A Large-scale Study on Predicting and Contextualizing Building Energy Usage

J. Zico Kolter
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139

Joseph Ferreira Jr.
Department of Urban Studies and Planning
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

In this paper we present a data-driven approach to modeling end user energy consumption in residential and commercial buildings. Our model is based upon a data set of monthly electricity and gas bills, collected by a utility over the course of several years, for approximately 6,500 buildings in Cambridge, MA. In addition, we use publicly available tax assessor records and geographical survey information to determine corresponding features for the buildings. Using both parametric and non-parametric learning methods, we learn models that predict distributions over energy usage based upon these features, and use these models to develop two end-user systems. For utilities or authorized institutions (those who may obtain access to the full data) we provide a system that visualizes energy consumption for each building in the city; this allows companies to quickly identify outliers (buildings which use much more energy than expected even after conditioning on the relevant predictors), for instance allowing them to target homes for potential retrofits or tiered pricing schemes. For other end users, we provide an interface for entering their own electricity and gas usage, along with basic information about their home, to determine how their consumption compares to that of similar buildings as predicted by our model. Merely allowing users to contextualize their consumption in this way, relating it to the consumption in similar buildings, can itself produce behavior changes to significantly reduce consumption.

Introduction

In the effort to build a sustainable society, energy issues play a crucial role. Humans consume an average of more than 16 terrawatts of power and growing, 86% of which comes from (unsustainable) fossil fuels (Multiple 2009). In the United States, 41% of all energy is consumed in residential and commercial buildings, mainly in the forms of electricity and natural gas. Reducing these consumptions in particular will play a large role in reducing our overall energy dependence.

As a whole, however, end users receive relatively little information about their energy usage. Most of the feedback we receive about our energy consumption comes via monthly electricity and gas bills, which provide little information or context to our usage other than a dollar amount. Crucially, energy bills provide no mechanism to determine things such as how a building's consumption compares to similar buildings, whether retrofits or upgraded appliances are financially reasonable, what portion of the consumption is simply due to location (cold locations will inevitably require more heating, for example) versus personal behavior, etc. Moreover, simply providing people with feedback about their energy use can itself produce behavior changes that significantly reduce energy consumption (Darby 2006; Neenan and Robinson 2009). Recent research has specifically highlighted the value of normative energy feedback, showing users how their usage relates to that of their peers and neighbors (Cialdini and Schultz 2004; Allcott 2010).

In this paper, we present a study of energy consumption in a large number of buildings in Cambridge, Massachusetts, and develop models than can begin to answer such questions. We use a data set consisting of monthly electrical and gas bills, collected by a utility, for approximately 6,500 separate buildings in the city. We integrate this usage information with publicly available tax assessor and geographical survey information to correlate energy usage with features such as living area, building value, building type, etc,. We perform a preliminary analysis of the data set, focusing on the value of using logarithmic scaling for energy usage, and we use feature selection methods to determine the most relevant features for predicting consumption. We then apply both parametric and non-parametric learning algorithms to predict distributions over building consumption. Finally, we use these models to develop EnergyView, a system that allows both utilities (at a city-wide scale) and end users (at a single building scale) to view and compare their energy usage to that of similar buildings as predicted by the models.

Related Work

This paper builds upon a number of works, both from the energy sector and the statistics and machine learning communities. Since this is primarily and application and analysis paper, we mainly use existing algorithmic approaches, but with a focus on recently developed methods for regression under non-Gaussian likelihoods (Kotz and Nadarajah 2004; Vanhatalo, Jylanki, and Vehtari 2009).

From the energy community, our work builds most di-

rectly upon the studies, mentioned above, that highlight the importance of normative feedback for improving energy efficiency (Cialdini and Schultz 2004; Allcott 2010). Indeed, we are aware of several companies that work in this area, but as they do not share the details of their models or data, it is difficult to know what methods they employ.

Several academic studies exist that examine individual home residential energy usage at a high resolution, (e.g., (Berges et al. 2009)), but this work is roughly orthogonal to this paper, as we here consider much lower resolution data, but for many more houses. Likewise, there exist several studies on building energy consumption at a highly aggregate level (Berry 2005), but again these differ from this work as we consider a data set obtained directly from a large number of individual buildings.

Thus, in relation to this past work, this paper makes several contributions. We analyze a large-scale real-world energy usage data set, illustrate several interesting characteristics of the data, use recent machine learning techniques to develop predictive models, and present a public end-user interface for obtaining contextual information about ones own energy use. To the best of our knowledge, this represents one of the largest-scale, publicly-available studies of its kind, conducted on real data, and the EnergyView tool represents one of the first tools of its kind where the algorithms behind its predictions are fully described.

