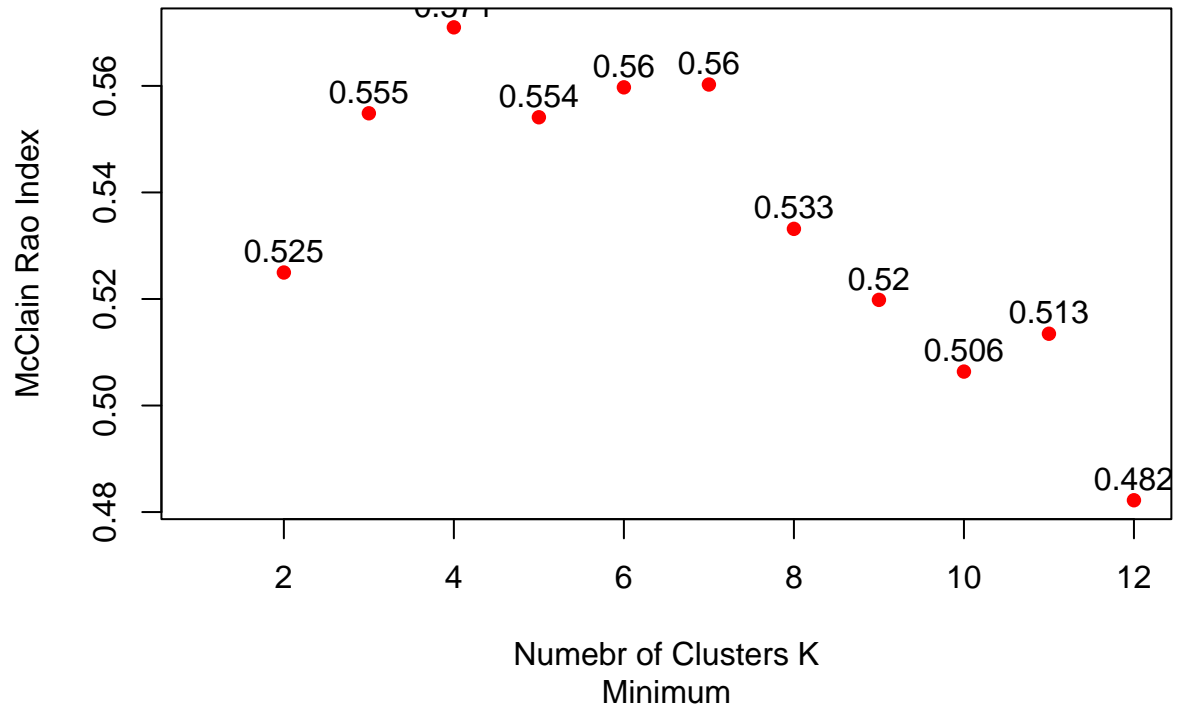
PBM Index of each cluster number

SD Index of each cluster number

Number of Clusters K
Minimum

Plot of 2 clusters

