WattHome: A Data-driven Approach for Energy Efficiency Analytics at City-scale

Srinivasan Iyengar, Stephen Lee, David Irwin, Prashant Shenoy, Benjamin Weil
University of Massachusetts Amherst

ABSTRACT

Buildings consume over 40% of the total energy in modern societies and improving their energy efficiency can significantly reduce our energy footprint. In this paper, we present WattHome, a data-driven approach to identify the least energy efficient buildings from a large population of buildings in a city or a region. Unlike previous approaches such as least squares that use point estimates, WattHome uses Bayesian inference to capture the stochasticity in the daily energy usage by estimating the parameter distribution of a building. Further, it compares them with similar homes in a given population using widely available datasets. WattHome also incorporates a fault detection algorithm to identify the underlying causes of energy inefficiency. We validate our approach using ground truth data from different geographical locations, which showcases its applicability in different settings. Moreover, we present results from a case study from a city containing >10,000 buildings and show that more than half of the buildings are inefficient in one way or another indicating a significant potential from energy improvement measures. Additionally, we provide probable cause of inefficiency andfi nd that 41%, 23.73%, and 0.51% homes have poor building envelope, heating, and cooling system faults respectively.

CCS CONCEPTS

• **Mathematics of computing** → *Bayesian computation*; Markov-chain Monte Carlo methods; • **Computing methodologies** → Anomaly detection; • **Hardware** → Energy metering;

KEYWORDS

Energy efficiency;Bayesian inference;Automated fault detection

ACM Reference Format:
Srinivasan Iyengar, Stephen Lee, David Irwin, Prashant Shenoy, Benjamin Weil. 2018. WattHome: A Data-driven Approach for Energy Efficiency Analytics at City-scale. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219825

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219825

1 INTRODUCTION

Buildings constitute around 40% of total energy and 70% of the overall electricity usage in the United States [1]. Consequently, building energy-efficiency has emerged as a significant area of research in smart grids. A typical city comprises a large number of buildings of different sizes and age. In general, the building stock in many North American and European cities tend to be old—while some are recently constructed, the majority were built decades ago. Moreover, it is not uncommon for buildings to be over a hundred years old [1]. Technological advances in building construction have yielded better-insulated envelopes as well as more energy-efficient air-conditioning, heating furnaces, and appliances, which can reduce the total energy consumption of a building. While newer buildings, as well as older ones that have undergone renovations, have adopted such efficiency measures, most are yet to benefit from such efficiency improvements. Since roughly half of a building's energy usage results from heating and cooling, opportunities abound for making efficiency improvements in cities around the world.

Since a city may consist of thousands of buildings, an essential first step for implementing energy-efficiency measures is to identify those that are the least efficient and thus have the greatest need for energy-efficiency improvements. Interestingly, naive approaches such as using the age of the building or its total energy bill to identify inefficient buildings do not work well. While older buildings are usually less efficient than newer ones, the correlation is shown to be weak [11]. Thus, *age alone is not an accurate indicator of efficiency*, since older buildings may have undergone renovations and energy improvements. Similarly, the total energy usage is not directly correlated to energy inefficiency. First, larger buildings will consume more energy than smaller ones. Even normalizing for size, greater energy usage does not necessarily point to inefficiencies. For example, a single-family home will have a higher energy demand (possibly due to the in-house washer, dryer, and water heater) compared to an identically sized apartment home. Thus,finding truly inefficient buildings requires more sophisticated methods.

In this paper, we present a data-driven approach for determining the least efficient buildings from a large population of buildings within a city or a region using energy data in association with other external public data sources. Such buildings can then become candidates for energy efficiency measures including targeted energy incentives for improvements or upgrades. So far, lack of granular city-wide datasets prevented large-scale energy efficiency analysis of buildings. However, with increasing smart meter installations across a utilities' customer base, energy usage information of buildings is readily available. By 2016, the US had more than 70 million installed smart meters (>700M worldwide) [2]. Also, real estate information describing a building's age, size, and other characteristic are public records in many countries. Further, weather

conditions can be accessed through REST APIs. Reliance on such readily available datasets make our approach broadly applicable.

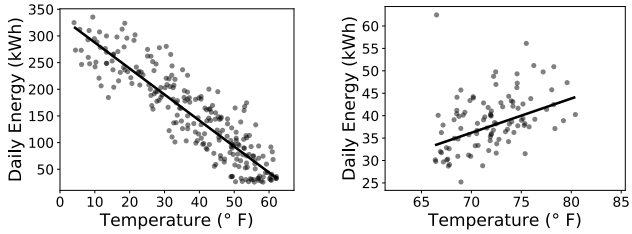
Given these datasets, our approach assumes it is possible to model a building's total energy usage as a sum of *weather-dependent* and *weather-independent* energy components. The weather-dependent component captures the heating and cooling energy usage, which is typically a function of the external temperature, while the weather-independent component captures the energy use from all other activities. Using this approach, we can then extract the parameter distributions that govern these energy components and identify causes of energy inefficiency by comparing them to those of other homes in a given population. For example, a model's parameter that is more sensitive to external temperature is indicative of inefficient heating or cooling. We also develop algorithms that use these comparisons to determine the probable causes of energy inefficiency.

While building energy models have been extensively studied in the energy science research for many decades [6, 15, 28], and practitioners such as energy auditors routinely use them to analyze a building's energy performance, there are important differences between current approaches and our technique. First, current models employ several important parameters that are often chosen manually, based on rules of thumb [21]. However, using manually chosen parameters may lead to incorrect analysis [10]. On the other hand, our technique determines a custom parameter distribution of the building model, and we experimentally show its efficacy over manual approaches. Second, the current energy models are based on least square regression analysis that provides point estimates. In contrast, our approach provides Bayesian estimates to determine building parameter distribution that captures the stochasticity in energy use. Third, current approaches need manual intervention to varying degrees to interpret model parameters and determine likely efficiency issues. Clearly, this does not scale to thousands of buildings across a city. Our technique automates this process by comparing model parameters with similar homes from the population and makes it feasible to perform large-scale analysis. Thus, we go beyond determining which buildings are inefficient by also designing algorithms that determine its probable causes.

