Applied Energy 202 (2017) 685–699

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Constructing large scale surrogate models from big data and artificial intelligence

Richard E. Edwards [a], Joshua New [b,*], Lynne E. Parker [a], Borui Cui [b], Jin Dong [b]

[a] University of Tennessee, Knoxville, TN 37996, USA
[b] Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

HIGHLIGHTS

- World's fastest supercomputer used for construction of benchmark datasets.
- Lasso regression and feed forward neural networks are scaled.
- Machine learning instances are evaluated for prediction vs. runtime performance.
- 3 datasets with 7–156 software inputs are used to predict over 3,000,000 outputs.
- Surrogate building energy model of EnergyPlus runs 60× faster.

ARTICLE INFO

Article history:
Received 30 January 2017
Received in revised form 25 April 2017
Accepted 22 May 2017

Keywords:
Machine learning
EnergyPlus
Building simulation
Energy modeling
Surrogate model

ABSTRACT

EnergyPlus is the U.S. Department of Energy's flagship whole-building energy simulation engine and provides extensive simulation capabilities. However, the computational cost of these capabilities has resulted in annual building simulations that typically requires 2–3 min of wall-clock time to complete. While EnergyPlus's overall speed is improving (EnergyPlus 7.0 is 25–40% faster than EnergyPlus 6.0), the overall computational burden still remains and is the top user complaint. In other engineering domains, researchers substitute surrogate or approximate models for the computationally expensive simulations to improve simulation and reduce calibration time. Previous work has successfully demonstrated small-scale EnergyPlus surrogate models that use 10–16 input variables to estimate a single output variable. This work leverages feed forward neural networks and Lasso regression to construct robust large-scale EnergyPlus surrogate models based on 3 benchmark datasets that have 7–156 inputs. These models were able to predict 15-min values for most of the 80–90 simulation outputs deemed most important by domain experts within 5% (whole building energy within 0.07%) and calculate those results within 3 s, greatly reducing the required simulation runtime for relatively close results. The techniques shown here allow any software to be approximated by machine learning in a way that allows one to quantify the trade-off of accuracy for execution time.

* Corresponding author.
E-mail address: newjr@ornl.gov (J. New).

http://dx.doi.org/10.1016/j.apenergy.2017.05.155
0306-2619/

1. Introduction

1.1. Background of research

A central challenge in building energy efficiency is to realistically model the energy-related physics of an individual building. This capability is necessary to reliably project how specific policy decisions or retrofit packages would help meet national energy targets or maximize return-on-investment. This challenge is complicated by the fact that individual buildings, unlike cars or airplanes, are manufactured in the field and vary greatly from what may be considered a prototypical building. Since most whole-building simulation engines, such as EnergyPlus, have thousands of very specific required inputs, most of these engines suffer greatly from the user expertise, time, and associated costs required to create an accurate virtual model of a real-world building. Moreover, this manual process of tuning a model to measured data is neither repeatable nor transferable.

EnergyPlus is currently DOE's flagship whole-building energy simulation engine developed with active involvement by many participating individuals and organizations since 1995, and is posted open-source on GitHub [2]. EnergyPlus consists of 1.2 million lines of code with the core consisting of 748,731 lines of C++ code. It uses a more extensible, modular architecture than DOE-2, the previous and still widely used simulation program, to perform the energy analysis and thermal load analysis for a building. The computational costs of these capabilities has resulted in annual building simulations that, depending on the complexity of the building information, often requires 5+ minutes ($10\times$–$100\times$ slower than DOE-2 [3]) of wall-clock time to complete. Simulation runtime of this program is practically important as it is used internationally to help create new buildings that are energy efficient, define optimal retrofit of existing buildings, helps define building codes, and is increasingly used by utilities in energy efficiency and demand side management programs.

