Energy and Buildings 119 (2016) 363–376

Contents lists available at ScienceDirect

Energy and Buildings

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

Identifying informative energy data in Bayesian calibration of building energy models

Wei Tian [a,b,*], Song Yang [a], Zhanyong Li [a,b], Shen Wei [c], Wei Pan [d], Yunliang Liu [a]

[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] Building Performance Analysis Group, Plymouth University, Plymouth, UK
[d] Department of Civil Engineering, The University of Hong Kong, Pokfulam, Hong Kong, China

ARTICLE INFO

Article history:
Received 31 January 2016
Received in revised form 16 March 2016
Accepted 17 March 2016
Available online 18 March 2016

Keywords:
Bayesian computation
Cluster analysis
Model calibration
Building energy

ABSTRACT

Bayesian computation has received increasing attention in calibrating building energy models due to its flexibility and accuracy. However, there has been little research on how to determine informative energy data in Bayesian calibration in building energy models. Therefore, this study aims to determine and choose informative energy data using correlation analysis and hierarchical clustering method. A case study of retail building is used to demonstrate the proposed methods to infer four unknown input parameters using EnergyPlus program. The results indicate that the different combinations of energy data can provide various levels of accuracy in estimating unknown input variables in model calibration. This suggests that Bayesian computation is suitable for inferring the parameters when there are missing energy data that can be treated as uninformative output data. The proposed method can be also used to find the redundant information on energy data in order to improve computational efficiency in Bayesian calibration.

1. Introduction

The building sector accounts for about 35% of global final energy consumption in 2010 [1]. Therefore, buildings play a significant role in substantially reducing energy consumption and carbon emissions. Additionally, more than half of existing buildings will be still in use in 2050 [1]. Hence, more attentions should be paid to retrofitting existing building stocks with respect to their energy performance [2].

The calibrated simulation approach is an effective method to analyse the potential effects of various energy efficiency measures for existing buildings [3–6]. When using this method, the inputs of a building energy model are often tuned in order to achieve a close match between the measured and the modeled energy data, and this tuning procedure is often named as calibration [7]. This calibrated model can then be used to estimate the energy saving potentials of various retrofit measures. The calibration approaches of energy models can be either deterministic or probabilistic. The unknown parameters using the deterministic method are treated as fixed point values [3], whereas the probabilistic method would lead to probability density functions for the calibrated parameters by considering uncertainty of these unknown variables [8]. In the field of building energy analysis, most of existing studies are based on the deterministic approach [9–11], which applies a manual trial-and-error method to obtain unknown parameters. As a result, this deterministic method is time-consuming and the results are much dependent on the analyst. In building energy models, a large number of combinations of input values may result in the same predicted energy use [3]. Moreover, this deterministic method can only test a limited number of possible solutions, so they may not represent the actual operation conditions with an inherently stochastic nature in a building. Hence, the estimations of energy savings are unreliable based on these limited samples. In contrast, the probabilistic approach can deal with this problem in an efficient and effective way by treating the unknown input parameters as random variables with probability density functions (PDFs). These PDFs can adequately quantify input uncertainties in actual buildings. This probabilistic calibration method (also called Bayesian analysis) has a great potential of helping to better analyse building energy consumption for energy retrofit projects [12]. It should be emphasized that the estimated energy data from energy conservation measures using Bayesian calibration would also have a probabilistic nature with probabilistic density functions. This is different from

* Corresponding author at: College of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin 300222, China.
E-mail addresses: tjtianjin@gmail.com, tjtianjin@126.com (W. Tian).

http://dx.doi.org/10.1016/j.enbuild.2016.03.042
0378-7788/

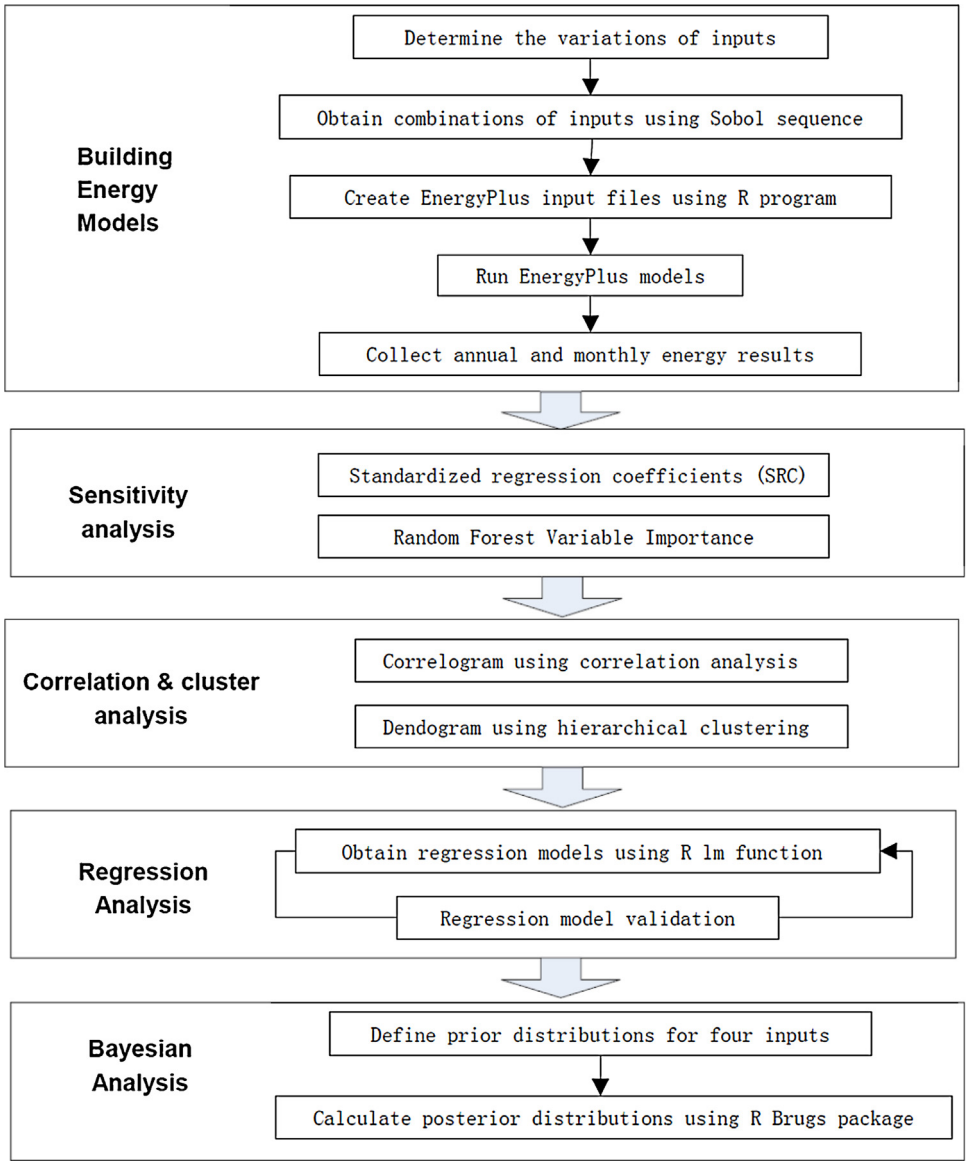

Fig. 1. The flow chart for input parameter estimation used in this research.

