Improving Energy Use Forecast for Campus Micro-grids using Indirect Indicators

Saima Aman
Department of Computer Science
University of Southern California
Los Angeles, CA
saman@usc.edu

Yogesh Simmhan
Department of Electrical Engineering
University of Southern California
Los Angeles, CA
simmhan@usc.edu

Viktor K. Prasanna
Department of Electrical Engineering
University of Southern California
Los Angeles, CA
prasanna@usc.edu

***Abstract*—The rising global demand for energy is best addressed by adopting and promoting sustainable methods of power consumption. We employ an informatics approach towards forecasting the energy consumption patterns in a university campus micro-grid which can be used for energy use planning and conservation. We use novel indirect indicators of energy that are commonly available to train regression tree models that can predict campus and building energy use for coarse (daily) and fine (15-min) time intervals, utilizing 3 years of sensor data collected at 15min intervals from 170 smart power meters. We analyze the impact of individual features used in the models to identify the ones best suited for the application. Our models show a high degree of accuracy with CV-RMSE errors ranging from 7.45% to 19.32%, and a reduction in error from baseline models by up to 53%.**

***Keywords*-energy forecast models; energy informatics**

I. Introduction

One of the critical challenges confronting modern societies is the need to attain energy sustainability. Buildings account for about 40% of the total urban energy consumption worldwide [6] and electricity forms 38% of total energy usage in the US [2]. Adoption of energy- efficient measures in buildings and institutional campuses can significantly contribute to energy conservation. The rollout of smart grids with the capability for real-time electricity usage sensing and bi-directional communication with power consumers provides electric utilities opportunities to better manage available capacity and curtail its usage during peak demand periods using pricing incentives. Reliable building energy forecast models can help predict energy use over short and long time durations, and inform residents and facility managers in planning electricity usage and facility improvements with an eye on reducing their energy footprint and power usage costs.

Energy analysis modeling of buildings is either based on steady state or dynamic conditions that physically characterize a building, or are based on measured building performance data. Smart meters have made available a large corpus of electricity consumption data at fine granularities [1]. Statistical analysis and machine learning methods can be used to mine sensor data and extract forecast models.

In this paper, we analyze the use of machine learnt models for energy use forecast at the University of Southern California (USC) in Los Angeles, which is a mini-city in the diversity of its buildings and number of occupants, and is a micro-grid test-bed for the DoE sponsored Los Angeles Smart Grid Demonstration Project [11]. These model predictions are meant both for fine and coarse granularities of time and space: at the building and campus levels, and for daily and 15-min time periods, to assist with different planning goals and provide insight into daily load requirements, peak demand periods for possible curtailment, and energy consumption drivers, such as seasonal variation, usage patterns, building types and functions.

Our work is novel in utilizing both direct and indirect indicators of energy use that are commonly available, such as energy usage information from smart meters, attributes from the universitys academic calendar that indicate occupancy patterns, static knowledge of buildings such as surface area, and historical weather information as attributes in our prediction models. Our study uses real world datasets from an operational campus to analyze the impact of different features in improving the model accuracy, and evaluates the efficacy of using global and bespoke forecast models for different building categories.

We make the following specific contributions.

1) We train regression tree models using electricity usage data and other features collected between 2008 and 2009 for 170 buildings on USC campus in Los Angeles to predict energy usage for the campus. The **campus-scale models** are tested for daily and 15-min energy use prediction for the year 2010, and compared against baseline predictions made using energy use averaging techniques. The results are analyzed and improvement from using *novel indirect energy use indicators* is evaluated.
2) We also develop **building-scale models** for predicting daily energy use for a representative group of 23 individual buildings from three different usage categories. Energy usage data and other features of these buildings for the years 2008 - 2009 are used to train the model and evaluate it by predicting the daily energy use for these buildings for the year 2010. In addition, we compare the effectiveness of *global building-scale models* and *local models* for individual building types.

II. Related Work

A. Modeling Methods

Several researchers have studied the problem of the modeling and prediction of building energy consumption. An energy prediction model is based on several parameters that are estimated using existing data that typically include energy consumption and temperature measurements recorded in the past. Prediction models proposed in literature belong to three categories: regression models, artificial neural network models and time-series models [14]. Each of these modeling approaches has its advantages and disadvantages, and the choice of a model is often application dependent.

Regression models have been found suitable for predicting average consumption over longer periods such as days or months [14]. These models can be developed quickly as they require calculation of only a few parameters. Models based on Artificial Neural Networks (ANN) have also been effective for building energy predictions as demonstrated by Olofsson and Andersson [9] to predict building energy use for both short and long term periods and for hourly energy use. Dong, Cao, and Lee [5] have used Support Vector Machines for building load forecasting. Yu et. al. [14] have used decision trees to model building energy use intensity (EUI) levels. In our work, we have used regression tree models that are a type of decision tree model.

B. Data used in Models

Besides training models using static data, some researchers have also focused on models that are capable of adapting to changing patterns of incoming data streams for real-time, on-line energy prediction [13].