Data Collection and Analysis

The primary data set we build our model upon is a collection of monthly electricity and gas bills, collected over several years from buildings in Cambridge, MA, and obtained from NStar, the electricity and gas utility in Cambridge. The data consists of electricity and gas account numbers, their corresponding street addresses, and monthly electricity and gas meter readings for each account over some period of time, typically two to three years. Electricity usage is given in kilowatt-hours per month, while gas usage is given in therms per month. To convert these to equivalent units, we use the conversion factor $(1/3)29.3$ therms/kwh, where 29.3 is the standard conversion rate, and 2/3 represents the average conversion loss for electricity generation in the United States; this also corresponds to the relative pricing of gas and electricity, an average of \$1.05/therm and \$0.11/kwh respectively (Multiple 2009). The primary goal of our algorithms will be to predict this total usage as a function of building features (though we also apply the same techniques to estimating monthly electricity and gas usage separately in order to provide more detailed feedback).

In addition to the energy data, we use publicly available tax assessor records[1] and a Geographic Information System (GIS) database[2] compiled by the city. The tax assessor records contain detailed features about every registered property and building in Cambridge, listing for each address features such as the value of the building, the property class (condominium, single family home, retail store, etc), the square footage, the year a building was built, and other similar features. The GIS database consists of polygonal outlines for parcels and buildings in the city, plus estimated roof heights for buildings (obtained via an aerial lidar scan). We aggregate features at the *whole building* level, so that, for example, multiple units at a given street address are aggregated into one building; this is a necessity since the utility data does not include unit numbers, but only a street number and street name. After correlating addresses between the energy usage, tax assessor, and GIS databases, and removing any entries for which there are less than a year of full gas and electricity readings, there are a total of 6,499 unique buildings that we include in our data set. This represents slightly more than half of the 12,792 unique addresses in the Cambridge tax assessor records (the main reason for omitting a building is that one or more of the correspond energy accounts don't span a full uninterrupted year in the utility data we have access to).

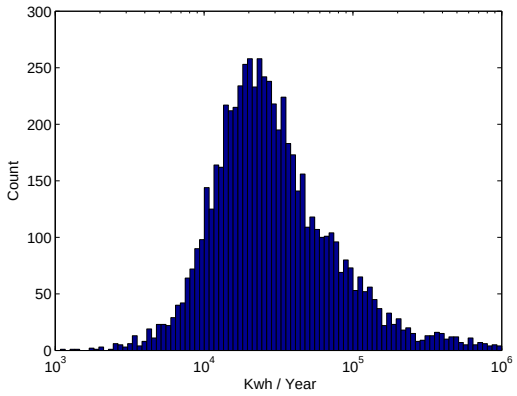

Figure 1: A histogram of total energy consumption per building, for the 6,499 buildings considered. Note the logarithmic scale on the x axis.

Logarithmic Energy Scaling

A preliminary analysis of the data illustrates several interesting features, which we build upon to develop models in the subsequent section. Figure 1 shows a histogram of total energy consumption per year for the different buildings in our data set. The data appears roughly Gaussian (although slightly skewed), but the important fact here is that the x axis is *logarithmic*, implying that total energy consumption roughly follows a log-normal distribution. This means that energy consumption varies *multiplicatively* between different houses: a house in the 80th percentile of energy consumption uses about 3 times as much energy as a house in the 20th percentile. Further, even the log-normal distribution in fact underestimates the spread of the data; the data is heavy-tailed, such that a log Student-t distribution (Cassidy,

[1]Tax assessor records for Cambridge are available via a web interface at `http://www2.cambridgema.gov/fiscalaffairs/PropertySearch.cfm`. We know of no source where the database can be downloaded directly, and instead had to use an automated tool to parse and import the records via this interface.

[2]GIS databases for Cambridge can be ordered at `http://www.cambridgema.gov/gis.aspx`.

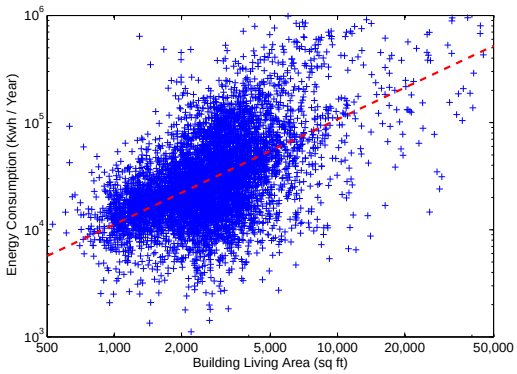

Figure 2: A plot of building square footage versus total yearly energy consumption, both in logarithmic scales. The dotted line shows the least-squares fit.

Hamp, and Ouyed 2010) fits the energy consumption better, as we will show quantitatively in the next section.

