Building and Environment 92 (2015) 61–74

Contents lists available at ScienceDirect

Building and Environment

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

Importance analysis and meta-model construction with correlated variables in evaluation of thermal performance of campus buildings

Wei Tian [a, b, *], Ruchi Choudhary [c], Godfried Augenbroe [d], Sang Hoon Lee [e]

[a] College of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin 300222, China
[b] Tianjin Key Laboratory of Integrated Design and On-line Monitoring for Light Industry & Food Machinery and Equipment, Tianjin 300222, China
[c] Engineering Department, University of Cambridge, Cambridge CB2 1PZ, United Kingdom
[d] College of Architecture, Georgia Institute of Technology, Atlanta, USA
[e] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 3 January 2015
Received in revised form
17 April 2015
Accepted 18 April 2015
Available online 30 April 2015

Keywords:
Building performance simulation
Meta-model construction
Variable importance
Building stock
Correlated variables

ABSTRACT

Statistical energy modelling & analysis of building stock is becoming mainstream in the context of city or district scale analysis of energy saving measures. A common aspect in such analyses is that there is generally a set of key explanatory variables – or the main inputs – that are statistically related to a quantity of interest (end-use energy or CO_2). In the context of energy use in buildings, it is not uncommon that the explanatory variables may be correlated. However, there has been little discussion about the correlated variables in building stock research. This paper uses a set of campus buildings as a demonstrative case study to investigate the application of variable importance and meta-model construction in the case of correlated inputs when quantifying energy demand of a building stock. The variable importance analysis can identify key factors that explain energy consumption of a building stock. To this end, it is necessary to apply methods suitable for correlated inputs because the observational data (inputs) of buildings are usually correlated. For constructing statistical energy meta-models, two types of regression models are used: linear and non-parametric models. The results indicate that the linear models perform well compared to the complicated non-parametric models in this case. In addition, a simple transformation of the response, commonly used in linear regression, can improve predictive performance of both the linear and non-parametric models.

1. Introduction

The building sector, including both residential and commercial buildings, accounts for around one-third of both global final energy use and CO_2 emissions [1]. Under business-as-usual scenario, the energy consumed by the built environment is expected to increase by 55% by 2050 [1]. This increase is mainly attributed to growth in the number of households, the built floor area, and power demand from new appliances and power-hungry services. In China, the building sector accounts for around 25% of total energy use [2] and the energy use in cooling-dominated areas is expected to rise due to climate change [3]. Moreover, around 60% of current building stock in some areas, such as the European Union, Russia, and the United States, will still exist in 2050 [4]. Any significant reductions in energy consumption therefore necessitate large-scale energy efficiency measures. Indeed, in recent years, there has been an increasing amount of studies on energy analysis of building stock as a whole [5–11]. Compared to research on energy performance of individual buildings, there are fewer studies on how to model and analyse the collective energy performance of a building stock – the area of work being relatively new [12].

Abbreviations: BTGP, Bayesian treed Gaussian process; CAR, correlation-adjusted correlation; DECM, domestic energy and carbon model; EPSCT, energy performance standard calculation toolkit; Fline, full linear model; GP, Gaussian process; GT, Georgia Institute of Technology; Lasso, least absolute shrinkage and selection operator; MARS, multivariate adaptive regression splines; NEP, network energy performance; PCR, principal component regression; PLS, partial least square; RMSE, root mean square error; RSS, sum of squared residuals; Srline, linear model with square root response; SVM, support vector machine; UPENN, University of Pennsylvania.

* Corresponding author. College of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin 300222, China. Tel./fax: +86 (022) 60273495.
E-mail address: tjtianjin@gmail.com (W. Tian).

http://dx.doi.org/10.1016/j.buildenv.2015.04.021
0360-1323/

Research on energy performance of building stock can be classified across three spatial scales: country, city, and district. The energy performance of the building stock at national scale has been the subject of analysis in different countries, such as Ireland [6], Greece [7], Portugal [5], UK [13,14], USA [15,16]. For example, Deru et al. [16] developed 16 reference commercial building type models to represent approximately 70% of the commercial buildings in USA by using EnergyPlus simulation program. Hughes et al. [14] implemented global sensitivity analysis and identified the wall insulation and demand (setpoint) temperature as the most significant uncertain parameters of an England housing energy model. For city-scale models, researchers have used various methods to study characteristic patterns of energy use in different cities, such as London [17,18], Stockholm [19], New York [20]. Tian et al. [17] applied spatial analysis methods to identify spatial patterns of domestic energy use in London. Howard et al. [20] applied a robust multiple regression method to estimate energy use intensity by end use in terms of building function in New York City. Tian and Choudhary [21] implemented a MARS (Multivariate adaptive regression splines) meta-model obtained by using EnergyPlus engineering models to estimate the energy consumption of secondary schools in London. This is an example of a sector specific analysis at city-scale. Examples of district-scale models are the analysis of commercial areas in city centre [22,23], and analysis of campus buildings [24,25]. Note that there may be studies that cover more than one spatial scale, or fall in-between two scales of analysis. For example, Cheng and Steemers [26] developed the Domestic Energy and Carbon Model (DECM) to predict domestic energy consumption at both national and district levels. They found that the dwelling types and socio-economic classes can account for over 85% of the variations of domestic energy use. Lee [24] developed the NEP (network energy performance) model to retrofit existing buildings and shared systems for large scale energy systems at a campus-scale.