Cluster 1 of 2 — Average Euclidean Distance is 17.95

Cluster 2 of 2 — Average Euclidean Distance is 18.927
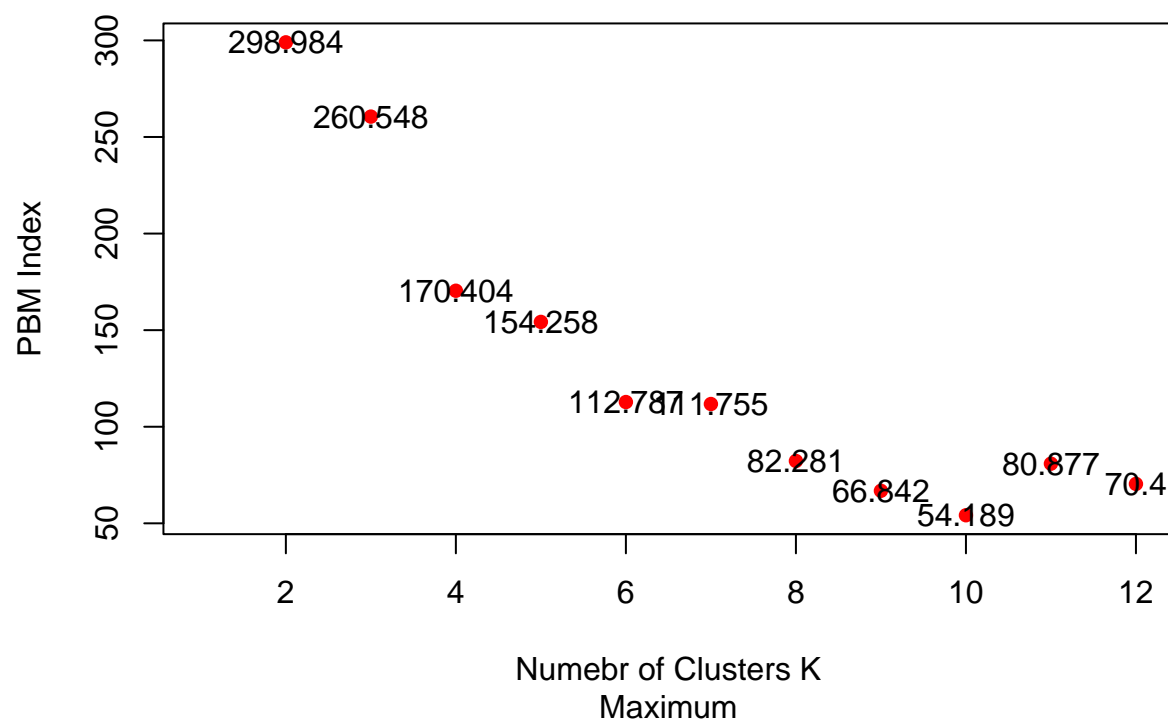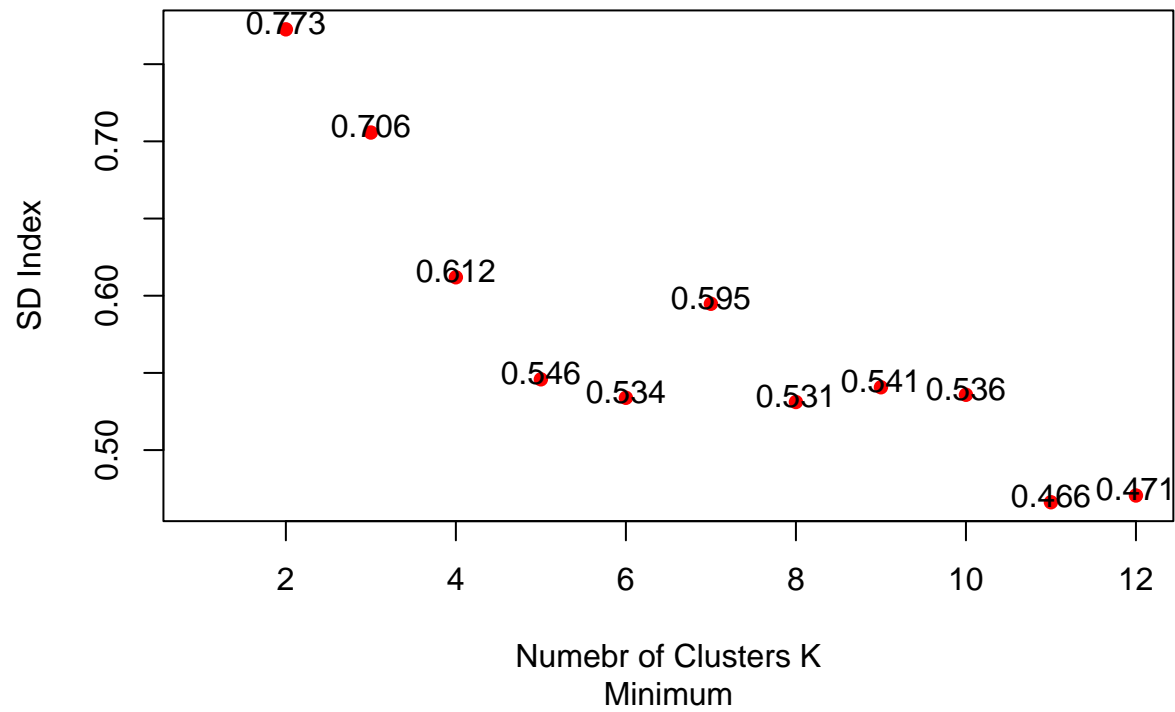