RC25165 (W1106-008) June 1, 2011
Other

IBM Research Report

Statistical Modeling for Anomaly Detection, Forecasting and Root Cause Analysis of Energy Consumption for a Portfolio of Buildings

Fei Liu, Huijing Jiang, Young M. Lee, Jane Snowdon
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

Michael Bobker
Building Performance Lab
CUNY Institute for Urban Systems
New York, NY USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

STATISTICAL MODELING FOR ANOMALY DETECTION, FORECASTING AND ROOT CAUSE ANALYSIS OF ENERGY CONSUMPTION FOR A PORTFOLIO OF BUILDINGS

Fei Liu[1], Huijing Jiang[1], Young M. Lee[1], Jane Snowdon[1], Michael Bobker[2]
[1]IBM Thomas J. Watson Research Center, Yorktown Heights, NY, U.S.A.
[2]Building Performance Lab, CUNY Institute for Urban Systems, New York, NY, U.S.A.

ABSTRACT

This paper describes the statistical analytics technology being developed to help K-12 public schools in New York City reduce the energy consumption. A multi-step statistical analysis procedure is proposed, to assess energy consumption and to identify energy saving opportunities for large portfolios of buildings such as the NYC K-12 public school buildings. The method borrows strength from and makes integrated use of the Variable Base Degree Day (VBDD) regression model, multivariate regression model and the Auto Regressive Integrated Moving Average (ARIMA) model. In the first step, we build a regression model which correlates the energy consumption with building characteristics for the whole portfolio of buildings. The energy related building characteristics are then identified through the stepwise variable selection technique. The results are valuable in providing building energy performance scores for the whole portfolio and benchmarking. Additionally, it offers insights for the energy consumption level of new buildings. In the second step, to accommodate building heterogeneity, we build the VBDD regression models separately for each building in the portfolio. These models are used to separate the base load energy consumption from the weather dependent usage. The results in this step consist of the base temperature estimates, as well as the estimated coefficients for the weather dependent variables, i.e., Heating Degree Days (HDD) and Cooling Degree Days (CDD) for all buildings. In the third step, we further conduct root cause analysis, by building the multivariate regression models for the base load and coefficient for HDD and CDD resulting from VBDD model, from which the performance scores can be derived for base load, heating, and cooling. Finally, in the last step, we model the dependent error structure through the ARIMA model. We also include seasonal factors in the model. The analytical method provides useful information to track and forecast the energy consumptions of the building portfolio, which will help facility staff and property managers achieve significant energy savings, greenhouse gas emission reductions and cost savings.

INTRODUCTION

Saving energy, improving efficiency of energy consumption, lowering energy cost, and reducing greenhouse gas emissions are key initiatives in many cities, municipalities and for building owners and operators. According to the World Business Council for Sustainable Development, buildings account for 40% of the worlds total energy consumption and, in 2005, nine gigatons of global carbon dioxide (CO2) emissions, well ahead of transportation and industry (WBCDS, 2009; DOE, 2008a). In the United States alone, commercial and residential buildings account for 38% of all CO2 emissions and 72% of electricity consumption according to the U. S. Department of Energy (DOE, 2008b,c). Furthermore, buildings use 13.6% of all potable water, or 15 trillion gallons per year, and 40% of raw materials globally (3 billion tons annually) (USGS, 2000; Roodman and Lenssen, 1995). Much of the energy consumption by commercial buildings is spent on lighting (twenty-six percent), followed by heating and cooling (thirteen percent and fourteen percent, respectively) (DOE, 2007).

With the U. S. building sector's energy consumption expected to increase by 35% between now and 2025 and commercial energy demand projected to grow at an average annual rate of 1.6% reaching 25.3 quads [1], or equivalently 25.3×10^{15} British thermal units (Btu), in 2025, a critical need exists to develop and deploy emerging energy-efficient technologies that can deliver reliable energy demand reductions throughout a building's lifespan while simultaneously satisfying the building occupants comfort, satisfaction and productivity (LBNL, 2009). Investing in energy efficient light bulbs and insulation materials and in automated shading has proven to reduce the energy demands on cooling and lighting (Lee et al., 2007). However, incremental improvements achieved by implementing individual energy efficient technologies alone are not sufficient to the successful achievement of the challeng-

[1]A quad is a unit of energy equal to 1015 BTU (British Thermal Units). The quad is commonly used when describing national or global energy budgets. A quad is approximately equal to $293,071,000,000$ kwh. http://www.aps.org/policy/reports/popa-reports/energy/units.cfm

ing objectives set forth by the Intergovernmental Panel on Climate Change (IPCC) and other directives issued by cities, for example PLANYC 2030 in New York City (NYC) (IPCC, 2007; PLANYC, 2007).