Approximately log-normal distributions of energy usage may initially seem rather surprising, but these are actually quite reasonable given other phenomena in economics and social science. In particular, many observed phenomena naturally follow roughly log-normal distributions (Limpert, Stahel, and Abbt 2001), including factors that would be expected to influence energy consumption, such as income, property sizes, and (in our own data set) building square footage. Furthermore, *power-law* behaviors, defined as a linear relationships between the logarithms of input and output variables (Newman 2005), are ubiquitous in many empirical domains, including domains related to city consumption and scaling (Kuhnert, Helbing, and West 2006). Since the Gaussian and multivariate T distributions are preserved by linear transformations (Kotz and Nadarajah 2004) (i.e., a linear function of a multivariate Gaussian or multivariate T distribution will also be Gaussian or multivariate T), log-normal or log-T input variables, together with a power-law relationship, would imply the same distribution for the output. Figure 2 illustrates this type of relationship for our data set, plotting building living area as square footage against total energy consumption on a log-log scale. While there is naturally a great deal of noise, there is also a fairly clear linear relationship, indicating an intuitive power-law relationship between square footage and energy consumption. While the relationships are more difficult to display when additional variates are introduced, the chief goal of the next section will be to exploit these types of relationships to derive predictive models of energy consumption based upon known features.

Data Features and Feature Selection

Finally, before discussing the modeling methods we use for this data, we discuss the actual features used to predict energy consumption, as well as feature selection procedures that identify the most relevant features for this task. We extract a total of 35 features from the tax assessor and GIS data set, all of which are shown in Table 1. Many of the features

Feature	% RMSE Reduction	Individual Correlation
Building Value	*24.721 %*	0.659
Num Electric Meters	*13.438 %*	0.647
Property Class	*12.729 %*	0.633
Living Area	*2.760 %*	0.611
Num Gas Meters	*2.517 %*	0.626
Heat Fuel	*2.241 %*	0.480
Building Style	*1.432 %*	0.632
Heat Type	*0.826 %*	0.431
Central AC	*0.749 %*	0.558
Overall Grade	-0.045 %	0.346
Roof Material	0.105 %	0.595
Exterior Wall Type	-0.300 %	0.597
Occupancy Type	-0.656 %	0.623
Laplacian Eigenvectors	0.138 %	0.180
GIS Parcel Area	0.164 %	0.381
Total Rooms	0.074 %	0.049
Residential Exemption	0.037 %	0.207
Fireplaces	0.014 %	0.092
Overall Condition	-0.068 %	0.234
Full Baths	0.000 %	0.101
Roof Type	-0.024 %	0.557
Assessed Value	-0.010 %	0.615
GIS Parcel Perimeter	0.001 %	0.371
Land Area	-0.007 %	0.352
Land Value	-0.202 %	0.377
Half Baths	-0.023 %	0.022
Garage Parking	-0.039 %	0.229
Year Built	-0.013 %	0.095
GIS Building Perimeter	-0.014 %	0.104
GIS Building Area	0.033 %	0.143
GIS Building Volume	0.016 %	0.112
Num Stories	-0.041 %	0.064
Open Parking	-0.011 %	0.069
Covered Parking	-0.013 %	0.071
Residential/Commercial	-0.553 %	0.272

Table 1: All the building features extracted from the tax assessor and GIS data, ranked in the order that they are selected in greedy forward feature selection. The first nine features (above the line) decrease cross-validation RMSE (second column), to a statistically significant degree ($p < 0.01$ in a pairwise t-test). Also shown is the correlation coefficient between total energy and the feature in isolation.

are real-valued, in which case we include their logarithm (owing to the discussion above) directly in the final feature vector. For discrete features, we use a standard binary encoding of the feature, i.e., for a k-valued discrete value we add the representation

$$\phi(x_i) \in \{0,1\}^k = (\mathbf{1}\{x_i = 1\}, \ldots, \mathbf{1}\{x_i = k\}) \qquad (1)$$

to the feature vector, where all the entries are zero if a discrete value is previously unseen or if the feature value is unknown. All the features listed in Table 1 are extracted directly from the GIS or tax assessor database with the exception of the "Laplacian Eigenvectors" feature, which

uses a standard spectral clustering procedure (Chung 1997). Briefly, we construct this feature by building a (symmetric) k-nn graph on the 2D building locations, then looking at the principle eigenvectors (here corresponding to the *lowest* eigenvalues) of the discrete *normalized Laplacian* operator on this graph, defined as

$$\mathcal{L} = I - D^{-1/2} A D^{-1/2} \tag{2}$$

where D is a diagonal matrix of node degrees in the graph and A is an adjacency matrix. These features provide an optimal orthogonal basis for reconstructing smooth functions on the graph's manifold, and thus in our setting correspond to smoothly spatially varying functions (in terms of the arrangement of the building locations).