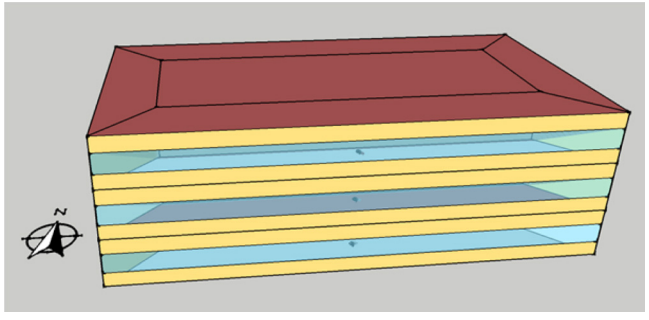

Fig. 2. The 3-D EnergyPlus models for a retail building.

the deterministic approach that only has fixed energy data predicted for various energy conservation measures.

Researchers in the field of building energy analysis have shown an increasing interest in using the Bayesian analysis to calibrate building energy models [13–17]. Heo et al. [8] implemented a Bayesian calibration method for an UK campus building. Their research is based on Kennedy and O'Hagan's framework [18] to create a Gaussian processes model firstly and then use Bayesian calibration for this statistical model because the Gaussian process formulation makes the computation of likelihoods feasible. Tian and Choudhary [14] proposed a probabilistic, engineering-based building stock model to estimate the unknown parameters in a specific building stock using Bayesian inverse methods. The main advantage of their method accounts for the variability of input parameters influencing energy performance among buildings. Riddle and Muehleisen [19] provided a good guidance on how to implement Bayesian calibration in building energy analysis.

However, there is still an unsolved issue that is whether all the energy data are equally important when inferring unknown parameters of a building energy model using Bayesian method. The intuitive thinking is that combing both monthly heating and electricity data can provide more reliable calibration results, in comparison with using only monthly electricity data. This is because the combination of both heating and electricity data can reflect the different patterns of energy use in buildings, which contains more information compared to only using electricity data.

Table 1
R packages used in the study.

Step	Package	Descriptions	Reference
1: energy models	randtoolbox	Sobol sequence generation	[31]
2: sensitivity analysis	sensitivity	Standardized regression coefficient	[38]
	randomForest	Random forest relative importance	[39]
3: correlation & clustering	ggplot2	Correlation mapping	[41]
	stats	Hierarchical clustering	[24]
4: regression analysis	stats	Linear regression	[45]
5: Bayesian analysis	BRugs	R interface to OpenBUGS software	[47]

Table 2
Main characteristics of a retail building.

Component	Item	Parameters
Envelope	Floor area (m^2)	2400
	Floor levels	3
	Window-wall ratio	0.40
	Zone number	15
	Wall U-value (W/m^2K)	0.36
	Roof U-value (W/m^2K)	0.28
	Window U-value and SHGC (solar heat gain coefficient)	See Table 3
	Infiltration rate (air changes per hour)	0.4
Internal heat gains	Lighting power density (W/m2)	16
	Daylighting	Lights will be off when daylighting is above threshold 400 lux
	Equipment gain, occupant density	See Table 3
	Hourly schedules for occupants, lights, and equipment	Retail buildings in China standard [25]
HVAC	–	Fan coil system with hot water and chilled water coils, heating provided by district heating system, cooling from centrifugal chiller and water-cooled cooling tower, ventilation requirements: 20 m^3/(h person) for retail

The next natural question to ask is how to identify key informative energy data in Bayesian calibration of building energy models. The method of solving this issue can bring at least two benefits. The first one is whether Bayesian calibration approach can be used when dealing with missing energy data, which often occurs due to occasional equipment failure. If available energy data contains sufficient information and missing energy data contains only redundant information, Bayesian analysis can provide reliable results in the case of missing energy data. The second one is to expedite the calculation by choosing only informative outputs in Bayesian calibration. This is because Bayesian calibration involves a large number of simulation runs with MCMC (Markov Chain Monte Carlo) and the calculation time can be reduced significantly by using only informative outputs [20].

Therefore, this study aims to determine informative energy data when using Bayesian calibration of building energy models. The informative energy data here refers to the data that can be used to accurately infer the values for unknown parameters of an energy model using Bayesian approach. Accordingly, these values would not be improved significantly by adding more energy data (named as uninformative output data) in Bayesian calibration. A retail building is used as a case study to infer the unknown input values for building energy models. To assess the accuracy of estimated values for building parameters, the true values of calibrated parameters have been pre-determined using the forward simulation method [3] in which the specified input values are used to obtain energy data with an energy simulation program. This method is similar to what Heo et al. [21] used by comparing the calibrated values with the known true values.

This paper is structured as follows. Section 2 describes the methodology used in the study, based on the five steps as shown in Fig. 1. Section 3 discusses the results from these statistical analysis approaches, containing sensitivity analysis, correlation/clustering analysis, regression models and Bayesian computation. At the last part, conclusions were made for the study and potential further work has been pointed out.

2. Methodology

The analysis carried out in this study is based on the procedure shown in Fig. 1, including five steps that will be separately described in the following five sub-sections. The first step is to create the energy models of buildings using energy simulation programs and run these models to obtain a matrix of inputs and outputs. A large number of energy simulation models are required for this step to provide sufficient data for the following steps. The second step is to implement sensitivity analysis in order to rank the importance of variables influencing energy performance. The ranking results from this step can explain the degree of uncertainty for inferred unknown parameters from the last step. The third step is to apply correlation and cluster analysis for the energy data to explore the characteristics of these data. The purpose of this step is to find the energy data that have similar trends, which indicates informative or uninformative data in calibrating energy models. The fourth step is to create a statistical energy model using linear or nonlinear models. The aim of this step is to provide reliable statistical energy models for model calibration because most of dynamic building energy models are infeasible to obtain the likelihood that is required in Bayesian inference [22,23]. The last step is to estimate the unknown parameters from Bayesian computation and compare these inferred results using either informative or uninformative outputs that have been identified from the third step.

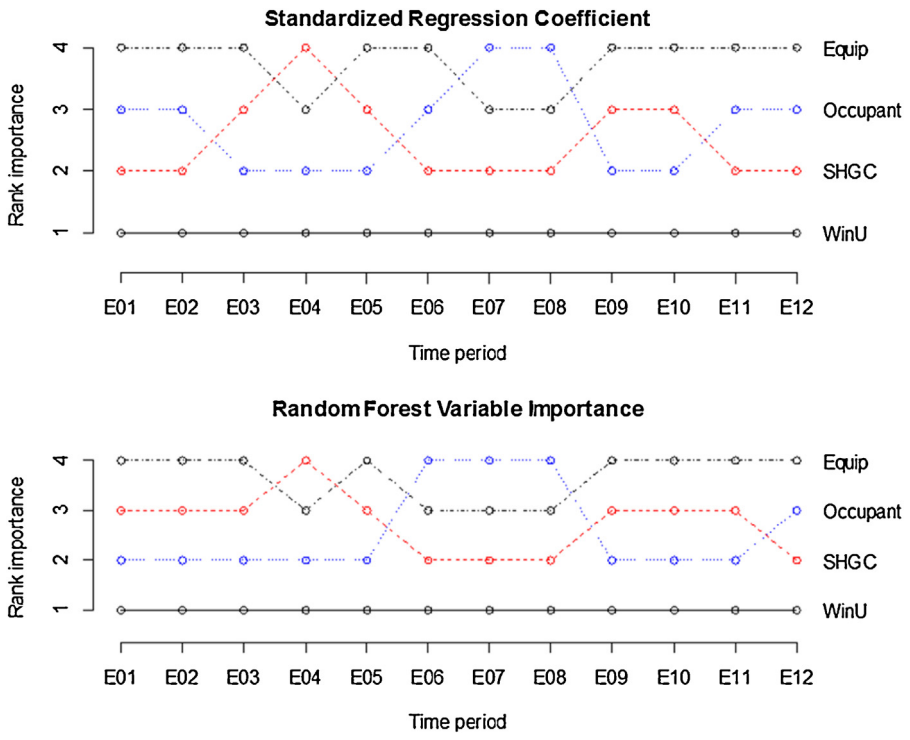

Fig. 3. Results from sensitivity analysis for monthly electricity in the retail building (E01 refers to the electricity use in January, and so forth).