PLANYC aims to reduce the city government's energy consumption and CO2 emissions by 30% by 2030 from 2005 levels. New York City's government spends over $1 billion a year on energy on their approximately 4,000 buildings (e.g. public schools, prisons, court houses, administrative buildings, waste water treatment plants, etc.). NYC plans to invest, each year, an amount equal to 10% of its energy expenses in energy-saving measures over the next 10 years. The largest segment of NYC government buildings are the 1,400 K-12 public schools, serving 1.1 million students and covering about 150 million square feet. The New York City Department of Education was interested in understanding how energy efficient their buildings are, what factors contribute to inefficiencies, what are the opportunities for improvement given budget constraints, and how and how much can they contribute to saving energy and reducing GHG emissions toward NYC's PlaNYC initiative.

As an important component of the IBM *smarter planet*$^{\mathrm{TM}}$*initiatives* (IBM, 2010b), the focus area of the *smarter buildings*$^{\mathrm{TM}}$is the development of new technologies that may help us to improve building energy efficiency and reduce greenhouse gas emissions. According to IBMs Smarter Planet$^{\mathrm{TM}}$Primer (IBM, 2010a), "A smarter building integrates major building systems on a common network. Information and functionality between systems is shared to improve energy efficiency, operational effectiveness, and occupant satisfaction." A smarter building is a complex system of systems that span heating and air conditioning, lighting, security, access control, entertainment, people movers, water, and monitoring and control and maintenance systems. Together, these systems have well managed and integrated physical and digital infrastructures that make the building safe, comfortable, and functional for its occupants and sustainable for the environment. A smarter building uses sensors, digital smart meters, digital controls, and analytic tools to automatically monitor and control services for its users. Thus, a smarter building is transforming into an instrumented, interconnected, and intelligent energy system which will help enable greenhouse gas reductions and lower costs while empowering building users, facility managers and building owner/operators. The advantages of installing smarter buildings on a massive scale are tremendous given that buildings account for 40% of the world's total energy consumption.

Developed along this effort is the *IBM Building Energy and Emission analytics* (*i-BEE*$^{\mathrm{TM}}$) Toolset, a new analytical tool which assesses, benchmarks, diagnoses, tracks, forecasts, simulates and optimizes energy consumption in building portfolios. Our focus in this paper is the statistical methodology in *i-BEE*$^{\mathrm{TM}}$, which is developed for detecting anomalies, forecasting and root cause analysis of monthly electricity, gas and steam consumption.

The problem of analyzing and monitoring building energy performance is a key step to improve energy efficiency and to reduce environmental impact and cost. As an initial effort of this initiative, IBM collaborates with the City University of New York (IBM, 2011) to analyze the energy use in the portfolio of K-12 public school buildings in New York City. We use this building portfolio as our test bed example in this paper. The building portfolio consists of about 1400 public school buildings, covering 150 million square feet. In addition, we collect relevant information such as weather, energy and building characteristics. Our objective is to develop a statistical methodology to help understand the energy use patterns throughout the school portfolio.

We develop a multi-step statistical analysis procedure, which combines the multivariate regression model, the Variable Base Degree Day (VBDD) regression model (Kissock et al., 2003) and the Auto Regressive Integrated Moving Average (ARIMA) model, to assess energy use and identify energy saving opportunities for large portfolios of buildings. In the first step, we build a regression model which correlates the energy consumption with building characteristics. The energy related building characteristics are then identified through the stepwise variable selection technique. The results are valuable in providing building energy performance scores for the whole portfolio. Additionally, it offers insights for the energy consumption level of new buildings. In the second step, to accommodate building heterogeneity, we build VBDD regression models separately for each building. These models are used to separate the base load energy consumption from the weather dependent usage. The results in this step consist of the base temperature estimates, as well as the estimated coefficients for HDD and CDD for all buildings. In the third step, we further conduct root cause analysis, by building the multivariate regression models for the results from VBDD model, from which the performance scores can be derived for base load, heating, and cooling. The VBDD regression model is a popular approach to analyze energy consumption, which assumes an independent error structure for the regression model. The assumption may not be realistic in practice because serial correlations exist

for building energy time series data, especially for our application with a large portfolio of buildings. To overcome this shortcoming, in the last step, we model the dependent error structure through the ARIMA model. We also include seasonal factors in the model. From our experience, the VBDD model, combined with the ARIMA model for the error structure, typically provides improved statistical performance compared to using VBDD alone. The results are used for detecting abnormal energy use and forecasting energy consumption for a portfolio of buildings.

The proposed technique provides an integrated analysis for building heterogeneity, the weather dependent patterns and the temporal dependent patterns. It has wide applicability in anomaly detection, forecasting and root cause analysis for building energy portfolios. In the remainder of this paper, we will first describe the general modeling framework, followed by the application of using the test bed example of the NYC school building portfolio.