In this paper, we introduce WattHome, a data-driven approach to determine the most inefficient buildings present in a city or a region. Our contributions are as follows:

Bayesian Estimation Approach. WattHome improves over prior work that provides point estimates and uses bayesian inference to capture the building model parameter distributions that governs the energy usage of a building. These building parameter distributions are compared using *second-order stochastic dominance* to create a partial order among buildings. Further, we propose a fault analysis algorithm that utilizes these partial orders to report outlier buildings and their probable causes. Moreover, we implement our approach as an *open source* tool that enables determining inefficient buildings at scale and is applicable to other regions or cities.

Model Validation and Analysis. We evaluate WattHome using energy data from two different cities in geographically diverse regions of the US. In particular, we show that our approach can disaggregate the buildings' energy usage into different components with high accuracy and tighter bounds on the model parameters — an improvement over the two popular baselines. In addition to disaggregation, our approach identifies buildings that have possible energy inefficiencies. In comparison to manual audit reports our approach correctly identified faults in nearly 95% of the cases.



(a) Winter Months (b) Summer Months

Figure 1: Linear relationship between energy consumption and ambient temperature for a Single Family home.

Real-world case study analysis. We examine our approach on energy usage from smart meters deployed in 10,107 buildings in a city. WattHome reported more than half of the buildings in our dataset as inefficient, which indicates a significant scope for making energy improvements in several cities. Further, our results indicate poor building envelope as a major cause of inefficiency, which accounts for around 41% of all homes. Heating and cooling system faults comprises 23.73%, and 0.51% of all homes respectively.

2 BACKGROUND

In this section, we present background on energy efficiency in buildings and techniques used to model a building's energy usage.

2.1 Energy Efficiency in Buildings

Energy usage in residential buildings has different sources such as heating and cooling, lighting, household appliances etc. There can be many causes of inefficiencies in each of these components, such as the use of inefficient incandescent lighting and the use of inefficient (e.g., non-energy star) appliances. Studies have shown that heating and cooling is the dominant portion of a building's energy usage, comprising over half of the total usage [1, 24], and it follows that the most significant cause of inefficiency lies in problems with heating and cooling. Two factors determine heating and cooling efficiency of a building: (1) the insulation of the building's external walls and roof ("building envelope") and their ability to minimize thermal leakage, and (2) the efficiency of the heating and cooling equipment. Recent technology improvements have seen advancements on both fronts. New buildings are constructed using modern methods and better construction materials that yield a building envelope that minimizes air leaks and thermal loss through better-insulated walls and roofs and high-efficiency windows and doors. Similarly, new high-efficiency heating and AC equipment are typically 20-30% more efficient than equipment typically installed in the late 1990s and early 2000s.

Unfortunately, older residential buildings and even ones built two decades ago do not incorporate such energy efficient features. Further, the building envelope can deteriorate over time due to age and weather and so can mechanical HVAC equipment. Consequently, an analysis of a building's heat and cooling energy use can point to the leading causes of a building's energy inefficiency.

2.2 Inferring a Building Energy Model

One approach to modeling a building's heating and cooling usage is to model its dependence on weather [31]. For example, a building's heating and cooling usage can be modeled as a linear function of external temperature. To intuitively understand why, consider cooling energy usage during the summer. The higher the outside temperature on hot summer days, the higher the AC energy usage. Since the difference between outside and inside temperatures is large, there is more thermal gain, which requires longer duration of cooling to maintain a set indoor temperature. Thus, there is a linear relationship between heating/cooling energy use and outside temperature (see Figure 1(a) and (b)). Given the linear dependence, linear models are commonly used within the energy science research [15, 25], to capture the relationship between energy use and outside temperature. However, most of the prior approaches do not consider uncertainties that are associated with indicators of building performance. Primarily, these models do not capture the stochastic variations in heating and cooling as well as the weather-independent energy usage resulting from day to day variations in human activities inside a home. As seen in Figure 1, such energy variations exist and our approach uses Bayesian inference to determine the distributions of the building parameter that models these uncertainties in energy use.

2.3 Problem Formulation

Consider a large population of buildings in a city. We assume that a trace of the total daily energy usage is available for each building. We also assume building characteristics, such as age, size, and type (Single Family, Apartment etc.) for each building along with the daily outdoor temperature data are available.

Let B be the set of all residential buildings containing information on building characteristics in a city. Further, $b_i \in B$ denotes the i^{th} residential building defined by a tuple $\langle E^{total}_{i,[1\ldots D]}, Age_i, Size_i, Type_i \rangle$. Here, $E^{total}_{i,[1\ldots D]}$ is the energy usage recorded by smart meters for a period of D days. Moreover, $T_{[1\ldots D]}$ is the external ambient temperature for the city during the D days. Thus, given $b_i \in B$ and $T_d \forall d \in D$, our problem is to determine $(a_1, \ldots, a_m)_i \in \{False, True\}^m$, where $a_1, \ldots, a_m$ are the m possible faults associated with the residential buildings.

3 WATTHOME: OUR APPROACH

In this section, we describe the details of our data-driven approach. WattHome's approach is depicted in Figure 2 and it involves three key steps: (i) Learn a *building energy model* for each home from energy usage data, (ii) Create a *partial order* of buildings using its parameter distribution from the building model, andfi nally (iii) Detect *building faults* causing energy inefficiency. Below, we discuss each step in detail.

3.1 Building Energy Model

Wefi rst provide the intuition behind our approach. Heating and cooling costs for a building can be understood using elementary thermodynamics. Typically, in colder months, the outside ambient temperature is colder than the inside building temperature, resulting in a net thermal loss where the inside heatfl ows outside through

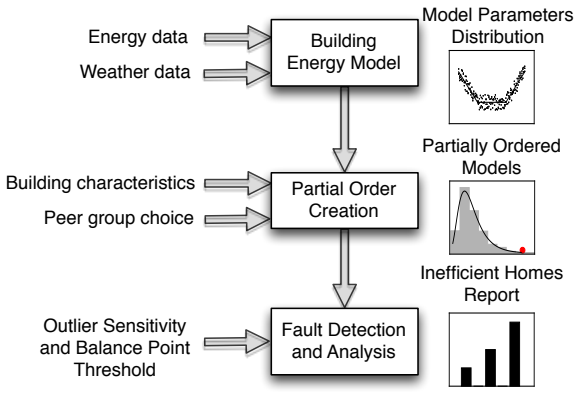


Figure 2: Overview of WattHome approach.

