Generating Load Profiles from Building Characteristics

Robert Clain
Stanford University
rclain@stanford.edu

***Abstract*—The goal of this project was to create an electrical load profile generator for commercial buildings that would be as accurate as possible given minimal input. Building characteristic data came from a CBECS survey, and sample profiles came from NREL simulations. A scoring method for produced profiles combining mean, spread, and correlation was created. Baseline performance was determined using DC component regression and nearest neighbor models, each rated using an average score of 100 randomly left out samples. A multivariate regression model, fit in the frequency domain, was created to reduce the bias of the DC model and improve correlation; however, this model was not better than using nearest neighbor due to phase mismatch. A clustering approach to increase the bias of nearest neighbor, grouping by medoids, improved correlation score. An added scaling factor further increased the total score. Lastly, matrix completion was used for profile generation. Results were better than those from regression, but worse than those from clustering.**

I. Introduction

In community planning and electrical system simulation, it is often desirable to have accurate profiles of loads, typically at the building level. For example, one may want to know if a proposed facility, consisting of multiple buildings, can meet its own energy needs throughout the year with solar panels. For another case, a building load profile may be needed for simulating the performance of some energy consumption optimization routine. If a building is metered, this data will be available, but the building may not be metered due to expense. The building may not exist at the time of study or will never actually exist at all. While it is possible to create a fairly accurate profile using simulation (for example, with the software EnergyPlus), this process can be time consuming and the detailed information required to make the model may not be available. Because of these issues, the goal of this project was to create a quick and fairly accurate method for generating a load profile given a subset of building characteristics.

Several models were trained in this study, using a combination of survey data collected by the DOE (Department of Energy) and simulation data from NREL (National Renewable Energy Laboratory). The survey data is from the 2003 CBECS (Commercial Buildings Energy Consumption Survey) [1], and the simulation data comes from EnergyPlus models created by NREL using the information from this particular CBECS survey [2]. The CBECS survey is conducted approximately every 4 years, and includes building information such as square footage, building use, occupancy, and total energy used per year. The simulation study [2] includes yearly load profiles with 1 hour intervals, in Joules, for most buildings from the survey.

Analysis of CBECS data [3] and load profile generation [4], [5] have been done before to some degree. Unlike these studies, the whole yearly profile is being generated for any general building.

II. Data Collection

The selected features from the CBECS fall under the category "General Building Information and Energy End Uses." These survey results were chosen since it is desirable to create profiles using information that can be easily obtained. The survey results were further paired down by removing variables with over 20% missing values, those with single categorical response, and those not useful for prediction (such as survey respondent ID, variance, etc.). The final subset of features is shown in Table I.

TABLE I. General Building Features

REGION8	Census region	CENDIV8	Census division
SQFT8*	Square footage	PBA8	Principal building activity
ELUSED8	Elec. used	NGUSED8	N. gas used
FKUSED8	Fuel used	PRUSED8	Propane used
STUSED8	Steam used	HWUSED8	Hot water used
CLIMATE8	Climate Zone	WLCNS8	Wall material
RFCNS8	Roof material	GLSSPC8	Percent glass
EQGLSS8	Equal side glass	BLDSHP8	Building shape
NFLOOR8*	# of floors	YRCON8*	Year constructed
GOVOWN8	Gov. owned	OWNER8	Owner
OWNOCC8	Owner occup.	NOCC8*	# businesses
MONUSE8*	Months used	PORVAC8	Vacancy
OPEN248	Open 24hr/day	OPNMF8	Open M-F
OPNWE8	Open weekend	WKHRS8*	Weekly operating hrs.
NWKER8*	# employees main shift	HT18	Energy use heating
HT28	Energy 2nd heating	COOL8	Energy use cooling
WATR8	Energy water heating	COOK8	Energy use cooking
MANU8	Energy use manuf.	GENR8	Energy use generation
ADJWT8*	Building weight		

Every asterisked variable is continuous, while the others are categorical. The NREL simulation results (existing stock only) were aggregated and then matched by respondent ID to the CBECS features to create the full dataset, containing survey results and load profiles for 4,820 buildings. Buildings with poor survey responses were discarded (mostly NA values), leading to an eventual 4,220 samples.

