AUTOMATICALLY CALIBRATING A PROBABILISTIC GRAPHICAL MODEL OF BUILDING ENERGY CONSUMPTION

Daniel Tarlow[1], Andrew Peterman[2], Benedict R. Schwegler[2,3] and Christopher Trigg[3]

[1] Department of Computer Science, University of Toronto, Canada
[2] Research and Development, Walt Disney Imagineering, Glendale, California
[3] Center for Integrated Facility Engineering, Stanford University

ABSTRACT

We introduce a framework and proof of concept for estimating building energy consumption that probabilistically combines a model of building physics with observed occupancy and detailed operations data, automatically learning a physically plausible model of the energy consumption.

Our framework has several desirable properties: data about one building can automatically be used to improve energy use estimates for other similar buildings; input fields can be left blank or specified approximately; and the output of our model is not only an estimate of energy usage, but a probability distribution over possible values.

We describe an initial implementation of our framework and present experimental results showing that this is a promising direction for future building simulation research.

INTRODUCTION

The Walt Disney Company (TWDC) has aggressive energy reduction goals for its worldwide parks and resorts. The ability to accurately estimate energy consumption for a wide variety of existing and not yet constructed buildings would allow facility managers to predict the optimum mix of capital improvements, occupant behavior incentives and operational efficiency measures needed to meet these energy reduction goals.

TWDC traditionally has used energy modeling tools such as Trane's TRACE and DOE-2 to size mechanical equipment for current and future needs. Unfortunately, TWDC has encountered difficulty in creating energy models that accurately predict energy consumption in a particular building in absolute terms—mainly, TWDC's current simulation techniques give accurate predictions of peak loads, but fail to predict daily energy consumption well enough to make energy efficiency recommendations. Especially in its theme parks, TWDC's buildings have unusual operational patterns and policies that traditional energy simulation tools have difficulty representing faithfully without extensive calibration. For example, building occupancy in Disney theme parks varies complexly as a function of day of week, season, weather, and several other factors. In addition, operational conditions such as near continuous open doors in retail and food service locations can be difficult to model using traditional simulation tools. However, within a theme park, the operational patterns of buildings share many similarities, so learning how one building deviates from typical operations provides information about how other buildings within the theme park are likely to deviate.

Gathering detailed information about building specifications, occupant behavior, weather, and climate can be difficult. Buildings inherently possess inconsistencies due to material aging effects, remodeling, and post-design construction changes or defects. Total park occupancy can be measured easily, but the distribution of people within each building in the park is more difficult. Weather is often not measured directly on site, and weather stations can have different weather patterns than building locations, even if they are only a few miles away. Current models often fail to capture the certainty, effect, and/or the degree to which these factors deviate from design conditions.

In lieu of precisely knowing input values, an engineer using a traditional building simulator is forced to make ad hoc estimates of input values based mainly upon years of experience. No confidence specifications are given in these estimates, potentially leading to the introduction of unquantified, arbitrarily large errors into the simulation process. Indeed, it is well known that it is a significant challenge to get energy simulator estimates to match collected real-world data (Maile et al. 2007).

Improved calibration can mitigate some errors for existing instrumented buildings, but it is often not feasible to install instruments, perform proper calibration and verify results for the large number of buildings found in organizations like TWDC.

The key idea of our work is to build a model that has two complementary components: first, we follow in the spirit of traditional simulators by encoding logical knowledge about the form of physical interactions that affect energy consumption in buildings; second, we specify the parameters governing the physical interactions probabilistically. These specifications—representing our prior beliefs about the physics of the world—are then rigorously combined with collected real-world data using a Bayesian model. We present an algorithm that chooses the model that maximizes the posterior probability of the model and the data– that which best balances physical plausibility with the ability to explain observed data.

Our work deviates from traditional deterministic methods in that we specify our model with *less* precision. We make up accuracy by letting collected data automatically influence our belief about the correct internal settings of the model.

Unlike many other statistical or artificial intelligence-based models of energy usage, we express our model of building physics and operation in terms of interpretable physical quantities, such as the cooling load per people-hours of occupancy. We can interpret—and thus express engineering knowledge and judgment about—many of the internal parameters. Other example parameters in our model are the percentage of people entering a retail shop who make a purchase and the number of air exchanges per hour that the ventilation system of the building uses. This is in contrast to other regression or artificial neural network models, where a coefficient of some predictor or weight on some connection usually does not have a natural interpretation.