Since we are in a data-rich setting, where we have significantly more features than examples, it is possible to simply pass all these features to our learning algorithms; even the algorithms we consider, which can in theory be sensitive to irrelevant features, do not perform significantly worse when provided with all the features. Nonetheless, for the sake of model simplicity and intuition, it is very useful to determine *which* of the features we extract are most useful for predicting energy consumption. To this end, we employ a simple greedy forward feature selection procedure that sequentially adds features based on how much they decrease training root mean squared error (RMSE) of a linear regression predictor. Table 1 shows a list of the features in the order they are selected by the greedy procedure. The table also shows how much adding the feature decreases the RMSE as measured by cross validation; thus, while adding a feature by definition will always decrease the training RMSE to some extent, only the first 9 features decrease the RMSE as measured via cross validation to a statistically significant degree ($p < 0.01$ in a pairwise t-test). These also correspond to features that we would expect to have a large impact on energy consumption: building value; square footage; number of electric/gas accounts (recall that since we are computing energy usage on a *per-building* basis, the number of electric and gas accounts serve as a rough proxy for the number of separate units); building class (a discrete feature that designates the building as a condominium, single family home, multi-family home, retail store, office building, etc); heat fuel (oil, gas, or electric); heat type (forced air, hot water, electric radiant, etc), and whether or not the house has central AC.

Equally interesting are the features that do *not* lead to a significant decrease in RMSE. For example, including the eigenvectors of the graph Laplacian does not significantly improve performance. This suggests that spatially varying attributes are not especially prevalent in this data set: while there is spatial correlation in total energy consumption, once we regress on other features that also are spatially correlated (such as building value), there is little added benefit to including purely spatial features. To illustrate this, we also include in Table 1 the correlation coefficient between each of the different features and the total energy consumption, which captures how correlated each feature is with the total energy independent of any other features. In cases where this correlation is high, yet the feature is ranked low in Table 1, then the feature is also highly correlated with the previously selected features, such that adding the feature to a linear regression model gives little added benefit.

Although the forward selection procedure that we use is well-known to be overly greedy in certain cases (Tibshirani 1996) (this had lead to a variety of alternative procedures such as LASSO-based feature selection (Efron et al. 2004)), the approach is often very effective both in practice and in theory (Zhang 2009). In previous experiments we found virtually no difference in the features selected when considering more complex procedures such as the LASSO. Additionally, since we are performing feature selection based on linear regression, there may be some concern of throwing out features that are potentially of interest to the non-linear regression algorithms we consider later. However, this has not been the case in the experiments we have considered, and we will discuss this further in the next section.

Modeling and Evaluation

In this section we present learning methods for predicting energy usage given known features of a building. As we don't expect to predict the energy usage exactly (certainly energy consumption depends on many variables that are not known, including end-user preferences), and since, as we will discuss more in the section, we are concerned with providing users with information about where they lie in the distribution of energy consumption, our focus is on *probabilistic* methods that return a distribution over possible energy consumption levels. Formally, our predictors will all have the form

$$y = f(\mathbf{x}) + \epsilon \tag{3}$$

where $y \in \mathbb{R}^m$ denotes the predicted energy usage (or rather, based on the discussion above, the logarithm of the predicted energy usage), $\mathbf{x} \in \mathbb{R}^n$ denotes a vector of inputs describing known features of the house (that we delineate below), and ϵ denotes a (possibly input dependent) zero-mean error term. Although we focus in this section on real-valued regression (predicting just the total energy consumption), we apply these same methods to multivariate predictions, for instance to predict the energy consumption in each month as shown later in the paper. We focus on two well-known probabilistic regression techniques: linear models and Gaussian process regression. However, given the above discussion regarding the log T distribution fitting the data better than log normal distributions, we focus also upon more recent work using non-Gaussian likelihoods.

Linear Regression

In the linear regression case,

$$f(\mathbf{x}) = \theta^T \mathbf{x} \tag{4}$$

for parameters $\theta \in \mathbb{R}^n$, and the error term is given by some input-independent distribution $p(\epsilon)$, often referred to as the likelihood function. Based upon the discussion in the previous section, we consider the standard normal error term (leading to ordinary least squares), a Student-t distributed error term, and a Laplace distributed error term (another heavy-tailed distribution commonly used in robust regression). The densities and notation for the normal, Laplace,

and T distribution are given respectively by

$$
\begin{aligned}
p(\epsilon\,;\sigma) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \\
p(\epsilon\,;\sigma) &= \frac{1}{2\sigma}\exp\left(-\frac{|\epsilon|}{\sigma}\right) \\
p(\epsilon\,;\sigma,\nu) &= \frac{\Gamma((\nu+1)/2)}{\Gamma\nu/2\sqrt{2\pi}\sigma}\left(1+\frac{\epsilon^2}{\nu\sigma^2}\right)^{\frac{-\nu+1}{2}}
\end{aligned}
\tag{5}
$$