Several experiments and analyses are based on synthetic data using building energy simulation programs, such as DeST and EnergyPlus [1] . In some studies, utility bills have been used to establish a baseline model of energy consumption [5]. In our experiments, we use real world energy use data and other publicly available attributes that are indirect indicators of energy use for both training and evaluation.

Researchers have tried different categories of attributes to characterize training data for building energy use models. These features typically belong to the following categories:

1) *Weather Data*: These include temperature and humidity measurements, heating degree days (HDD), cooling degree days (CDD), and temperature difference between indoors and outdoors.
2) *Building Data*: Studies have included a variety of physical building data, including wall insulation thickness, heat transfer coefficient of external walls and roof, orientation of windows, window to wall ratio in each orientation, building shade coefficient, and shading coefficient of windows, solar absorption.
3) *Occupancy Data*: With improvement in the quality of thermal properties of buildings due to energy regulations, the energy use contributed by building characteristics is decreasing and making the role of the occupants more significant [9]. Occupant behavior can cause variation in energy consumption in different building units. However, this data is not easily available as not all buildings are equipped with sensors to monitor occupant's energy usage activity. It is also difficult to predict occupant behavior at the design stage of a building. Yalcintas [12] and Agarwal et. al. [4] have incorporated occupancy data using sensor data and heuristics, such as open/closed state of the doors.

In our experiments, we make use of several additional indirect indicators of energy use attributes, such as the academic schedule, that reduces our dependence on specialized data collection and increase the relevance of our models. These are described in more detail in Section IV.

Our work is unique in the following aspects:

1) We build a single, unified building energy use prediction model that is applicable to diverse buildings on an academic campus, both current and planned. Other research work has focused exclusively on homogeneous buildings, such as commercial or residential. Such a uniform model allows us to eventually extrapolate to a city scale where building diversity is common.
2) We take an information-driven approach that leverages diverse sources and indirect indicators that complement domain-specific attributes. Using data that is often available publicly makes our models more globally and easily applicable.
3) We identify attributes that can be applied during the design as well as the operation phase of the buildings, allowing modeling of the impact of current and future buildings.

III. Regression Tree Learning

The regression tree method generates a decision tree with leaves of the tree ending in a regression function. In a decision tree, a path from the root to a leaf node describes the sequence of tests that are performed in arriving at the decision present in the leaf node. Decision trees are generated in a top-down fashion by choosing the most likely attribute for decision-making at each level. Each attribute that is chosen partitions the remaining training data into subsets depending on the value of the decision made. This method of recursive partitioning leads to smaller regions where simple models can be applied.

One of the advantages of the decision tree based methods is their flowchart style tree structure that is easy to interpret from a domain perspective (as opposed to, say, ANN).

[1] http://apps1.eere.energy.gov/buildings/tools_directory/doe_sponsored.cfm

Another advantage is that once the model is trained, making predictions is fast, as it just requires looking up the values in the tree. For many application areas, such as real-time energy predictions, a fast operation is of significance since the models operate on streaming sensor data for online predictions for operational use. Regression trees have been successfully applied in several domains, including atmospheric science [8], public health [7] and environment [10], etc.

IV. Campus and Building-Scale Energy Use Models

A. CampusEnergy Use Dataset

We use the energy-use data collected every 15 minute by the Facilities Management Services (FMS) from the smart meters installed in 170 buildings at USC campus for 3 years[2]. For daily energy consumption, we aggregate the 15-min data for each day for each building. For campus-level energy consumption, we aggregate the data for individual buildings at 15-min and day-level granularities. As with any other sensor network data stream, there are missing values in the dataset. We adopt interpolation and extrapolation methods to fill in the missing values. We observe $<5\%$ of data points requiring such corrections.

For the building-scale energy use models, we identify and select 23 representative buildings on the campus. Energy use of these buildings from 2001-2009 was studied as part of the USC's Greenhouse Gas Emissions (GHG) report [3]. The report categorizes buildings into three groups: Academic, Residential, and Other. The buildings we chose for our study exhibit different characteristics: 9 buildings have predominantly classrooms and offices, 6 are student residential buildings and dormitories, and 8 buildings of the type Other include Auditorium, Library, Parking Structures, and Gymnasium buildings.

B. Selection of Indirect Energy Use Indicators

Several aspects affect the energy consumption patterns in buildings. The other datasets that we use in our models are:

1) *Weather conditions*: We used maximum and average temperature data as well as humidity data to model weather conditions. We obtain weather data from Weather Underground website[3] using a web form that returns a CSV file for the given date ranges. The weather and humidity data is available at one-hour granularity. For our fine-grained models, we interpolated the values to 15-min durations.
2) *Building-specific parameters*: We use Gross Area, Net Area in use, and Year of construction to characterize each building. This data is publicly available from the USC website[4]. In addition, we also use a categorical attribute: Type of Building, which is obtained from the USC's GHG report. The year of construction for the buildings used in our experiments varies from 1919 to 2006.
3) *Occupancy Patterns*: To incorporate occupancy related information, we used three categorical attributes: *day of the week, semester, and holiday*, which are indirect indicators of energy use in a campus environment and publicly available from the universitys academic calendar for the past three years. To highlight the relevance of this information, Figure 1 shows the fall in demand for every weekend across the whole year while Figure 7 shows the fall in demand during the summer months for a residential building when only a few students remain on campus. It is also evident from the plots that energy demand falls during the holidays, such as the Spring Break in March and Thanksgiving in November.