We emphasize that we do not know the values of the internal model parameters precisely, but we can express reasonable prior beliefs about their values. Our algorithm intelligently searches for a setting of parameters that is simultaneously consistent with our beliefs about the physics of the building operation *and* that is able to explain data we have collected about building operations and energy consumption.

Finally, when parameters are expressed in easily interpretable terms, we can meaningfully specify which parameters are shared between buildings. Buildings within the same resort or theme park will share weather and occupancy patterns; buildings constructed at the same time with the same specifications will share similar construction details and likelihood of defects; and buildings with the same layout (e.g. copies of a building built both in Hong Kong and Florida) will share process energy loads, building layouts, and potentially even some occupancy patterns. When parameters are shared across buildings, learning from data about one building can be used to make inferences about other similar buildings.

BACKGROUND

Bayesian networks are perhaps the best known variant of a family of models known as *probabilistic graphical models*—an elegant framework able to represent probabilistic interactions along with structure about the form of interactions. They are popular tools that have seen success modeling many complex real-world phenomena in many fields including artificial intelligence, computer vision, and computational biology.

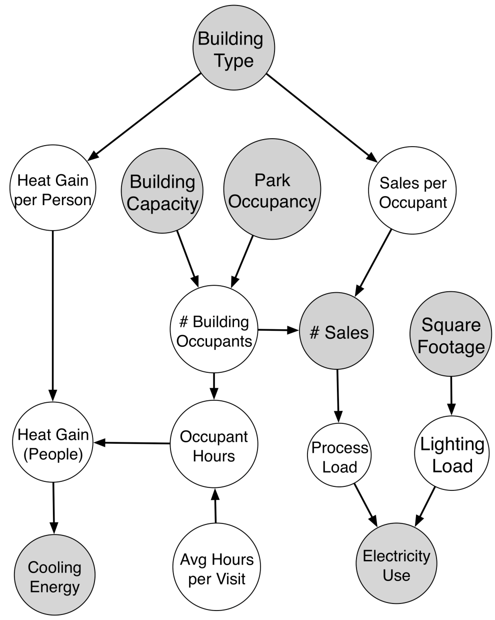

Figure 1 A Bayesian network structure for the occupancy component of our model

Bayesian Networks

Formally, a Bayesian network represents a joint probability distribution P over a set of random variables $X = \{X_1, \ldots, X_n\}$. A Bayesian network consists of two parts:

- A directed, acyclic graph structure G over nodes X.
- A conditional probability distribution for each node: $P(X_j \mid U_j)$ where U_j is the set of nodes that are parents of X_j.

Each conditional probability distribution is governed by a set of discrete or real-valued parameters, which we will refer to as *model parameters*, θ.

When model parameters are fixed, the full joint probability distribution is given as a product of local interactions:

$$P(X;\theta) = \prod_j P(X_j \mid U_j;\theta_j)$$

When the number of parents of each node is relatively small, this representation is significantly more compact than a naïve representation of the full probability distribution, which can be prohibitively large in most non-trivial applications.

Example 1: Figure 1 shows a portion of the Bayesian network structure that models building occupancy. The variable `Number of Building Occupants` (B) is a parent of `Number of Sales` (S), indicating that the probability distribution over the number of sales that are made in a building on a given day is conditional on the number of people in the building. If we assume temporarily that B has no parents and is drawn from a normal distribution, this interaction might take the form

$$B \sim Normal(\mu_B, \sigma_B)$$
$$S \sim \alpha B$$

In this case, μ_B and σ_B are model parameters representing the mean and standard deviation of the number of people in the building, and α is a model parameter representing the percentage of people in the building that make a purchase. By properties of normal distributions, S would then also be distributed normally with mean $\alpha\mu_B$ and standard deviation $\alpha\sigma_B$.

Given a Bayesian network structure and model parameters, probabilistic queries can be efficiently computed. The two most popular queries are finding a posterior probability distribution over values of missing variables given a subset of observed variables; and finding the mode of the full joint probability distribution, possibly conditioned on a subset of other variables. In this work, we are interested only in the former.

Much of the power of the Bayesian network representation is in its ability to robustly deal with missing data. In Figure 1, we have collected at least some data for each of the shaded nodes, while we were unable to measure the values of the unshaded nodes. However, because we are able to roughly specify the form of interactions between variables, we can infer beliefs over the values of missing data, and we can use these beliefs to estimate unknown model parameters.