DEVELOPING THE STATISTICAL TOOLKIT

To motivate the approach we take to model energy use of building portfolios, it is useful to begin at the end, and consider the type of outputs that will result from the methodology. From the statistical toolset to be developed, we wanted to be able to answer the following questions:

1. Which building parametric data (e.g., building characteristics, operational activities and occupant behavior) is the most useful for predicting building energy use?
2. How can we benchmark the relative building energy performance within the portfolio?
3. What percentages of the total energy use are due to base load, heating use and cooling use, respectively?
4. What are the potential improvement opportunities / root causes for less efficient buildings?
5. How can we offset the weather dependent factors, and perform improvement tracking and energy savings from retrofit activities?
6. How can we detect abnormal energy use in the historical energy use data?
7. How much energy do we expect to use in the future?

To address these questions, we develop a multistep statistical modeling strategy. The statistical models utilize typical data collected about the building energy portfolio, such as

- energy use data for each building;
- building characteristics such as the gross floor area (GFA), age of the building, occupant density, and number of each equipment type (e.g., refrigerator, freezers, etc);
- building operation and activity;
- weather data such as outside temperature and relative humidity.

The statistical modeling strategy we developed consists of the following three major modules

- Variable Based Degree Day (*VBDD*) model with building effect for each building;
- Multivariate regression models (*Multi-regress*): one for the overall energy use of the whole portfolio, and ones for base load, heating, cooling which utilize the outputs from *VBDD* of each building;
- Time series models (*TS-model*), which utilize the outputs from *VBDD*.

The system is best described by the schematic given in Figure 1. We will discuss the modeling details in the rest of this section. We note that these three modules can be integrated, in order to answer the aforementioned questions, as follows.

- *Multi-regress* module is used to answer questions 1, 2.
- *VBDD* module is used to answer questions 3, 5.
- *VBDD* module and *Multi-regress* module are combined together to answer question 4.
- *VBDD* module and *TS-model* module are combined together to answer questions 6 and 7.

BUILDING EFFECTS VBDD MODEL

To better manage the energy portfolio of the New York public school buildings, it is very important to first understand the energy usage patterns for all buildings. The overall energy consumption for commercial buildings like the NYC public school buildings can typically be divided into the following three categories of usage: base load, heating and cooling. Here, the base load refers to the energy consumption that does not depend on outside temperature. Typical usage that falls into this category includes cooking, lighting and hot water usage, and plug loads such as computers. In contrast, the heating and cooling usage depend on the outside temperature. Specifically, there exists some balance-point temperature such that the space-heating energy usage increases as the outdoor temperature decreases below the balance-point temperature, whereas the space-cooling energy use increases as the outdoor temperature increases above the balance-point temperature. We use the following notations to describe the model development. Denote the total number of buildings in the portfolio by n, the total number of months of the billing cycle by m, and the number of days in month t by d_t, respectively. Let T_{itd} be the average outdoor temperature for building i on day d of month t, $i \in \{1, \ldots, n\}$, $t \in \{1, \ldots, m\}$,

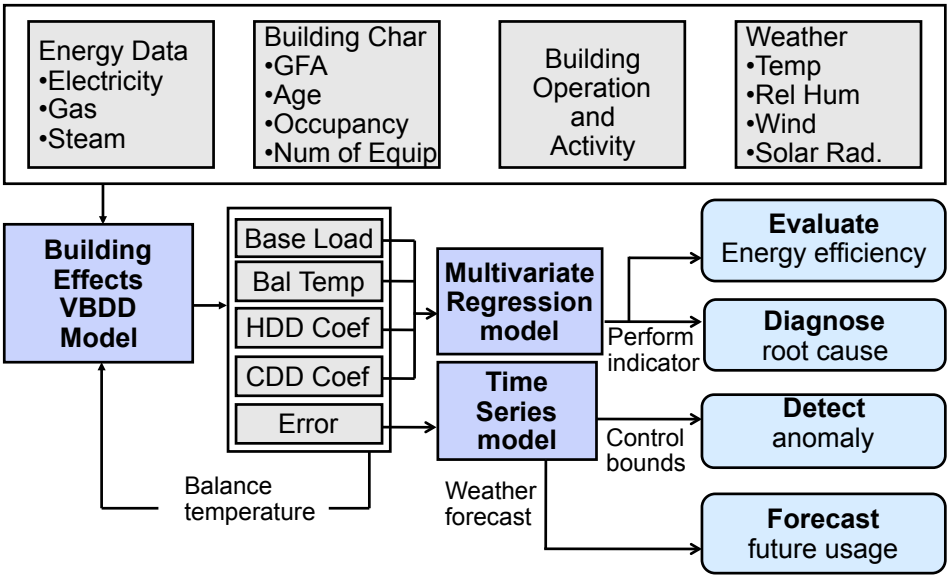

Figure 1: Building energy portfolio analysis system

$d \in \{1, \ldots, d_t\}$. Denote the balance-point temperature for building i by $T_i^{(b)}$. Since only the total monthly energy usage data is available, following Kissock et al. (2003), we define the heating degree day (HDD) and the cooling degree day (CDD) for building i in month t as