Table 3
Variables used in regression analysis.

Variable	Short names	Unit	Range
Window U-value	WinU	W/m^2K	1.5–3.5
Window SHGC (solar heat gain coefficient)	SHGC	–	0.3–0.7
Occupant density	Occupant	m^2/person	3–5
Equipment peak heat gain	Equipment	W/m^2	11–15

All the statistical analysis is implemented using R program, a computational environment for statistical computing and graphic with advanced functions [24]. Table 1 lists all the R packages used in this research.

2.1. Building energy models

Fig. 2 depicts a 3-D energy model for the case study used in this analysis. Table 2 summarizes the main characteristics for this building. Table 3 lists the values of input parameters, and these values are changeable in the energy models. The ranges of these parameters (named as prior information as will be described in Section 2.5) could be obtained using field knowledge or site survey in order to define the possible values for these unknown parameters in a specific building. The retail building is a three-storey rectangular building with a floor area of 2400 m^2. The parameter values related to thermal performance are mainly derived from Chinese design standard for non-residential buildings (GB50189-2005) [25]. The building is assumed to be located in Tianjin, China and the corresponding weather data is downloaded from the EnergyPlus website [26].

The simulation is carried out using the EnergyPlus program, developed by the Department of Energy, USA [26]. EnergyPlus is a whole building energy simulation program to model both energy use and process loads, such as heating, cooling, ventilation, lighting, and water use in buildings. It can use sub-hourly time steps to explore the interactions between thermal zones and the environment or between thermal zones and HVAC systems in order to model systems with fast dynamics. A number of advanced functions (including latest fenestration models [27], illuminance and glare calculation [28], functional Mockup interface [29]) are also available. Moreover, this program has been extensively tested and validated using both analytical and comparative tests [26]. Hence, it has been widely used in the field of dynamic building energy simulation [5,30,31].

The advantage of using EnergyPlus, especially for this research, is that the input data file (the IDF file) for EnergyPlus models uses text format [32]. Thus, the model definitions can be easily edited for parametric analysis, since a large number of EnergyPlus models are required for all the next four steps (shown in Fig. 1) in this research. The EnergyPlus models are automatically created using the R statistical program, so the building energy simulation is directly linked to the statistical analysis used in this study [24].

The combinations of inputs (as defined in Table 3) are constructed using the Sobol sequence, one of quasi-random low-discrepancy sequences (LDS). LDS (also called quasi-random numbers) are designed to create uniform sample points [33]. The Sobol sequence can meet three requirements: (1) best uniformity as the number of sampling iterations approaches infinity; (2) good distributions for small initial sets; and (3) fast computation [33]. Therefore, this sampling method is adopted in this study to obtain the combinations of four parameters with the uniform distributions (Table 3). The sampling number is chosen as 40 based on the recommendation from Levy and Steinberg [34]. Additionally, the extra 40 simulation runs (i.e. testing data set) are performed as well to validate the regression energy models created from the first 40 runs (i.e. training data set). In total, the 80 EnergyPlus simulation models are used in this study. The R randtoolbox package is used for generating the Sobol sequence in this analysis [35].

Two energy performance indicators used in this analysis are annual heating and electricity use per floor area (unit: kWh/m^2 year) from this retail building. The R program is also used to automatically collect monthly energy data from these simulation results. Note that the heating data is only available for five months

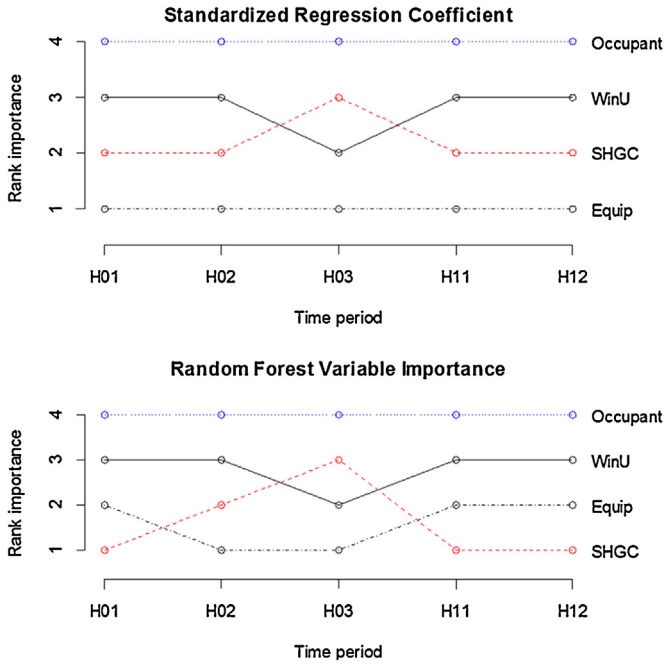

Fig. 4. Results from sensitivity analysis for monthly heating energy in the retail building (H01 denotes the heating energy in January and so forth).

(November, December, January, February, and March) due to the regulation of district heating used for this retail building from local government in Tianjin, China.

2.2. Sensitivity analysis

The sensitivity analysis is to identify important variables that affect the energy performance of buildings. In this study, the main purpose of this step is to obtain the ranking results of variable importance, which can explain the degree of variations of estimated parameters using Bayesian calibration. Furthermore, the results from this step can help better understand the complicated relationships between input factors interested in a project and energy performance of buildings. A number of sensitivity analysis approaches are available in building energy assessment [36–39]. This paper uses two fundamentally different approaches to provide robust analysis results: SRC (standardized regression coefficient) and random forest variable importance. A brief description of them is presented here. For detailed information, please refer to [40,41]. These two sensitivity analysis methods are carried out using R sensitivity [42] and random Forest [43] packages, respectively.

The SRC is the most popular method in the field of building energy analysis since it is available from almost every statistical program and is also easy to interpret [36]. Compared to conventional regression coefficients, the effect of units of input variables for the SRC is removed by normalizing their variances to one. Hence, the SRC can measure the relative importance of independent variables. The SRC ranges from −1 to +1. A minus SRC indicates that the input and output would tend to move in the opposite directions. The higher the absolute SRC value is, the more important the input variable is. Note that this method is not suitable for non-linear relationships between inputs and outputs.