Bayesian Bayesian Networks

Bayesian networks are named so because of their use of Bayes' rule to make inferences, and—somewhat conterintuitively—it is common to use them in conjunction with frequentist methods. However, Bayesian networks are well suited for Bayesian modeling.

To make Bayesian networks Bayesian, we specify a prior probability distribution, $P(\theta)$, over model parameters. In this way, we are able to quantify our prior beliefs about the physics of the world and how buildings operate. In the case of Example 1, we might specify a prior distribution that says α is roughly .5, but there is a significant standard deviation—say .2—in our belief. This might be formalized by setting $P(\alpha) = Normal(\alpha; .5, .2)$. We emphasize that this is very different from forcing α to take on a value of .5 in our model, because a wide range of values are possible for α, the data we have collected will help us choose the most likely setting. The purpose of the prior distributions is to encourage the final model to be physically plausible, not to assert over-confident influence on the model.

The Expectation Maximization Algorithm

The expectation maximization (EM) algorithm is used to estimate model parameters from data when some variables—such as the number of building occupants—are not observed (Neal & Hinton, 1998). Formally, given a data set of N data instances, $D = \{X^{(i)} \mid i = \{1, \ldots, N\}\}$, where $X^{(i)} = \{X_1^{(i)}, \ldots, X_n^{(i)}\}$, our goal is to find model parameters, θ, that maximize the *posterior* probability of the data and the model parameters:

$$P(D,\theta) = P(\theta) \cdot \prod_i P(X^{(i)} \mid \theta)$$

Let $X^{(i)}$ be divided into hidden and visible variables, $h^{(i)}$ and $v^{(i)}$, respectively. If we integrate over the hidden variables according to the (unknown) posterior distribution of the hidden variables conditioned on the visible variables, then we get the posterior likelihood. In log terms:

$$\log P(D,\theta) = \log P(\theta) + \sum_i \int_{h^{(i)}} \log P(h^{(i)}, v^{(i)} \mid \theta) dh^{(i)}$$

Even though we do not know the conditional distribution over the hidden variables, we can choose an *arbitrary* distribution, $Q(h^{(i)})$, and use Jensen's inequality to lower bound the posterior probability:

$$\geq \log P(\theta) + \sum_i \int_{h^{(i)}} Q_i(h^{(i)}) \log \frac{P(h^{(i)}, v^{(i)} \mid \theta)}{Q_i(h^{(i)})} dh^{(i)}$$

We will refer to this lower bound as the (negative) *free energy* due to its use in statistical physics. Q is parameterized by a set of real numbers that we will refer to as *variational parameters*. Maximizing the negative free energy (or minimizing the free energy) over Q is easier computationally than directly maximizing the log likelihood, and it provably increases a lower bound on the likelihood. The optimal distribution, Q^*, can be shown to be $P(h \mid v, \theta)$, and at this point the bound is tight (Neal & Hinton, 1998).

Intuitively, Q is our probabilistic belief about the unknown values of variables in our data set—if we cannot directly measure a variable's value, then we will maintain a probability distribution over its possible values. In the E step, we revise our beliefs about the missing values. In the M step, we use observed data and our beliefs about missing values to estimate model parameters. This optimization proceeds in an alternating fashion until convergence.

MODEL

Our model is focused on capturing the major contributions to cooling loads and overall building electricity use due to lighting, fans, and process loads. We work at a daily time granularity, performing summations and averages as necessary. Using finer time granularity lends statistical strength to the model, giving us more observations of energy consumption under different occupancy and environmental conditions. We are interested in tackling the challenges associated with collecting finer-grained time granularity, but doing so would require changes in data collection practices and non-trivial changes to the model formulation.

We emphasize that we are not trying to model energy consumption with the precision of complex thermodynamics simulators; we are sacrificing some precision in exchange for accuracy gained by having our model automatically tune itself using collected data. Finally, a direction of future work is to refine the structure of our model via discussion with more experts.

To construct our model, we begin by manually specifying a Bayesian network structure, along with prior distributions over model parameters. In addition, we specify hard constraints about the legal range of some parameters to maintain physical sanity. Nodes in the network represent variables or model parameters, which are either real-valued (e.g., Ventilation Flow Rate or Outdoor Temperature) or discrete (e.g., Date or Building Type). There is one instantiation of each variable for each combination of day and building, while model parameters are shared across all buildings and days. Details about parameters and prior distributions are given in Table 1.