III. Error Estimates

A custom measure of load profile similarity was created for calculating training and test scores in order to capture the qualities of the desired output. These qualities are generally not present in a standard metric. For example, a sum of squares difference may prefer a flat signal that is closer to the true one than one that oscillates in tandem. Here it was assumed that there are three desirable traits:

- The generated and true profile are comparable in magnitude.
- The generated and true profiles have similar variation in electricity use over the domain.
- The patterns present in the true profile are present in the generated profile, and they align properly.

Since there exist individual standard metrics for these properties, a weighted combination of measures was used (albeit with a somewhat subjective choice of weights, which may change depending on application). The final summed measure

comparing the real and predicted profiles, denoted $M(S_R, S_P)$, is:

$$w_1 \frac{1}{1 + NAE(\mu)} + w_2 \frac{1}{1 + NAE(s)} + w_3 dCorr(S_R, S_P) \quad (1)$$

All terms are between 0 and 1. The first is a transformed normalized absolute error of the difference of means between signals, normalized with respect to the true signal. The second term is a similarly transformed normalized absolute error of the difference of spread between signals ($maximum - minimum$), and the third is the distance correlation, measuring the linear relationship between signals. The weights chosen were $1/6$, $2/6$, and $3/6$ respectively to get a result between 0 and 1. A perfect match yields a 1, while the worst match yields a 0.

The predictive power of each model was estimated using an average of scores from 100 randomly left out samples (left out one at a time from the training set, then tested). While this was not ideal and a more robust cross-validation is desired, such computation is prohibitively expensive given the nature of the similarity metric and the size of the data. The average of each term in (1) and the average final measure have been reported for comparison.

IV. Baseline Models

Two simple models were chosen to compare against others. One is k-nearest neighbors (with $k = 1$), which simply picks the closest profile from the dataset. Another is DC regression, which regresses on a single frequency component of the response, even though many are present.

A. Nearest Neighbor

The distance measure used for knn was non-obvious, as there is a mixture of continuous and categorical variables in the dataset. Two different methods were used, both of which first scaled the data to be centered and have unit variance. For the first, the nearest neighbor was determined using the continous variables (Euclidean distance) with unlikely ties broken by the number of identical categorical variables (Hamming distance), as many of the important variables found in [3] are continuous. The second method used all of the variables with dummy variables introduced for categorical features. The score results are summarized in Table II and Figure 1 shows the scores of the second method.

TABLE II. Nearest Neighbor Score

Method & Set	Mean Score	Spread Score	Correlation Score	Total Score
1 Training	1	1	1	1
1 Test	0.6554176	0.5772903	0.6322919	0.6178123
2 Training	1	1	1	1
2 Test	0.5762448	0.5739004	0.6762918	0.6254868

Though the resultant score seems very good, the method is highly variable, as evidenced by the histograms and difference in training/test scores.

B. DC Component Regression

A regression approach was also used as a baseline as it is more biased than nearest neighbor and has the ability to generate a profile outside those seen in the training data. As the true responses are high in dimension, regression was performed on the simplest component, the frequency 0 amplitude. Two methods were used here as well. The first method is a pure linear regression, while the second is a random forest regression. The second approach was chosen since unlike linear

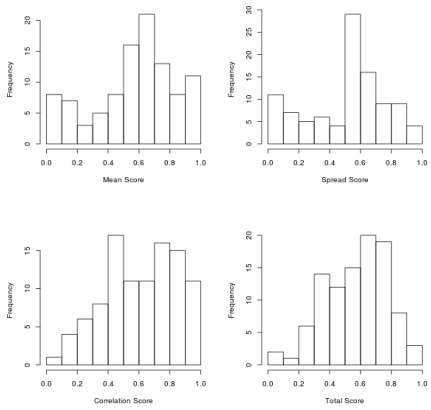

Fig. 1. Nearest Neighbor Score Histograms

regression, results are within the minimum and maximums in the dataset. Any negative values predicted by the first method were set to equal 0.