The random forest method is a popular choice of machine learning algorithm for both classification and regression [44]. The random forest is similar to the boosting method that builds a large number of de-correlated trees and obtaining average values [41]. Another useful feature from this method is to assess the relative importance of variables using a decrease in predictive accuracy for various models. Firstly, a reference accuracy is created as a normal tree is grown in a model. Secondly, the values for the *i*th variable are randomly permuted in the analysis and a new predictive accuracy is obtained. Apparently, this new accuracy would be decreased, compared to the reference accuracy due to the permutation. Then, the decrease in accuracy for the *i*th variable is averaged over all the trees. If the same procedure is used for all the variables, the decreases in accuracy for all the variables can evaluate the importance of these variables. A larger decrease in accuracy means that variable is more important in the analysis.

2.3. Correlation and cluster analysis

The correlogram and clustering analysis used in this section aim to explore the data structure of monthly energy performance in buildings. The correlation analysis is focused on the bi-variable analysis, whereas the hierarchical clustering can explore the similarity of multiple variables. These two approaches are also applicable for small time scales, such as daily or hourly energy data.

The correlogram (also called corrgram) is a color-mapped matrix of correlation coefficients in multivariate analysis [45]. This method is very useful to show the correlation structure of energy data in buildings. The Pearson product-moment correlation coefficient is defined as the covariance of the two variables divided by the product of their stand deviations.

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \quad (1)$$

where r is the Pearson correlation coefficient between two variables (x and y), n is the number of observations, $\bar{x}$ is the mean of the first variable x, $\bar{y}$ is the mean of the second variable y, s_x is the standard deviation of the first variable x, and s_y is the standard deviation of the second variable y. This correlation coefficient is used in this paper to assess the relationship between two different monthly energy data.

The hierarchical clustering analysis is to build a hierarchy of clusters based on data similarities [41]. Compared to a more commonly used clustering method (K-means), the advantage of hierarchical clustering is that there is no need to choose the number of clusters before performing clustering analysis. The results from this method have hierarchical nature in which the clusters at each upper level of the hierarchy tree are composed of the clusters from the lower level. As a result, there is only one large cluster including all the energy data at the highest level, whereas each energy data is one smallest cluster at the lowest level. Therefore, the energy data in different middle groups can be regarded as having different characteristics of energy use, while the energy data from the same middle group can be treated as the similar energy trends. The choice of which hierarchical level used for determining informative or uninformative energy data should be dependent on specific context of a building project. For instance, analysts should choose lower level number if there are a small number of energy data. It also depends on the degree of dissimilarity for energy data from hierarchical clustering. If there are significant differences of two groups, then the energy data in these two groups can be regarded as informative energy data. The corresponding diagram from hierarchical clustering is called dendrogram to illustrate the clusters of energy data. The dendrogram provides an interpretable description figure of hierarchical clustering results, a feature of using this hierarchical clustering approach as will be described in Section 3.2. Note that this hypothesis of informative or uninformative energy data should be verified as will be described in Section 3.4.

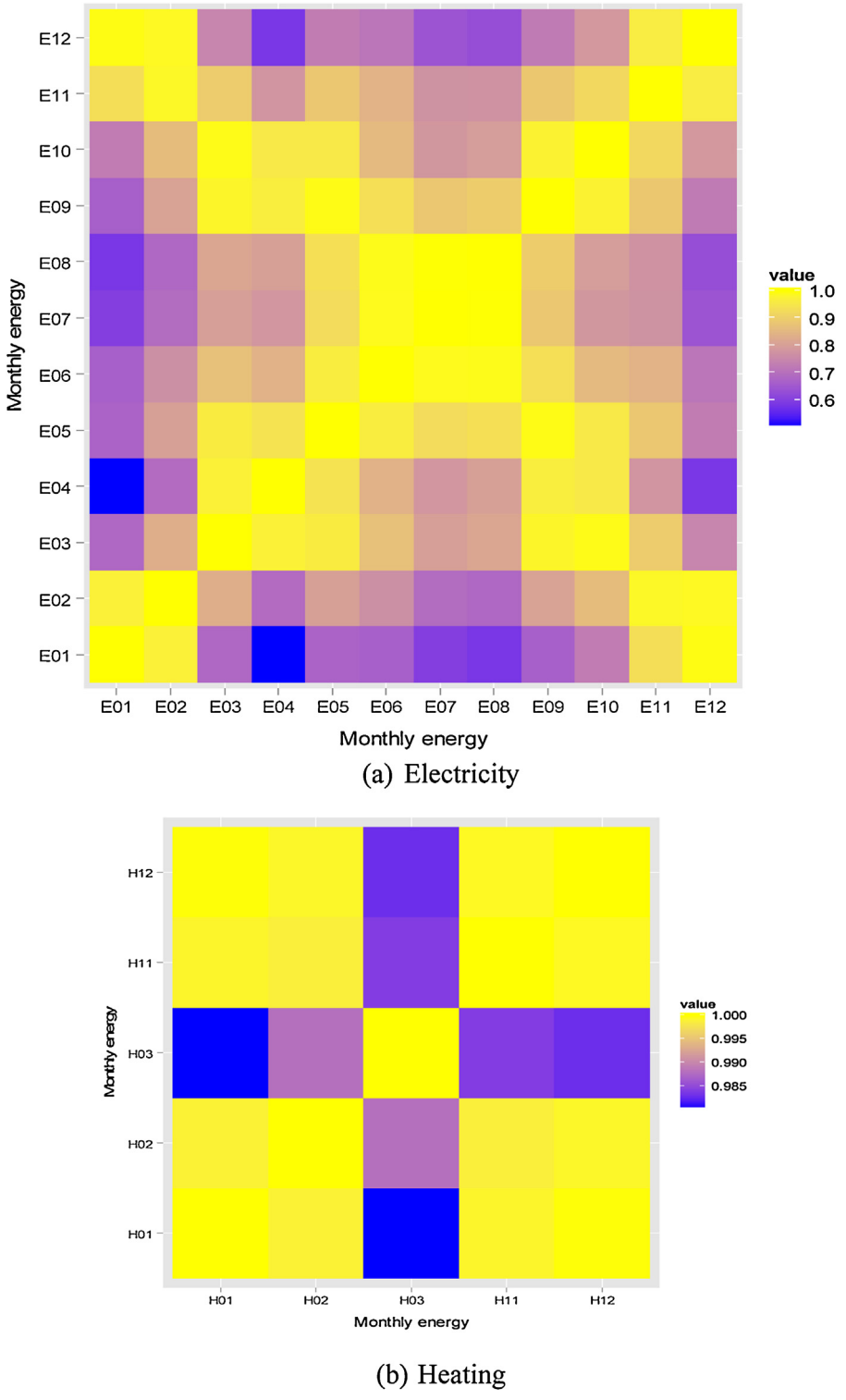

Fig. 5. Correlogram of monthly electricity and heating for the retail building (E01 refers to the electricity use in January, and so forth; H01 denotes the heating energy in January and so forth).

2.4. Regression models

After the annual and monthly energy simulation results are collected, various types of regression models can be developed, depending on the number of input variables and model accuracy. These regression models can be divided into parametric and non-parametric models [46]. The parametric models include linear and non-linear predetermined regression forms, whereas the non-parametric regression method is more flexible without a formulaic way of defining the relationship between inputs and outputs [46].

