Data-Driven Forecasting Algorithms for Building Energy Consumption

Hae Young Noh\textsuperscript{a} and Ram Rajagopal\textsuperscript{b}

\textsuperscript{a}Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA;
\textsuperscript{b}Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, 94305, USA

ABSTRACT

This paper introduces two forecasting methods for building energy consumption data that are recorded from smart meters in high resolution. For utility companies, it is important to reliably forecast the aggregate consumption profile to determine energy supply for the next day and prevent any crisis. The proposed methods involve forecasting individual load on the basis of their measurement history and weather data without using complicated models of building system. The first method is most efficient for a very short-term prediction, such as the prediction period of one hour, and uses a simple adaptive time-series model. For a longer-term prediction, a nonparametric Gaussian process has been applied to forecast the load profiles and their uncertainty bounds to predict a day-ahead. These methods are computationally simple and adaptive and thus suitable for analyzing a large set of data whose pattern changes over the time. These forecasting methods are applied to several sets of building energy consumption data for lighting and heating-ventilation-air-conditioning (HVAC) systems collected from a campus building at Stanford University. The measurements are collected every minute, and corresponding weather data are provided hourly. The results show that the proposed algorithms can predict those energy consumption data with high accuracy.

Keywords: building energy consumption, forecasting, autoregressive model, Gaussian process, distribution forecasting, covariance function, nonparametric density estimation, Bayesian analysis

1. INTRODUCTION

The operation of modern society and individual life heavily depends on technologies that constantly require consuming energy. Therefore, it is critical to secure reliable and sufficient supply of energy in order to assure the functionality and safety of the society. One of the most important issues to secure a reliable and sufficient energy supply in the grid is the ability to estimate and predict power demand. Among various sources, buildings account for about 20–40\% of energy consumption.\textsuperscript{1} As more entities are equipped with smart meters, we can collect energy consumption data on a very fine resolution; however, proper aggregation and organization of data including data cleansing and preprocessing, and efficient data analysis to extract necessary information for pattern classification, forecasting, and optimal control are still under development. Challenges with processing high-resolution smart meter data come from improper calibration, lossy communication, large size of data, and complex and vague underlying dynamics of demand that are highly non-stationary and user-dependent.

Previous studies on forecasting energy consumption often involve complicated models for point-wise predictions. Such models include artificial neural network\textsuperscript{2–4} and support vector machine.\textsuperscript{5} Additive models were also applied to decompose the consumption data into several functions of different characteristics. Galiana et al.\textsuperscript{6} applied to half-hourly and hourly data sets an additive model of periodic functions and a autoregressive model with exogenous input to consider time periodicity, load autocorrelation, and temperature effects. Park et al.\textsuperscript{7} also used an additive model that decomposes the load time-series into nominal, type, and residual loads, where

Further author information: (Send correspondence to Hae Young Noh.)
Hae Young Noh: E-mail: noh@cmu.edu, Telephone: 1 412-268-5253
Ram Rajagopal: E-mail: ramr@stanford.edu, Telephone: 1 650-725-4268
Kalman filter, exponential smoothing, and autoregressive model are used to model them, respectively. Charytoniuk and Kotas used a nonparametric density estimation to find the distribution of energy consumption for particular value of temperature and time, which can be quite cumbersome to compute for fine-resolution data with high-dimensional features.

In this paper, we introduce two simple energy forecasting methods that use smart meter measurements and weather data to predict building energy consumption, such as lighting loads and heating-ventilation-air-conditioning (HVAC) systems. Hereafter, energy consumption data are referred to as outputs and weather data as inputs. The first method involves adaptive time-series modeling of the output using exogenous inputs. This method is sensitive to high-frequency variation of the data, and thus suitable for forecasting short-term patterns, such as an hour-ahead prediction. In contrast, the second method involves modeling the output as a Gaussian process (GP) and characterizes its overall behavior through covariance functions. The GP method has been widely used in engineering and science for regression and classification by characterizing the overall structure of the data. Hence, this method is appropriate for a longer-term profile prediction, such as a day-ahead or longer profile prediction. In addition, this method provides distribution predictions, instead of point predictions. In other words, the GP method can represent the uncertainties about the predicted values through their joint probability distribution.

The rest of the paper is organized as follows. Section 2 introduces two forecasting algorithms for forecasting. In section 3, the energy consumption data in a campus building are described, and the forecasting algorithms are applied to them. Finally, section 4 provide conclusions.

2. FORECASTING ALGORITHMS

2.1 Data preprocessing
Smart meter data often include improperly calibrated values, missing values, or erroneous values. Before applying forecasting algorithms, the data need to be properly cleaned and imputed so that all the data are comparable to each other. To address these problems, various data cleansing and preprocessing techniques have been developed in signal processing. Among them, we applied spline smoothing and interpolation to remove noise. For imputation, the 1-nearest neighbor method is applied to estimate missing values by finding the previous time of the day with the most similar consumption pattern.