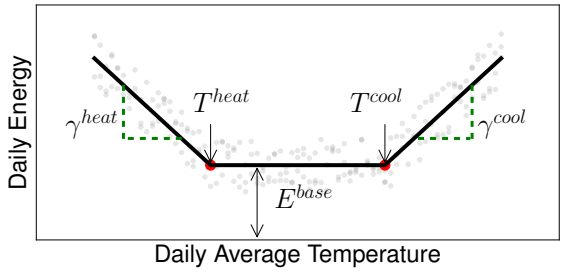


Figure 3: Energy usage versus outdoor temperature.

the building envelope, causing the inside temperature to drop. In warmer months, the opposite is true. The building experiences a net heat gain where the heatfl ows inside, causing the building temperature to rise.

It follows that every home has a specific temperature T_b, where there is neither thermal loss nor thermal gain i.e. the thermodynamic equilibrium. When the outside temperature is above T_b, there is a need for AC to cool the home. Conversely, when the temperature is below T_b, there is a need for a heater to heat the home. This temperature T_b is called the balance point temperature of the building. The rate of thermal loss or thermal gain depends on the degree of insulation, airtightness of the building envelope and surface area exposed to outside elements. Better the insulation and airtightness, smaller the rate of loss or gain for a given temperature differential relative to T_b. The difference between the outside temperature and the balance point temperature T_b is also referred as the *degree-days* — an indication of how many degrees warmer or colder is the outside weather relative to the building's balance point.

Based on this intuition, we now describe our building energy model. Any energy load in a building can be classified as weather independent and dependent. A weather independent load is one where the energy consumed by the device is uncorrelated to the outside temperature — consumption from loads such as lighting, electronic devices, and household appliances depend on human activity rather than outside weather. Heating and cooling equipment

constitute weather dependent loads, as their consumption linearly dependent on the outside temperature relative to the balance point.

If we assume that weather independent loads are distributed around a constant value (also called the base load); then the total energy consumed is the sum of the base load and the weather dependent loads (heating and cooling loads) and defined as:

$$E_d^{total} = E_d^{heat} + E_d^{cool} + E^{base} \qquad \forall d \in D \tag{1}$$

where E_d^{total} denotes the total energy used by a building on day $d \in D$. E_d^{heat} and E_d^{cool} denote the energy used for heating and cooling, respectively, on day d, while E^{base} denotes the energy usage of base load appliances. Thus, given a series of observations of the total energy usage and the outside ambient temperature, it is possible tofi t a regression and learn thefi xed weather independent component (base load) and the temperature dependent component (heating and cooling). This forms the basis for inferring our weather-aware building energy model.

Figure 3 illustrates the relationship between outdoor temperature and the energy consumption of a building. The individual data points represent the daily energy usage (along the Y-axis) for a given average outdoor temperature (along the X-axis) of a building. Thefi gure shows that the building has two balance point temperatures — a heating balance point temperature T^{heat}, below which heating units are turned on, and a cooling balance point temperature T^{cool}, above which air-conditioning is turned on. Further, thefi gure also shows a piecewise linearfi t over the daily energy usage. When the outdoor temperature is between the two balance points, the building consumes energy that is distributed around a constant value defined as the *base load* E^{base} energy consumption. The weather dependent components, i.e. the heating E^{heat} and cooling E^{cool} energy consumption, are a function of ambient outdoor temperature T_d and are defined as:

$$E_d^{heat} = \gamma^{heat}(T^{heat} - T_d)^+ \quad \forall d \in D \tag{2}$$

$$E_d^{cool} = \gamma^{cool}(T_d - T^{cool})^+ \quad \forall d \in D \tag{3}$$

where γ^{heat} and γ^{cool} are the heating and the cooling slope in the above linear equations and represent a positive constant factor indicating the sensitivity of the building to temperature changes; and $()^+$ indicates the value is zero if negative and ensures either energy from heating or cooling is considered. Using (2) and (3), energy model in (1) can be represented as a piecewise linear model:

$$E_d^{total} = E^{base} + \gamma^{heat}(T^{heat} - T_d)^+ + \gamma^{cool}(T_d - T^{cool})^+ \forall d \in D \tag{4}$$

The model in (4) is known as the *degree-day* model [25] and forms our base energy model for estimating the building parameters.

3.1.1 Bayesian Inference Parameter Estimation. While methods like Maximum Likelihood Estimation (MLE) or Maximum a posteriori estimation (MAP) can be used for determining the building parameters, they provide point estimates that can hide relevant information (such as not capturing the uncertainties in human energy usage). To capture human variations, we require probability density function of the parameters. Thus, we use Bayesian inference approach, which provides the posterior distribution of parameters.

Prior
$E^{base} \sim \mathcal{N}(20, 20)$, $\gamma^{heat} \sim \mathcal{N}(0, 4)$, $\gamma^{cool} \sim \mathcal{N}(0, 4)$
$T^{heat} \sim \mathcal{U}(32, 100)$, $T^{cool} \sim \mathcal{U}(32, 100)$, $\sigma \sim Cauchy(0, 5)$
Regression Equation
$\mu_d = E^{base} + \gamma^{heat}(T^{heat} - T_d)^+ + \gamma^{cool}(T_d - T^{cool})^+ \; \forall d \in D$
Model Likelihood
$E_d^{total} \sim \mathcal{N}(\mu_d, \sigma^2)$
Parameter Bounds
$E^{base}, \gamma^{heat}, \gamma^{cool} >= 0$ and $T^{heat} <= T^{cool}$

Table 1: Bayesian formulation of our building energy model.

We model (4) using a bayesian approach and assume the error process to be normally distributed ($\mathcal{N}(0, \sigma^2)$). Thus, the daily energy consumption E_d^{total} is normally distributed with parameters mean (μ) and variance (σ^2), where μ is equal to the right hand side of (4). Note that energy consumption E_d^{total} is known and so is the independent variable i.e. ambient temperature T_d. However, the building parameters (γ^{heat}, γ^{cool}, T^{heat}, T^{cool}, and E^{base}) are unknown. Using Bayesian inference, we can then compute a *posterior* distribution for each of these parameters that best explains the *evidence* (i.e. the known values for E_d^{total} and $T_d \forall d \in D$) from initially assuming a *prior* distribution.