The analyst should choose full linear models for calibration problems in the first place. There are at least three reasons for this choice. First, simple linear models are more robust and have less overfitting compared to non-parametric models [47]. Second, the calibration problems for building energy models often contains a limited number of input parameters. Therefore, it is unlikely to exist

a strong non-linear relationship between these input factors and the energy use of a specific building [16,48]. If there are a large number of parameters, the analyst should firstly run sensitivity analysis to choose several important factors influencing building energy use for calibration study. Third, linear models have many variations, including the transformation of both inputs and outputs. For example, the interaction and second-order terms can be added to increase the predictive capability if the full linear model has high prediction errors. Many other transformations may be also used, such as log, inverse, square functions [37,49].

The accuracy of regression models is evaluated by using two metrics: R2 and RMSE (root mean square error). R2 is the coefficient of determination to indicate the proportion of the total variations of response (building energy in this case) explained by the regression models. The higher R2, the more accurate the regression model is. RMSE is the root mean square error to indicate the average errors of regression models. This value has the same unit as the building energy consumption in this study. A lower RMSE indicates a better model. It should be emphasized that the R2 or RMSE from the training data set (the first 40 simulation models in this case) cannot be used to accurately assess the predictive performance of regression models for the future data set. The training errors consistently decrease with an increase in model complexity, possibly due to overfitting. However, the testing error may increase at some point. Therefore, the extra 40 simulation runs regarded as a testing data set are carried out in order to test the predictive performance of the regression models developed from the first 40 building energy simulation runs.

2.5. Bayesian analysis

In this study, the four input parameters (listed in Table 3) are estimated using monthly energy data for heating and electricity. To assess the accuracy of calibration method, it is necessary to know the true values of input parameters. In actual buildings, there are usually very high uncertainties on the values of these input parameters. Hence, this research is designed to evaluate the reliability of the estimation method by using the specified inputs and the energy consumption simulated from EnergyPlus models. Therefore, it is assumed that the energy data are available and the input parameter will be inferred using Bayesian computation. As a result, these inferred inputs can be compared to the true values in order to accurately assess the validity of the estimation method [21]. Note that the 40 data sets used for creating regression models described in the last section cannot be used in this Bayesian analysis. This is because the accuracy of both the regression models and the Bayesian computation will be evaluated. In this research, we randomly choose three data points of the 40 extra data set. These three random data set are used for Bayesian computation to provide robust analysis.

In Bayesian statistical analysis, all the uncertainty should be regarded as probabilities [12]. In this case, the input parameters in building energy models can be modeled using probabilities, named as prior and posterior distributions. The prior distributions of building input parameters can be obtained from previous documents or expert knowledge. Then the new posterior distributions will be computed based on the Bayes Theorem to combine the data and prior distributions [50],

$$p(\theta|y) = p(y|\theta)\,p(\theta)/p(y) \quad (2)$$

where $p(\theta|y)$ is the posterior distribution containing the updated knowledge, $p(y|\theta)$ is the likelihood, $p(\theta)$ is the prior distribution, and $p(y)$ is the marginal likelihood. Therefore, the posterior probability is proportional to the prior probability times the likelihood. If there is a large data sample, then the influence of prior becomes smaller. In contrast, for a small data sample, the prior distribution tends to have a predominant effect on the posterior distribution. In the case of calibrating the building energy models, this means the calculated building inputs would more depend on the prior information (i.e. how much we know about the building investigated in the project) when the number of available energy data is small due to missing data or other issues.

Modern Bayesian analysis is usually performed to obtain the posterior distributions using the Markov chain Monte Carlo (MCMC) method. This is because the integrals cannot be evaluated using calculus in most of actual problems and the numerical approximations become necessary, especially MCMC simulation method. The MCMC method has become the main computational approach in Bayesian analysis because this method can draw samples from high-dimensional posterior densities [22].

Three parameters need to be determined to obtain results in this case. The three parameters are burn-in length, iteration number, and standard deviation for error distribution. The first parameter, burn-in length, is to describe initial non-stationary portion of the chain that needs to be deleted for final analysis. This number is taken as 20,000 in this study, which is sufficient to avoid the influence of initial values. The second parameter, iteration number, is to make sure the final chain has reached its equilibrium. A substantial sample size is required to this end and 100,000 is used in this study. Preliminary study indicates that the results would very similar after this larger number of simulation runs. A two-chain run is also implemented to detect the convergence of results. The third parameter, standard deviation of error distribution, is another key value in this case. After running several values for this parameter, 10 is taken as a final value to show the spread of residuals. A large standard deviation of error distribution would result in larger variations of estimation unknown inputs, whereas a small value for this parameter would lead to no sufficient data for final analysis, may even crash for this Bayesian model since it cannot find proper solutions in the case of small errors.

The R BRugs package is used for Bayesian analysis in this research [51]. This package is an interface of the OpenBUGS program for Bayesian analysis using MCMC sampling. OpenBUGS software is an open source Bayesian analysis environment. The advantage of using the BRugs package is to make full use of advanced statistical functions in the R environment. For example, the non-standard semi-parametric regression can be implemented using this package [52].

Table 4
Results for sensitivity analysis of annual electricity and heating use from SRC and random forest methods in the retail building.

Energy	Input	Method	
		SRC	Random forest
Electricity	Equipment heat gain	0.821	97.442
	Solar heat gain coefficient	0.459	51.258
	Occupant density	−0.403	47.775
	Window U-value	0.001	0.957
Heating	Occupant density	−0.961	122.022
	Window U-value	0.231	34.895
	Solar heat gain coefficient	−0.185	23.663
	Equipment heat gain	−0.091	10.435

3. Results and discussion

3.1. Results from sensitivity analysis

Table 4 lists the results from the sensitivity analysis for annual electricity and heating energy in the case study building. Figs. 3 and 4 show the sensitivity analysis results for monthly electricity and heating in the case study building. The analysis is based

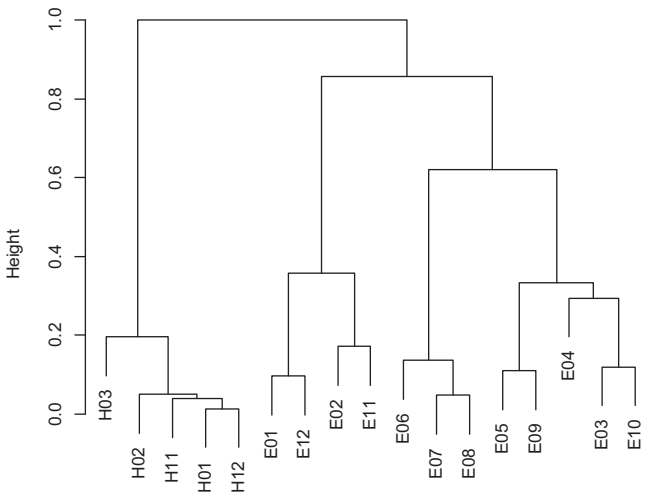

Fig. 6. Hierarchical cluster analysis for monthly heating and electricity energy in the retail building (E01 refers to the electricity use in January, and so forth; H01 denotes the heating energy in January and so forth).