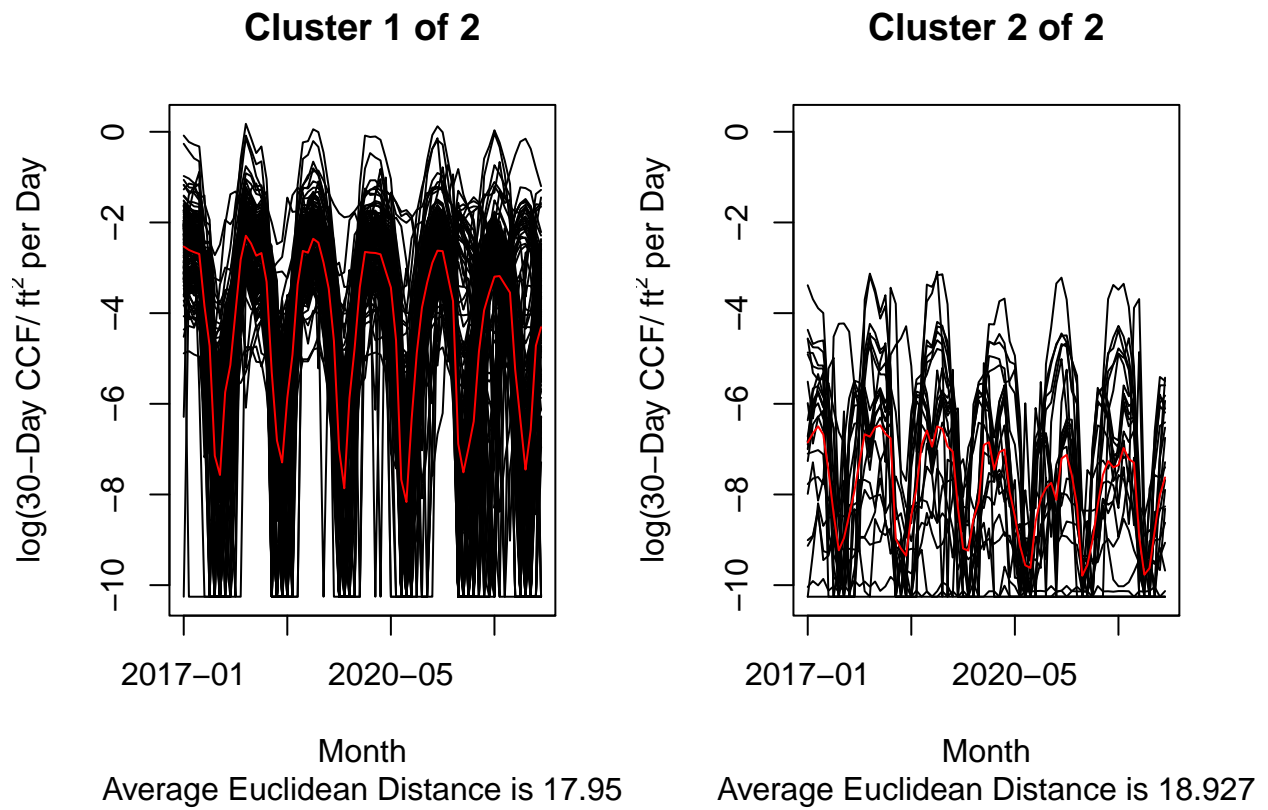
```
## [1] "The Silhouetteindex is 0.445. The number of cluster is 2"
```

```
## [1] "The C-index is 0.094. The number of cluster is 2"
```