where σ is a scale parameter and ν is a degree of freedom parameter for the T distribution. Given a data set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, we can compute maximum likelihood estimates of θ and the relevant distribution parameters for each of these models. Defining the design matrices

$$
\mathbf{X} \equiv \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}, \quad \mathbf{y} \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \tag{6}
$$

then for the normal likelihood the ML estimates are

$$
\begin{aligned}
\hat{\theta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
\hat{\sigma}^2 &= \frac{1}{n}(\mathbf{y}-\mathbf{X}\hat{\theta})^T(\mathbf{y}-\mathbf{X}\hat{\theta}).
\end{aligned}
\tag{7}
$$

For the Laplace likelihood

$$
\begin{aligned}
\hat{\theta} &= \arg\min_{\theta} \|\mathbf{y}-\mathbf{X}\theta\|_1 \\
\hat{\sigma} &= \frac{1}{n}\|\mathbf{y}-\mathbf{X}\hat{\theta}\|_1
\end{aligned}
\tag{8}
$$

where $\|\cdot\|_1$ denote the ℓ_1 norm (the sum of the absolute values of a vector), which makes the optimization over θ a convex optimization problem (solved, for example, via linear programming (Boyd and Vandenberghe 2004)). For the T distribution there is no closed form estimate for the parameters, but they can be obtained via the EM algorithm, which amounts to iterating the following updates until convergence

$$
\begin{aligned}
\mathbf{s}_i &\leftarrow \frac{1}{\hat{\sigma}^2}(y_i-\hat{\theta}^T\mathbf{x}_i)^T(y_i-\hat{\theta}^T\mathbf{x}_i) \\
\mathbf{w}_i &\leftarrow \frac{\hat{\nu}+1}{\hat{\nu}+\mathbf{s}_i} \\
\hat{\theta} &\leftarrow (\mathbf{X}^T\mathrm{diag}(\mathbf{w})\mathbf{X})^{-1}\mathbf{X}^T\mathrm{diag}(\mathbf{w})\mathbf{y} \\
\hat{\sigma}^2 &\leftarrow \frac{1}{1^T\mathbf{w}}(\mathbf{y}-\mathbf{X}\hat{\theta})^T\mathrm{diag}(\mathbf{w})(\mathbf{y}-\mathbf{X}\hat{\theta}) \\
\hat{\nu} &\leftarrow \arg\min_{\nu} \sum_{i=1}^N \log p(y-\hat{\theta}^T\mathbf{x}_i\;;\;0,\hat{\sigma},\nu)
\end{aligned}
\tag{9}
$$

where probability in the optimization over ν is the density of the T distribution, and this step is performed via numerical optimization.[3] Because the log likelihood is not convex with respect to the parameters, there is no guarantee that this procedure will find the global optimum but in practice if we initialize $\hat{\theta}$ and $\hat{\sigma}^2$ to be, for example, the least squares estimates, then this procedure is quite robust.

[3]Regression with a T distribution is not often done in this manner (for example, the presentation in (Kotz and Nadarajah 2004) does not describe this procedure), but this a straightforward extension of ML estimation for the T distribution to the regression setting, and seems to be well known.

Gaussian Process Regression

Gaussian process (GP) regression (Rasmussen and Williams 2006) provides a non-parametric regression method, and allows for much richer representations than in linear regression. In GP regression, the underlying regression function f is modeled as a Gaussian process, a stochastic process where the distribution of $f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots$ for any set of inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots$ is *jointly Gaussian* with some mean (which we will assume to be zero), and covariances

$$
\mathrm{Cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = K(\mathbf{x}_i, \mathbf{x}_j) \tag{10}
$$

for some positive definite *kernel function* $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Overloading notation slightly, we write this as

$$
\mathbf{f}|\mathbf{X} \sim \mathcal{N}(0, \mathbf{K}(\mathbf{X},\mathbf{X})) \tag{11}
$$

and where in the following we will simply use $\mathbf{K} \in \mathbb{R}^{N\times N}$ to denote $\mathbf{K}(\mathbf{X},\mathbf{X})$ (the kernel matrix formed between all the training examples). As before, we assume that $y_i = f(\mathbf{x}_i) + \epsilon$, where ϵ is again a zero-mean error term with one of the three distributions described above. When ϵ is Gaussian with variance σ_ϵ^2, the distribution $p(f(\mathbf{x}')|\mathbf{X},\mathbf{y},\mathbf{x}')$ (the distribution value of $f(\mathbf{x}')$ for some new input $\mathbf{x}'$) is also Gaussian, and can be computed in closed form

$$
\begin{aligned}
f(\mathbf{x}') \sim \mathcal{N}\big(&\mathbf{K}(\mathbf{x}',\mathbf{X})(\mathbf{K}+\sigma_\epsilon^2 I)^{-1}\mathbf{y}, \\
&K(\mathbf{x}',\mathbf{x}') - \mathbf{K}(\mathbf{x}',\mathbf{X})(\mathbf{K}+\sigma_\epsilon^2 I)^{-1}\mathbf{K}(\mathbf{X},\mathbf{x}')\big).
\end{aligned}
\tag{12}
$$