[2]The data is available for use in academic projects on request from the university's Sustainability Office.

[3]http://www.wunderground.com

[4]http://fmsmaps.usc.edu/mapguide6/upcmaps/Web/cfm/bl_list_no.cfm

C. Model Training and Testing Framework

We build this model using MATLABs Statistics toolbox package. We process and load the input training and test data values as comma separated value (csv) files and the output prediction results are also saved to csv files. The scripts take between a few seconds to less than 2 minutes to complete the training and testing. We use MATLABs classregtree function that builds regression trees and provides numerical output as required for our experiments. The classregtree function creates a decision tree based on the training dataset and predicts response values using regression. The tree is a binary tree and each decision node evaluates an attribute of the data. The scripts are run on a Macbook Pro laptop with a 2.03GHz processor and 4GB RAM, running OS X. We use MATLAB version R2010a.

Generally, the performance of regression methods is evaluated using the coefficient of variation of the root mean squared error (CV-RMSE) and mean bias error (MBE) [4, 11]. The CV-RMSE is similar to an R^2 error and gives a measure of the degree of scatter. It is given by:

$$CV - RMSE = \frac{\sqrt{\sum_{i=1}^{n} \frac{(o_i - p_i)^2}{n}}}{\overline{o}}$$

where, o_i is the i^{th} observed value, p_i is the i^{th} predicted value, n is the number of values being predicted, and o is the mean of the n observed values.

The MBE gives a measure of the bias in the model. A positive MBE indicates a model that tends to over-predict while a negative MBE indicates under-prediction. We use CV-RMSE as a measure of the precision of our model and also study the role of bias in our models.

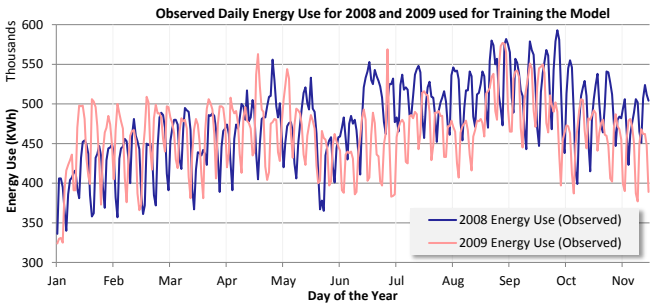

Figure 1. Campus daily energy use observed for 2008 and 2009. Used as training data for model.

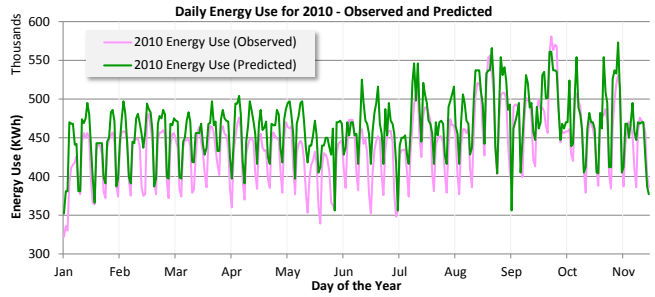

Figure 2. Campus daily energy use for 2010. Both observed and predicted values are shown. CV-RMSE =7.45%, r = 0.816, $p < 0.001$.

V. Experiments

A. Campus-Scale Daily Energy Use Model

The campus-scale daily energy use model is used to predict the energy consumed (KWh) each day for the entire campus. This is useful for long term planning of energy use (e.g. pricing levels to sign up with the electric utility, purchasing energy storage, etc.) and for day ahead planning (e.g. provisioning backup generators). The model uses Day of week, Semester, Maximum Temperature, and Holiday as features of a given day for the prediction. The model is trained using these features and the observed daily energy usage for the years 2008 and 2009, and it is used to predict the daily energy usage for 2010. Fig. 1 plots the daily energy usage for 2008 and 2009 used for training, and Fig. 2 shows the observed and predicted energy use for 2010 (till Nov, 2010 for which data was available).

As we can see from the training data, a cyclic high and low energy use trend is observed during the weekdays (Mon-Fri) and the weekends (Sat-Sun), and similarly during the semester (Jan-May, Aug-Dec) and during the summer and winter breaks. However, there is significant variability in the energy used even for similar weeks of the year in 2008 and 2009. For e.g., there is higher energy consumption during the month of Jan, 2009 compared to Jan, 2008, and lower during the month of Jun-2009 compared to Jun-2008.

The prediction from the regression tree model we train closely matches with the observed energy usage. The X-Y scatter plot between observed and predicted values shows

Table I
PREDICTION ERRORS FOR BASELINES AND REGRESSION TREE MODELS FOR DAILY CAMPUS ENERGY USE FORECAST

Model Used	CV-RMSE
Annual Mean	11.32%
Day of Week Mean	14.39%
Day of Year Mean	12.62%
Regression Tree	7.45%

a Pearsons correlation coefficient r = 0.816, a confidence interval greater than 99.99% for a two-tailed test (i.e. $p < 0.001$), and a slope of 1.051 when the intercept is set to (0,0). The CV-RMSE for the prediction is 7.45%.