Walls, Window, Roof, and Ventilation

Space precludes us from giving every conditional probability distribution in the network, but the form of interactions follows standard practice such as described in common references (Tao & Janis 2001; ASHRAE 1997). The difference is that deterministic computations are replaced with probabilistic interactions. For example, a typical form for estimating the conductive heat gain through a wall is $H = U \cdot A \cdot TETD$, where H is the heat gain, U is the U value of the wall, A is the area, and TETD is the total equivalent temperature difference. In our model, U, A, and TETD are variables specific to a building-day combination. H is represented as being distributed normally with a mean equal to the product of the three variables and unknown variance. TETD is an unobserved variable influenced by surface properties of the wall, the insolation upon the vertical surface, and the difference between outdoor and indoor temperatures. We split walls and windows up by orientation and represent solar heat gain factors for each orientation using collected data for horizontal insolation, data on the solar zenith angle, and geometry.

We follow similar practices for conductive and radiative heat gain through windows, conductive heat gain through the roof, and heat gain due to ventilation. See Figure 2 for the network structure.

Occupancy and Heat Gained via Internal Loads

Our model of occupancy and internal loads is illustrated in Figure 1.

We do not directly measure the number of occupants inside a building over the course of the day, but we have detailed records of overall park occupancy and transactions in an individual building. We model the number sales as the product of the number of building occupants times the model parameter for sales per occupant. The number of building occupants is itself distributed conditionally on the overall park occupancy on a given day and the individual building capacity.

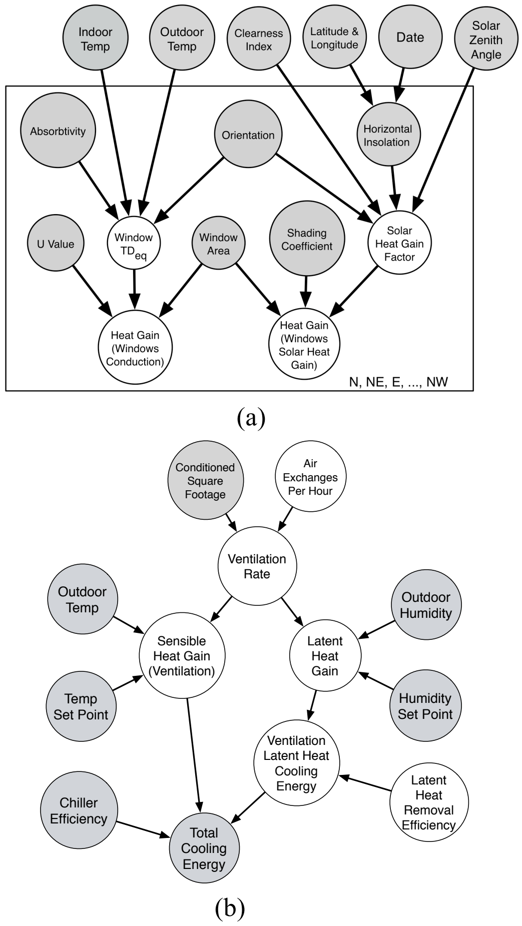

Figure 2 Bayesian network structures for window and ventilation heat gains

Electricity Use

Electricity use is mostly described in Figure 1, but we also model a separate intensity per square foot for kitchen areas, and we include a fan electricity term that is distributed conditionally upon the total cooling energy for the day. Total energy use is the sum of electricity use and cooling energy.

EXPERIMENTAL EVALUATION

We set up experiments to evaluate the ability of our model to learn from real world data then make generalizations about energy consumption in buildings previously unseen by the model.

We chose to work with food service and retail buildings due to their generality, wider applicability to the field, and because we have detailed records of the number of transactions at each location–Disney's worldwide industrial engineering teams maintain detailed occupancy and transaction data at hourly intervals.