The training and test score results are summarized in Table III. These values for linear regression actually represent the scores from fitting on purely SQFT8, as it was found it gives an R^2 of 0.828 compared to $R^2 = 0.8841$ when using all of the predictors. A histogram of scores for the linear regression is shown in Figure 2.

TABLE III. DC Regression Score

Method & Set	Mean Score	Spread Score	Correlation Score	Total Score
1 Training	0.5610864	0.5000000	0.0000000	0.2601811
1 Test	0.5433592	0.5000000	0.0000000	0.2572265
2 Training	0.5395164	0.5000000	0.0000000	0.2565861
2 Test	0.5512175	0.5000000	0.0000000	0.2585363

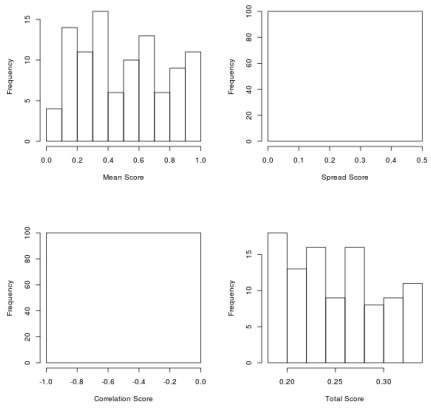

Fig. 2. DC Regression Score Histograms

Only the first score component is affected by DC regression, as a flat signal has no spread or movements to correlate. This value is close to that achieved from nearest neighbor. The regression is very biased, hence the similarities in training and test scores.

V. Frequency-Domain Regression

The first attempt at a more complicated and more accurate model was to create a multivariate multiple linear regression on the amplitude and phase of different frequency components of the responses using a truncated Fast Fourier Transform

(FFT) using all of the predictors. The decision to truncate the transform was based on the notion that while more frequencies may generate a better profile, if the sinusoids are not aligned the resultant signal may look like garbage (due to incorrect constructive and deconstructive interference). Since each load profile has a well defined minimum energy usage, and truncated Fourier series tend to have too much swing beyond this point, the minimum was also included in the model as a lower bound.

For each signal, an FFT was performed and the K frequencies with the largest amplitudes were stored. Afterwards, the K most common frequencies stored across all signals were chosen to represent every response signal. Increasing K had diminishing returns on improving similarity to the true responses. By visual inspection, it was determined that $K = 15$ yielded fairly good results while still being small enough in value. A histogram of the common frequencies for this K is shown in Figure 3, and a sample truncated series is shown in Figure 4.

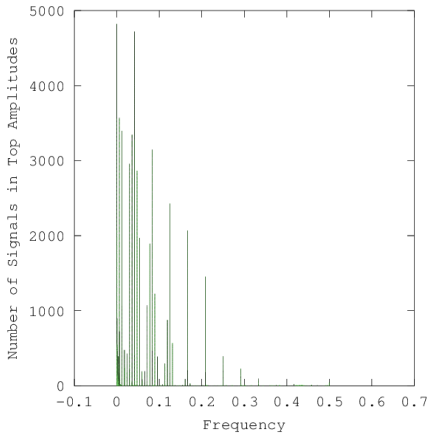

Fig. 3. Common Frequencies with Large Amplitude

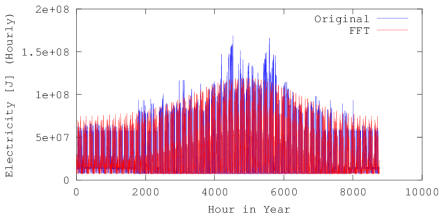

Fig. 4. Sample Transformed Response

Prior to regression, the phase components were also altered to lie in the range 0 to 2π. The resulting scores are shown in Table IV. A second method, using a multivariate regression tree, was also shown as this was done for DC regression. The histogram of these scores for the linear regression method is shown in Figure 5.