Plot of 3 clusters

[1] "The Silhouetteindex is 0.304. The number of cluster is 3"

[1] "The C-index is 0.08. The number of cluster is 3"

Cluster 1 of 3

log(30−Day CCF/ ft² per Day)

Month
Average Euclidean Distance is 14.97

Cluster 2 of 3

log(30−Day CCF/ ft² per Day)

Month
Average Euclidean Distance is 18.927

Cluster 3 of 3

log(30−Day CCF/ ft² per Day)

Month
Average Euclidean Distance is 18.3

9

Plot of 4 clusters

Cluster 1 of 4

log(30−Day CCF/ ft² per Day)

Month
Average Euclidean Distance is 14.998

Cluster 2 of 4

log(30−Day CCF/ ft² per Day)

Month
Average Euclidean Distance is 18.927

Cluster 3 of 4

log(30−Day CCF/ ft² per Day)

Month
Average Euclidean Distance is 16.161

Cluster 4 of 4

log(30−Day CCF/ ft² per Day)
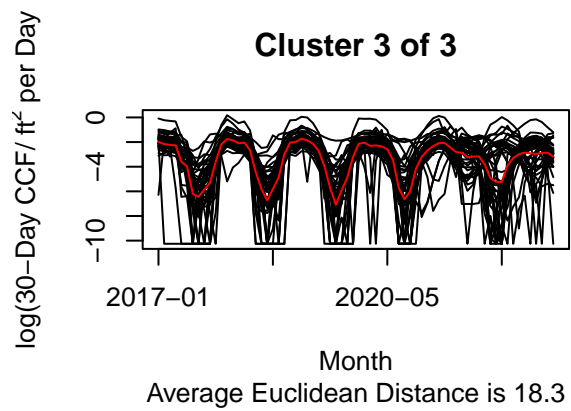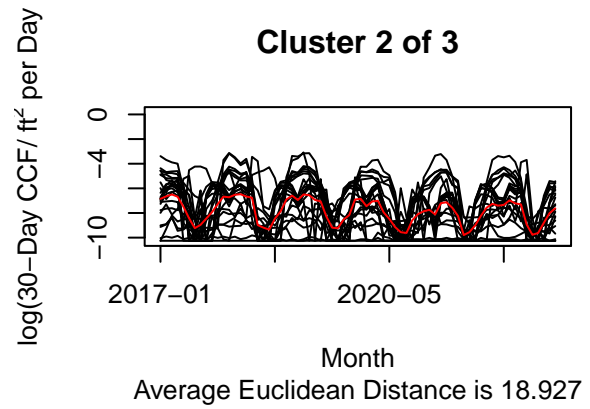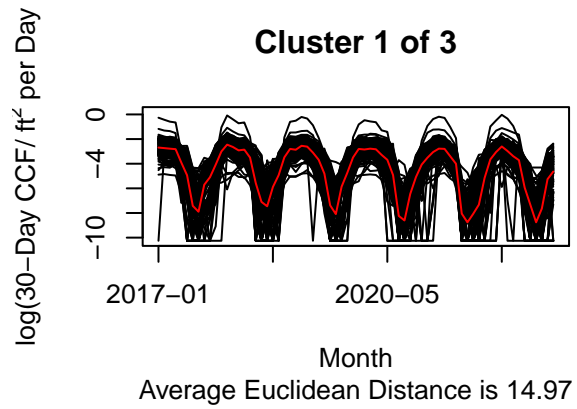
Month
Average Euclidean Distance is 14.089

```
## [1] "The Silhouetteindex is 0.258. The number of cluster is 4"
```

```
## [1] "The C-index is 0.094. The number of cluster is 4"
```

Plot of 5 clusters

[1] "The Silhouetteindex is 0.253. The number of cluster is 5"

[1] "The C-index is 0.072. The number of cluster is 5"

Cluster 1 of 5

Month
Average Euclidean Distance is 13.983

Cluster 2 of 5

Month
Average Euclidean Distance is 13.118

Cluster 3 of 5

Month
Average Euclidean Distance is 16.867

Cluster 4 of 5

Month
Average Euclidean Distance is 14.577

Cluster 5 of 5

Month
Average Euclidean Distance is 17.711

Plot of 6 clusters

Cluster 1 of 6

log(30−Day CCF/ ft² per Day)

2017−01   2020−05

Month
Average Euclidean Distance is 14.013

Cluster 2 of 6

log(30−Day CCF/ ft² per Day)

2017−01   2020−05

Month
Average Euclidean Distance is 13.118

Cluster 3 of 6

log(30−Day CCF/ ft² per Day)

2017−01   2020−05

Month
Average Euclidean Distance is 13.668

Cluster 4 of 6

log(30−Day CCF/ ft² per Day)