Table 1 A Subset of Model Parameters and Priors

Model Parameter	Prior Mean	Prior Std. Dev.	Constraints
Hours / visit	.5	.25	>= 0
Sales / visitor	.75	.5	>= 0
Heat gain (Wh) / person-hour	73	15	>= 0
Total electricity / kWh cooling	1	.5	>= 0
Kitchen electricity per m^2 intensity	.1	.1	>= 0
Retail & dining electricity per m^2 lighting intensity	.05	.05	>= 0
Chiller efficiency	.75	.1	>= 0
Latent heat removal efficiency	-	-	>= 0
Outside wall temperature difference / insolation (kWh)	-	-	>= 0
Air exchanges / hour	1	1	>= 0
Shading coefficient	.85	.15	>= 0
Diffuse radiation coefficient	.1	.1	>= 0

Disney's other main building type—ride or attraction buildings—possesses specific process loads and occupancy patterns not applicable outside the theme park industry. The chosen building types are relatively large energy consumers with the greatest perceived potential for energy efficiency measures.

Data Collection

The Theme Park and Resorts division of TWDC maintains an internal utility reporting system to track, report, and record the status and consumption of electrical and mechanical utility systems on property in California, Florida, Paris, and Hong Kong. For the purposes of this initial study, we identified three retail and food service buildings, all contained in Florida's Walt Disney World Resort, possessing occupancy, energy, and climate data at the proper granularity for our study. Relatively complete data sets of utility, climate, and occupancy exist for the period of approximately two years between 2006 and 2008; however, an important note is that nearly all sources of data contain missing data and outliers. In order to work with the data in more interpretable units, we performed some conversions and energy consumption calculations using standard practices and mechanical efficiency values determined by internal Disney Imagineers—the engineering design division of TWDC (Tao & Janis 2001; ASHRAE 1997).

We used design drawings for each building to determine the wall, window, roof, and room schedules. Rooms were divided into retail, dining, and kitchen areas. We used material properties from Disney design standards and construction documentation.

All-sky insolation—direct normal to the surface of the earth, clearness index, clear sky insolation, and solar zenith angle data at the theme park's longitude and latitude were provided by NASA Langley Research Center Atmospheric Science Data Center. Historical temperature, humidity, and cloudiness data were accessed from the Weather Underground public historical archives on their website.

Variational EM Algorithm

To learn the values of the model parameters, we use a variant of the EM algorithm known as a *fully factorized variational EM algorithm* (Frey & Jojic 2005). Specifically, Q is restricted to take the form $Q(h) = \prod_i Q(h_i)$. This is a common variant, chosen primarily for improved computational efficiency. In the E step, we analytically solve for variational parameter updates by taking partial derivatives of the free energy with respect to each variational parameter. We find a local optimum by setting the partial derivative to zero and solving algebraically for each parameter. In the M step, we simultaneously optimize over all model parameters using a gradient-based constrained optimization routine based on Newton's method.

We initialize the algorithm randomly but within the range of reasonable values dictated by our prior distributions, then alternate E and M step updates until the free energy does not improve further. Typically, most of the gains are made in the first few iterations, then convergence is reached within 100 iterations. Run times are on the order of several minutes for the large data sets and around a minute for the smaller data sets.

Across Building Parameter Estimation

We split 900 days worth of energy use data for each of our three buildings into a training set and a test set. The training set has one-fourth of all cooling energy and electricity use values hidden (held out) for all buildings. All values for a single building are also held out. Points not included in the training set are put in the test set. The algorithm is given the training set and asked to make predictions about the values of data in the test set without looking at held out values.

When no energy use values for a building are given to the algorithm, we call the building the *held out building*. We made three data sets and ran an experiment using each: in experiment 1, building 1 is

Table 2
Total Energy Use Prediction Error (in kWh)

EXP.	LINEAR REGRESSION	PHYSICAL MODEL (PRIOR)	BAYESIAN NETWORK
1	3009.4	1755.6	**1522.2**
2	2892.9	1712.7	**1119. 9**
3	4262.3	**732.1**	1384.2

Table 3
Cooling Energy Prediction Error (in kWh)

EXP.	LINEAR REGRESSION	PHYSICAL MODEL (PRIOR)	BAYESIAN NETWORK
1	1535.7	1163.8	**656.2**
2	1332.5	729.6	**661.6**
3	2063.4	**280.4**	443.0

the held out building; in experiment 2, building 2 is the held out building; and in experiment 3, building 3 is the held out building.

We ran three algorithms:

- An intelligent linear regression model, where the independent variables are products of relevant physical quantities, such as U x A x TETD.
- Our model, but with nearly all points held out for *all buildings*. In this case, the specification of the prior model dominates the prediction. This lets us see the beliefs we have encoded in the prior before the data plays much role in adapting the predictions.
- Our full model.