2.2 Forecasting algorithms
To develop forecasting methods for energy consumption data, we developed two different approaches. The first approach defines a set of functions that defines the relationship between inputs, such as time and weather information, and outputs, such as energy consumption. This approach is simple and intuitive but requires prior knowledge and/or assumption about the input-output relationship. In the absence of accurate prior knowledge, the prediction will be poor. To address this problem, multiple classes of functions can be considered to model the relationship; however, increases in the class size and complexity (or flexibility) cause an over-fitting problem, and thus, a proper model selection scheme needs to be applied. In this study, we represented the data as autoregressive processes and then incorporated a moving window to increase the flexibility of the model by making it adaptive. On the other hand, the second approach assigns a prior probability distribution over all possible functions to represent the data and then updates this distribution according to the observed data. This approach provides a greater flexibility to the model because an infinite set of functions is considered for modeling the data, but it can be computationally very intensive. For this reason, a Gaussian process is adopted, which imposes preference bias using a Gaussian distribution to keep the computation simple.

2.2.1 Short-term forecasting: Adaptive Autoregressive Model
An autoregressive time-series model (AR) predicts the function output as a linear function of previous output observations and potentially exogenous inputs if available. Let $f(t)$ and $u_i(t)$ ($i = 1, \ldots, N$) denote the function output and $i^{th}$ exogenous input at time $t$. The AR model of order $p$ is given as

$$f(t) = \sum_{k=1}^{p} \alpha_k f(t-k) + \sum_{i=1}^{N} \beta_i u_i(t) + \varepsilon_i(t)$$ (1)
where $\alpha_k$ is the $k^{th}$ AR coefficient, $\beta_i$ is the $i^{th}$ input coefficient, and $\epsilon_i(t)$ is the residual. We can determine these coefficients using the training data and predict $f(t+1)$ on the basis of estimate coefficients. Multi-step ahead predictions can be made from consecutively making the next-step prediction or directly computing the autoregressive relationship between $f(t)$ and $\{f(t-n), f(t-n-1), \ldots, f(t-n-p+1)\}$, where $n$ is the prediction horizon.

In addition to the AR model, we applied a moving window to the training set. One of the concerns with imposing a particular model is its consistency with observed data. If the model is too rigid to incorporate the pattern of the data, the accuracy of the prediction is low due to a large bias. Therefore, we assumed that the data have a linear structure at least within a short time window and thus added an adaptive nature to the model. In other words, the model coefficients are locally estimated using a subset of the training data within a short time window of width $w$ and then this window shifts in time as we make the next prediction.

### 2.2.2 Long-term forecasting: Gaussian Process Regression

A Gaussian process (GP) is a stochastic process for a function the joint distribution of any finite collection of whose outputs is a Gaussian distribution. A GP is a nonparametric generalization of the Gaussian probability distribution: as the Gaussian probability distribution assigns likeliness or preference to random variables or random vectors of finite dimension, the Gaussian process describes the probability of functions, $f(x)$. Because the function can take continuous values for inputs and outputs, the Gaussian process is defined for infinite dimensional functions. Unlike a parametric model, which is completely described by a fixed and finite number of parameters, a nonparametric model involves infinite number of parameters or the number of its parameter grows as more data are observed.

A GP has a much greater flexibility than parametric models because it does not compress the data into a fixed number of parameters. In fact, many machine learning algorithms or statistical models are special cases of the GP with particular form of covariance functions. For example, linear regression and logistic regression using Gaussian priors for parameters are simple examples of the GP with degenerate covariance functions. A degenerate covariance function has a finite number of non-zero eigenvalues. Actually, a GP can be interpreted as a linear regression in the Bayesian framework using infinitely many basis functions of inputs, which can be eigenfunctions of the covariance function. Moreover, Neal (1996) found that a certain type of a neural network with one hidden layer converges to a GP. In general, GPs are computationally efficient for regression and prediction and provides an entire predictive joint density instead of giving a single prediction value with no information about its uncertainty.

Just like a Gaussian distribution, a GP is defined by a mean function, $m(x)$, and a covariance function, $k(x, x')$,

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$ (2)

where

$$m(x) = \mathbb{E}[f(x)],$$ (3)

$$k(x, x') = \exp[(f(x) - m(x))(f(x') - m(x'))].$$ (4)

In general, a constant or a simple polynomial form, of one or two degree, can be used as a prior mean function. Since the posterior will be computed based on the data, imposing a simple form to a prior mean function is not too restrictive. The covariance function characterizes the properties of the functions to be considered in the regression. For example, we can assign a higher probability to stationary and slowly varying functions by selecting a particular family of covariance functions and fine-tune the degree of change and stationarity by setting its parameters, called hyper-parameters. We can also define other characteristics such as sudden spikes or periodicity in the functions. Therefore, training the GP model mainly involves inferencing the covariance function. We can select a family or a combination of families of covariance functions and set hyper-parameters heuristically by examining observations. We can approach it more systematically using cross-validation methods or a hierarchical Bayesian model selection by defining prior distributions for candidate models and corresponding
parameters. The prior for models are often taken to be flat, so the posterior distribution of the model is proportional to the marginal likelihood, the likelihood marginalized over all hyper-parameters.

