Stanford | ENGINEERING

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

Carol Hsin, Stanford University

Summary

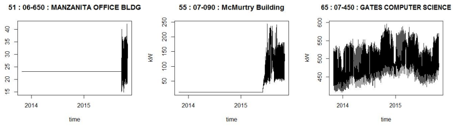
The objective of this study is to return a probability distribution of the expected electricity load in kiloWatts (kW) given a building and a date and/or time. The raw data consists the mean kW of 15 minute intervals for two years from 122 meters corresponding to 109 buildings at Stanford resulting in 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.

Background

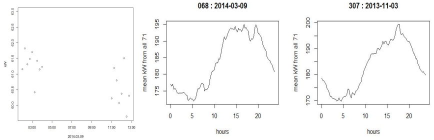
Electricity isn't easily stored, so producers must meet the maximum demand at any time or outages will ensue. Producers want to reduce operational costs by keeping just minimum the generation reserves, but it would also be problematic for outages to occur due to inadequate supply. Thus, accurate electrical forecasts are essential to energy management because they allow dispatchers to make decisions on the optimal, real-time scheduling of electricity production between power plants or generation units

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.

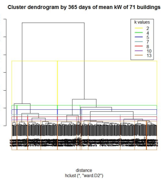


In the data cleaning process, meters servicing the same buildings were merged and exploration showed no meters serviced more than one building. This process reduced the data from 122 to 93 variables. All 93 remaining were plotted and analyzed to remove anomalies, especially buildings with missing data 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.

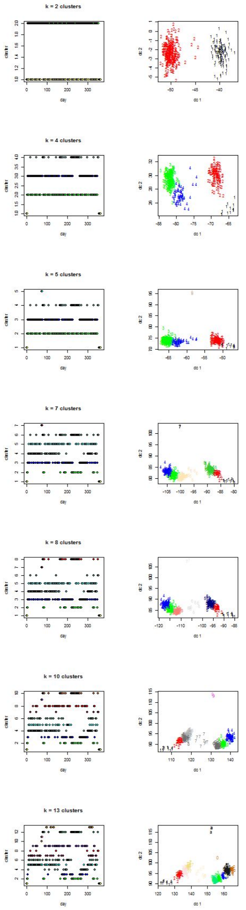


The resulting matrix of 71x70176 was then put through further processing, with two clustering steps, one for clustering days and one for clustering buildings. For both, the data was divided so only the first year was used, matrix of 71x35040 from "2013-10-28 PDT" to "2014-10-27 23:45:00 PDT". The data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

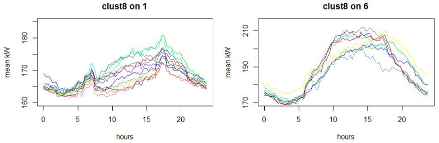
$$\begin{bmatrix} 1 & \cdots & 71 \\ \vdots & \ddots & \vdots \\ 35040 & \cdots & 35040\text{x}71 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ \vdots \\ 35040 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \cdots & 96 \\ \vdots & \ddots & \vdots \\ 365 & \cdots & 365\text{x}96 \end{bmatrix}$$



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.

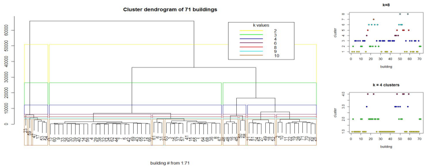


K=8 was chosen by the typical distance metric and looking at the reasonableness of the general patterns of days when clustered, e.g. cluster 1 and 6.



$$\begin{bmatrix} 1 \\ \vdots \\ 35040 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \cdots & 96 \\ \vdots & \ddots & \vdots \\ 365 & \cdots & 365\text{x}96 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \cdots & 96 \\ \vdots & \ddots & \vdots \\ 8 & \cdots & 8\text{x}96 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ \vdots \\ 768 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \cdots & 768 \\ \vdots & \ddots & \vdots \\ 71 & \cdots & 71\text{x}768 \end{bmatrix}$$

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

AR Building 27		
date	trainRMSE	testRMSE
001	3.11562406136734	5.92908086302396
243	2.80726350414573	27.1727170432979
107	3.57288754308325	25.0050035248015
337	3.60722058939789	19.6730977413011
341	3.64815270136514	6.34749946969175
298	3.17871587441246	13.7552846364742
076	3.00282373966765	16.4055959407633
259	2.86685920192124	22.5292421013259

AR Building 50		
date	trainRMSE	testRMSE
001	3.38363367869690	11.8094774861282
243	3.73769993365999	17.0612639134373
107	4.57362220540115	37.5793096579895
337	3.94042377491766	48.1890094398847
341	3.88495198175363	10.8304380342995
298	4.21250373061003	14.1253591604769
076	3.97807505678671	17.8563097039936
259	3.56124275581864	36.8535230182934

AR Building 41		
date	trainRMSE	testRMSE
001	11.3137171108784	1.63099169210129
243	1.74704263141441	16.5290857191283
107	15.8048459682757	68.6245047572578
337	16.1212470934128	71.9370965560381
341	16.0362318459376	53.1374392548315
298	16.6870934592130	40.6767609910458
076	16.1701762503520	37.6412048858712
259	5.48500264308082	50.8484470522624

AR Building 53		
date	trainRMSE	testRMSE
001	15.9652386786304	25.3668404088917
243	10.8554679708035	54.8878136779890
107	16.6909585241200	61.2952412778726
337	13.8077246997190	59.4570004960659
341	15.3236243694047	51.9888134588392
298	12.7511734577707	20.4150748457536
076	11.3992393682145	36.7204665273363
259	11.7719471437517	107.842217763288

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.

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).