Figure 3 (a) – (c) show the linear regression reconstruction of the held out points for the completely held out building. (a) is the predicted energy consumption by building 1 from experiment 1, (b) is the predicted energy consumption by building 2 from experiment 2, etc.

The linear regression is well able to capture the variation in buildings for which it is provided data (not shown). However, since it has no prior beliefs about the plausibility of any model, it is equally happy with any model that has the same squared error over the observed data points. In each case, we see the consequence of this agnosticism in the inability to generalize to the held out building.

Figure 3 (d) – (f) show our specification of the model's prior beliefs. There is a fair amount of variance in the predictions, but they are within the range of reasonable values.

Figure 3 (g) – (i) show the result of applying our Bayesian network model to the same data. It is able to tune itself on data from two buildings in order to improve the predictions on the third building.

Quantitatively, Table 2 and Table 3 show the (square root of the) average squared difference between the model's predictions and true held out values for total energy use and cooling energy, respectively. In all cases, the Bayesian network model is superior to the linear regression by a significant margin, and in all but one case the information from the two other buildings is helpful in improving estimates over the prior for the third.

DISCUSSION

It is well understood in the building simulation field that there are uncertainties in the simulation process that must be dealt with in order to produce accurate energy consumption estimates. As such, there are several schools of thought on how to best approach the problem.

Perhaps the most common method of reducing uncertainty is to make a more concerted effort to measure inputs. At the most basic level, industry standard methods such as the International Performance and Measurement Verification Protocol (IPMVP) call for the use of weather, operations, and energy consumption data collected on-site to update the inputs of pre-construction models.

These industry standard methods have the benefit of being direct. However, obtaining more accurate estimates of input values often requires special equipment and manual effort, making the calibration of a large number of buildings expensive and time consuming. Though errors are less likely when more care is put into measuring inputs, the same shortcomings of the original model are still potentially present, since not all inputs can be measured accurately and precisely.

Yang et al. (2005) calibrate a traditional thermodynamic model (DOE-2) using artificial neural networks, and the Commercial Building Incentive Program online tool (Hepting et al. 2000) uses a regression model to estimate energy savings (but the supervisory data in this case comes from a traditional simulator). These works are similar to ours in that they automatically use data to refine output estimates. The drawback of these methods is that the parameters in neural networks, for example, are notoriously difficult to interpret, so it is difficult to express constraints on the model ensuring physical plausibility. If the model chooses parameters that capture some specious anomaly in the data, then—as we see with our linear regression model—the ability of the model to generalize to new settings will be compromised.

The work that is closest to ours is that of Reddy et al. (2007), which uses a Monte Carlo simulation to calibrate a DOE-2 model. Like us, the authors report results as a distribution over energy consumption values. This work represents uncertainty not only in the operational estimates, but also—more generally—in the model parameters themselves, using several plausible settings.