TABLE IV. FFT REGRESSION SCORE

Method & Set	Mean Score	Spread Score	Correlation Score	Total Score
1 Training	0.5229658	0.5711416	0.5074720	0.5312775
1 Test	0.5017372	0.5386633	0.4586032	0.4924789
2 Training	0.4242117	0.5231067	0.5058988	0.4980203
2 Test	0.4129745	0.5262464	0.5305609	0.5095250

While the FFT regression did increase the correlation score compared to DC regression, performance still fell short of the nearest neighbor score. This is because while the amplitude

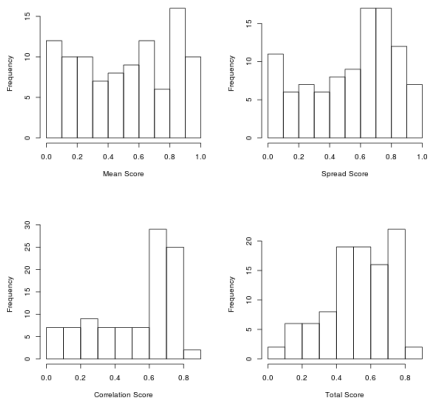

Fig. 5. FFT Regression Score Histogram

predictions were quite reasonable, the phases did not match up, making the recombined signal appear distorted. On closer inspection, it appeared that the phases congregate around certain values in the dataset. Early attempts to address this issue involved using k-means to cluster the phase values (using minimum angular distances) and predict them separately from amplitudes. This method had poor results. Another attempt to fix the issue removed close frequencies (taking frequencies based on power spectra instead of FFT). This did not seem to improve results much.

VI. MEDOID CLUSTERING

Clustering was pursued further due to the poor regression results in comparison to nearest neighbor. The clustering approach selected, k-medoids, is similar to k-means except values are selected from within the dataset. In this case k representative signals were selected from the dataset and group labels were assigned. This was preferred since the method of generating the centroid is again non-obvious. Since correlation is the most heavily weighted term and the full similarity metric is computationally inefficient, a similarity matrix was generated using Pearson's correlation distance, somewhat related to the distance correlation. The number of groups was selected from a finite set by choosing the result with the lowest test error for the third term in the similarity measure. This value was selected to be 14. Classification for the group labels used a support vector machine with a linear kernel, with error being calculated between the true signal and the reference signal for the predicted label. Results are presented in Table V and Figure 6.

TABLE V. MEDOID CLUSTERING SCORE

Set	Mean Score	Spread Score	Correlation Score	Total Score
Training	0.2854639	0.2974438	0.7697043	0.5315774
Test	0.3810489	0.3931529	0.7532599	0.5711890

While the correlation score was improved upon over nearest neighbor, the mean and spread scores were severely reduced. This is because they were not taken into account in generating the similarity matrix.

VII. MEDOID SCALING

Since it appeared there was a common pattern to the mean and spread scores for which medoid clusterig poorly predicted, a second stage prediction for mean value was carried out within the predicted group using a random forest. The

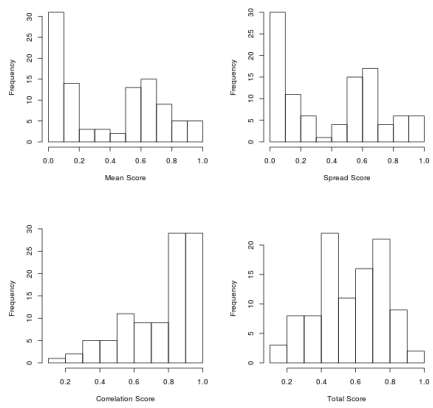

Fig. 6. Medoid Clustering Score Histogram

mean was then divided by the mean of the centroid to get an appropriate scaling factor. This method had very good results. The correlation score was very similar to what was achieved earlier and the mean/spread scores improved. Results are shown in Table VI and Figure 7.