2017−01   2020−05

Month
Average Euclidean Distance is 12.286

Cluster 5 of 6

log(30−Day CCF/ ft² per Day)

2017−01   2020−05

Month
Average Euclidean Distance is 17.711

Cluster 6 of 6

log(30−Day CCF/ ft² per Day)

2017−01   2020−05

Month
Average Euclidean Distance is 17.206

```
## [1] "The Silhouetteindex is 0.246. The number of cluster is 6"
```

```
## [1] "The C-index is 0.078. The number of cluster is 6"
```

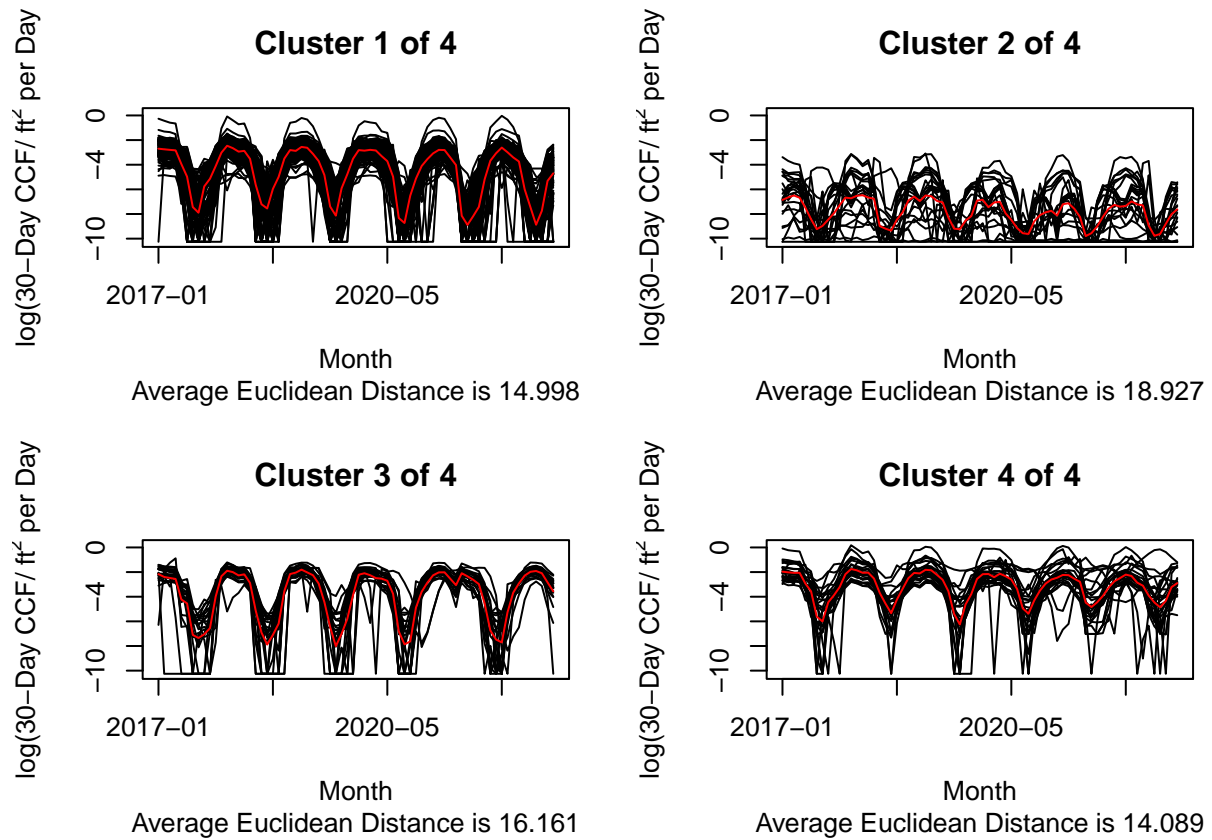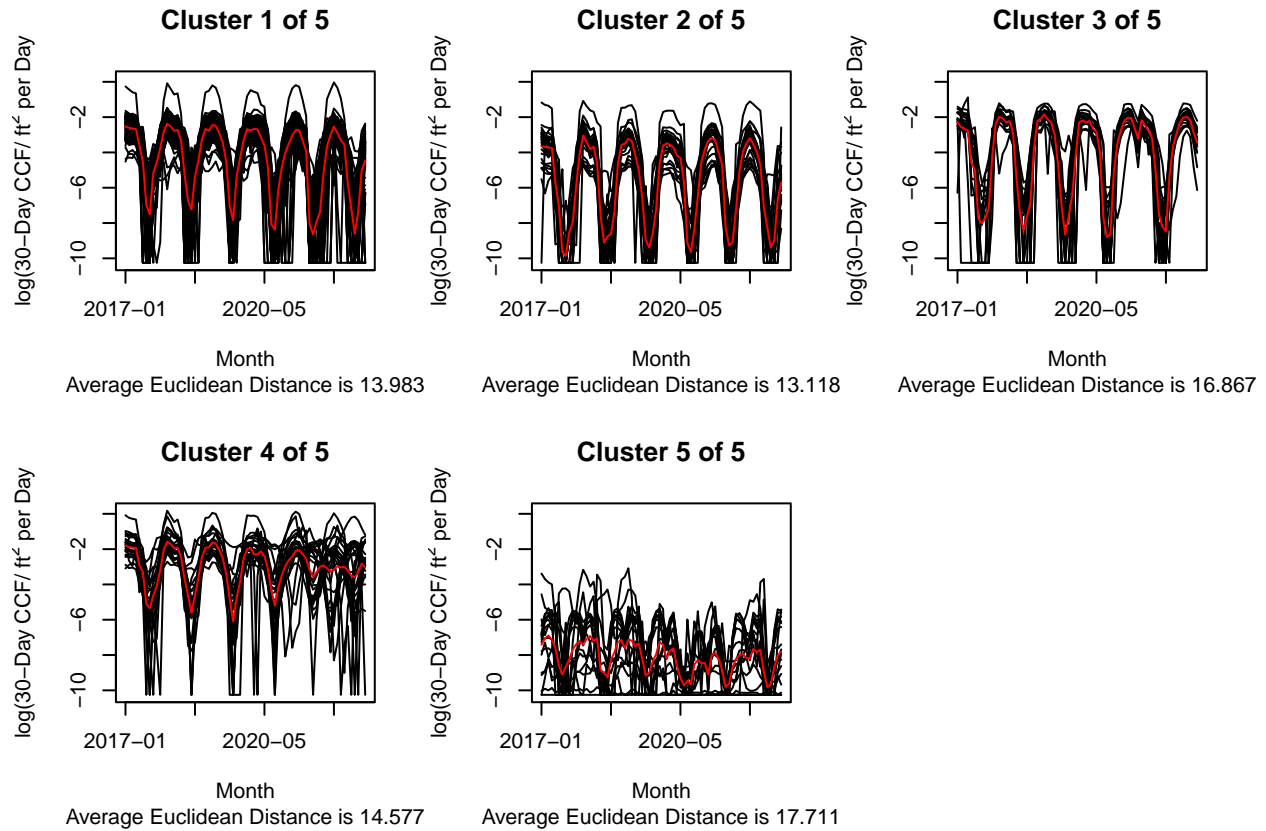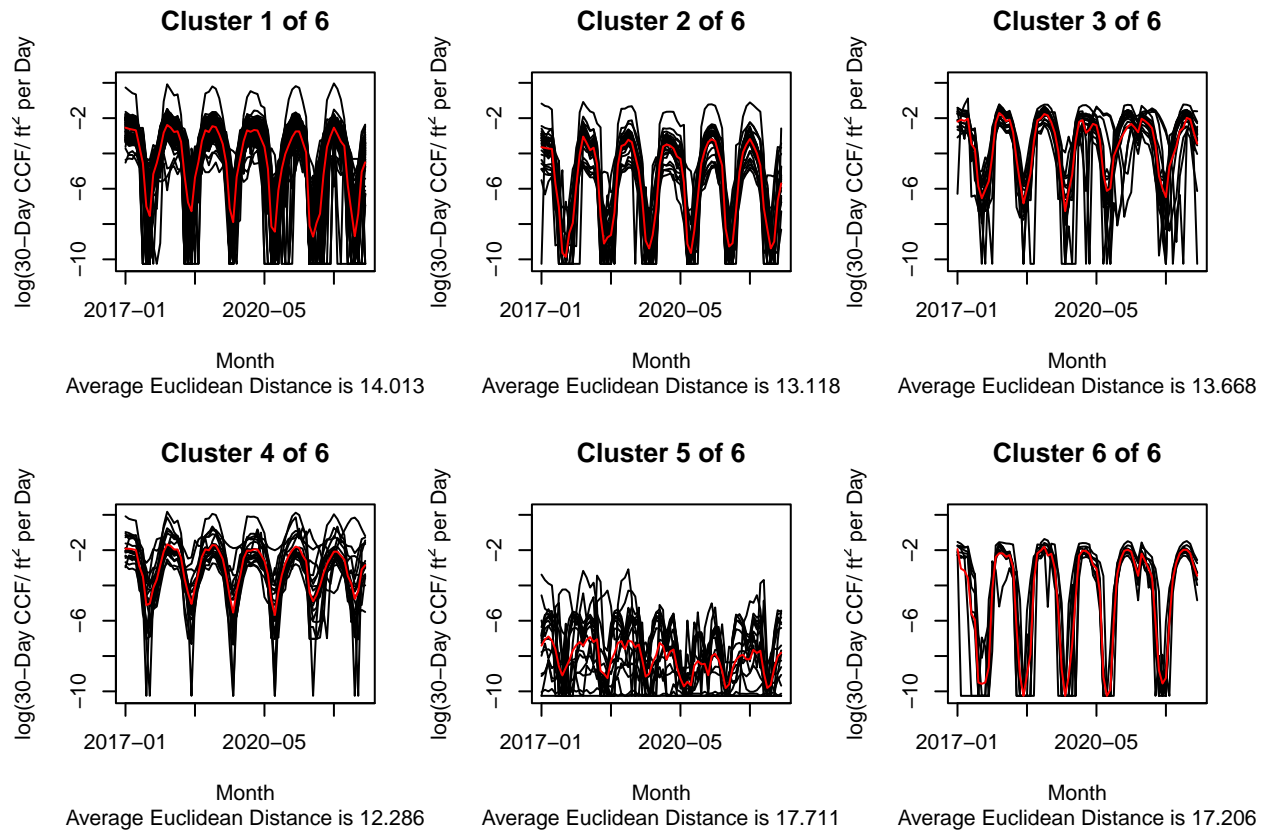# Plot of 7 clusters