$$\text{HDD}_t(T_i^{(b)}) = \sum_{d=1}^{d_t} (T_i^{(b)} - T_{itd})^+ ,$$

and

$$\text{CDD}_t(T_i^{(b)}) = \sum_{d=1}^{d_t} (T_{itd} - T_i^{(b)})^+ .$$

Here, $\text{HDD}_t(T_i^{(b)})$ and $\text{CDD}_t(T_i^{(b)})$ are the cumulative heating and cooling energy usage for month t when the balance point temperature is set to be $T_i^{(b)}$. The total monthly energy usage data for building i, y_{it}, can be further modeled as

$$y_{it} = b_i + c_i \text{CDD}_t(T_i^{(b)}) + h_i \text{HDD}_t(T_i^{(b)}) + \epsilon_{it} , \quad (1)$$

where b_i is the base load usage, c_i is the cooling coefficient, h_i is the heating coefficient, and ϵ_{it} are the error terms reflecting the month-to-month variations that can not be explained by base, heating or cooling usage. We further restrict that $b_i > 0$, $c_i > 0$ and $h_i > 0$.

The model in (1) is also known as the "four-parameter change point model" and is proposed in Kissock et al. (2003) to measure retrofit savings. As noted by Kissock et al. (2003), this model is particular appropriate for modeling the heating and cooling energy use in buildings with high latent loads (energy usage that cannot be calibrated) like the NYC public school buildings. Despite the same modeling strategy, we utilize the results from model (1) in broader ways compared to Kissock et al. (2003). In addition to the quantification of retrofitting savings as in Kissock et al. (2003), we can further conduct peer comparison on the base loads, heating coefficients and cooling coefficients. To be specific, we develop multivariate regression models on the estimated base loads, heating and cooling coefficients with respect to the school characteristics, the result of which can be utilized for root cause analysis, providing insights for root cause of energy consumption and energy savings from retrofit. The model in (1) also serves as our first attempt to remove the energy use trend that is dependent of the outdoor temperature. As we shall see, after removing such trend, we can further model the error terms ϵ_{it} by more sophisticated statistical models such as the auto-regressive integrated moving average models (ARIMA), which allows for more flexible data structures.

To fit model (1), we first select a set of possible values for the balance point temperature through the whole portfolio, denoted by $\mathcal{T} = \{T_1, \ldots, T_q\}$. For the NYC school building portfolio, we find that $\mathcal{T} = \{55, 56, \ldots, 70\}$ is a reasonable range for the balance point temperatures (measured in Fahrenheit) throughout the portfolio. For each building, we then iterate through $T \in \mathcal{T}$, and estimate the parameters by the ordinary least squares (OLS) estimate, subject to the constraint $b_i \geq 0$, $c_i \geq 0$, $h_i \geq 0$, as follows. For a given $T_k \in \mathcal{T}$, we can further represent model (1) in matrix-vector form,

$$\boldsymbol{Y}_i = \boldsymbol{X}_{ik} \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i ,$$

where $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{im})'$, $\boldsymbol{\beta}_i = (b_i, c_i, h_i)'$, $\boldsymbol{\epsilon}_i =$

$(\epsilon_{i1}, \ldots, \epsilon_{im})^{'}$, and

$$\boldsymbol{X}_{ik} = \begin{pmatrix} 1 & \text{CDD}_1(T_k) & \text{HDD}_1(T_k) \\ 1 & \text{CDD}_2(T_k) & \text{HDD}_2(T_k) \\ \vdots & \vdots & \vdots \\ 1 & \text{CDD}_m(T_k) & \text{HDD}_m(T_k) \end{pmatrix}.$$

We can then obtain the OLS estimate for $\boldsymbol{\beta}_i$ under $T_i^{(b)} = T_k$ as

$$\widehat{\boldsymbol{\beta}}_i^{(k)} = (\boldsymbol{X}_{ik}^{'}\boldsymbol{X}_{ik})^{-1}\boldsymbol{X}_{ik}^{'}\boldsymbol{Y}_i\,,$$

and we set the final estimates to be equal to 0 if the OLS estimate is negative. We then calculate the R^2 under $T_i^{(b)} = T_k$ as

$$R_{ik}^2 = 1 - \frac{\sum_t (y_{it} - \hat{y}_{it}^{(k)})^2}{\sum_t (y_{it} - \bar{y}_i)^2}\,,$$

where $\hat{y}_{it}^{(k)}$ is the fitted value for the y_{it} under T_k, and $\bar{y}_i$ is the sample mean of y_{it}. We hence select the balance point temperature according to the best fit to the data, i.e., the temperature which is associated with the largest R^2.

We summarize the algorithm for fitting model (1) as follows.