TABLE VI. SCALED MEDOID CLUSTERING SCORE

Set	Mean Score	Spread Score	Correlation Score	Total Score
Training	0.7436180	0.6385262	0.7703260	0.7219414
Test	0.8275045	0.6839815	0.8003134	0.7660679

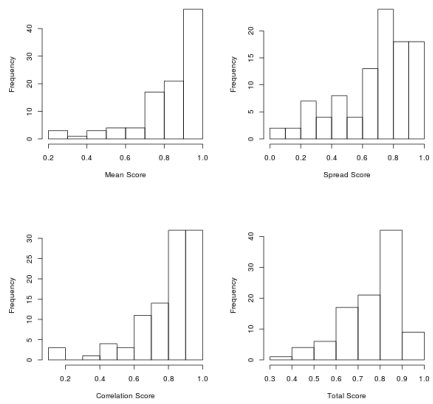

Fig. 7. Scaled Medoid Clustering Score Histogram

VIII. MATRIX COMPLETION

Lastly, matrix completion was used in an attempt to create a prediction signal. Matrix completion attempts to make the dataset matrix, with the predicting features and NA responses tacked on, low rank. As attaining low rank implies matching correlations between data, and the signal quality is primarily scored by correlation, it was thought this method may give good results. Specifically, alternating least squares was used to fill in the missing values. Results are shown in Table VII and Figure 8.

TABLE VII. MATRIX COMPLETION SCORE

Set	Mean Score	Spread Score	Correlation Score	Total Score
Test	0.5564556	0.5553247	0.5287055	0.5422036

These scores are better than those from regression, yet worse than those from nearest neighbors and clustering. The

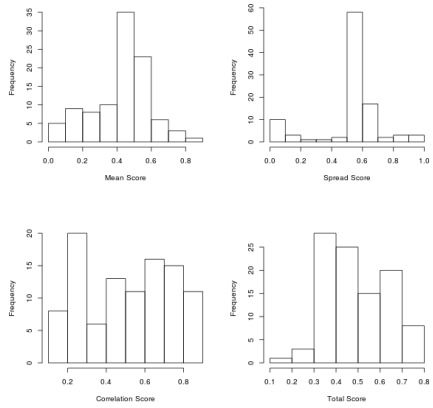

Fig. 8. Matrix Completion Score Histogram

profiles, upon examination, looked much better than those from regression but contained a lot of noise. In addition, they seemed to be somewhat restricted in the types of profiles that could be produced (for example, all had a central peak). The means and spreads are of correct proportion, but typically a few orders of magnitude off of the true values (leading to many mean and spread scores near 0.5).

IX. CONCLUSIONS

The small number of predictors and size of the output made it very difficult to generate a representative signal. Clustering was found to give the best performance. It was the only method found to give better results than the baseline nearest neighbor. While regression by itself did not seem to work well, clustering was noticeably improved by regressing on metadata extracted from the profiles (creating a scaling factor from predicted mean) after they had been grouped.

The profiles from this study are much less homogeneous than those in the references, which may be why generative approaches failed. A GAM seemed to work well in [5], but that regression was for much more similar buildings (homes) and used data purposefully left out of this study (such as total yearly consumption).

Future work could improve upon a number of aspects of this study. First, a more computationally efficient metric calculation would better facilitate cross-validation, which was limited here to fewer random samples. The measures for mean and spread could also be adjusted to use a log scale for representing order of magnitude difference, as many of the values were close to 0.5 (which represent 100% difference). A more robust feature selection, where terms from other parts of the survey were selected and some used here were discarded would also be useful if it would limit the information required as input for the profile generation or improve performance.

REFERENCES

[1] "Commercial buildings energy and consumption survey (cbecs), 2003," http://www.eia.gov/consumption/commercial/data/2003.

[2] "National renewable energy lab simulation data," http://en.openei.org/datasets/node/41.

[3] T. Sharp, "Energy benchmarking in commercial office buildings," Oak Ridge National Laboratory, Tech. Rep., 1992.

[4] R. A. A.-A. Abdelbaset Ihbal, HS Rajamani and M. Jalboub, "The generation of electric load profiles in the uk domestic buildings through statistical predictions," *Journal of Energy and Power Engineering*, pp. 250–258, February 2012.

[5] S. Heunis and M. Dekenah, "A load profile prediction model for residential consumers," *Energize*, pp. 46–49, May 2010.

APPENDIX

Sample Results - Below are profiles generated for the first 10 samples in the dataset using the various approaches. This subset does not represent profiles related to testing scores (in fact, for this small subset, nearest neighbors scored abnormally high for the correlation term and medoid clustering scored lower), but is meant to demonstrate some of the issues encountered.