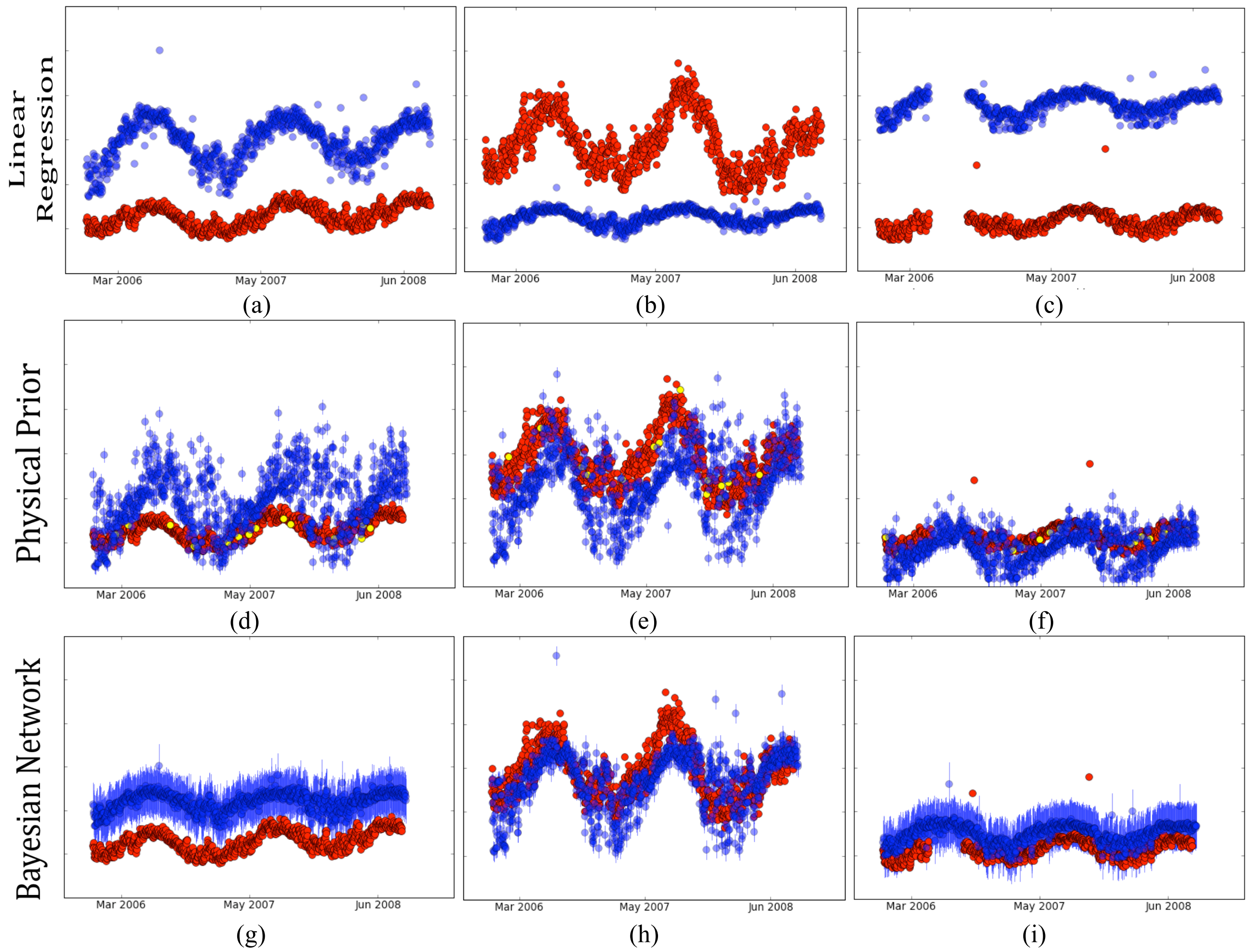

Figure 3 Results for across building generalization total energy use experiments. Yellow points are observed energy use data, red are held out points, and blue are model predictions (all in kWh vs date, best viewed in color).

In its implementation, there are several shortcomings to Reddy's approach (2007). First, each variable is restricted to take on only one of a small number of discrete values, which greatly limits the type of interactions between variables that can be represented. We do not discretize any continuous variables, instead choosing to represent relationships by the true physical equations that govern them. Second–and perhaps most seriously–the DOE-2 simulator is treated as a black box within the Monte Carlo routine. The outer calibration procedure has no knowledge about *why* the inner simulation procedure is producing its results. As a result, the Monte Carlo method employed is quite inefficient, essentially employing a strategy of trying all combinations of parameter settings.

We are instead focused on developing an approach that is *massively* scalable, able to simultaneously calibrate and share information between thousands or tens of thousands of buildings at once, using a rich physical model and data from several years for each building. To accomplish this, we suggest that the knowledge in the thermodynamic physics models such as DOE-2 and EnergyPlus needs to be encoded in a way that the simulator procedure and the calibration procedure can automatically share information and operate as one joint procedure. In our representation, this knowledge is encoded in the Bayesian network structure and parameters, and it is shared between our simulation and calibration procedure when we differentiate the free energy with respect to model and variational parameters.

Undoubtedly, building information models (BIM) will play a significant role in the development of a data-driven approach to energy-use simulation. BIM-based generation of drawings provides more complete data sets about building construction, tolerances, materials, and physical attributes before, during and after construction (Ibrahim 2002). As BIM technology continues to permeate the Architecture, Engineering, and Construction (AEC) industry, we anticipate more robust capabilities to predict energy usage by utilizing data from BIM design and operation of BIM compliant buildings.

