

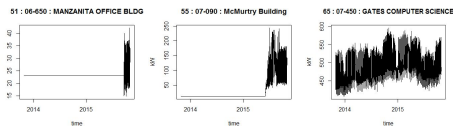
Stanford ENGINEERING GPR and K-means for VST load forecasting of individual buildings at Stanford



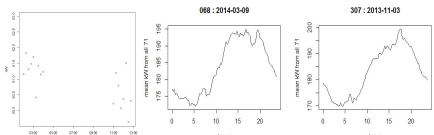
Carol Hsin, Stanford University

Summary

The objective of this study is to return a probability distribution of the expected electricity load in kiloWatts (kW) given a building and a date and/or time. The raw data consists of the mean kW of 15 minute intervals for two years from 122 meters corresponding to 109 buildings at Stanford resulting in a data set of 70176 observations. The data was cleaned and culled into a 70176x71 matrix. K-means clustering was ran on a year of data with analysis resulting in 8 clusters for days (1:365) and 4 for buildings (1:71). The clustering was used to select test dates and test buildings to be culled from the second year. The baseline model is linear regression using only the vector from each building and time features. Gaussian process regression models are work on progress and most of future experimentation will be due to tuning the kernel/covariance matrix and hyperparameters for the Gaussian process regression models.

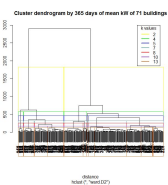


In the data cleaning process, meters servicing the same buildings were merged and exploration showed no meters serviced more than one building. This process reduced the data from 122 to 93 variables. All 93 remaining were plotted and analyzed to remove anomalies, especially buildings with missing data that was replaced with a number as a stand-in. This reduced the data to 71 variables from 122. Missing daylight savings points were addressed by repeating means and multiple points merged with averaging.

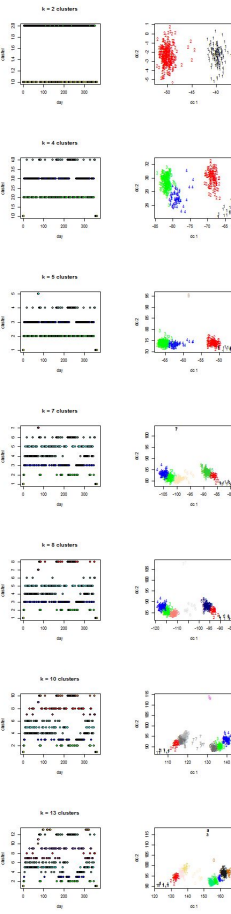


The resulting matrix of 71x70176 was then put through further processing, with two clustering steps, one for clustering days and one for clustering buildings. For both, the data was divided so only the first year was used, matrix of 71x35040 from "2013-10-28 PDT" to "2014-10-27 23:45:00 PDT". The data was aggregated and transformed into a 365x96 matrix for kmeans clustering by the rows to yield clustered days.

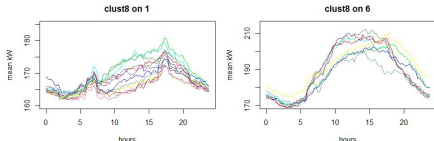
$$\begin{bmatrix} 1 & \dots & 71 \\ 35040 & \dots & 35040 \times 71 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ \vdots \\ 35040 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \dots & 96 \\ 365 & \dots & 365 \times 96 \end{bmatrix}$$



The resulting hierarchical cluster dendrogram revealed 7 possible K-values based on euclidean distances calculated. For some K-values, it was clear which days were being clustered, e.g. winter break in K=4 cluster 4. Since each point is in 96 dimensional space, visualizing the clustering usual discriminant coordinates.

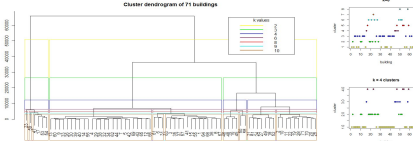


K=8 was chosen by the typical distance metric and looking at the reasonableness of the general patterns of days when clustered, e.g. cluster 1 and 6.



$$\begin{bmatrix} 1 \\ \vdots \\ 35040 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \dots & 96 \\ \vdots & \dots & \vdots \\ 365 & \dots & 365 \times 96 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \dots & 96 \\ \vdots & \dots & \vdots \\ 8 & \dots & 8 \times 96 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ \vdots \\ 768 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \dots & 768 \\ \vdots & \dots & \vdots \\ 71 & \dots & 71 \times 768 \end{bmatrix}$$

For clustering of buildings, each 35040 vector was transformed into a 365x96 matrix from which mean profiles for the clustered days were created for a matrix of 8x96 and then the matrix was unwrapped and inserted into the larger 71x768 matrix for clustering. The resulting cluster dendrogram showed 7 possible k-values.



K=4 was chosen by distances and because it was the highest number given distances. It's not clear why certain buildings were clustered together, but the clusterings were very similar.

Results

All Building 27				All Building 50			
year	clusterID	meterID	mean kW	year	clusterID	meterID	mean kW
2013	1	13152823871267	13.962876401292	2013	1	33638326794009	13.962714201292
2014	1	14967262810181	25.172737642029	2014	1	12359883800000	13.962876401292
2013	1	13578927838026	25.076923120815	2013	1	45782226258115	13.962876401292
2013	1	14052220000000	19.076478102811	2013	1	13682226258115	13.962876401292
2013	1	14484120213024	8.1478940010175	2013	1	38484928457303	13.962876401292
2014	1	1327512714124	13.752240800742	2014	1	42222270824011	13.962876401292
2013	1	14052220000000	19.076478102811	2013	1	37878787052671	13.962876401292
2014	1	14052220000000	19.076478102811	2014	1	34222270824011	13.962876401292

Test buildings and test dates were chosen based on the clusterings, for 4 test buildings with 8 test dates (7008 data points) each. Baseline AR models were run on the test buildings for each test date.

Data and Experiments

The raw data obtained from Stanford Facilities consists of the mean kW of 15 minute intervals for two years from 122 meters corresponding to 109 buildings at Stanford resulting in a data set of 70176 observations. Each building may be serviced by multiple meters and each meter may be servicing multiple buildings. The data ranges from October 28, 2013 at 12:00:00 AM to October 28, 2015 at 10:30:00 PM. Preliminary data exploration revealed there are 8 missing and 8 repeated data points within the period being examined. They correspond to daylight savings times, so the data for that period is missing because the hours don't actually exist and this was confirmed because Nov 3 has extra data points. This caused problems to be addressed in later sections.

Future Work

Most of the work completed consists of gathering and cleaning data, and then performing the clustering analysis to determine how best to split the data and as preparation for future models. The baseline AR model was done. The models planned involve using Gaussian Process Regression with to predict the loads of the test days per test building based on historic clustered days, historic clustered days and data from clustered buildings, and just based on 60 days prior to the test date (can't do more because GPR is nonparametric, so will store the covariance matrix per run and if all data were used, errors appear due to RAM limitations).