When ϵ is Laplace or T distributed, there is no longer a closed form expression for the distribution over $f(\mathbf{x}')$, and we must resort to approximate methods such as Expectation Propagation, Variational Bayes, or the Laplace approximation (Rasmussen and Williams 2006; Vanhatalo, Jylanki, and Vehtari 2009). A description of these methods is beyond the scope of this paper, but several freely available software packages exist that implement such approximate methods for GP regression with non-Gaussian likelihood.[4]

For our task we use the squared exponential kernel function, with independent length scales, given by

$$
K(\mathbf{x},\mathbf{x}') = s^2\exp\left(\sum_{i=1}^n \frac{(x_i-x_i')^2}{2\ell_i^2}\right) \tag{13}
$$

where the the magnitude $s \in \mathbb{R}_+$ and length scales $\ell_i \in \mathbb{R}_+$ are free parameters. As is common practice, we estimate these parameters by maximizing the *marginal likelihood* of the training data, given by

$$
p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f}. \tag{14}
$$

As before, when the likelihood is Gaussian this term (and its derivatives with respect to the free parameters) can be computed analytically. When the likelihood is not Gaussian, approximations are again needed, but again the software packages mentioned above provide methods for doing so.

[4]In this work we use the GPML package (Rasmuseen and Hickish 2011) for the Gaussian and Laplace likelihood GP regression, and the GPstuff package (Vanahtalo et al. 2011) for the T likelihood GP regression.

Method	Log Likelihood	RMSE	No Log RMSE ($\times 10^5$)
Output only, Normal likelihood	-1.484 (-1.482)	1.066 (1.065)	11.106 (11.105)
Output only, Laplace likelihood	-1.413 (-1.412)	1.079 (1.079)	11.110 (11.110)
Output only, T likelihood	-1.399 (-1.399)	1.074 (1.074)	11.109 (11.109)
Linear regression, Normal likelihood	-0.813 (-0.788)	0.545 (0.532)	9.581 (9.231)
Linear regression, Laplace likelihood	-0.710 (-0.685)	0.549 (0.537)	9.422 (9.397)
Linear regression, T likelihood	-0.695 (-0.674)	0.547 (0.536)	9.488 (9.402)
GP regression, Normal likelihood	-0.782 (-0.747)	**0.531 (0.503)**	**9.212 (8.016)**
GP regression, Laplace likelihood	-0.660 (-0.620)	0.535 (0.495)	9.704 (7.609)
GP regression, T likelihood	**-0.629** (-0.557)	0.543 (0.502)	9.746 (5.928)
GP regression, all features	-0.786 (-0.710)	0.533 (0.485)	9.243 (6.313)
Linear regression, no log	-15.240 (-14.962)	1.738 (1.732)	9.260 (7.566)
GP regression, no log	-15.874 (-90.589)	2.775 (0.624)	11.240 (2.889)

Table 2: Cross validation performance (and training set performance in parenthesis) of the different algorithms evaluated by three metrics: log likelihood of the data, root mean squared error (RMSE), and RMSE on the data before logarithmic scaling. Items in bold indicate the best performing method, statistically significant in a pairwise t-test with $p < 0.01$.

Experimental Setup and Results

We evaluated the performance of the algorithms described above using 5 fold cross validation: we divided the 6,499 data points randomly into 5 approximately equal-sized groups, trained the above regression methods on the union of 4 of these groups, then tested on the held out data; we repeated this for each of the 5 groups, and report the average error and log likelihood over all the held-out examples. To report training errors, we trained and tested on the entire data set. For the GP models, since hyperparameter optimization is quite computationally intensive, we optimized these parameters by maximizing marginal likelihood only on a random subset of 700 of the training examples for each cross validation fold; once we learned these hyperparameters, however, we used the entire training set to predict the held-out data.