1) Comparison with Baseline Models: We compare our regression tree model against three baseline models that are commonly used and are constructed using the historical energy use values: annual mean, day of week mean (DoW) and day of year (DoY) mean. Annual mean is a single number that is the average of the daily energy use on all days in 2008 and 2009. This value is: 11.32%. DoW mean averages the energy use for each day of the week (Sun, Mon, ..., Sat) across all two years, and uses the resulting 7 values as the predictors. DoY mean averages each day in the year (1...365) across the two years and uses these 365 values as the predictor. Table I shows the CV-RMSE of the predictions made by the baseline models and our regression tree model.

As we can see, the regression tree model performs significantly better than the other models and reduces the CV-RMSE of the next best model (Annual mean) by 34%. The reason the other model make reasonable predictions can also be explained. DoY mean captures the energy use variability due to the seasonal temperature, and may thus be able to partially capture the "Temperature" feature used in our regression tree model. It can also partially capture the "Semester" variation and to a limited extent, the "Holiday" feature, when they occur on the dame day each year. The DoY mean captures the "Day of Week" feature we use and can account for the class schedules since the class pattern within a week repeats across all weeks of the semester. But the schedule change across semesters is not capture. By explicitly adding these features to our regression tree model, we are able to better predict the daily energy use that is affected by them.

2) Discussion of Outliers: There are several outliers in the prediction that can be interpreted using available knowledge of the domain. Some of the outliers in prediction can be explained in terms of certain events/dates of the year, which mainly deal with holidays and transitions between semesters. For instance, if we look at the top two absolute prediction errors at the campus level, they occurred on 30 May, 2010 and 5 Sept, 2010, which precede Memorial Day and Labor Day holidays, respectively. Even though the holiday information is captured as a feature of our model, the prediction error is high. This indicates additional handling

Table II
PREDICTION ERRORS FOR BASELINES AND REGRESSION TREE MODELS FOR DAILY CAMPUS ENERGY USE FORECAST

Weekday	Semester	Temperature	Holiday	CV-RMSE
♦	♦	♦	♦	7.40%
♦		♦	♦	7.60%
♦		♦		7.95%
♦	♦	♦		8.05%
♦	♦		♦	8.37%
♦	♦			8.54%
♦				8.86%
		♦		10.48%
	♦			11.05%
			♦	11.54%

that may be necessary for holidays and transition points within features.

3) Relative Impact of Features: One novel aspect of our modeling approach is to use indirect indicators of energy use that are unique to a campus environment to help make better predictions. We evaluate the impact that each of these features has on the accuracy of model prediction by training and testing the regression tree model for all combinations of the features and comparing their CV-RMSE values. Table II highlights some of these 15 combinations that were trained and tested, sorted in the order of lowest to highest errors. Rows shaded gray indicate the use of just a single feature by the model. The row that is underlined indicates the use of all features in the model.

We see that the use of the day of the week feature provides consistently lower error values by the models, making it the most important feature that affects the daily energy use prediction. In fact, using weekday just by itself in the model provides an error of 8.86%, which is only greater by 1.46% than a prediction model that uses all four features with an error of 7.40%. This is understandable given the campus environment where weekends and weekdays have sharply different energy use profiles. The next important feature that lowers the error is temperature a hot day ends up causing higher energy use due to increased effort by A/C units in cooling the campus buildings.

While temperature has been conventionally used in energy use forecast models, we find that the use of additional indirect indicators by our model reduces the error when just using temperature from 10.48% to 7.40% a 29% improvement in error. However, temperature is still a useful feature for the daily energy use predictions since it improves a model that does not use temperate, but all other features, from an error of 8.37% to 7.40%. We see that the use of semester and holiday attributes also help improve a model.

4) Possible Corrections for Over Estimation: Of the 325 days we predict for the year 2010 using our machine learnt regression tree model, we observe that the predictions are predominantly overestimating the daily energy use for 279 days and underestimating it for 46 days. This indicates that the regression tree model has potential of improvement. As a naive attempt at improving the model by reducing the overestimation skew, we apply a negative correction factor to the predicted results as a post processing step of the model prediction. We use two methods: one reduces the predicted energy use value by a constant value of 24,000KWh that was chosen by trial and error, while the other uses a relative correction that reduces each predicted value by 7.40% that corresponds to the CV-RMSE that is seen. (For context, the daily energy use average that is observed is 460,154KWh over the 3 years.)

We see that the constant correction of -24,000KWh reduces the CV-RMSE to 5.10%, with over and underestimations becoming 152 and 166, while using a fractional correction of -7.40% gives a CV-RMSE of 5.63%, with over and underestimations becoming 94 and 231. While these were ad-hoc methods to improve the prediction, it suggests that the machine learnt regression tree model that we use can be further improved by automatically determining constant or relative corrections to account for over or underestimation skews.