Reducing the runtime of EnergyPlus is the top priority of the development team with EnergyPlus 7.0 being 25–40% faster than EnergyPlus 6.0 [4]. But even with a 40% reduction in runtime, manually tuning EnergyPlus building models to align with utility data so that one creates a legally-useful software model of a building is still a slow and tedious process. For example, an engineer manually tuning a simulation is not likely to wait the 3–7 min required to run an EnergyPlus simulation before proceeding to the next tuning step; likewise, the Autotune methodology [5] runs 1024 simulations, which at only 3 min per simulation would require over 2 days. One solution is to construct surrogates to reduce the overall computational burden. Surrogates, which are generally statistically generated models, are built to provide rapid approximations of the original model, and require less computational resources [6].

In addition to the significant computational load issue, another main concern is the accuracy of the simulation engines for realistically modeling a virtual building such that matches a real-world building. A 2008 study [7] found 190 Home Energy Saver, REM/Rate, and SIMPLE residential simulation models had 25.1–96.6% error compared to actual monthly electrical energy usage. Another 2012 study [8] found that 859 residential models across Home Energy Saver, REM/Rate, and SIMPLE simulation engines had a mean absolute percent difference of 24% from actual monthly electrical energy usage and 24–37% from actual natural gas use for a sample of 500 houses. It should be noted that all of these studies use comparisons to monthly utility bill data; the challenge of accurately matching hourly or 15-min data for dozens of submetered data channels is significantly more difficult.

The challenge for simulation accuracy can be reduced to two primary issues: (1) a gap between the as-modeled and as-built structure, and (2) limitations of the modeling engine's capabilities. Gaps between as-modeled and as-built structures have many sources, but ultimately the fault lies in inaccurate input files rather than simulation engine itself. For example, infiltration, the rate at which air and the energy in it flows through the building envelope is not currently able to be cheaply tested despite its importance for energy efficiency. Blower-door tests can determine infiltration rate at a given pressure (usu. 50 Pascals) but is a 1-time measurement that, in reality, experiences significant variances as a function of temperature, wind speed, and wind direction. A second issue is the schedule for building usage, which includes number of occupants, times of occupancy, heating, ventilation and air-conditioning (HVAC) set-points, operations schedule, and other factors. For many of these, cost-effective sensors simply do not exist or are not typically deployed in a building. In many cases, occupancy schedules and relatively static set-point temperatures are estimated and then used later to "tune-up" a simulation to match whole-building data without regard to the accuracy of the actual HVAC thermostat set-points.

1.2. Literature review

Statistical energy models have been widely used for energy prediction [9,10], and energy optimization [11,12]. Building energy models calibration is critical in bringing simulated energy use closer to the actual consumption [13]. Researchers have shown an increasing interest in using various statistical tools for building energy models calibration [14–22]. Though many statistical energy models have been proposed for building energy analysis, they can be divided into two categories: data-driven models when detailed engineering energy models are available, and surrogate model-driven when only computationally cheap models are provided. There have also been attempts to combine data from both field measurement and computer simulations for calibration of building energy simulation models [16]. In contrast to simple linear regression, Gaussian process (GP) models [15] are used to capture the features of complex nonlinear and multivariable interactions of building energy behavior. Correlation analysis and hierarchical clustering has been utilized [19] to determine and choose informative energy data. The Bayesian technique becomes popular in this area since it is capable of parameter estimation even when there are missing energy data which are considered as uninformative output data. Bayesian technique based model can be used for multiple purposes, e.g. retrofit analysis, model-based optimal controls and energy diagnostics [23]. Provided a case without complete or a sufficiently large dataset, bootstrap is a powerful statistical tool to assess the accuracy of an estimator by random sampling with replacement from an original dataset [18].

Uncertainties and sensitivity analysis in building energy simulation has been investigated [24–29]. Uncertainty analysis (UA) takes into account uncertainties due to inherent simplifications of any model and lack of information with regard to input data. Understanding how uncertainties in energy use predictions from simulation software is important to achieve more effective energy efficiency upgrade packages and operational strategies for buildings [30]. On the other hand, sensitivity analysis (SA) consists of modifying model inputs in order to explore the relationship between input parameter variations and overall energy performance of the building [31]. The sensitivity analysis can also identify the most influential parameters to determine which should be tuned at high priority [32]. Both UA and SA should be integrated within calibration methodologies since they play an important role in building model accuracy [33]. To overcome the difficulties of getting information from SA using detailed models, macroparame-