[1] "The Silhouetteindex is 0.214. The number of cluster is 7"

[1] "The C-index is 0.079. The number of cluster is 7"

Cluster 1 of 7

Average Euclidean Distance is 13.887

Cluster 2 of 7

Average Euclidean Distance is 13.118

Cluster 3 of 7

Average Euclidean Distance is 12.198

Cluster 4 of 7

Average Euclidean Distance is 11.855

Cluster 5 of 7

Average Euclidean Distance is 17.711

Cluster 6 of 7

Average Euclidean Distance is 18.027

Cluster 7 of 7

Average Euclidean Distance is 20.089

Plot of 8 clusters

[1] "The Silhouetteindex is 0.213. The number of cluster is 8"

[1] "The C-index is 0.069. The number of cluster is 8"

Cluster 1 of 8 — Average Euclidean Distance is 9.936
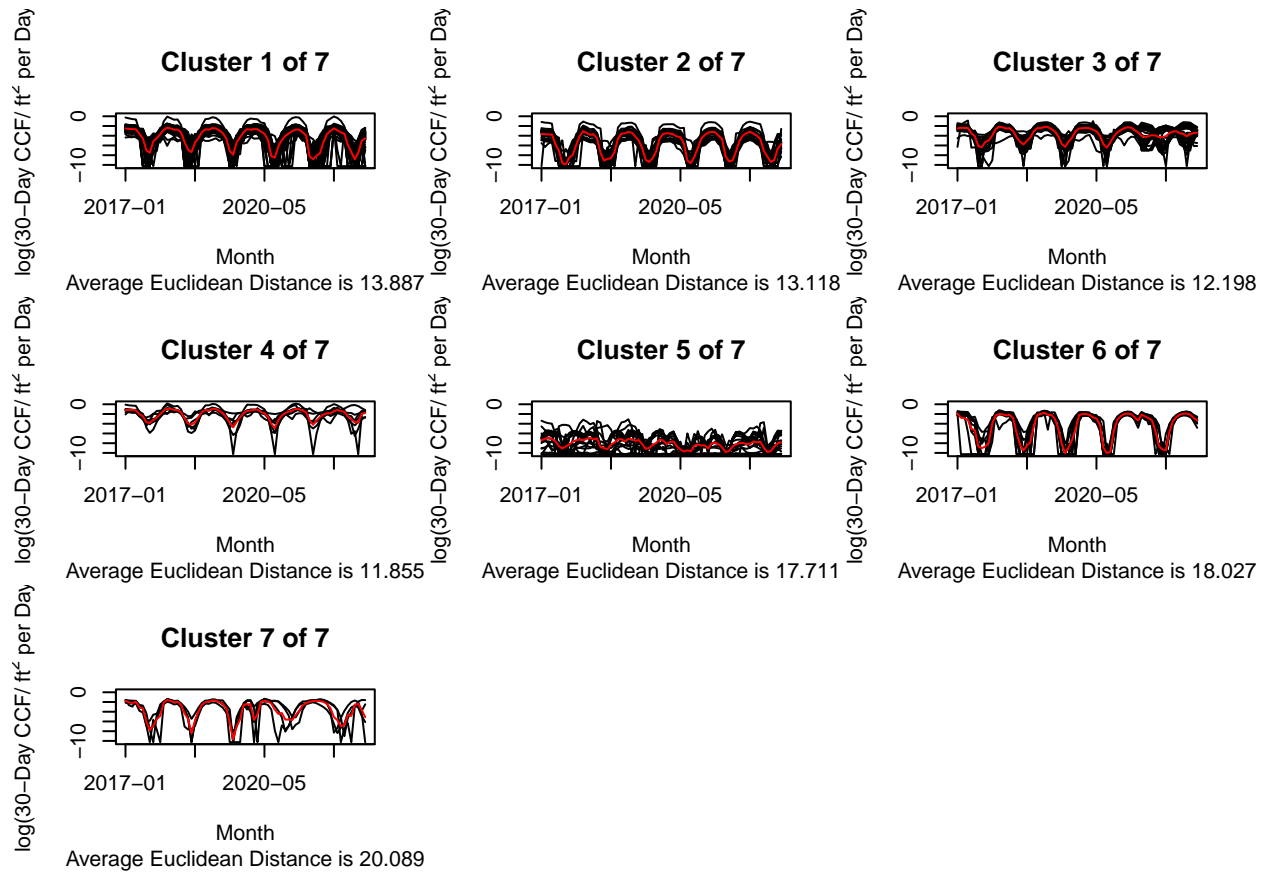
Cluster 2 of 8 — Average Euclidean Distance is 15.07

Cluster 3 of 8 — Average Euclidean Distance is 9.393

Cluster 4 of 8 — Average Euclidean Distance is 12.353

Cluster 5 of 8 — Average Euclidean Distance is 17.711

Cluster 6 of 8 — Average Euclidean Distance is 17.206

Cluster 7 of 8 — Average Euclidean Distance is 20.089

Cluster 8 of 8 — Average Euclidean Distance is 12.81

Plot of 9 clusters

Cluster 1 of 9

Month

Average Euclidean Distance is 9.936

Cluster 2 of 9

Month

Average Euclidean Distance is 15.076

Cluster 3 of 9

Month

Average Euclidean Distance is 9.393

Cluster 4 of 9

Month

Average Euclidean Distance is 12.353

Cluster 5 of 9

Month

Average Euclidean Distance is 17.711

Cluster 6 of 9

Month

Average Euclidean Distance is 17.206

Cluster 7 of 9

Month

Average Euclidean Distance is 20.089

Cluster 8 of 9

Month

Average Euclidean Distance is 14.492

Cluster 9 of 9

Month

Average Euclidean Distance is 12.81

```
## [1] "The Silhouetteindex is 0.206. The number of cluster is 9"
```

```
## [1] "The C-index is 0.061. The number of cluster is 9"
```