B. Campus-Scale 15-min Energy Use Model

The campus-scale 15-min energy use model is used to predict the energy consumed during short intervals of time during a day. Such predictions are useful for programming the duty cycles of A/C and ventilator units in a building management system or responding to load curtailment signals sent to the campus by a power utility. The model uses Day of Week, Semester, Maximum Temperature, and Holiday features as in the daily energy use model discussed before, and in addition adds Humidity for the campus as a feature. The model is trained on these features using 15-min interval data from the years 2008 and 2009, and used to predict the 15-min energy use for the year 2010. Figure 3 plots the observed and predicted energy usage at 15-min intervals for the first week 2010, with 96 values per day. We omit showing the full years prediction for brevity. Instead, the X-Y scatter plot of all observed and predicted 15-min energy use values for the year 2010 (till Nov, 2010 for which data was available) are plotted in Figure 4.

The observed energy use in Figure 3 shows the energy use cycling between low, high and low from midnight to noon and back to midnight. We also see that the holiday (1 Jan, 2010) and weekend (2-3 Jan, 2010) consumer lower energy than the weekdays (4-7 Jan, 2010). There are also sharp spikes in energy use just before midnight on the weekdays that indicate pre-cooling of chilled storage units to make use of lower pricing at nights.

The 15-min energy use predictions made by our regression tree model tracks the observed values. The Pearsons correlation coefficient between observed and predicted values is r = 0.637, with a confidence interval greater than 99.99% for a two-tailed test (i.e. $p < 0.001$), and a slope of 1.069

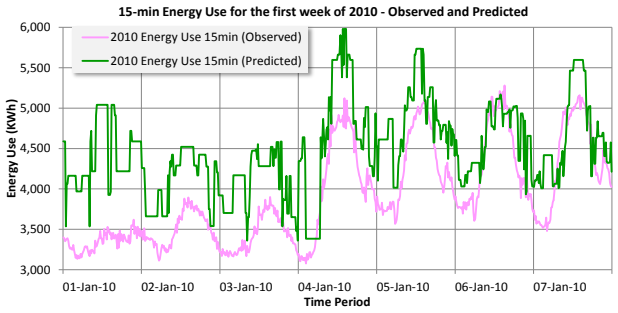

Figure 3. Campus 15-min energy use for the first week of 2010. Both observed and predicted values are shown. CV-RMSE =13.70%, r = 0.637, $p < 0.001$.

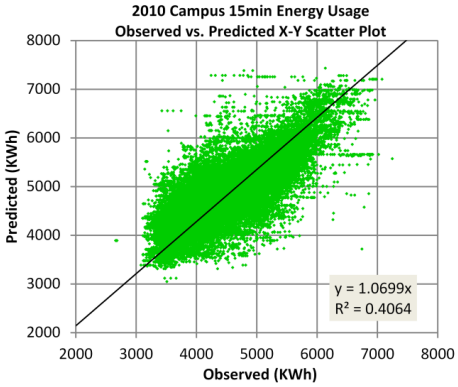

Figure 4. X-Y scatter plot of observed and predicted 15-min energy use values for Jan-Nov, 2010. 31,200 points are plotted, with a trend line set to pass at (0,0).

when the intercept is set to (0,0). The CV-RMSE for the prediction is 13.70%.

1) Comparison with Baseline: The regression tree model that we train is compared with three baseline models that just use historical energy usage information to make predictions. The annual 15-min mean model uses the average energy used over all 15-min periods for 2008 and 2009 as a single value prediction. The Time of Week (ToW) model uses the average energy used during each 15-min time period in a week as a predictor (7 days * 24 hours * 4 periods). The Time of Year (ToY) model uses the average time during each distinct 15-min time period in the year, averaged over two years (365 days * 24 hours * 4 periods) as a prediction for the corresponding time periods in future years. Table III shows the CV-RMSE of the predictions made by the baseline models and our regression tree model.

The regression model we train performs relatively better than the baseline models with an improvement in the absolute errors of between 1.37% and 3.67%. The baseline models also perform respectably. Intuitively, the ToW model is able to account for the weekend/weekday feature used in our regression tree model, as well as the variation in temperature during the day. Similar days of the week and time periods during these days will correspond to similar classes being scheduled and rooms being occupied. Likewise, the

Table III
PREDICTION ERRORS FOR BASELINES AND REGRESSION TREE MODELS FOR 15-MIN CAMPUS ENERGY USE FORECAST

Model Used	CV-RMSE
Annual 15-min Mean	17.37%
Time of Week Mean	16.00%
Time of Year Mean	15.07%
Regression Tree	13.70%

temperature variation during the day is also a function of the time of the day. The ToY model can capture the seasonal as well as daily temperature variations, as also the holidays that occur on the same days across different years (e.g. July 4) and the semester. This logical overlap with most features that is considered by our regression tree model causes the ToY model to better the ToW and Annual mean models, and fall close to the performance of our model. As we shall see, the use of Humidity in our model actually reduces our model prediction power: using attributes other than humidity allows us to reduce the error to 13.02%, which is better than the ToY model by 13.6