To determine the posterior distribution of the individual parameters, we use the Markov chain Monte Carlo (MCMC) method that generates samples from the posterior distribution by forming a reversible Markov-chain with the same equilibrium distribution. We introduce a prior distribution that represents the initial belief regarding the building parameters. For example, the two balance point temperatures will be between a wide range of 32°F and 100°F. This belief can be represented using a uniform prior with the said range. Similarly, the baseload, heating slope and cooling slope can be drawn from a weakly informative gaussian prior having non-zero values. This is because baseload, a unit of energy, cannot be negative. Similarly, slope values must be positive as they represent increase in energy per unit temperature. The parameters of the gaussian priors are scaled to our setting and selected based on the recommendations provided by Gelman et al. [17]. To simplify our building model, we assume that the parameters are independent, i.e., the heating, cooling and the baseload parameters do not affect one another.

Several MCMC methods leverage different strategies to lead from these priors towards the target posterior distribution. We employed No-U-turn sampler, a sophisticated MCMC method, which has shown to converge quickly towards the target distribution. Thus, after an initial *burn in* samples, we can draw samples approximating the true posterior distribution. From these post-burn-in samples, a posterior distribution for the individual building parameters can be formed. Our complete Bayesian model is defined in Table 1.

Since buildings are of different sizes, simply comparing the parameters in absolute terms is not meaningful. To enable such comparison, we initially normalize the energy usage by building size before the Bayesian inference. Hence, in our case, E^{base} represents base load energy use per unit area. Similarly, heating slope γ^{heat}

and cooling slope γ^{cool} gives change in energy per degree temperature per unit area. Thus, the balance point parameters (T^{heat} and T^{cool}) are not normalized as they are unaffected by the size of the house. We construct a cumulative distribution ($F_{\gamma^{heat}}$, $F_{\gamma^{cool}}$, $F_{E^{base}}$) for each of the building model parameter (γ^{heat}, γ^{cool}, E^{base}) from their respective density functions (posterior) obtained after the inference. For the balance point parameters (T^{heat} and T^{cool}), we only use its mean values as they tend to remainfi xed for a given building irrespective of human variation. This completes our approach for creating the building energy model.

3.2 Partial Order Creation

Rather than relying on rule-of-thumb measures to interpret model parameters that change with geography and many other building characteristics, we propose comparing them with those of similar homes from a given population. Given the above model, we create a partial order of buildings as follows. Wefi rst create *peer groups* using the building's physical attributes (e.g., age of the building, building type etc.). Next, within each peer group we create a *partial order* of the buildings for each building parameter distribution. Below, we describe each step in detail.

3.2.1 Peer groups creation. To enable a meaningful comparison, we compare the building model parameters only within their cohort. We use three building attributes for peer group creation namely: (i) property class (e.g., single family, apartment, etc.), (ii) built area (e.g., 2000 to 300 sq.ft.), and (iii) year built (e.g.1945 to 1965). For instance, buildings constructed in different years adhere to different energy regulations and standards, and thus, it is not meaningful to compare them. Similarly, building types and age group have different characteristics and it would be unreasonable to compare them. Hence, our approach allows the creation of peer groups to enable comparison within a cohort to determine inefficient homes.

3.2.2 Stochastic order of building parameters. Since the building model parameters are probabilistic distributions, we cannot simply compare these uncertain quantities and create a *total ordering*. Statistics, such as mean, median or mode, provide a single number to capture the behavior of the whole distribution. While these *point estimates* can be used to compare two distributions, they typically hide useful information regarding their shape and may not account for any heavy-tailed nature that is present in a building parameter distribution. Hence, we use *second order stochastic dominance*, a well-known concept in decision theory for comparing two distributions [26], to create a partial order of the building parameters within a peer group.

The main idea behind determining *second order stochastic dominance* is that for a given building model parameter p, if distribution F_p dominates G_p i.e., $F_p \succeq_2 G_p$, then the area enclosed between F_p and G_p distribution should be non-negative up to every point in x:

$$\int_a^x (G_p(t) - F_p(t))dt \geq 0 \qquad \forall x \in [a, b] \tag{5}$$

Figure 4 depicts stochastic ordering of two distribution F_p and G_p where; (i) F_p does not dominate G_p i.e. $F_p \not\succeq_2 G_p$ and (ii) F_p dominates G_p i.e., $F_p \succeq_2 G_p$. The area shaded in green shows the region where F_p dominates G_p, and the red region shows G_p dominates F_p. In Figure 4(a), we observe that $F_p \not\succeq_2 G_p$, since there are no green area greater or equal to the left of the red area. In contrast, Figure 4(b) and (c) shows F_p dominates G_p because for every red area, there exists a larger green area located to its left.

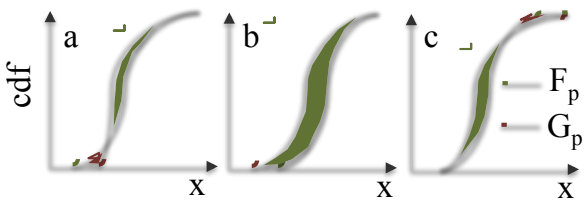


Figure 4: Stochastic ordering of two distributions F_p and G_p. (a) F_p does not dominate G_p. In (b) and (c) F_p dominates G_p.

Indicator Characteristics	Probable Building Faults
High Heating Slope	Inefficient Heater, Building Envelope
High Cooling Slope	Inefficient AC, Building Envelope
High Heating Balance Point	High Set point, Poor Building Envelope
Low Cooling Balance Point	Low Set point, Poor Building Envelope
High Base load	Inefficient Appliances

Table 2: Indicator building model characteristics and associated probable building faults.

To intuitively understand the implications of stochastic dominance in our scenario, let us consider two distributions F_p and G_p of a building parameter p from two separate buildings A and B respectively. As noted earlier, building parameters influences energy usage, such that higher parameter values implies higher energy usage, and vice-versa. Let us assume that building A's normalized energy usage is greater than building G's normalized energy usage, such that distribution F_p dominates G_p i.e., $F_p \succeq G_p$. Clearly, the building parameter distribution F_p for building A will lie on the right-side of distribution G_p as A has higher energy usage. In fact, since $F_p \succeq G_p$, by definition, the distribution F_p will be on the right of G_p for a majority of the region. However, homes may have similar building parameter distribution, i.e the distribution has similar shape and tendency. In such cases, it is possible that neither home will dominate the other. Stochastic dominance thus enables interpretation of the building parameter distribution with respect to one another, with higher energy usage buildings having a tendency to lie on the right side of the population. This allows separation of homes with dominant distributions from non-dominant ones. We run a pair-wise comparison of all buildings within a cohort for each building model parameter p. This gives us the partial order for all pairs and parameters, which we use to detect inefficient homes.