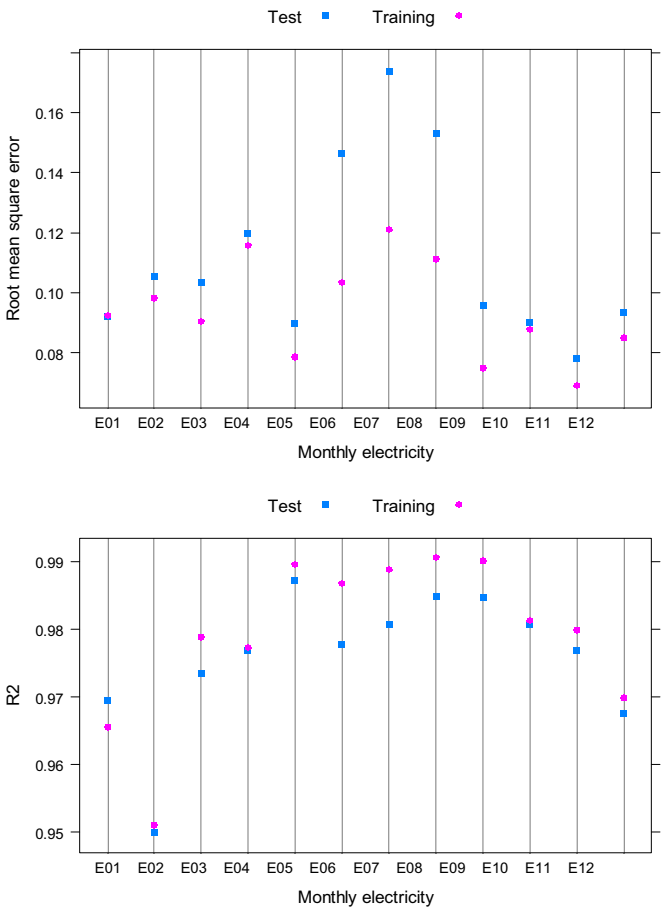

Fig. 7. Comparison of regression results for monthly electricity in terms of RMSE (root mean square error) and R2 (coefficient of determination) in the retail building (E01 refers to the electricity use in January, and so forth).

on two sensitivity analysis methods: standardized regression coefficient (SRC) and random forest, as described in Section 2.2.

The sensitivity results are consistent from two methods for both annual electricity and heating energy as shown in Table 4. As explained in Section 2.2, a higher absolute SRC value means a more important variable and a larger Random forest value indicates a more influential factor. The most important factor influencing annual electricity use is the equipment heat gain (unit: W/m^2), while the dominant variable affecting annual heating energy is the occupant density (unit: m^2/person) in this retail building.

Fig. 3 reveals that the important factors influencing heating energy would significantly change with different months as can be seen from both two sensitivity analysis approaches. In cold months (such as January and December), the heat gain from equipment is the most important factor, whereas the occupant density becomes more important in summer. This is because the main end-use energy for electricity would change from equipment to cooling energy use as weather warms up. The occupant density has an apparent influence on cooling energy consumption due to ventilation loads. The effect of the window SHGC (solar heat gain coefficient) on electricity is more significantly in the transition season between winter and summer because of the use of daylighting. As for the window U-value, this variable is the least important factor in all the months for electricity use.

Fig. 4 presents the ranking importance affecting monthly heating energy use in this building. It is apparent that occupant density is the most important factor in all the months due to the direct relationship between occupant density and ventilation heating loads. The next important factor is the U-value of windows except for March. There are some discrepancies for the ranking of equipment heat gain and window SHGC from the SRC and random forest method.

As discussed above, the sensitivity results from annual energy data would present an overall trend of relative importance for input variables. However, more detailed information can be only shown from monthly sensitivity analysis. For instance, the most influential factor in summer is very different from the winter as shown in Fig. 3, which would provide more insights on comprehending the characteristic of energy consumption in buildings. For the purpose of energy model calibration, this suggests that the collecting information on equipment heat gain is more important to provide a better match on simulation and monitoring data in winter, whereas the reliable data on occupant density would be useful for a better prediction on electricity use in summer.

3.2. Results from correlation and cluster analysis

Fig. 5 shows the correlograms for monthly electricity and heating energy in this case study building. Fig. 6 illustrates the result from the hierarchical cluster analysis for monthly heating and electricity use in this retail building.

The color change from yellow to blue indicates a decrease of correlation coefficient in Fig. 5. As might be expected, the monthly electricity use has an apparent seasonal pattern. The correlation coefficients for electricity consumption are high, which occurs in three types of months: cold, hot, and transition months. The cold months include November, December, January, and February. The hot months are June, July, and August. The transition months contain March, April, May, September, and October. This suggests that the energy data can be grouped into three types: summer, winter, and transition months. This finding is in agreement with the trends from Fig. 6.

Fig. 6 shows the results from the hierarchical cluster analysis for monthly energy data. At the lowest level, each cluster contains a single observation, i.e. monthly energy data in this research. At the highest level, there is one cluster including all of the energy data. The heating and electricity data are divided in the second level, which means the patterns of energy use from heating and electricity are very different. For electricity use, there are three clusters (hot, cold, and transition months), which is the same as the results from Fig. 5.

For monthly heating energy use, two groups can be identified from both Figs. 5 (b) and 6 . The first group contains four months: November, December, January, and February. As can be seen from Fig. 6, more similarity of heating use can be found for January and December in this group. The second group only has March.

Based on the discussion above, the combinations of different output data are used to determine whether the calibration results from the informative outputs identified from Figs. 5 and 6 are better than the uninformative outputs. The informative outputs here refer to the monthly energy data in different groups/clusters that have high dissimilarities based on correlation/cluster analysis. The uninformative outputs are the monthly energy data with high similarity as can be seen from Fig. 5 and 6. For example, the monthly heating energy data in January and December are uninformative outputs since they are in the final cluster as shown in Fig. 6. In contrast, the monthly heating data in January and March are regarded as informative outputs since they are in different clusters (Fig. 6) and the correlation coefficient is low between them (Fig. 5).

Table 5 lists all the six cases used in this study in order to determine whether the results for informative and uninformative outputs from Bayesian calibration are different. For the cases of 1-1 and 1-2, the same number of output data (four for these two cases) is used and the difference is the energy data in various months. For the cases of 2-1 and 2-2, the eight output data are used for Bayesian computation, but the months for energy data are different. For the 3-1 case, the number of output data is seventeen, while the 3-2 case has ten energy data. The purpose of these two last cases is to compare whether the results from Bayesian analysis have significant differences from less outputs with rich information and all the outputs data available in building energy assessment.

It should be emphasized that the combination of both sensitivity analysis (Section 3.1) and correlation/cluster analysis (Section 3.2) can be used together to explore the characteristics of monthly energy use in buildings. The advantage of using sensitivity analysis is to analyse inputs and outputs at the same time. Therefore, the results of sensitivity analysis can guide analysts on which factors they should spend more time in collecting information. In contrast, the correlation & cluster analysis shows there is redundant information on the patterns of energy use for the same group of data identified in buildings.

3.3. Results from regression analysis

Figs. 7 and 8 show the results from regression analysis for monthly electricity and heating in this building. As shown in these two figures, the differences between training and testing data sets are not significant in terms of both the root mean square errors (RMSE) and the coefficients of determination (R^2). Hence, the regression models obtained from this analysis are accurate enough for Bayesian analysis, which will be implemented in the next section. In the case of large difference between training and test data sets, the regression models should be improved to avoid over-fitting.