2) Discussion of Outliers: We plot the absolute error percentage (i.e *(observed - predicted)/ observed*) for each 15-min energy use prediction by our model against the actual observations for the year 2010 in Figure 5 using a "radar" chart for compact representation. This shows the clustering of errors and helps analyze unique causes of errors. As can be seen, most of the errors fall within the 20% error circle. There are a number of places where the error values for concurrent time periods increase (or decrease) sharply. These can be seen as streaks of dots moving away from the center, like the spokes of a wheel. We can correlate these with the actual and predicted values (e.g. in Figure 3) and see that these are regions where our regression tree model provides a single value for a period of time while the actual energy use during that period is increasing (or decreasing). This is illustrated for the time period around noon of 2 Jan, 2010 in Figure 3 where the prediction is flat at 4500KWh while the actual observation is increasing from 2500KWh to 3900KWh and down again. This corresponds to the streak at the 12'o clock position in Figure 5 that reaches the 40% error circle.

There are other outliers that are the result of the predictions failing at crossover points in the parameter space.

3) Relative Impact of Features: The five features used by the regression tree model for predicting 15-min energy use contribute disproportionately to the prediction error of the model. We analyze their relative impact by training and testing the regression tree model using all combinations of the five features, as before, using years 2008 and 2009 as training data and year 2010 as test. Table IV shows some of the key results from the 31 models that were built, sorted by low to high CV-RMSE values. The rows that are shaded in

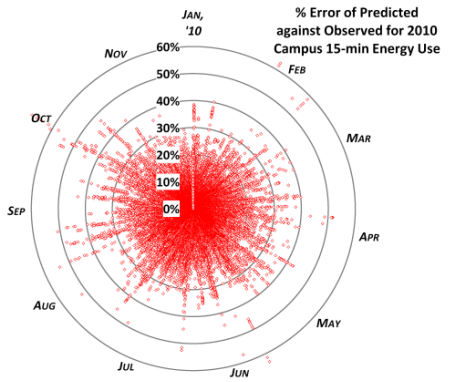

Figure 5. Absolute error % of all the campus 15-min energy use predictions for 2010 as compared to their observations. Circumference is the time axis while the radius of the circles is the absolute error %. 31,000 points are shown.

gray represent those using just a single feature while the row with the error underlined is the model that uses all features. Temperature is the most consequential feature and appears consistently in the models with the least error. While the temperature used in the campus daily energy use model was the maximum daily temperature, the temperature used here is the hourly maximum temperature. It is also, along with humidity, the feature whose value changes the most within a day as the others are constant for a given day. So it proves effective in capturing the energy use variation within a day.

Weekday is the next best predictor for reasons similar to the campus daily energy use model. In fact, just weekday and temperature are able to provide a marginally better prediction than all the features together. But this is due to the fact that humidity has a negative influence on the model prediction. This is unlike the campus-scale daily energy use model where adding a feature improved the model prediction. We see a CV-RMSE of 13.02% using features other than humidity while it is 13.70% when using it. This disproves our initial expectation that humidity will affect campus energy use.

When compared to traditional energy use prediction models that reply exclusively on temperature for forecasting, our use of the additional indirect indicators weekday, semester and holiday helps reduce the error from 14.87% to 13.03%.

4) Possible Corrections for Over Estimation: We make some simple attempts to examine if the regression tree model can be improved to provide more accurate results. Specifically, of the 31,200 15-min time periods we predict the energy use for in 2010, we observe that the predictions are predominantly overestimating for 24,667 periods and underestimating for 6,531 periods. The observed values have a mean of 4,535KWh for each 15-min period in 2010. Subtracting a constant correction value of 360KWh from all prediction results reduces the CV-RMSE from 13.70% to 11.34%, provides a more even distribution of over and underestimations of 15,300 and 15,897, and improves the correlation coefficient r from 0.637 to a more accurate 0.679 leaving the precision unchanged. While this is an arbitrary correction value arrived at by trial and error, it offers potential for improving the model we use.

Table IV
PREDICTION ERRORS FOR BASELINES AND REGRESSION TREE MODELS FOR 15-MIN CAMPUS ENERGY USE FORECAST

Weekday	Semester	Temperature	Humidity	Holiday	CV-RMSE
♦	♦	♦		♦	13.02%
♦		♦		♦	13.30%
♦	♦	♦			13.39%
♦		♦			13.57%
♦	♦	♦	♦	♦	13.70%
♦		♦	♦	♦	14.22%
♦	♦	♦	♦		14.42%
	♦	♦		♦	14.65%
		♦		♦	14.72%
			♦		14.87%
♦					15.96%
♦	♦		♦	♦	16.59%
				♦	16.69%
	♦				17.30%
				♦	17.63%

C. Building-Scale Daily Energy Use Models

Building-scale daily energy use forecast models help identify total power consumed by individual buildings on campus. Such models can be used by facility managers to plan building level operations such as pre-cooling of rooms before working hours to minimize peak power usage, for scheduling classes to ensure uniform energy use, and identify buildings that have low carbon footprint per occupant.