Table 2 shows the performance of the different algorithms on this data set. We evaluated the algorithms via three metrics: log likelihood of the data, root mean squared error (RMSE) on the logarithmically scaled outputs, and RMSE on the original energy consumptions (without logarithmic scaling, noted in Table 2 by "No-Log"). For comparison, we also present "Output only" results, which simply involve fitting the T, Laplacian, and normal distributions directly to the log of the energy data (without any regressors). Summarizing the results briefly, the best-performing model we obtain is able to explain about 75% of the variance (in the logarithmic scale) using the features described above. This naturally leaves a great deal of variance unexplained, but of course this is expected, since we must imagine that some elements of the energy usage are simply behaviorally based and cannot be predicted in the normal sense; indeed, these are precisely the situations where we want to present such information to the user.

In greater detail, as seen in the table, the GP methods obtain the best overall performance, with the normal likelihood performing best as measured by the RMSE and the T likelihood performing best in terms of the log likelihood. However, while the GP methods do perform better than the simple linear regression models, we argue that the simple linear models are in some respects preferable for this domain. The linear regression methods all obtain RMSE that is only marginally worse than the GP methods, they allow for simple descriptions of energy usage in terms of power-law relationship, and they provide much more succinct, computationally efficient, and interpretable models. We also look at alternative regression approaches, such as including all the extracted features for the non-linear GP, or using the data without logarithmic scaling to make predictions. In both cases, the resulting approaches perform no better, and in the case of omitting the log transformation, the resulting methods can perform much worse, both in terms of the log-based RMSE and the RMSE in the original scale (log likelihood terms are not directly comparable, as the data is not on the same scale before the log transformation).

EnergyView

Based upon the models from the previous section, we have developed an end-user application, EnergyView, that lets companies or individuals view and compare their energy usage, based upon how their true usage compares to the model's predictions. The system consists of two parts: for the utilities or authorized organizations (anyone who may obtain access to the individual energy records), we have developed a graphical interface that layers energy consumption over a map of the area. The interface allows users to quickly determine outliers according to the model: building that use significantly more (or less, though presumably little needs to be done in this case) energy than predicted by the model (i.e., even accounting for all the observed features of the building). Thus, these are prime candidates for buildings where energy usage could be reduced by behavior changes, retrofits, etc. Presenting this information to a community organization aimed at energy efficiency, for example, could greatly help such groups decide where to focus resources. A (simulated) screenshot of this system is shown in Figure 3. Due to privacy considerations, such a system is unlikely to become publicly available (though users could opt in to allow public display of their information), but could still be highly valuable for authorized groups.

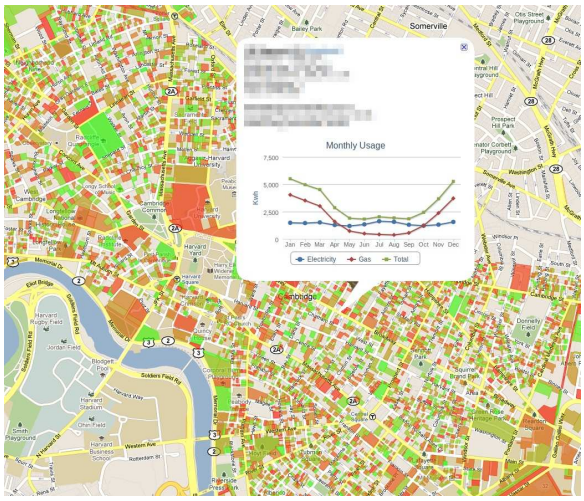

Figure 3: Example image of the city-level EnergyView tool. While the image here is indicative of how the interface looks, for privacy reasons the color codes and energy consumption in this picture are generated via random sampling, and do not correspond to the actual building consumptions in Cambridge.

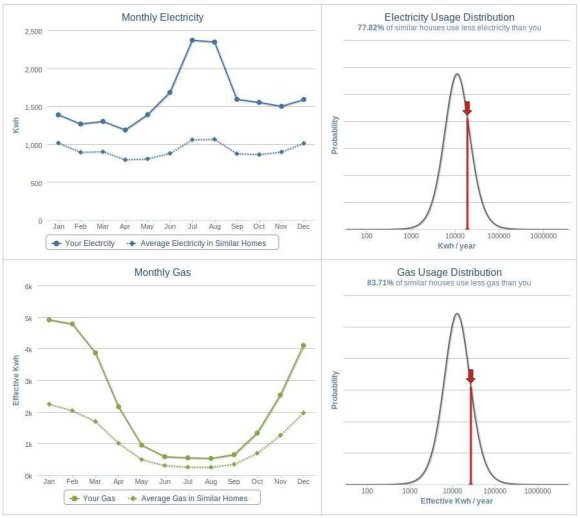

Figure 4: Image of the home-level EnergyView analysis.

