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

J. Zico Kolter  
*CS and AI Lab (CSAIL), MIT*

Joseph Ferreira Jr.  
*Urban Studies and Planning, MIT*

AAAI Special track on Sustainability and AI, 2011

Presented by  
*Saima Aman*
Abstract

A data-driven approach to modeling end user energy consumption in residential and commercial buildings (based upon features, such as living area, building value, building type, etc.).

Data set
• Monthly electricity and gas bills for 6500 buildings (approx) in Cambridge, MA. (several years)
• Tax assessor records and geographical survey information for building features (publicly available)

EnergyView System
1) For utilities/authorized institutions: A system to visualize energy consumption for each building in the city
   Advantage: Quickly identify outliers; target homes for potential retrofits/ tiered pricing
2) For other end users: An interface for entering their info (electricity & gas usage; home info)
   Advantage: Compares their consumption to that of similar buildings.

Motivation
Allowing users to contextualize their consumption; relating it to the consumption in similar buildings, can produce behavior changes to reduce consumption.
Background

• 86% of the total energy consumed worldwide comes from \textit{(unsustainable)} fossil fuels
• In the US, 41% of all energy (electricity & natural gas) is consumed in residential and commercial buildings

\textbf{Shortcomings of current electricity and gas bills}
• little information or context to usage other than a dollar amount.
• no comparison with of a building’s consumption with similar buildings
• no info about the financial feasibility of retrofits/ upgraded appliances
• no info about the portion of consumption due to location (eg., cold locations) versus personal behavior
Contribution

• They analyze a large-scale real-world energy usage data set
• Illustrate several interesting characteristics of the data
• Present an end-user interface for obtaining contextual information about ones own energy use.

• One of the largest-scale, publicly-available studies of its kind, conducted on real data
• EnergyView tool represents one of the first tools of its kind where the algorithms behind its predictions are fully described.
Data Collection and Analysis

Electric and Gas bills
• Provided by NStar – electricity and gas utility in Cambridge, MA.
• Data fields:
  • Account numbers
  • Corresponding street address
  • Monthly electricity and gas meter readings (typically, for two to three years)
  • Electricity usage (in kilowatt-hours per month)
  • Gas usage (in therms per month – converted into kWh to predict TOTAL usage)

Tax Assessor and GIS Data
• Both publicly available
• Tax Assessor data fields:
  • value of the building
  • property class (condominium, single family home, retail store, etc)
  • square footage
  • building construction year, etc.
• GIS data fields:
  • polygonal outlines for parcels and buildings in the city
  • estimated roof heights for buildings (obtained via an aerial lidar scan)
Logarithmic Energy Scaling

- 6,499 unique buildings (with complete uninterrupted data for at least a year)
- More than half of the 12,792 unique addresses in the Cambridge tax assessor records

Histogram of total energy consumption per building for 6499 buildings

- The data appears roughly Gaussian
- The x axis is logarithmic => total energy consumption follows a log-normal distribution
Logarithmic Energy Scaling (contd.)

- Many observed phenomena naturally follow roughly log-normal distributions
- These include factors that influence energy consumption, such as income, property sizes, and building square footage.

Plot of building square footage versus total yearly energy consumption (log-log scale)

- Besides great deal of noise, there is a fairly clear linear relationship, indicating an intuitive power-law relationship between square footage and energy consumption.
  - (Power-law behavior – linear relation between log of input and output variables)

- Exploit these types of relationships to derive predictive models of energy consumption
Feature Selection

• For real-valued attributes, their logarithm was included in the feature vector
• For discrete features, a standard binary encoding of the feature was included

• **Selection**: For the sake of model simplicity and intuition, it is useful to determine which features are most useful for predicting energy consumption.

• Used a **greedy forward feature selection** procedure that sequentially adds features based on how much they decrease training root mean squared error (RMSE) as measured by *cross validation* of a linear regression predictor.