Algorithm: Fitting the VBDD model in (1)

- *Input*: monthly energy usage data y_{it}, $i = 1, \ldots, n$ and $t \in \{1, \ldots, m\}$, and outdoor temperature T_{itd}, $d \in \{1, \ldots n_t\}$.
- For $i \in \{1, \ldots n\}$ and $T_k \in \mathcal{T}$,
  1. Calculate CDD_t and HDD_t.
  2. Estimate b_i, c_i, h_i by OLS, subject to $b_i > 0$, $c_i > 0$ and $h_i > 0$. Denote the resultant estimates by $\hat{b}_i^{(k)}, \hat{c}_i^{(k)}, \hat{h}_i^{(k)}$.
  3. Calculate R_k^2.
  4. Select k such that $R_k^2 = \max_{l \in \{1,\ldots,q\}} R_l^2$, and set $\hat{T}_i^{(b)} = T_k$.
- *Output*: $\hat{T}_i^{(b)}$, $\hat{b}_i^{(k)}, \hat{c}_i^{(k)}, \hat{h}_i^{(k)}, \hat{\epsilon}_{it}^{(k)}$, $i = 1, \ldots, n$, $t = 1, \ldots, m$.

MULTIVARIATE REGRESSION MODELS

For large building portfolios like the NYC school system, the amount of data is usually huge. With the energy data being collected cumulatively over time for each building, a quick overview of the energy performance throughout the portfolio is critical for successful management. The U.S. Environmental Protection Agency (EPA) Energy Star Performance Rating EPA (2009) provides a valuable management tool to assess the building energy efficiency, relative to similar buildings and climate zones nationwide. In addition to the nationwide comparison, the assessment of energy efficiency, relative to peer buildings within the portfolio, which provides local and more detailed information, is also of great value for the building managers. This calls for the development of a "local" performance indicator, solely based on the data within the portfolio. In this section, we will develop multivariate regression models, which lead to a building performance indicator for local portfolio assessment.

We first describe the general method utilized to derive the building performance indicator. Let y_i be the quantity of our interest, typically referred to as the response variable, and $x_{i1}, \ldots, x_{ip}$ be the p predictor variables. The multivariate regression model takes the form

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \epsilon_i\,, \qquad (2)$$

where $\beta_1, \ldots \beta_p$ are the regression coefficients and ϵ_i is the error term. As in EPA (2009), we utilize the multivariate regression model in (2) to normalize the response variable according to the predictor variables.

In the building portfolio context, the predictor variables consist of the building characteristics and operational activities, for example, gross square feet, number of windows and number of operating hours. Some of the predictors are not directly energy related. We thus perform a stepwise variable selection procedure, to remove the redundant variables from the model.

The multivariate regression model in (2) also serves as a reference to the energy use for the general population in the building portfolio. As a result, while fitting the model, we remove data that cannot represent the general population (e.g., outliers). In practice, a data point is identified as an outlier and be removed from the subsequent analysis if its absolute value of the standardized residual is larger than 2. After removing the outliers, we fit the model again and use the resultant estimates for further analysis.

The expected value of the response variable, for a specific building with given characteristics, can be immediately calculated as

$$\hat{y}_i = \text{E}(y_i \mid x_{i1}, \ldots, x_{ip}) = x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + \ldots + x_{ip}\hat{\beta}_p\,. \qquad (3)$$

The standardized residual, $\hat{z}_i$, can be then calculated as

$$\hat{z}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}}\,,$$

where $\hat{\sigma}$ is the standard error. We note that for a portfolio with many buildings, the $\hat{z}_i$ approximately follows the standard normal distribution. Matching z_i to the standard normal curve, we can calculate the probability that buildings with the same characteristics consume more energy than building i as

$$\text{P}(Z > \hat{z}_i) = 1 - \Phi(\hat{z}_i)\,,$$

where Z is a standard normal random variable and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. We further define the performance indicator for building i as

$$\text{PI}_i = 100 \times \text{P}(Z > \hat{z}_i) = 100 - 100\Phi(\hat{z}_i)\,.$$

The resulting performance indicator is a number between 0 and 100, with larger values indicating better efficiency. In addition, the interpretation of the performance indicator is very intuitive. By definition, the performance indicator is the percentage of similar buildings that consume more energy than building i in the portfolio. We illustrate this calculation in Figure 2.