The second component of the EnergyView system is an end-user tool that lets an owner enter the relevant information about their own home or building, plus gas and electrical consumption, in order to find where they lie in relation to the model's prediction (we refer to this process as *contextualizing* energy usage, as it puts it in the context of similar homes). Because such a tool does not disclose identifying information, it can be released publicly and used by building owners or efficiency groups alike. A screenshot of the resulting analysis page is shown in Figure 4, showing a building's consumption versus expected consumption for each month and displaying where the building lies in the overall distribution. A web version of the tool is available `http://people.csail.mit.edu/kolter/energyview`. Although, because our approach is fully data-driving, the validity of the model is specific to the general area of Cambridge, MA, the techniques are quite general, and the same system could be set up for any location given access to similar data.

Conclusion

In this paper we presented a study and analysis of building energy consumption in a large urban environment. To the best of our knowledge, this represents one of the largest academic studies of individual home and building energy consumption using real data for a single city. We have analyzed the distributions of this data, and used these insights to learn models that can predict expected energy usage given features of the home, and which are able to explain roughly 75% of the observed variance in energy consumption according to our logarithmic scaling of the data. Finally, we describe and release EnergyView, a tool that uses these models to visualize both city-level information, and building-level information, allowing end-users to see where their home lies within the expected distribution.

Looking forward, there are numerous possible directions for future work. We are continuing to look for additional sources of data that could be incorporated into the model, such as home survey data or image data. Despite the fact that spatial features did not help the predictions in our current data set, location data seems still to be a promising source of information, and there is a need for algorithms or features that can better capture such relationships. We are working to expand the EnergyView application to allow for opt-in data sharing and incorporation of entered data into the internal models. More broadly speaking, the models and EnergyView tool we present here represents just one facet of the building energy problem; equally important is how we use these models to produce actionable information to the proper groups that can actually lead to energy savings. Understanding how to best achieve these savings using the model's predictions and determining how the predictions correlate with changeable behaviors in a community, remains a crucial question for further work.

Acknowledgements

We thank Harvey Michaels for his assistance in obtaining the NStar data. J. Zico Kolter is supported by an NSF Computing Innovation Fellowship.

References

Allcott, H. 2010. Social norms and energy conservation. Available at `http://web.mit.edu/allcott/www/papers.html`.

Berges, M.; Goldman, E.; Matthews, H. S.; and Soibelman, L. 2009. Learning systems for electric consumption of buildings. In *ASCI International Workshop on Computing in Civil Engineering*.

Berry, C. 2005. Residential energy consumption survey. Available at http://www.eia.doe.gov/emeu/recs/.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press.

Cassidy, D. T.; Hamp, M. J.; and Ouyed, R. 2010. Pricing european options with a log student t distribution: A gosset formula. *Physica A: Statistical Mechanics and its Applications* 389(25).

Chung, F. 1997. *Spectral Graph Theory*. Americal Mathematical Society.

Cialdini, R., and Schultz, W. 2004. Understanding and motivating conservation via social norms. Technical report, William and Flora Hewlett Foundation.

Darby, S. 2006. The effectiveness of feedback on energy consumption. Technical report, Environmental Change Institute, University of Oxford.

Efron, B.; Johnstone, I.; Hastie, T.; and Tibshirani, R. 2004. Least angle regression. *Annals of Statistics* 32(2):407–499.

Kotz, S., and Nadarajah, S. 2004. *Multivariate t distributions and their applications*. Cambridge University Press.

Kuhnert, C.; Helbing, D.; and West, G. B. 2006. Scaling laws in urban supply networks. *Physica A: Statistical Mechanics and its Applications* 363(1):96–103.

Limpert, E.; Stahel, W. A.; and Abbt, M. 2001. Log-normal distributions across the sciences: keys and clues. *BioScience* 51(5).

Multiple. 2009. Annual energy review 2009. Technical report, U.S. Energy Information Administration.

Neenan, B., and Robinson, J. 2009. Residential electricity use feedback: A research synthesis and economic framework. Technical report, Electric Power Research Institute.

Newman, M. E. J. 2005. Power laws, pareto distributions and zipf's law. *Comtemporary Physics* 46:323–351.

Rasmuseen, C. E., and Hickish, H. 2011. Gpml matlab code. Available at `http://www.gaussianprocess.org/gpml/code/matlab/doc/`.

Rasmussen, C. E., and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statisitcal Society B* 58(1).

Vanahtalo, J.; Riihimaki, J.; Hartikainen, J.; and Vehtari, A. 2011. Bayesian modeling with gaussian processes using the matlab toolbox gpstuff. *in submission*. Code available at `http://www.lce.hut.fi/research/mm/gpstuff/`.

Vanhatalo, J.; Jylanki, P.; and Vehtari, A. 2009. Gaussian process regression with student-t likelihood. In *Neural Information Processing Systems*.

Zhang, T. 2009. On the consistency of feature selection using greedy least squares. *Journal of Machine Learning Research* 10:555–568.