As may be expected, the model accuracy from training data set is higher than that from test data set for both the RMSE and R^2. This is because a statistical model may include random error or noise instead of underlying relationship in obtaining this model using train data that can lead to overfitting (also called over-optimistic). Hence, the prediction errors using this statistical model are larger than the training errors. Note that the over-fitting is more severe in summer in comparison with winter since the differences between training and test data sets are larger. This is because the combination of cooling energy and daylighting technology makes the patterns of summer electricity more complicated, comparted to winter in this building.

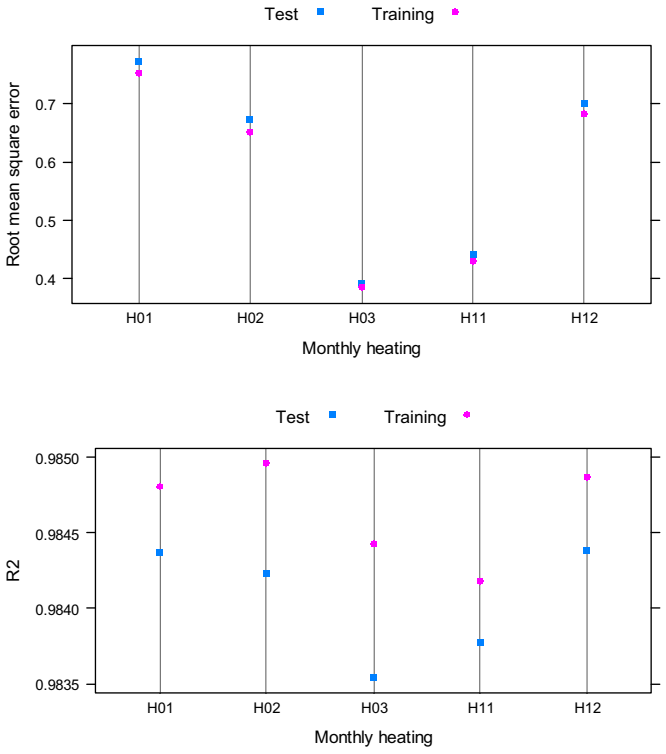

Fig. 8. Comparison of regression results for monthly heating in terms of RMSE (root mean square error) and R2 (coefficient of determination) in the retail building (H01 denotes the heating energy in January and so forth).

3.4. Results from Bayesian analysis

Three different data points (named as data A, B, and C) are used for Bayesian analysis to provide thorough analysis. Fig. 9–11 show the calibration results from the six different cases as listed in Table 5 for the first data point (data A). Tables 6 and 7 are obtained using data B and C, respectively.

The black vertical lines in Figs. 9–11 are the true values used for computing the output values. As a result, if the distributions of estimated inputs are closer to the corresponding vertical lines in terms of both central tendency and spread of the data, the results from Bayesian analysis are more accurate. The central tendency can be observed using the difference between the median of the distribution and the true values as listed in Tables 6 and 7. The spread of data is determined using the IQR (interquartile range) of the distribution, which is a measure of statistical dispersion between the upper (75th percentile) and lower (25th percentile) quartiles. A smaller IQR of inferred results indicate the more accurate input values from Bayesian analysis.

Fig. 9 illustrates the results from Bayesian calibration using four uninformative and four informative outputs, i.e. case 1-1 and 1-2 as listed in Table 5. The input values estimated from the case 1-1 have larger variations comparted to the case 2-1, which can be confirmed from Table 6 and 7 from the other two data points in terms of IQR. This indicates that the estimated inputs are more reliable based on the informative outputs if using the same number of output data (energy use in this study). The same conclusion can be also obtained based on Fig. 10. The computed inputs are more accurate from informative outputs than those from uninformative outputs in terms of the central tendency and the variations of input distributions. As for window U-value, the results from the case 2-2 are apparently better than those from the case 2-1 (as shown in Figs. 9 and 10) since more heating data is used in Bayesian computation. Note that the window U-value is ranked as the least important

Table 5
Cases of Bayesian analysis for three data points (A, B, and C).

No.	Case	Output number	Detailed outputs
1	1-1 Uninformative	4	Electricity: May, Sep Heating: Jan, Dec
	1-2 Informative	4	Electricity: Jan, Sep Heating: Jan, Mar
2	2-1 Uninformative	8	Electricity: Mar – Oct Heating: None
	2-2 Informative	8	Electricity: Jan, Apr, Jun, Sep, Nov Heating: Jan, Feb, Mar
3	3-1 All outputs	17	All the outputs (12 monthly electricity and 5 monthly heating data)
	3-2 Informative	10	Electricity: Feb, Apr, Jun, Aug, Sep, Oct, Dec Heating: Jan, Mar, Nov