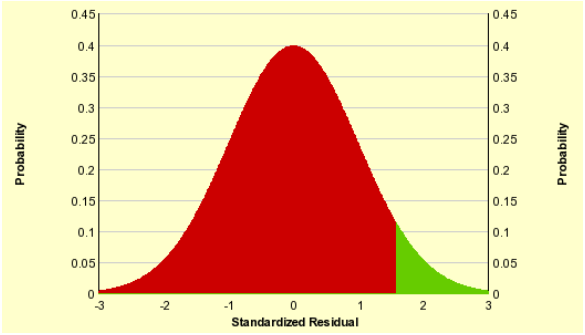

Figure 2: An illustration of the performance indicator. The bell curve is the density function for the standard normal distribution. The corresponding value at the boundary between the red area and the green area is the standardized residual for the specific building. The green region represents buildings that consume more energy than the specific building. The red region represents buildings that consume less energy than the specific building. The performance indicator is defined as the area of the green region multiplied by 100*, which is 10 for this example.*

To assess the overall energy use efficiency, we can build a multivariate regression model for the average energy use for each building in the portfolio. In addition, we use the multivariate regression models in conjunction with the building effects VBDD models described in the previous Section. Indeed, by using $\hat{b}_i$, $\hat{c}_i$, $\hat{h}_i$ obtained from the building effects VBDD models as the response variables, we obtain three additional performance indicators, which reflect the energy efficiency of the base load, heating use and cooling use. These three additional performance indicators provide valuable insights for root cause analysis of building energy performance.

Finally, we note that the multivariate regression model not only provides a valuable normalization tool, but also serves as an energy use evaluator in the stage of budget planning and operational scheduling for a new building.

TIME SERIES MODELS

Recall that $\hat{\epsilon}_{it}$, $t = 1, \ldots, m$, are the estimated error terms from the VBDD models. In the context of building energy analysis, these error terms typically exhibit marked seasonal and dynamic patterns, which may be human behavior related and cannot be explained solely by the weather. In the NYC school system, for example, the error terms typically have drops in summer and holiday seasons, due to less energy demand during those periods. In addition, the data are also dynamically structured, with potential auto-correlations between the errors at consecutive time points. In this section, we develop time series models to further investigate the underlying temporal dependent structures.

The time series modeling, which consists of two major steps, is conducted for each individual building independently. In this first step, we remove the seasonal patterns by using a regression model, with $\hat{\epsilon}_{it}$ being the response variable and the 12 monthly seasonal factors being the predictor variables. To avoid the collinearity issue, we set the monthly seasonal factor for December equal to 0. We denote the error terms after removing the seasonal patterns by $\tilde{\epsilon}_{it}$.

In the second step, we model the dynamic structure of $\tilde{\epsilon}_{it}$ by the autoregressive integrated moving average (ARIMA) model. The ARIMA model is developed to model time series data, for better understanding of the present data and accurately forecasting future data points. See Brockwell and Davis (2006) for a complete view of the statistical developments on an ARIMA model. Despite its popularity in statistical literature, the ARIMA model has been rarely used in the context of building energy, partly because of its complex modeling schemes. Nevertheless, the ARIMA model provides a more flexible, possibly non-stationary structure to model building energy patterns, which is essential for simultaneously modeling of a large number of buildings.

Let p, d, q be non-negative integers. $\{\tilde{\epsilon}_{it}, t = 1, \ldots, m\}$ is said to follow an ARIMA(p, d, q), if

$$(1 - \sum_{\ell=1}^{p} \phi_{i\ell} \mathrm{L}^{\ell})(1 - \mathrm{L})^{\ell} \tilde{\epsilon}_{it} = (1 + \sum_{\ell=1}^{q} \theta_{i\ell} \mathrm{L}^{\ell}) \eta_{it}\,,$$

where L is the lag operator, $\mathrm{L}\tilde{\epsilon}_{it} = \tilde{\epsilon}_{i,t-1}$; p, d, q are the orders of auto-regressive, integrated, and moving average parts of the model; $\{\phi_{i\ell}, \ell = 1, \ldots, p\}$ and $\{\theta_{i\ell}, \ell = 1, \ldots, q\}$ are the parameters associated with the auto-regressive and moving average parts of the model; and η_{it} are mutually independent standard normal random variables. The ARIMA models are the most general class of models for forecasting a time series which can be stationarized by transformations such as

differencing. In fact, the order of the integrated part d reflects the trend of the data (e.g., $d = 0$ no trend, $d = 1$ linear trend, $d = 2$ quadratic trend, etc), while p and q control how fast the auto-correlation decays.

One practical issue in fitting ARIMA models for building portfolio is the model choice, i.e., the choice of p, d, q. With a large number of buildings under consideration, detailed modeling for each individual building is not feasible. To automatically choose the appropriate order for each building, we utilize the Bayesian Information Criterion (BIC) (Schwarz, 1978), which is defined as

$$\text{BIC} = -2\log(\text{Maximum Likelihood}) + k\log(n)\,,$$

where k is the number of parameters and n is the number of observations. BIC is a criterion for selecting the optimal model from a class of parametric models. It overcomes the overfitting problems by regularizing the number of parameters. We select the optimal model for building i as the one which results in the smallest BIC value.