3.3 Fault Detection and Analysis

Wefi rst discuss the causes of inefficiencies associated with the different model parameters. Later, we present our algorithm that identifies inefficient homes and its potential cause.

3.3.1 Parameter relationship with building faults. Heating slope γ_{heat} and heating balance point temperature T^{heat} are the two

parameters that enable our model to interpret the heating inefficiencies of a home. Buildings with high γ^{heat} lose heat at a higher rate, which in turn affects heating unit usage (i.e., consumes more power) to compensate for the high loss rate. A high energy loss rate can be attributed to poor building insulation, air leakages, or inefficient or heating unit. Separately, heating balance point temperature also indicates inefficiencies in the heating component of a home. A high balance point temperature suggests two possible inefficiencies: (i) high thermostat set-point temperature[1] and (ii) poor building insulation. If the set-point temperature is high during winters, heating units turn on more frequently to maintain the indoor temperature at set-point. In contrast, if building insulation is poor, more heat is lost through the building envelope. Thus, heating units will be turned on frequently to sustain the high heating balance point temperature. Similarly, we can interpret the cooling slopes γ^{cool} and cooling balance point temperature, which points to inefficiencies in cooling units or building envelope.

Homes with high E^{base} indicate high appliance usage or inefficient appliances. In such homes, energy retrofits may not help reduce energy consumption. However, these homes may benefit from replacing old appliances (water heater, dryer) with newer energy star rated ones. We summarize the association between probable causes of building faults and model parameter in Table 2.

3.3.2 Fault Analysis Algorithm. Wefi rst use the partially ordered set of buildings to determine the outliers for each parameter and then use the mapping in Table 2 to assign building faults. To determine outliers, note that the energy usage of an inefficient home would be high. Thus, the building parameter distribution of an inefficient home will tend to be *stochastically dominant* with respect to others in their peer group. However, among inefficient homes, the building parameter distribution may be similar, and thus their distributions may not be stochastically dominant to one another. Similarly, within energy efficient homes this distinction of dominance may not be apparent, as their distribution may be identical to one another. We use this insight to define a building as *inefficient* in a given model parameter, if it is stochastically dominant compared to a majority of the homes within its cohort. For instance, if a building's heating parameter distribution $F_{\hat{\gamma}^{heat}}$ is dominant across more than τ% of the buildings, we conclude that the building is inefficient and has a *high* heating slope. Here, τ is the sensitivity threshold for WattHome and provides thefl exibility to control the number of inefficient homes. The higher the threshold value, the higher the possibility of identifying an inefficient home. For all experiments, we chose this to be 75%. Thus, for each parameter, we determine whether a building is inefficient if its distribution is dominant beyond a certain threshold. We use a balance point threshold to determine buildings with high balance point temperature. Wefl ag buildings as inefficient if the mean value obtained after inference for heating (or cooling) balance point temperature T^{heat} (or T^{cool}) is greater than (less than) specific heating (or cooling) balance point threshold 70°F (55°F) — a common choice employed by expert auditors. We present the pseudo-code to determine inefficient buildings in Algorithm 1.

[1]Set point temperature and balance point temperature have a linear relationship

Algorithm 1 Fault Analysis Algorithm

Inputs: Sensitivity (τ), buildings (B)
procedure FindInefficientHomes(τ, B)
  count = {}; homes = {}
  for p in [$\gamma^{heat}, \gamma^{cool}, E^{base}$] **do**
    for ($b1$, $b2$) $\leftarrow$ $^{|B|}P_2$ **do** // all-pairs permutation
      if $F_p(b1) \geq_2 F_p(b2)$ **then**
        count[p, $b1$] +=1
    for $b \leftarrow B$ **do** homes[p, b] = count[p, b] $\geq \tau$
  for $b \leftarrow B$ **do** homes[T^{heat}, b] = $T_b^{heat} > 70°F$
  for $b \leftarrow B$ **do** homes[T^{cool}, b] = $T_b^{cool} < 55°F$
  return homes
Inputs: building (b), parameters (P), fault_map (M)
procedure GetRootCause(h, P, M)
  faults = []
  for $p \leftarrow P$ **do**
    if homes[p, b] **then**
      faults += $M[p]$ // append list
  return faults

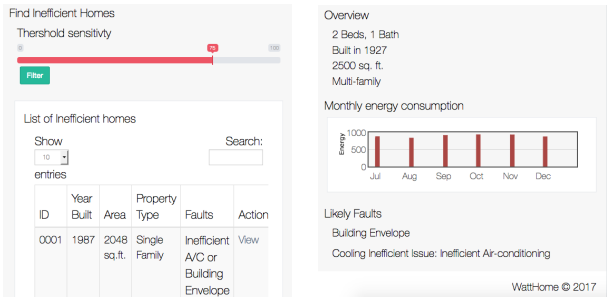
As noted earlier, each parameter in the building model affects an energy component defined in (4). Any irregularity in the building parameter, in comparison to its peer group, points to possible inefficiency in the said energy component. We outline our pseudo-code forfi nding root cause in Algorithm 1. First, we create a mapping of indicators of deviations in building model parameters to possible faults using Table 2. We provide the mapping as an input to our algorithm. Next, we associate a fault to a home if it wasfl agged inefficient for the given parameter p. For instance, if a home isfl agged as high base load, we say that the home has inefficient appliances. Similarly, an inefficient home with high heating slope is assigned faults related to heating inefficiencies. We then generate a report of the list of potential faults in a given home.

4 IMPLEMENTATION

We implemented WattHome as an open source tool. WattHome is split into two components — (i) a Unix-like command line tool [2] that uses PyStan, a statistical modeling library, to implement our bayesian model, and (ii) a web-based application interface implemented using Django framework for interacting with the command line tool. Users can interact with either component, and provide their energy traces and building information, to determine likely reasons of inefficiency.

Our system works as follows. When users provide their energy traces and building information (such as zip code, year built, etc.), WattHome builds a custom bayesian model of the home using the local weather data and the details provided by the user. The weather data of a nearby airport is used as a proxy for local weather conditions, and WattHome periodically fetches and updates this data from public APIs. Next, users provide a sensitivity threshold that is used to create a partially ordered set of inefficient homes. As utility companies may have a limited audit budget to manually

[2]We have publicly released the code and the tool. http://bit.ly/2nU7kA5



(a) Find Inefficient homes (b) Inefficiency Report

Figure 5: Screenshot of our implementation of WattHome.

