Schneider Electric

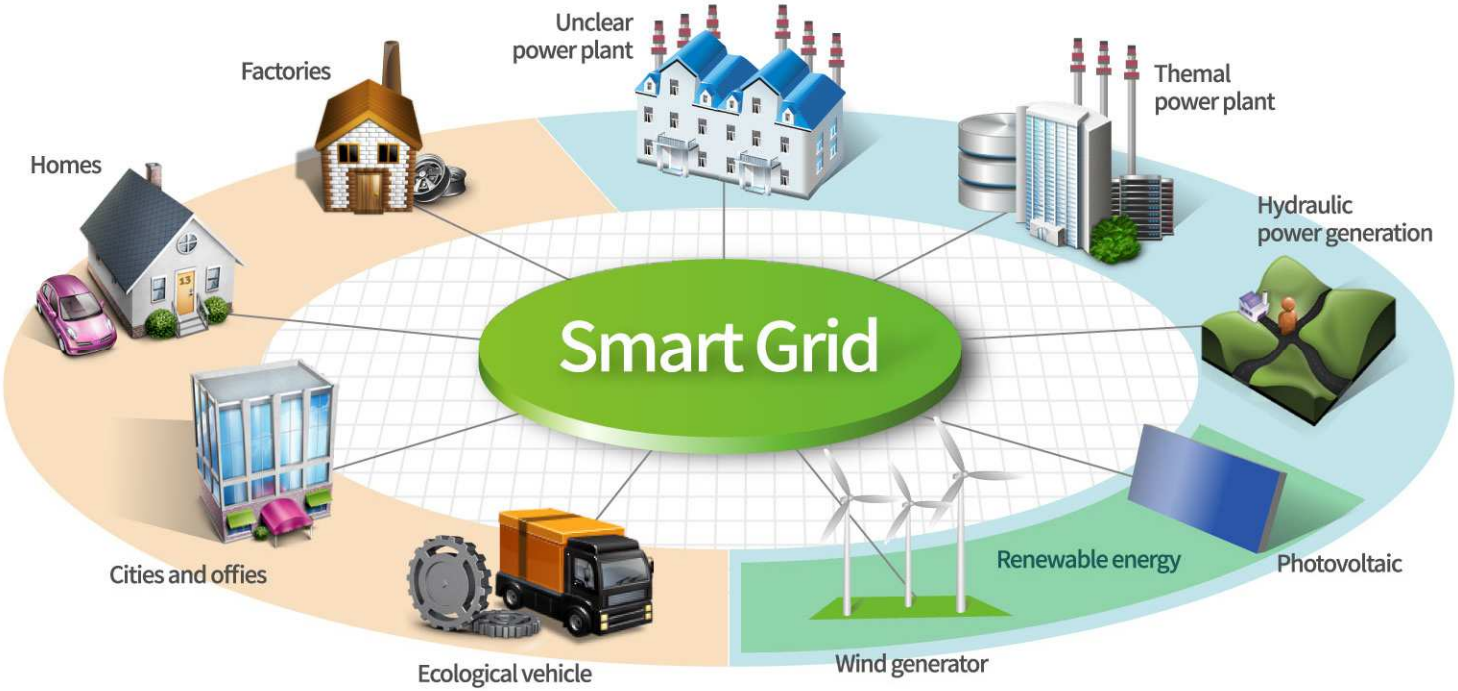
Schneider Electric DMS NS

The Choice of Metric for Clustering of Electrical Power Distribution Consumers

Nikola Obrenović, Goran Vidaković, Ivan Luković

iDSC 2017, Salzburg, Austria

Smart Grid

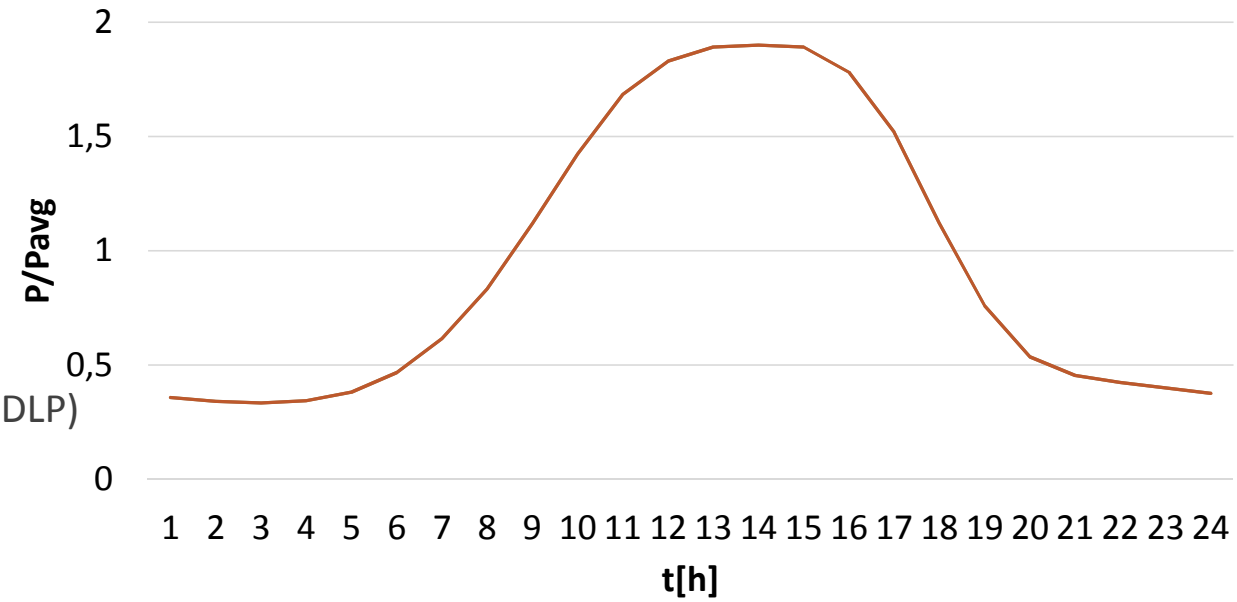

Unclear
power plant
Factories
Themal
power plant
Homes
Hydraulic
power generation
Smart Grid
Cities and offies
Renewable energy
Photovoltaic
Ecological vehicle
Wind generator

Power Distribution Management System

- Network monitoring
  - Load flow
- Network control
- Network planning

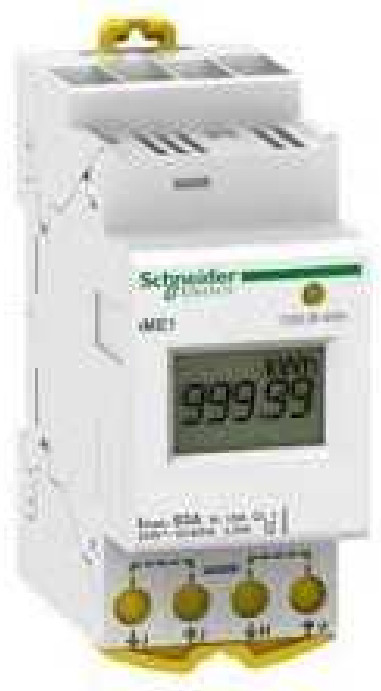
Load Model

- Model of an electrical consumer:
  - Annual average active power
  - Annual average reactive power
  - Load type
- Load type:
  - Set of normalized daily load profiles (DLP)
  - One per (season, day type)