The ARIMA models have two primary applications in the building energy management: anomaly detection and forecasting future energy use. To use the ARIMA model to perform anomaly detection, we first construct the 95% confidence bounds for the energy use history. A data point, which falls outside the confidence bounds, will be considered as an abnormal consumption. An alert will follow, requiring further investigation on the abnormal usage. In terms of forecasting, we first calculate the expected weather dependent energy use, where the weather data is estimated as the average of weather trajectory. We then add the seasonal factors, and the remaining terms forecast by the ARIMA models to forecast the overall energy use. We illustrate the applications of ARIMA model to detect anomaly and forecast in the context of the test bed example in the section to follow.

CASE STUDY

In this section, we demonstrate the method proposed in the previous section using the NYC K-12 public school building portfolio. The portfolio consists of four types of energy use data, i.e., electricity, natural gas, fuel oil and steam, from July, 2005 to September, 2010. We perform the analysis for all energy types, but here we focus on the electricity data for demonstration purposes. For building characteristics and operational activities, the data consists of information such as whether a building has cooking facilities, whether a building has a swimming pool, whether a building has mechanical ventilation, number of floors, whether a building is open during weekends, the number of months a building is in operation, weekly operating hours, percentage of area air conditioned, percentage of area heated, number of personal computers, number of students, student capacity, year a building is built and the gross floor area. We note that some characteristics are presently set to the default values throughout the whole data set and therefore are removed from consideration while we build the models. In fact, collecting accurate building characteristic information is still an ongoing effort. Nevertheless, the general modeling framework remains the same with the currently available data.

Exploratory analysis suggests we take the logarithm transformations on the overall energy use and the gross floor area while building the multivariate regression model for the overall energy use. In addition, to include the building age effects in the model, we note that the building ages are clustered, and we further define the following four age factors according to these clusters: (1) prior to 1915 (2) between 1916 and 1945 (3) between 1946 and 1985 (4) after 1986. The stepwise variable selection procedure suggests that the gross floor area, percentage of area air conditioned, number of students, number of personal computers, number of floors, whether a building has cooking facilities, whether a building was built after 1986 are related to electricity use. We thus include these variables in the final model. According to this model, we can further calculate the overall energy use efficiency for all buildings in the portfolio. In Figure 2, we show the building performance indicator for the overall energy use for a particular building. We refer to this building as the demo building. In the rest of this section, we will illustrate the rest results using the demo building.

We further develop the *VBDD* models for all buildings, about $1,400$ buildings, in the portfolio. In Table 1, we show the resultant estimates for balance temperature, base load (kBtu), cooling coefficient and heating coefficient for the demo building. We further calculate the cooling energy and the heating energy used for a particular year using the results in this table. Based on these results, we further divide the fiscal year energy use into base load, heating energy use, cooling energy use, and other use.

Table 1: Estimated Base load, cooling coefficient and heating coefficient from the VBDD model for the demo building. The estimated balance temperature is in Fahrenheit and the rest are in the unit of kBtu.

Balance Temperature	60
Base Load	521216.90
Cooling Coef	153.01
Heating Coef	103.88

To conduct the root cause analysis, we build the multivariate regression models for the estimated base load, cooling coefficients and heating coefficients from the *VBDD* models. We can then further rank the performance indicators from smallest to the largest, which corresponds to the retrofit priority for this building. Table 2 shows the performance indicators for the base load, the heating use and the cooling use for the demo building. As we can see from this table, the demo building has average performance for heating (performance indicator equals to 50), below average performance for base load (performance indicator equals to 25), and below average performance for cooling (performance indicator equals to 10). Since the cooling performance is the worst among all three perspectives, we assign the top priority to retrofit the cooling system of the building.

Table 2: The performance indicators to analyze the root cause for the demo building. First column: possible root cause. Second column: performance indicators. Third column: retrofit priority.

Root Cause	**PI**	**Priority**
Base Load	25	2
Cooling	10	1
Heating	50	3

Finally, we fit the ARIMA models to the error terms. For the demo building, the model that best fits the data is an ARIMA$(0, 0, 1)$. From this model, we further derive the predicted values, along with 95% confidence intervals from the ARIMA model for the error terms. Combining the predicted values and confidence intervals for the error terms with the base load, heating use and cooling use, we obtain the predicted values and 95% confidence intervals for the original energy use data. The 95% confidence intervals are defined as the control limits. Any data points that fall outside the control limits is defined as an anomaly. An alert is issued to call for further investigation whenever an anomaly occurs in the energy use history. *i-BEE* also provides the next 12 month energy forecast. To forecast future energy use, we first forecast the error terms using the ARIMA model. We then add the base load, heating use and cooling use to forecast the overall energy use. In generating the forecast for heating and cooling use, we use the average heating degree days and cooling degree days in the weather history. Weather forecast or user input can also be used in forecasting future energy use. We show a chart that describes the anomaly detection and forecast for the demo building in Figure 3. Two anomalies are illustrated in the chart, calling for further investigation.