Note that the mean and the covariance functions are properties of function output values, but they are defined by input values. The outputs are the variables we want to forecast, and the inputs are a subset of any available information that is useful for predicting the output values, such as any exogenous variables corresponding to an output observation, observation time, spatial information, and even previous output observations. In our case, outputs are energy consumption, and inputs are weather data and time.

A Gaussian process satisfies a consistency requirement, also known as a marginalization property because it is defined as a collection of function outputs (i.e., random variables). In other words, any partition of the collection of the random variables has a Gaussian distribution with corresponding partitions of the mean vector and the covariance matrix of the original Gaussian distribution:

Let the function output vector $f_X$ for a set of input values $X$ has a prior distribution according to the GP,

$$f_X \sim \mathcal{N}(m(X), K(X, X))$$  \hspace{1cm} (5)

Similarly, the joint prior distribution of the training outputs, $f_{tr}$, and the test outputs, $f_{te}$, is

$$\begin{bmatrix} f_{tr} \\ f_{te} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} m(X_{tr}) \\ m(X_{te}) \end{bmatrix}, \begin{bmatrix} K(X_{tr}, X_{tr}) & K(X_{tr}, X_{te}) \\ K(X_{te}, X_{tr}) & K(X_{te}, X_{te}) \end{bmatrix} \right)$$  \hspace{1cm} (6)

To obtain the posterior distribution for $f_{te}$, which is the predictive distribution of the test outputs given the training observations, we need to eliminate the functions that do not agree with the observations. The naive computational approach would be to draw samples from the prior and then keep only the ones that match the observations, but it is computationally inefficient to generate numerous functions and eliminate most of them. Fortunately, we can compute the analytical solution for the GP simply using the conditional probability of the joint Gaussian distribution:

$$f_{te} | f_{tr}, X_{tr}, X_{te} \sim \mathcal{N} \left( m(X_{te}) + K(X_{te}, X_{tr}) K(X_{tr}, X_{tr})^{-1} (f_{tr} - m(X_{tr})), \right.$$

$$\left. K(X_{te}, X_{te}) - K(X_{te}, X_{tr}) K(X_{tr}, X_{tr})^{-1} K(X_{tr}, X_{te}) \right).$$  \hspace{1cm} (7)

Equation 7 provides the predictive distribution of the function conditioned on the training data. The posterior mean is a linear function of training data, which is why the GP is referred to as a linear predictor. On the other hand, the mean can be seen as a linear function of covariance functions between the test input(s) and the training inputs. It explains why the covariance function characterizes the GP. The posterior covariance is a function of only the inputs and does not depend on the training outputs. The first term $K(X_{te}, X_{tr})$ is the prior covariance and the second term $K(X_{te}, X_{tr}) K(X_{tr}, X_{tr})^{-1} K(X_{tr}, X_{te})$ represents the uncertainty reduction from the training observations.

3. APPLICATION TO Y2E2 BUILDING ENERGY CONSUMPTION DATA

3.1 Description of data

The Jerry Yang and Akiko Yamazaki Environment and Energy (Y2E2) building in Stanford University is a 166,000 square foot, three-story building accommodating several departments and schools. The building includes several laboratories in the basement, a café on the first floor, individual offices, lecture rooms, and conference rooms on all three floors. It is equipped with over 2370 measurement points for heating, ventilation, and air conditioning (HVAC) systems collecting data every one minute. The measurements include both sensing and control points that involve various building operation and energy saving features. Our primary interest is on power/energy consumption measurements, including lighting, plug loads, and heating and cooling loads. Electrical power consumption for lighting and plug loads are recorded in watts. For heating and cooling purposes, Y2E2 uses hot water, measured in pounds, and chilled water, measured in tons. The Cardinal Co-generation Plant provides chilled water and hot steam to the campus, and hot water is produced from heat exchangers using the hot steam. Chilled water systems are used to cool down the components and equipment in the building, and hot water systems are used to heat the air-handling units, which supply heat to the building. For forecasting, electrical energy measurements are aggregated to 15 minute intervals and heating and cooling data are aggregated to hourly intervals because the measurements are repetitive and/or have high frequency noise.
3.2 Results

3.2.1 15 minute forecasting

For short-term forecasting, we investigated AR models with various sizes of training data. We assumed that within a short period of time, the data approximately follow an autoregressive process and applied the AR model with a moving window. After testing various orders of AR models, it was observed that lower orders performed better than higher orders. Figure 1 shows the root-mean-square-error (RMSE) of the lighting power consumption predictions normalized by the root-mean-square of the measurements for various prediction horizon and window sizes using the AR order 1. It is observed that smaller window size and shorter prediction horizon have lower prediction errors, which implies that the characteristics of the consumption pattern changes rapidly.