Characersitics	Dataset 1	Dataset 2
# of Homes	163	10,107
Duration	2013	2015
Built Area Range (sq.ft.)	758-6516	250-10,000
Year Built Range	1912-2014	1760-2013
Location	Austin, TX	A city in New England

Table 3: Key characteristics of Dataport and New England-based utility smart meter dataset

inspect homes, the threshold provides user thefl exibility to control the list of least efficient home. Figure 5(a) shows how users can adjust the sensitivity parameter to get inefficient homes. Finally, our WattHome generates a report listing inefficient homes and their likely faults. Figure 5(b) shows the inefficiency report for a single home listing likely faults.

5 EXPERIMENTAL VALIDATION

Wefi rst validate our model estimates against ground truth data from two cities and evaluate its efficacy.

5.1 Dataset Description

5.1.1 Dataset 1: Dataport (Austin, Texas). Ourfi rst dataset contains energy consumption information from homes located in Austin, Texas from the Dataport Research Program [3]. The dataset contains energy breakdown at an appliance level, which serves as ground truth to understand how our approach disaggregates energy components. We select a subset of homes (163 in total) from this dataset having HVAC, baseload appliances along withthe total energy usage information. Since most homes enrolled in the Dataport research program are energy-conscious homeowners, and have energy efficient homes, we use this dataset only for validating our energy disaggregation process.

5.1.2 Dataset 2: Utility smart meter data (New England). This dataset contains smart meter data for 10,107 homes from a small city in the New England region of the United States [20]. The dataset has energy usage (in kWh) from both electricity and gas meters. Each home may have more than one smart meter — such as a meter to report gas usage and another to report electricity usage. For homes with multiple meters (gas and electric), we combine

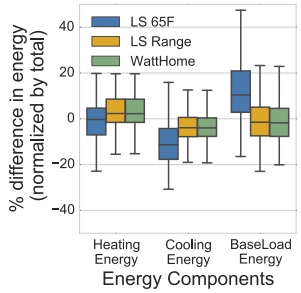


Figure 6: Validation of energy split using the two baselines and our model.

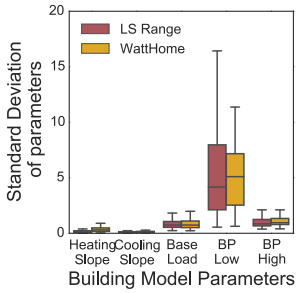


Figure 7: Comparison of the standard deviation of parameters.

their energy usage to determine the building's daily energy consumption for an entire year (2015). Apart from energy usage, the dataset also contains real estate information that includes building's size, the number of rooms, bedrooms, property type (single family, apartment, etc.). We also have manual audit reports for some of the homes. We use this as our ground truth data for validating our approach. Further, we have weather information of the city containing average daily outdoor temperature. We summarize the characteristics of both the datasets in Table 3.

5.2 Energy Split Validation

We now validate the efficacy of our model in disaggregating the overall energy usage into distinct energy components, i.e., heating, cooling, and baseload. For this experiment, we restrict our analysis to the 163 homes from the Dataport dataset.

We compare our technique with two baseline techniques (*LS 65F* and *LS Range*), commonly used in prior work, which use the degree-days model to provide point estimates of the individual building model parameters. Ourfi rst baseline technique, *LS 65F*, estimates the three building energy parameters (γ^{heat}, γ^{cool}, σ, E^{base}) using least-squaresfi t and assumes the balance point temperature to be constant (65°F). This is a widely used approach by energy practitioners around the US and recommended by official bodies such as ASHRAE [7]. Our second baseline technique, *LS Range*, estimates all thefi ve building energy parameters (γ^{heat}, γ^{cool}, T^{heat}, T^{cool}, and E^{base}) using the least-squaresfi t. Unlike the baseline approaches, WattHome estimates the parameter distribution and thus to compare we use the mean of the posterior distribution of the parameters to get thefi xed proportion of the energy splits.

Figure 6 shows the distribution of percentage difference in the energy usage with the ground truth for each energy component. While *LS Range* and *WattHome* have median error of ≈-1.6%, *LS 65F* have a median error of 10% for baseload energy. Unlike *LS 65F*, *LS Range* and *WattHome* do not assume a constant balance point temperature and thus have lower error. Figure 7 compares the standard deviation of the building parameters from the two approaches. In *WattHome*, the standard deviations are obtained from the parameter posterior distributions. Whereas, in case of *LS Range*, the standard deviations are calculated from the covariance matrix outputted by the least-squares routine. While the results for

the four parameters are similar, the spread of standard deviation for the lower balance point is much smaller in *WattHome* compared to *LS Range*. Thus, *WattHome* provides an equivalent or tighter bound compared to *LS Range*.

Summary: *Fixed parameters provide poor estimate of the building parameter. WattHome provides lower error and tighter parameter estimates compared to other baseline techniques.*

5.3 Faulty Homes Validation

We now examine the accuracy of our model in reporting homes with likely faults. We ran our algorithm on all homes in the New England dataset to generate a list of outlier homes for each of the parameter and then compare our results withfi ndings from manual energy audits (ground truth). Since manual audit reports contain faults related to building envelope and HVAC devices only, we only report these results and inefficiencies arising from base energy usage and faulty set points were not analyzed.

To determine the accuracy, we compare an inefficient building's parameter to the audit report conducted in the past and verify whether it has any building faults. The audit reports were manually compiled by an expert on-field auditor identifying and suggesting energy efficiency improvement measures. Wefi nd that WattHome reported 59 homes with building envelope faults, out of which 56 buildings were in the audit report, an accuracy of 95%. Moreover, wefi nd that 46 of the 56 homes with building envelope faults also had faulty HVAC systems.

Summary: *WattHome identified parameter related faults in a building with high accuracy. In particular, our approach correctly identified 95% of the homes that werefl agged by expert auditors as having either faulty building envelope or HVAC systems.*

6 CASE STUDY: IDENTIFYING INEFFICIENT HOMES IN A CITY

We conduct a case study on the New England dataset to determine the least efficient residential buildings in the city. In particular, we seek to gain insights on the following questions: (i) What percentage of the homes are energy inefficient? (ii) Which groups of homes are the most energy inefficient? (iii) What are the most common causes of energy inefficiency? Wefi rst provide a brief analysis of the distribution of the energy split.