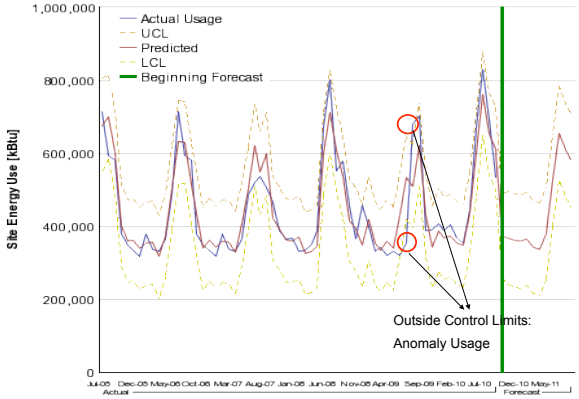

Figure 3: Anomaly detection and energy forecast for the demo building (x-axis: month, y-axis: site energy use in kBtu). Anomaly detection period starts from the beginning of the trajectory and ends before the start of the green vertical bar. Forecast period begins after the green vertical bar. In the picture, the blue line represents the actual energy use; dashed yellow lines represent the upper control limit (UCL) and lower control limit (LCL); the red line represents the predicted values for the energy use.

CONCLUSION

In this paper, we propose a multi-step statistical analysis toolkit to model and analyze the energy use in large portfolios of buildings. We develop and make integrated use of multivariate regression models, building effects VBDD models, and the ARIMA models. The results obtained using the proposed methodology provide valuable insights for anomaly detection, energy forecast and root cause analysis. We implemented our method for the NYC K-12 public school building portfolio and demonstrated its usefulness.

The research contents presented in this paper are our initial efforts to develop a statistics based toolkit for building energy portfolios. Another thread of research activities at IBM is to develop physics based models using inverse modeling and parameter estimation techniques. Directly based on heat and mass transport phenomena, these models provide estimates for heat transfer parameters with real physical meanings, such as the R-value for roof, R-value for wall, U-value for windows, and infiltration coefficient. It would be of interest to combine the strengths of statistical methods with physics based models through integrated modeling, where information produced from one type of modeling is used for the other type of modeling. We plan to pursue this direction in our future research and development.

ACKNOWLEDGEMENT

We wish to thank our colleagues contributing to this project: Paul Nevill, Estepan Melikse-

tian, Pawan Chowdhary, Lianjun An, Raya Horesh, Chandra Reddy, Young Tae Chae of IBM Reseach; Janine Belfest of Optimal Green Operations.

REFERENCES

Brockwell, P. J. and Davis, R. A. 2006. *Time Series: Theory and Methods*. Springer, 2 edition.

DOE 2007. Buildings energy data book.

DOE 2008a. http://www.eia.doe.gov/oiaf/1605/ggrpt/carbon.html.

DOE 2008b. Assumptions to the annual energy outlook.

DOE 2008c. Eia annual energy outlook.

EPA 2009. Energy star performance ratings: Technical methodology. *United States Environment Protection Agency*.

IBM 2010a. Ibm smarter planet primer. http://smarterplanet.tumblr.com/post/58347576/a-smart-planet-primer-to-view-full-screen-click.

IBM 2010b. Smarter Planet Initiatives. www.ibm.com/smarterplanet/global/files/us__en_us_buildings__green_buildings.pdf. [Online].

IBM 2011. City university of new york and ibm to reduce energy consumption in public school buildings. http://www.ibm.com/press/us/en/pressrelease/34080.wss.

IPCC 2007. Ipcc fourth assessment report: Climate change 2007: Mitigation of climate change. http://www.ipcc.ch/publications_and_data/publications_and_data_reports.htm#1.

Kissock, J. K., Haberl, J., and Claridge, D. E. 2003. Inverse modeling toolkit (1050rp): Numerical algorithms. *ASHRAE Transactions*, 109:425–434.

LBNL 2009. Daylighting the new york times headquarters building. Lawrence Berkeley National Lab, http://windows.lbl.gov/comm_perf/newyorktimes.htm.

Lee, E. S., Hughes, G. D., Clear, R. D., Fernandes, L. L., Kiliccote, S., Piette, M. A., Rubinstein, F. M., and Selkowitz, S. E. 2007. Daylighting the new york times headquarters building, final report: Commissioning daylighting systems and estimation of demand response. Lawrence Berkeley National Laboratory, Berkeley, CA.

PLANYC 2007. Planyc 2030: A greener, greater new york. http://www.nyc.gov/html/planyc2030/html/home/home.shtml.

Roodman, D. M. and Lenssen, N. 1995. *A Building Revolution: How Ecology and Health Concerns Are Transforming Construction*. Worldwatch Institute.

Schwarz, G. E. 1978. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

USGS 2000. Us geological survey 2000.

WBCDS 2009. Transforming the market: Energy efficiency in buildings.