We train the regression tree model for building-scale daily energy use prediction using the following features: Weekday, Semester, Holiday, (Maximum) Temperature, Average temperature, Gross Building Area, Net Area in Use, Year of Construction, and the Type of building. We use data for 23 buildings on campus for 2008 and 2009 for training and test it on data from 2010 (Jan-Nov). We build a single (global) model for all these 23 buildings. The CV-RMSE for the daily energy use predictions by our model is shown in Table VI under Global Model, while the plot of observations and predictions for 3 buildings, one each of academic (OHE), residential (DMT) and other (PSB) type, are shown in Figures 6 - 8.

1) Comparison with Baseline: We use three baseline models for evaluating the prediction power of our regression tree model. These are similar to the baseline models used for the campus daily energy use, with the difference being we use the means over each building rather than the entire campus. For brevity, we examine results for three buildings.

Table V lists the CV-RMSE for the predictions using the baseline and regression tree models for the three buildings. The regression tree model outperforms the baseline models

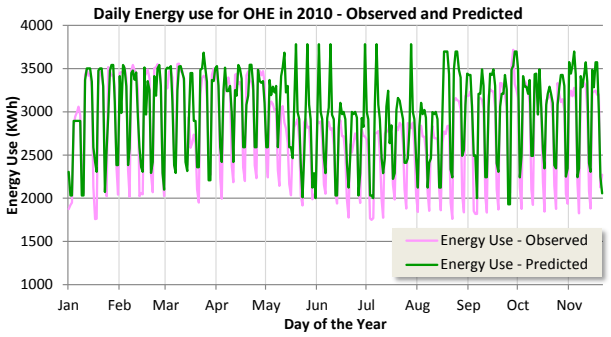

Figure 6. Daily energy use for the OHE academic building for Jan-Nov 2010. Both observed and global model predicted values are shown.

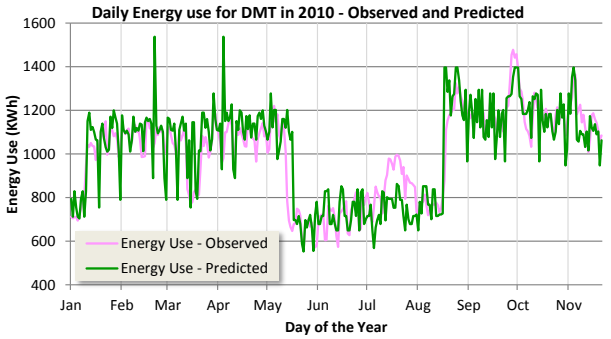

Figure 7. Daily energy use for the DMT residential building for Jan-Nov 2010. Both observed and global model predicted values are shown.

for the residential building, OHE, recording an 8% or better difference in absolute error compared to the next best baseline model. The improvement is less marked for the other building type, PSB, but still better at 19.32% as compared to next best at 23.39%. Despite the building features being intrinsic to the baseline model (since each model is for a specific building), the use of the indirect indicators of weekday, holiday, semester and temperatures help the regression tree model outperform the baselines. The only exception is the accuracy of prediction for the residential building, DMT, where the day of the year performs better at 9.62% error than our global regression tree at 11.77% error.

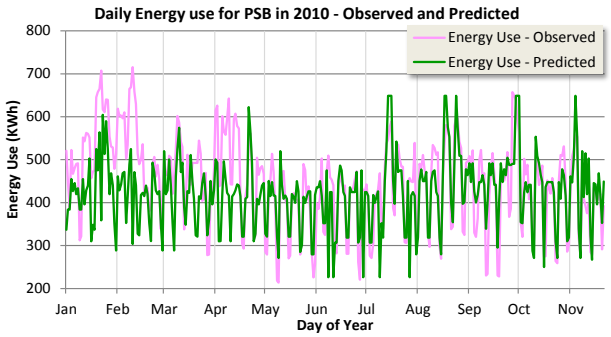

Figure 8. Daily energy use for the PSB other building for Jan-Nov 2010. Both observed and global model predicted values are shown.

Table V
PREDICTION ERRORS FOR BASELINES AND REGRESSION TREE MODELS FOR DAILY BUILDING ENERGY USE FORECAST

	OHE	DMT	PSB
Annual Mean	20.55	19.77	23.39
Day of Week	26.13	20.09	23.65
Day of Year	24.64	9.62	27. 48
Regression Tree (Global)	12.09	11. 77	19.32

This may be attributed to the fact that the baseline model is specific to the DMT building while the regression tree model cannot differentiate between residential buildings that may vary in their energy use.

2) Discussion of Outliers: There are several outliers in the prediction that can be explained based on additional domain insight. For example, if we look at the absolute prediction errors for the residential category, many of them occur in the period 13-16 May, 2010 which marks the transition between Spring and Summer semesters. Also, in the academic category, among the top prediction errors is again a period of transition between the Summer and Fall semesters (17-18 Aug).

3) Comparison of global and local model for each building type: We build a single regression tree model (the Global Model) that works for all 23 campus buildings to perform daily energy use prediction for individual buildings. However, recognizing that building of different types (i.e. academic, residential and other) may have unique characteristics, we also build one regression tree model (the Local Model) for each building type for comparison. Note that the global building model does include building type as a feature. But the local model in a way forces the decision tree to be partitioned at the root based on the building type. Table VI lists the performance of the Local and the Global Models.