• 9 features decrease RMSE by a statistically significant degree ($p < 0.01$ in a pairwise t-test)
• Correspond to features expected to have a large impact on energy consumption:
  • building value; square footage
  • number of electric/gas accounts (~ number of separate units)
  • building class (condo, single-family home, multi-family home, retail store, office, etc.)
  • heat fuel (oil, gas, or electric)
  • heat type (forced air, hot water, electric radiant, etc)
  • whether or not the house has central AC.
Feature Selection (contd.)

• Features from Tax Assessor and GIS Records

<table>
<thead>
<tr>
<th>Feature</th>
<th>% RMSE Reduction</th>
<th>Individual Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building Value</td>
<td>24.721%</td>
<td>0.659</td>
</tr>
<tr>
<td>Num Electric Meters</td>
<td>13.438%</td>
<td>0.647</td>
</tr>
<tr>
<td>Property Class</td>
<td>12.729%</td>
<td>0.633</td>
</tr>
<tr>
<td>Living Area</td>
<td>2.760%</td>
<td>0.611</td>
</tr>
<tr>
<td>Num Gas Meters</td>
<td>2.517%</td>
<td>0.626</td>
</tr>
<tr>
<td>Heat Fuel</td>
<td>2.241%</td>
<td>0.480</td>
</tr>
<tr>
<td>Building Style</td>
<td>1.432%</td>
<td>0.632</td>
</tr>
<tr>
<td>Heat Type</td>
<td>0.826%</td>
<td>0.431</td>
</tr>
<tr>
<td>Central AC</td>
<td>0.749%</td>
<td>0.558</td>
</tr>
</tbody>
</table>

• 35 attributes used from tax assessor and GIS records

• Ranked in the order that they are selected in greedy forward feature selection.

• Correlation coefficient between total energy and the features in isolation is also shown.
Modeling

• Not expected to predict the energy usage exactly
• Goal: provide information about where users lie in the distribution of energy consumption
• Focus is on probabilistic methods that return a distribution over possible energy consumption levels

Predictors will have the form:  
\[ y = f(x) + e \]

- \( y \) – predicted energy usage (log of predicted energy, in this exp.)
- \( x \) – a vector of inputs describing known features of the house
- \( e \) – error term
Modeling (contd.)

Linear regression
- The error term is given by input-independent distribution $p(e) \rightarrow$ likelihood function.
- Likelihood functions:
  - Standard normal error term (leading to ordinary least squares)
  - Student-t distributed error term
  - Laplace distributed error term

Gaussian process regression
- Non-parametric regression
- Used the GPML package for the Gaussian and Laplace likelihood GP regression
- Used the GPstuff package for the T likelihood GP regression.
Experiments

5 fold cross validation
• divided the 6,499 data points randomly into 5 equal-sized group
• trained regression methods on the union of 4 groups, and tested on the held out data
• repeated for all 5 groups and report the average error and log likelihood

Training errors
regression methods were trained and tested on the entire data set.

Hyperparameter optimization for GP models
• computationally intensive
• parameters optimized by maximizing marginal likelihood only on a random subset of 700 of the training examples for each cross validation fold

Evaluation Metrics
• log likelihood of the data
• root mean squared error (RMSE) on the logarithmically scaled outputs
• RMSE on the original energy consumptions (without log scaling)
Results

Cross-validation performance (training error in parenthesis):

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

- “Output only” results – involve fitting the T, Laplacian, and normal distributions directly to the log of the energy data (without any regressors).

- The best-performing model is able to explain about 75% of the variance (in the log scale).
- Some elements of energy usage are simply behaviorally based and cannot be predicted.
Results (contd.)

GP methods perform better than simple linear regression models

However, simple linear models are preferable:
• allow for simple descriptions of energy usage in terms of power-law relationship
• provide much more succinct, computationally efficient, and interpretable models.
• obtain RMSE that is only marginally worse than the GP methods
EnergyView

• For utilities/authorized organizations:

City-level Energy view
EnergyView

• Contextualizing Energy Usage:

End-user Energy view

• Available online: http://people.csail.mit.edu/ kolter/energyview