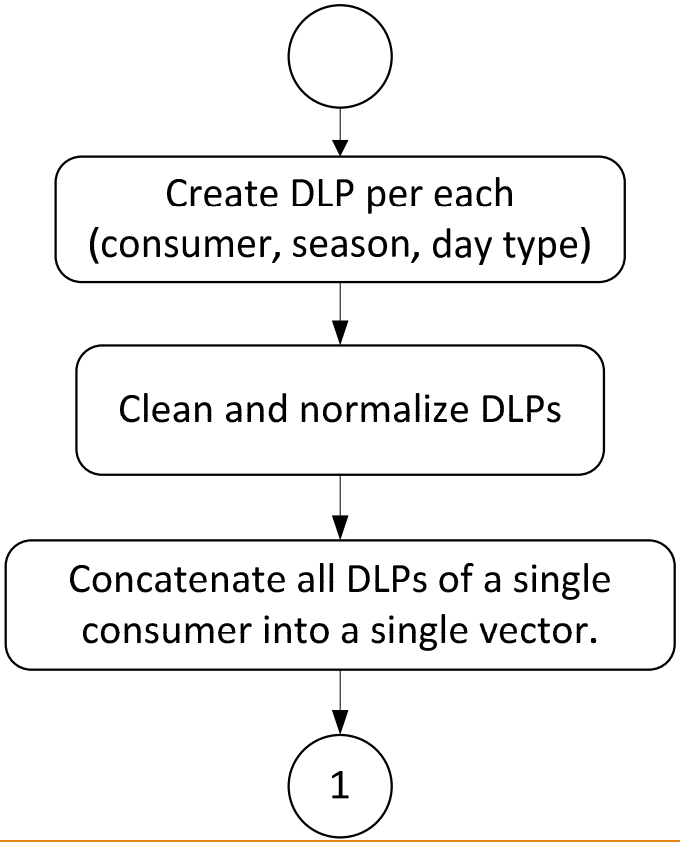


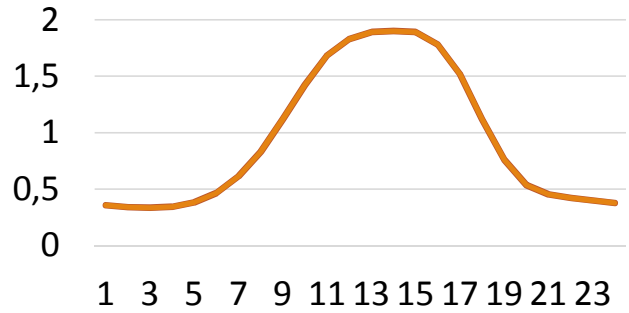
Load Type Creation Algorithm

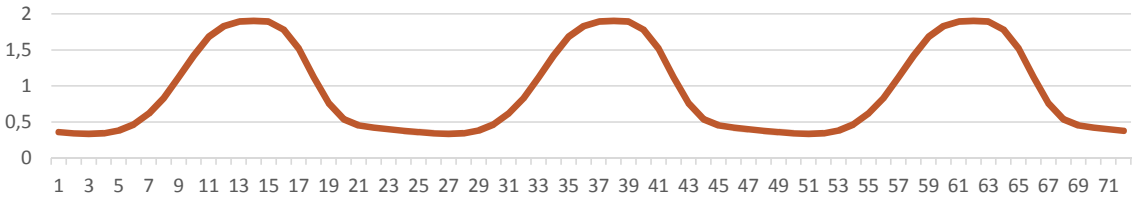
- Input
  - Measurements of P and Q for a year period – 15’ sampling rate
  - Seasons
  - Day types
  - Minimum and maximum number of load types (clusters)
- Output
  - Set of load types (clusters)
- Implemented in pure C#
- 7 Million consumers