We see that the global model performs better than the local model in a majority of cases (14 of the 22 buildings). This is particularly true for the academic and the other buildings where the local model performs better for only one in three cases. This indicates less homogeneity within the academic and other building types to warrant promoting building type to the root of the regression tree. This can be attested to given the widely varying sizes of the academic buildings and the different uses for the other building category, ranging from a library to parking garages. Thus, a local model may not necessarily be suitable based on the building type. However, there may be other aspects of the buildings that may motivate the need for local models for a specific class of buildings for better performance.

VI. Conclusion

We describe energy forecast models for a campus microgrid based on machine learnt regression tree models. Besides

Table VI
COMPARISON OF PREDICTION ERRORS FOR GLOBAL AND LOCAL BUILDING MODELS

Building Code	Type	CV-RMSE (Local Model)	CV-RMSE (Global Model)
ASC	Academic	11.65%	12.03%
EEB	Academic	8.46%	9.19%
JKP	Academic	13.63%	13.48%
LAW	Academic	8.39%	8.20%
OHE	Academic	12.63%	12.09%
RTH	Academic	45.48%	5.33%
THH	Academic	101.42%	22.28%
VKC	Academic	25.33%	26.26%
WPH	Academic	51.47%	30.50%
DMT	Residential	11.07%	11.77%
DXM	Residential	11.08%	11.75%
FLT	Residential	28.44%	29.55%
IRC	Residential	43.06%	38.59%
PTD	Residential	548.44%	26.67%
TRO	Residential	2.57e+4%	11.54%
ADM	Other	44.28%	44.22%
DML	Other	18.32%	20.43%
LRC	Other	19.70%	20.09%
PRB	Other	24.86%	9.38%
PSA	Other	120.18%	11.68%
PSB	Other	222.59%	19.32%
PSD	Other	61.79%	3.64%
PSX	Other	39.01%	5.31%

just energy use trends and common features used for forecast such as temperature, we introduce indirect energy use indicators such as the academic calendar and building attributes to provide better predictions.

We observe that there is scope for improving the models to provide more accurate and precise results. Our naive approaches for correcting overestimation of energy use values show potential for advances. Exploratory work on using ANN models also show promise, though their use will be limited to better forecasting rather than trying to understand the causes for energy use that a regression trees structure would provide.

As smart meters and other ambient sensors are widely deployed, mining the data collected by them for monitoring and forecasting of the energy footprint will be important to ensure energy conservation. Our regression tree prediction models demonstrate their usefulness for building and campus energy management to optimize the buildings daily operation, schedule its use for a semester, and effectively design and implement university energy policies. These models can also be applied to buildings in the design stage to perform cost and energy use trade-offs between constructing new buildings and retrofitting existing ones to make them energy-efficient. The goal of this exercise is to eventually provide machine learnt models that operate on a city scale and can dynamically adapt to changing conditions.

ACKNOWLEDGMENT

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000192. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

[1] FERS assessment of demand response and advanced metering, staff report, 2008.

[2] US EIA Annual Energy Review, 2009.

[3] University Greenhouse Gas Emissions: 2001-2009 Report, USC Office of Sustainability. USC, Los Angeles, CA, 2010.

[4] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng. Occupancy-driven energy management for smart building automation. In *Buildsys 2010 - 2nd ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Buildings*, 2010.

[5] B. Dong, C. Cao, and S. E. Lee. Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, 37:545–553, 2005.

[6] J. Laustsen. Energy efficiency requirements in building codes. In *Energy Efficiency Policies for New Buildings, IEA Information Paper*. International Energy Agency, 2008.

[7] S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann, and W. Rakowski. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. In *Annals of Behavioral Medicine*, volume 26, pages 172–181, 2003.

[8] J. Michaelsen, D. S. Schimel, M. A. Friedl, F. W. Davis, and R. C. Dubayah. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *Journal of Vegetarian Science*, 2009.

[9] T. Olofsson and S. Andersson. Long-term energy demand predictions based on short-term measured data. *Energy and Buildings*, 33(2):85–91, January 2001.

[10] J. J. Rothwella, M. N. Futterb, and N. B. Disea. A classification and regression tree model of controls on dissolved inorganic nitrogen leaching from european forests. *Environmental Pollution*, 156(2), 2008.

[11] Y. Simmhan, V. Prasanna, S. Aman, S. Natarajan, W. Yin, and Q. Zhou. Towards data-driven demand-response optimization in a campus microgrid. In *ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Buildings (BuildSys)*, 2011.

[12] M. Yalcintas. Energy-savings predictions for building- equipment retrofits. *Energy and Buildings*, 40(12):2111–2120, 2008.

[13] J. Yang, R. H., and R. Zmeureanu. On-line building energy prediction using adaptive artificial neural networks. *Energy and Buildings*, 37(12):1250–1259, December 2005.

[14] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino. A decision tree method for building energy demand modeling. *Energy and Buildings*, 42(10):1637–1646, 2010.