6.1 Energy Split Distribution Analysis

To get thefi xed proportion of the energy split, we use the mean of the posterior estimates to compute the disaggregated energy usage i.e. heating, cooling and base load components. To compare the energy components, we compute the *Energy Usage Intensity* (EUI), by normalizing the energy component with the building's built area. Figure 8(a) shows the heating, cooling, base load and total EUI distribution grouped by property type across all homes. The figure shows that the base load is the highest component of energy usage in most Mixed Use and Apartment property types followed by heating and cooling. However, for Single family homes, the heating cost is usually higher. The high base load can be attributed to lighting, water heating, and other appliances. Further, since the New England region has more winter days, homes require more heating, and thus expected to have a higher heating energy footprint compared to cooling. In particular, the average heating energy required is almost 20× that of average cooling energy. We also observe that the normalized total energy usage of single and multi family homes is the highest — presumably due to more number of appliances. The median energy EUI of the Single family home is ≈53 kBtu/sq.ft. (1 kW=3.412kBtu), which is almost twice that of Apartment homes (≈26.8 kBtu/sq.ft.).

Observation: *Heating energy consumption is 20× that of cooling energy on an average. Energy consumption among Single and Multi family homes is much higher than Apartment or Mixed use homes.*

Heating Outliers	Cooling Outliers	Base load Outliers	Overall Outliers
3162	1033	2016	5079

Table 4: Summary of all inefficient homes in the data set.

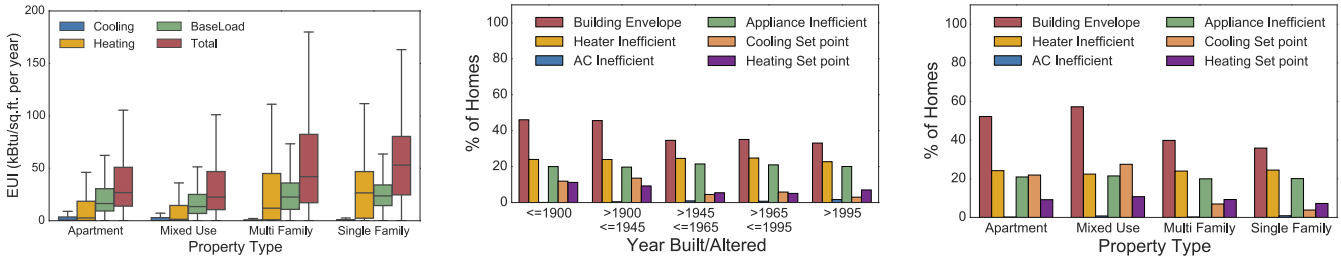
6.2 Efficiency Analysis

In this section, we analyze the results of our approach on the utility company's dataset described earlier. We created peer groups to identify inefficient homes in their respective cohort. To do so, we used three building attributes (property type, age, and area), which created 120 peer groups in total. Among these peer groups, we discarded groups with less than 20 homes, as it didn't have enough population size for a meaningful analysis. In all, 67 peer groups containing a total of 186 homes were discarded. Below, we present our analysis on the remaining 9,921 homes.

6.2.1 Identifying inefficient homes. We examine the number of homes that arefl agged as inefficient for each of the energy components using our approach. Table 4 shows the summary of inefficient homes across all peer groups. We note that a home may have multiple inefficiencies, such as inefficient heating and high base load and thus may be inefficient in several of the energy components. Our results show that the overall percentage of inefficient homes across all residential homes is 50.25%. Further, almost 62.25% of all inefficient homes have either inefficient heater or poor building envelope, and 4144 homes have either inefficient heating or cooling.

Observation: *More than half of the buildings in our dataset are likely to be energy inefficient, of which almost 62.25% homes have inefficient heating as a probable cause.*

6.2.2 Identifying faults in inefficient homes. We now analyze the cause for inefficiency in these inefficient homes. Figure 8(b) shows the percentage of inefficient homes within each building age group across all faults. Note that a home may have multiple faults. We observe that the building envelope fault is the major cause of inefficiency, followed by inefficiency in heaters and other base load appliances. Across all age groups, nearly 41% of the homes have building envelope faults, while 23.73% and 0.51% homes have heating and cooling system faults respectively. Thefi gure also shows that some homes might have set point faults. In particular, 18.06% of the homes have issues with either high heating or low cooling set point temperature. These faults indicate likely issues with thermostat setting. Adjusting the thermostat set point temperature in these home may likely improve its efficiency. As shown, homes



(a) Total Energy Split (b) By Building Age (c) By Property Type

Figure 8: (a) Disaggregated energy usage for all homes. (b) and (c) Possible fault types in different building groups.

built/altered before 1945 have a higher proportion of inefficient homes. However, the percentage difference with other age groups is <15%.

Figure 8(c) shows the percentage of inefficient homes within each building property type and faults. We observe that the building envelope faults are the most common faults across all building types. Further, wefi nd that except for HVAC appliance related faults, mixed use property type has the highest percentage of inefficiency in the remaining fault categories. After mixed use property type, apartments tend to have a higher percentage of inefficient homes followed by multi family and single family property types.

Observation: *Building envelope faults is one of the major cause for inefficiency and present in nearly 41% of homes. However, 18.06% of homes have thermostat set point faults. Changing their set-point may likely improve efficiency in these homes.*

7 RELATED WORK

Diagnosing and reducing energy consumption in buildings is an important problem [9, 16, 23, 32]. Various methods have been proposed to detect abnormal energy consumption in a building [13, 23, 27]. However, these methods focused on commercial buildings that require expensive building management systems [13, 27] or requires costly instrumentation using sensors for monitoring purposes [9, 22]. Sensors allowfi ne-grained monitoring of energy usage but are not scalable due to high installation costs. Unlike prior approaches, our model does not require building management systems or costly instrumentation and use ubiquitous smart meter data to determine energy inefficiency in buildings.

Prior work have also proposed automatic modeling of residential loads [5]. Studies have shown that compound loads can be disaggregated into basic load patterns. Separately, there has been studies on non-intrusive load monitoring (NILM), which allow disaggregation of a household's total energy into its contributing appliances, and does not require building instrumentation [8, 18]. However, most NILM techniques requirefi ne-grained datasets for training purposes and assume energy consumption patterns are similar across homes [8]. On the other hand, our approach makes no such assumption on energy consumption patterns and is applicable across multiple homes as it uses coarse-grained energy usage data that are readily available from utility companies [4].