Figure 2 shows the predicted and the measured values using the AR model of order 1 for prediction horizon of 15 minutes, and window size of one hour. Corresponding scatter plot of the predictions and measurements is shown in Figure 3. The correlation coefficient is 0.994, and the coefficient of determination is 0.987.
3.2.2 Day-ahead forecasting

Day-ahead forecasting method is applied to cooling energy consumption data using outside temperature and time as inputs. A combination of several covariance functions is used for the GP model of the energy consumption data. First, the data is periodic in time, thus the product of a periodic function and a squared exponential function is used to represent the varying amplitude of periodic data. Then, the squared exponential function is used for temperature input, which in turn is combined with a random noise function. The squared exponential function represents a smooth variation of consumption data as a function of temperature. The mean of the GP is defined as a constant value. These are represented as

\[
m(x) = a, \tag{8}
\]

\[
K(x, x') = \theta_1^2 \times \exp \left( -\frac{(x_1 - x'_1)^2}{2\theta_2^2} - \frac{2\sin^2 \left( \frac{\pi |x_1 - x'_1|}{\theta_3} \right)}{\theta_4^2} \right) + \theta_5^2 \times \exp \left( -\frac{(x_2 - x'_2)^2}{2\theta_6^2} \right) + \theta_9^2 \delta_{x, x'} \tag{9}
\]

where \(x_1\) and \(x'_1\) are times, \(x_2\) and \(x'_2\) are temperatures, and \(\delta_{i,j}\) is the Kronecker’s delta function for \(i\) and \(j\). To learn the hyper-parameters, \(\theta_i s\), of the combined covariance function, maximum likelihood estimations are obtained using conjugate gradient method with the Polak-Ribiere formula. The optimization is performed with ten different initial values, and the hyper-parameters with a maximum likelihood among the ten cases are used for prediction. Five days of data are used for training the model for forecasting a whole 24-hour cooling energy consumption profile. Figure 4 shows the consumption measurements and a day-ahead prediction values. The dotted line is the mean prediction, and the shaded area indicates two standard deviations above and below the mean. Small breaks between shaded areas separate different days at midnight. It was observed that the measurements are mostly within the two standard deviations from the mean prediction, which implies that the prediction can be made with about 95% confidence level.

3.3 Application of forecasting

Accurate and reliable forecasting of energy consumption provides numerous benefits. It allows us to plan ahead accordingly in order to secure enough energy supply and prevent electricity crisis. In particular, accurate peak demand prediction helps reducing energy cost by avoiding the last-minute use of expensive generators. We can also diagnose the building system performance by comparing forecasted values with measurements. Large deviation from the prediction may indicate malfunctioning of the system or an unpredicted event. In addition, because forecasting models represent the relationship between predictors (e.g., weather variables, human behavior, system parameters, and so on) and resulting predictions (e.g., energy consumption), we can develop control schemes to
improve energy efficiency. For example, a proper incentive can be designed to affect the input variables in such a way that the corresponding power consumption will decrease at particular times in order to shift aggregate power consumption peaks to off-peak hours.

4. CONCLUSIONS

This paper introduces two simple energy forecasting methods based on smart meter measurements and weather data to predict building energy consumption. The first method involves adaptive autoregressive (AR) modeling of the smart meter data with exogenous inputs. This method is appropriate for predicting short-term patterns, such as 15 minute or an hour ahead, because it is sensitive to high-frequency variation of the data. On the other hand, the second method uses a Gaussian process to characterize the overall dynamics of the smart meter data and provides a longer-term profile prediction, such as a day ahead. In addition, the Gaussian process results in the mean and the covariance of the predictions, which define the complete distribution of the predictions. These methods are applied to various energy consumption data from a campus building (Y2E2) at Stanford University. The results show that the adaptive AR model was able to predict the consumption data with $R^2$ value of 0.987, and the GP model resulted in most of the data to be within the 95% confidence intervals. These forecasting methods can improve the performance of smart grid systems by allowing reliable energy supply through accurate planning. Moreover, they can be used as a diagnostic and control method to detect malfunctioning of the system and enhance the energy efficiency.

ACKNOWLEDGMENTS

The research presented in this paper is partially supported by the National Science Foundation CIF Research Grant No. 1116377 and Powell Foundation Fellowship. Their support is greatly appreciated.

REFERENCES