Finally, we have chosen to work with Disney buildings because they give us flexibility to collect exactly the data that we need, but there are several other sources of data that would be interesting to pursue further. Examples include the Commercial Building Energy Consumption Survey (CBECS) data, which collects quadrennial data on U.S. commercial building types, building energy consumption, and energy-related building characteristics (EIA 2003), and Singapore's e-Energy (NUS)—an online spreadsheet tool containing a dataset of buildings in Singapore.

CONCLUSIONS & FUTURE WORK

We have presented a probabilistic, data-driven approach to estimating energy usage in buildings. The primary goal of our work is to accurately estimate energy use and express our confidence in those results for large organizations possessing large stocks of moderately similar buildings. We do so while avoiding the need for costly individual building calibrations. We believe there is significant potential to achieve these goals through further development of the data-driven concept presented in this paper.

An obvious drawback of our system is that it is unlikely to generalize well when buildings, systems, and environments are significantly different from the collected data set. Our model representing building physics is significantly less sophisticated than those in traditional simulators and therefore breaks down quickly where data are lacking. Undoubtedly, our model could benefit from (a) input from experts who are more knowledgeable about the inner-workings of buildings and the relationships among variables; and (b) from more data from a wide variety of buildings.

First, accurate daily predictions allow our operations staff to more effectively plan operations and control systems such that the building can dynamically respond to real-time use patterns.

On a broader level, we argue that when trying to estimate energy consumption over a large building stock, we should shift the focus to extracting the information contained in the data that is used for calibration in such a way that it can be used to improve real-world energy consumption estimates in many similar buildings. In order to do this on a large scale, we argue that we need to expose the inner workings of the building physics simulation to the calibration procedure. We do this by formalizing our beliefs about buildings physics in a Bayesian network representation. This and similar representations could further lead to exciting new approaches to problems such as automated building design, automatic model selection, and estimating the relative value of different data collection projects.

Finally, large-scale data collection and organization is not at all trivial, even within a single corporation. On top of the significant logistical concerns, there are sensor failures that lead to missing data and outliers, which requires ongoing monitoring and maintenance. A goal of this work, though, is to argue that the information contained within the data is valuable enough to warrant devoting substantial resources to its collection and management.

REFERENCES

American Society of Heating, Refrigerating, and Air-Conditioning Engineers. 1989. ASHRAE Handbook Fundamentals. Atlanta.

Tao, W.K.Y. and Janis, R.R. 2001. Mechanical and Electrical Systems in Buildings. New Jersey: Prentice Hall.

Energy Information Administration. 2003. Commercial Building Energy Consumption Survey (CBECS). Official Energy Statistics of the United States Government.

Frey, B. J., and Jojic, N. 2005. A Comparison of Algorithms for Inference and Learning in Probabilistic Graphical Models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27(9). 1392–1416.

Hepting, C., Ehre, D., and Mottillo, M. 2000. Web-Based Building Performance Assessment: The CBIP Screening Tool. Building Energy Simulation User News 21 (2000) 2-5.

Ibrahim, M. Krawczyk, R. & Schipporeit, G.: 2002, CAD Smart Objects: Potentials and Limitations, eCAADe 21.

International Performance Measurement and Verification Protocol Volume III, Chapter 4: Measurement and Verification Methods. 2003.

Maile, T., Fischer, M., Bazjanac, V. 2007. Building energy performance simulation tools – a life-cycle and interoperable perspective. Center for Integrated Facility Engineering (CIFE) Working Paper #107. Retrieved August 1, 2008. http://cife.stanford.edu/online.publications/WP107.pdf

National University of Singapore Center for Total Building Performance. e-Energy. http://www.bdg.nus.edu.sg/buildingEnergy/e_energy/index.html

Neal, R. M. and Hinton, G. E. 1998. A new view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan, editor, Learning in Graphical Models. Kluwer, Dordrecht, Netherlands.

Olofsson, T., Meier, A., and Lamberts, R. 2004. Rating the Energy Performance of Buildings. The International Journal of Low Energy and Sustainable Buildings 3 (2004).

Reddy, T., Maor, I. and Panjapornpon, C. 2007. Calibrating Detailed Building Energy Simulation Programs with Measured Data- Part 1. HVAC & R Research 13 (2007) 221-241.

Yang, Y., Rivard, H., and Zmeureanu, R. 2005. Building Energy Predictions with Adaptive Artificial Neural Networks. Proceedings of the IBPSA Building Simulation 2005 Conference.