A common aspect in the aforementioned studies is that there is generally a set of key explanatory variables — or the main inputs — that influence the quantity of interest (end-use energy or CO_2). Examples of key explanatory variables are: wall insulation, lighting demand, exposed wall area, operation schedules, etc. The resulting change in the quantity of interest due to a relative change in the value of one or more explanatory variables is generally the method by which a user can quantify energy saving measures. However, it is not uncommon that explanatory variables may be correlated. For instance, if both the occupancy and the built-up floor area are used as explanatory variables to quantify energy consumption of a district, it is likely that the two variables are correlated. Common regression-based methods (such as standardized regression coefficients) cannot be used to quantify the individual influence correlated variables on a given quantity of interest [27]. Indeed, correlated inputs in regression analysis lead to large variances of some estimated regression coefficients, and may lead to unstable regression equations [28]. Despite the fact that many building stock energy models are based on deriving some form of statistical relationship between the explanatory variables and a quantity of interest, little attention has been paid to the occurrence of correlated variables in such models [23,27,29].

Therefore, the purpose of this paper is to provide clear guidance on how to implement variable importance analysis and create statistical energy models in the presence of correlated variables for a building stock. Campus buildings are chosen as a demonstrative case study in this research because building stock information for campus buildings is more readily available than other types of buildings. For variable importance analysis, we concentrate on how to choose the appropriate methods in the case of correlated inputs and obtain the so-called probabilistic sensitivity index (not only point indicators) in order to correctly and thoroughly understand the influence of key variables on the energy consumption of a given building stock. For statistical energy models, we compare the characteristics of different regression models in the context of analysing large stocks of buildings. These type of statistical energy models are also called meta-models or surrogate models [27].

The paper is structured as follows. First, the input data and the building energy model of the campus buildings are described. Second, the statistical approaches used in this paper are introduced, including variable importance and meta-models (linear and non-parametric models). The results from variable importance are discussed to infer the main factors that influence the energy consumption of the campus buildings. The next section compares the surrogate models derived from eight different statistical regression approaches. Lastly, practical guidance is presented for applying variable importance and meta-model construction in the case of correlated factors for a building stock.

2. Data and building energy models

The data used in this analysis consists of 114 buildings in the University of Pennsylvania (called UPENN buildings) and 30 buildings in the Georgia Institute of Technology (called GT buildings) as collated and shown in Table 1 and Fig. 1. The energy simulation is carried out by using EPSCT (Energy Performance Standard Calculation Toolkit) [30]. More detailed information on building stock data and EPSCT will be described in this section.

2.1. Campus building data

2.1.1. Inputs variables from UPENN and GT buildings

The statistical summary of input variables collected from 114 UPENN buildings and 30 GT buildings have been listed in Table 1. The choice of these inputs is mainly based on heuristics: the potential factors significantly influencing energy use based on previous theoretical analysis or case studies [14,21,26,27].

The inputs shown in Table 1 can be described by three categories: building geometry, thermal properties of building envelope, and internal heat gains. The first one is related to the physical form of the building, containing 13 factors. Their short names are defined by starting with "G" (Table 1). These inputs mainly include total building area, building height, opaque (i.e. wall, but excluding window area) and window area (also called glazing in Table 1) in different orientations. The second one is related to thermal properties of building envelope, which contains 4 factors. Their short names begin with "S". These factors include the U-values of opaque and transparent elements, solar transmittance of windows. The third one is related to internal heat gains and occupants, which contains 11 factors. Their short names start with "I". These variables include the peak internal heat gains from equipment and lighting, metabolic rate, and occupancy density. Hourly schedules for occupant, equipment and lighting are summarized by six indicators: total occupancy hours; total equipment hours; total lighting hours; average weekday occupant; average weekday equipment; average weekday lighting. The first three factors denote total occupied hours in a typical week, while the last three factors represent the total daytime hours in a typical working day. The total hours for occupant, equipment and lighting are defined as adding five times of sum of weekday hourly values and two times of sum of weekend hourly values together in order to be consistent with five weekdays and two weekends in a week. These total hourly schedules represent the total hourly values from one typical week schedules. Average weekday daytime schedules for occupant, equipment and lighting are defined as