Various energy performance assessment methods exist to quantify energy use in buildings and identify energy inefficiency [19, 29, 30]. A common approach is to use degree-days method, a linear regression model, for calculating building energy consumption [14, 15, 25]. However, these approaches do not consider uncertainties that are associated with indicators of building performance. The idea of modeling uncertainties in thermal comfort is studied in [12]. However, it is restricted to a single office building with cooling and heating systems. Unlike previous studies, our approach can be used to identify least energy efficient home at scale without manual expert intervention. Further, we propose a novel Bayesian model to account for uncertainties arising from human factors. Finally, we use actual ground truth data to validate our approach and show its efficacy on a large scale city-wide data.

8 CONCLUSIONS

Improving efficiency of buildings is an important problem, and thefi rst step is to identify inefficient buildings. In this paper, we proposed WattHome, a data-drive approach to identify the least energy efficient homes in a city or region. We also implemented our approach as an open source tool, which we used to evaluate datasets from different geographical locations. We validated our approach on ground truth data and showed that our model correctly identified 95% of the homes with inefficiencies. Our case study on a city-scale dataset showed that more than half of the buildings in our dataset are energy inefficient in one way or another, of which almost 62.25% of homes with heating related inefficiencies as probable cause. This shows that a lot of buildings can benefit from energy efficiency improvements.

As part of future work, we intend to deliver individual inefficiency report generated from our web application to the different homeowners. These nudges can be used to motivate and incentivize homeowners towards energy efficiency measures.

ACKNOWLEDGMENTS

We thank all the reviewers for their insightful comments that helped us improve the paper. This research is supported by NSF grants IIP-1534080, CNS-1645952, CNS-1405826, CNS-1253063, CNS-1505422 and the Massachusetts Department of Energy Resources. Any opinions,fi ndings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] 2016. U.S. Energy Information Administration. (visited on May 2016). https://www.eia.gov/. (2016).

[2] 2017. Advanced Metering Infrastructure - AMI - is a fundamental part of the grid's evolution. https://goo.gl/CkXh92. (October 2017).

[3] 2017. Dataport dataset. https://dataport.cloud/. (2017).

[4] 2017. Green Button Initiative. http://www.greenbuttondata.org/. (2017).

[5] M. Aftab, C.K. Chau, and M. Khonji. 2017. Real-time Appliance Identification using Smart Plugs: Demo Abstract. In *Proceedings of the Eighth International Conference on Future Energy Systems*.

[6] J.C. Allen. 1976. A modified sine wave method for calculating degree days. *Environmental Entomology* (1976).

[7] FUNIP ASHRAE. 2013. Fundamentals handbook. *IP Edition* (2013).

[8] N. Batra, O. Parson, M. Berges, A. Singh, and A. Rogers. 2014. A comparison of non-intrusive load monitoring methods for commercial and residential buildings. (2014).

[9] G. Bellala, M. Marwah, M. Arlitt, G. Lyon, and C. Bash. 2012. Following the electrons: methods for power management in commercial buildings. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*.

[10] M A Brown, M Cox, B Staver, and P Baer. 2014. Climate change and energy demand in buildings. *Proceedings of the American Council for an Energy Efficient Economy (ACEEE) Summer Study on Energy Efficiency in Buildings*. (2014).

[11] William Chung, YV Hui, and Y Miu Lam. 2006. Benchmarking the energy efficiency of commercial buildings. *Applied energy* 83, 1 (2006), 1–14.

[12] S. De Wit. 1997. Influence of modeling uncertainties on the simulation of building thermal comfort performance. In *Building Simulation*.

[13] C. Fan, F. Xiao, and S. Wang. 2014. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy* (2014).

[14] H. Fei, Y. Kim, S. Sahu, M. Naphade, S.K. Mamidipalli, and J Hutchinson. 2013. Heat pump detection from coarse grained smart meter data with positive and unlabeled learning. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*.

[15] M. Fels. 1986. PRISM: An Introduction. *Energy and Buildings* (1986).

[16] R. Fontugne, J. Ortiz, N. Tremblay, P. Borgnat, P. Flandrin, K. Fukuda, D. Culler, and H. Esaki. 2013. Strip, bind, and search: a method for identifying abnormal energy consumption in buildings. In *Proceedings of the 12th international conference on Information processing in sensor networks*.

[17] Andrew Gelman et al. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis* 1, 3 (2006), 515–534.

[18] G. Hart. 1992. Nonintrusive appliance load monitoring. *Proc. IEEE* (1992).

[19] J.S. Hygh, J.F. DeCarolis, D.B. Hill, and S.R. Ranjithan. 2012. Multivariate regression as an energy assessment tool in early building design. *Building and Environment* (2012).

[20] S. Iyengar, S. Lee, D. Irwin, and P. Shenoy. 2016. Analyzing Energy Usage on a City-scale using Utility Smart Meters. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*.

[21] P. Jacobs and H. Henderson. 2002. State-of-the-art review of whole building, building envelope, and HVAC component and system simulation and design tools. *Architectural Energy Corporation* (2002).

[22] H. Janetzko, F. Stoffel, S. Mittelstädt, and D.A. Keim. 2014. Anomaly detection for visual analytics of power consumption data. *Computers & Graphics* (2014).

[23] S. Katipamula and M. Brambley. 2005. Review article: methods for fault detection, diagnostics, and prognostics for building systemsÑa review, Part I. *HVAC&R Research*.

[24] J. Kelso (Ed.). 2012. *Buildings Energy Data Book*. Department of Energy.

[25] J. Kissock, J. Haberl, and D. Claridge. 2002. *Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models*. Technical Report. Texas A&M University.

[26] H. Levy. 2015. *Stochastic dominance: Investment decision making under uncertainty*. Springer.

[27] J.E. Seem. 2007. Using intelligent data analysis to detect abnormal energy consumption in buildings. *Energy and buildings* (2007).

[28] HCS Thom. 1954. The rational relationship between heating degree days and temperature. *Monthly Weather Review* (1954).

[29] S. Wang, C. Yan, and F. Xiao. 2012. Quantitative energy performance assessment methods for existing buildings. *Energy and Buildings* (2012).

[30] C. Yan, S. Wang, and F. Xiao. 2012. A simplified energy performance assessment method for existing buildings based on energy bill disaggregation. *Energy and buildings* (2012).

[31] H. Zhao and F. Magoulès. 2012. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews* (2012).

[32] Q. Zhou, S. Wang, and Z. Ma. 2009. A model-based fault detection and diagnosis strategy for HVAC systems. *International Journal of Energy Research* (2009).