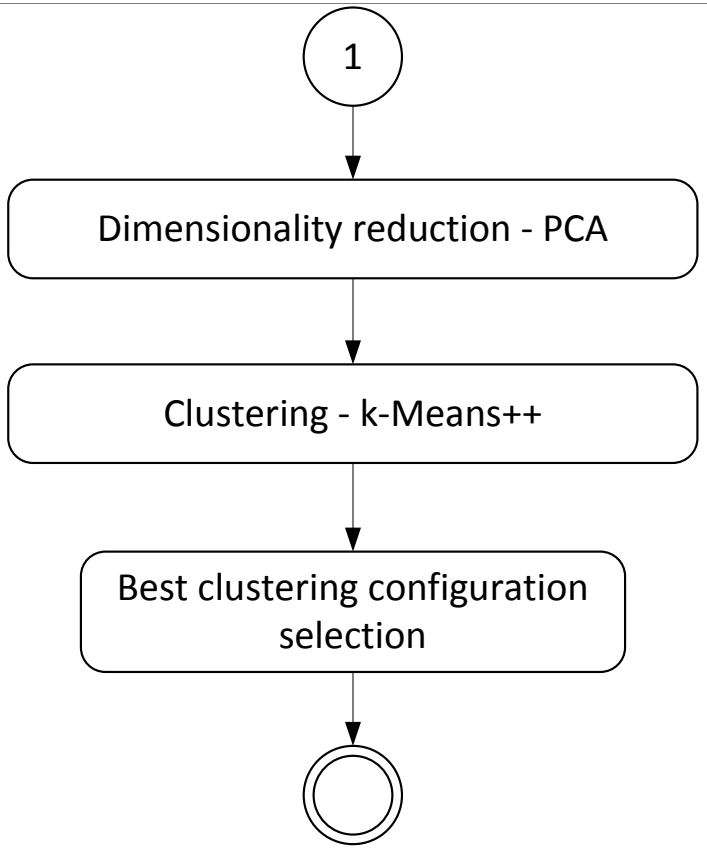
Load Type Creation Algorithm







Load Type Creation Algorithm



Analyzed Metrics

- **Minkowski distance**

$$d_M(x, y) = (\sum_{k=1}^{n} |x_k - y_k|^r)^{\frac{1}{r}}$$

- **Manhattan distance:** $r = 1$
- **Euclidean distance:** $r = 2$

x, y - consumer's vectors

x_k, y_k - value at index k of consumer's vectors

n – number of dimensions (consumer's vector length)

r – metric parameter

Analyzed Metrics

- **Cosine distance**

$$d_S(x,y) = 1 - \cos(x,y) = 1 - \frac{xy}{\|x\|\|y\|}$$

- **Cross Correlation distance**

$$d_{CC}(x,y,\alpha) = 1 - \frac{\sum_{k=1}^{n}[(x_k - \bar{x})(y_{k-\alpha} - \bar{y})]}{\sqrt{\sum_{k=1}^{n}(x_k - \bar{x})^2}\sqrt{\sum_{k=1}^{n}(y_{k-\alpha} - \bar{y})^2}}$$

- α - delay
- $\bar{x}$ - mean value of x
- $\bar{y}$ - mean value of y

Analyzed Metrics

- **Spearman's rank correlation coefficient**
  - Ranks each value in the time series
  - Time series normalized to the unitless domain

$$d_{SR}(x,y) = 1 - \left(\frac{6\sum_{k=1}^{n}(rank(x_k) - rank(y_k))^2}{n(n^2-1)}\right)$$

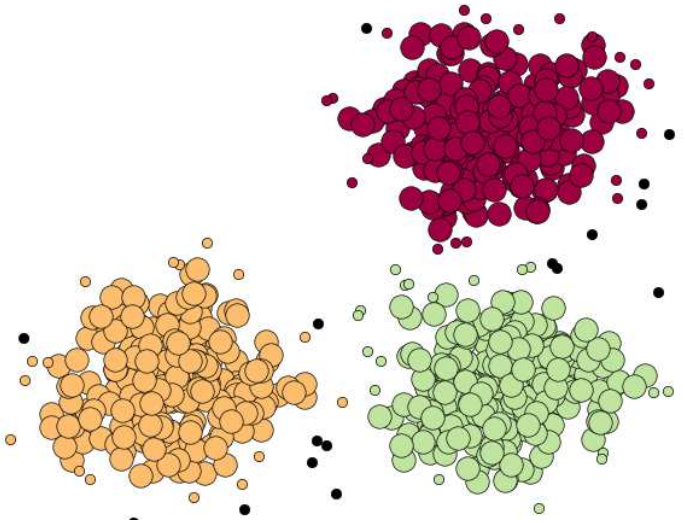
- **Curve Shape Distance**
  - Accounts for the curvature

$$d_{CS}(x,y) = d_E(x,y) + \sum_{k=1}^{n-1} |(x_{k+1} - x_k) - (y_{k+1} - y_k)|/\Delta t$$

Validity Indices

Validity assessment includes two measurement criteria:

- *Compactness:* The members of each cluster should be as close to each other as possible
  - A common measure: the variance
- *Separation:* The clusters should be widely separated
  - Distance between the closest member of the clusters
  - Distance between the most distant members
  - Distance between the centres of the clusters



Davies-Bouldin (DB) Validity Index

- DB index: the average of similarity between each cluster and its most similar one
- The lower DB index – the better cluster configuration

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$d_{ij} = d\left(v_i, v_j\right), \; s_i = \frac{1}{\left\|c_i\right\|} \sum_{x \in c_i} d\left(x, v_i\right)$$

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i$$

$$R_i = \max_{j=1 \ldots nc, i \neq j} \left(R_{ij}\right), \; i = 1 \ldots n_c$$

SD Validity Index

- SD validity index: the average scattering of clusters and inversed total separation of clusters
- The lower SD index – the better cluster configuration

$$Scatt = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\left\| \sigma(v_i) \right\|}{\left\| \sigma(x) \right\|}$$

$$Dis = \frac{\max\limits_{i,j=1\ldots n_c} \left(\left\| v_j - v_i \right\| \right)}{\min\limits_{i,j=1\ldots n_c} \left(\left\| v_j - v_i \right\| \right)} \sum_{i=1}^{n_c} \left(\sum_{\substack{j=1, \\ i \neq j}}^{n_c} \left\| v_j - v_i \right\| \right)^{-1}$$

$$SD = \alpha \cdot Scatt + Dis$$

Assessment Process

- Data set
  - European power network
  - 3 different locations: CR1, CR2, CR3
  - 2000 monitored consumers per location
- Input data
  - Metered active power
  - Metered reactive power
- Clustering with different metrics in each data set
  - From 2 to 20 clusters

Results with PCA

VALUES OF SD VALIDITY INDEX

	CR1	CR2	CR3
Euclidean (L2)	0.70	0.72	0.70
Cosine	1.00	1.02	0.80
Cross Correlation	1.00	1.18	0.96
Spearman	1.06	1.23	1.23
Curve Shape Dis.	0.84	0.70	0.63

VALUES OF DB VALIDITY INDEX

	CR1	CR2	CR3
Euclidean (L2)	1.52	1.59	1.25
Cosine	2.70	2.33	2.63
Cross Correlation	1.85	1.92	1.89
Spearman	4.35	4.76	4.76
Curve Shape Dis.	1.79	1.67	1.79

Results without PCA

VALUES OF SD VALIDITY INDEX

	CR1	CR2	CR3
Euclidean (L2)	0.79	0.72	0.69
Cosine	0.68	0.62	0.65
Cross Correlation	1.32	1.39	1.16
Spearman	1.25	1.25	1.19
Curve Shape Dis.	0.35	0.47	0.42

VALUES OF DB VALIDITY INDEX

	CR1	CR2	CR3
Euclidean (L2)	1.56	1.33	1.43
Cosine	1.59	1.61	1.59
Cross Correlation	2.22	2.17	2.17
Spearman	2.38	2.44	2.56
Curve Shape Dis.	1.28	1.47	1

Conclusions

- Data transformed with PCA ⇒ Euclidean metric
- Original (untransformed) data ⇒ Curve Shape Distance
- Overall best performance: original data and Curve Shape Distance

Future research paths

- Test other clustering algorithms
- Development of a consumption-based metric
- A larger number of data sets from different countries

Thank you!