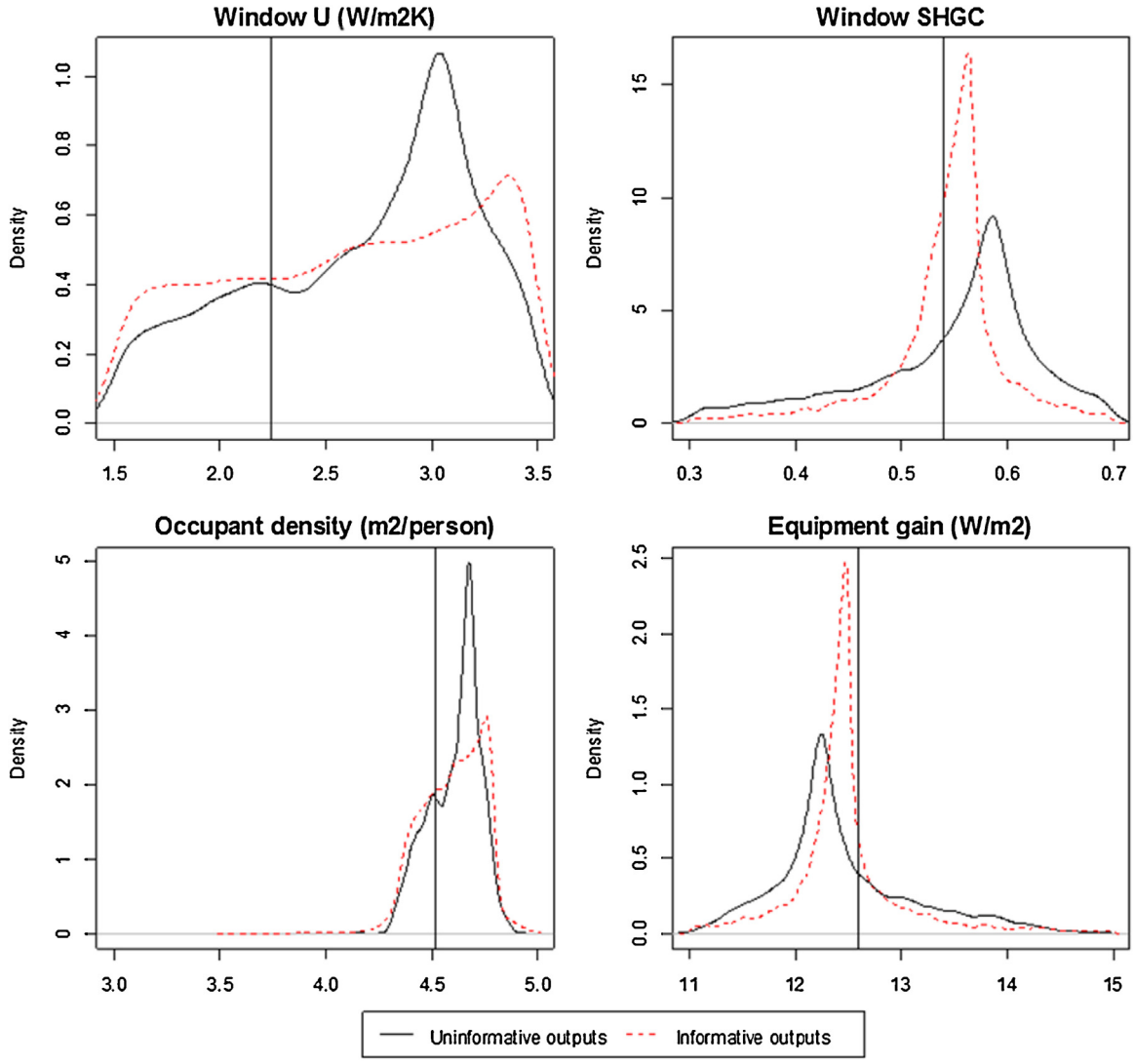

Fig. 9. Comparison of calibration results using both 4 uninformative and informative outputs in the retail building (for detailed descriptions, please refer to cases 1-1 and 1-2 in Table 5).

factor influencing electricity use in this case as described in Section 3.1.

Fig. 11 shows the comparison of calibration results using both ten informative outputs and all the seventeen output data for the retail building. In terms of the differences between the median of the inferred data and the true values, the results from two cases have the similar accuracy. As for the IQR, the variations from all the outputs are slightly better than the results from the 10 informative output data. The same conclusion can be drawn from the cases 3-1 and 3-2 (Tables 6 and 7). Therefore, the estimated input data can be very reliable if there are missing data, containing redundant information on energy patterns using Bayesian analysis method. For example, if only available energy data are the same as the informative output named as case 3-2 listed in Table 5 and remaining energy data (such as electricity use in January, March, May, and heating data in December) are missing, the results from Bayesian inference would be still very accurate.

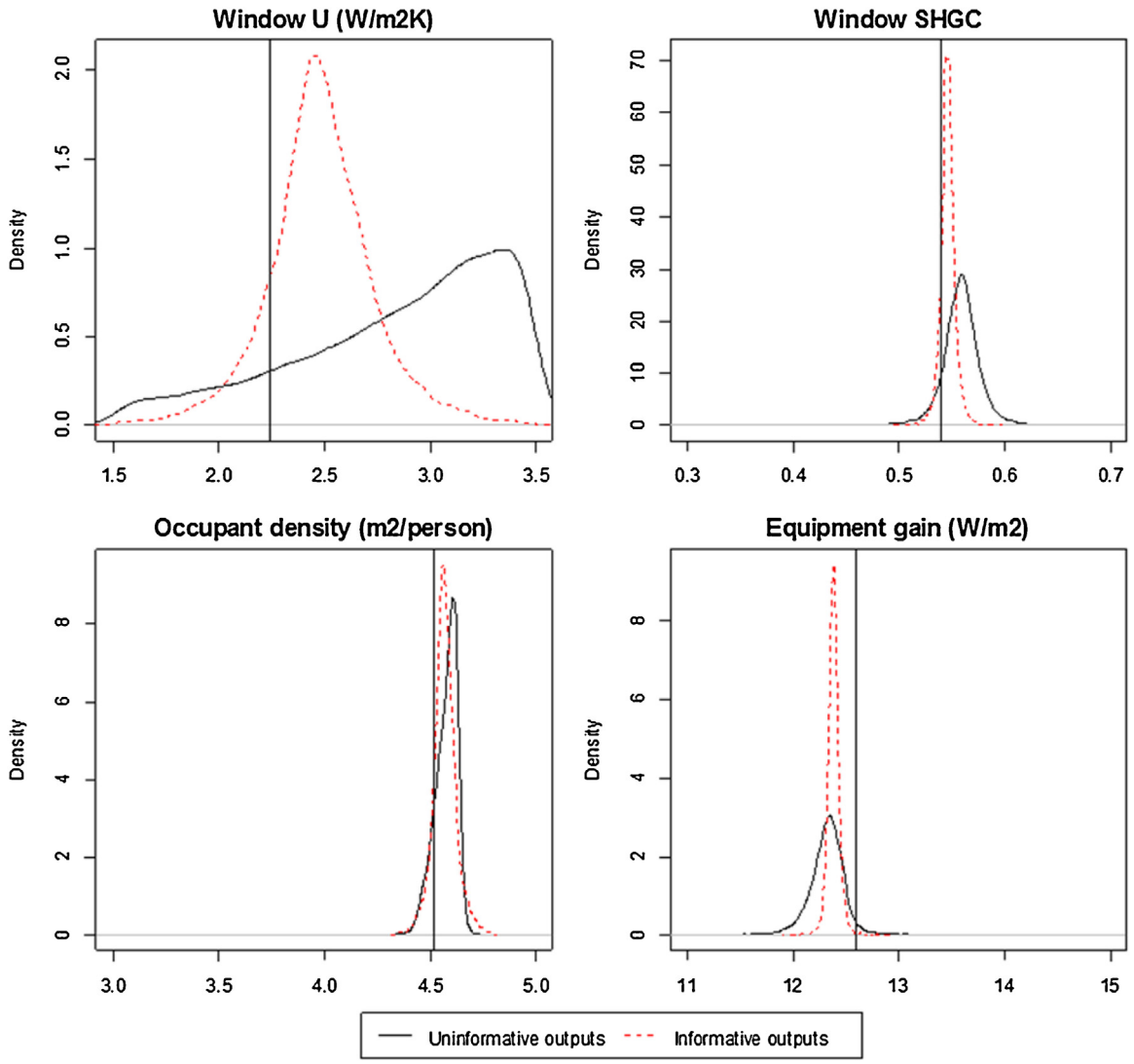

Fig. 10. Comparison of calibration results using both 8 uninformative and informative outputs in the retail building (for detailed descriptions, please refer to the cases 2-1 and 2-2 in Table 5).

Table 6
Summary of Bayesian computation from six cases for data B.

Case[a]	Item	WinU (W/m^{2}K)	SHGC	Occupant (m^2/person)	Equip (W/m^2)
	True values	2.927	0.624	3.171	12.344
1-1	Median[b]	2.918	0.536	3.107	13.04
	Difference[c]	0.009	0.088	0.064	0.696
	IQR[d]	0.500	0.170	0.119	1.301
1-2	Median	2.944	0.629	3.087	12.29
	Difference	0.017	0.005	0.084	0.054
	IQR	0.464	0.028	0.108	0.176
2-1	Median	3.083	0.666	3.052	12.08
	Difference	0.156	0.042	0.119	0.264
	IQR	0.462	0.023	0.047	0.188
2-2	Median	3.219	0.641	3.130	12.350
	Difference	0.292	0.017	0.041	0.006
	IQR	0.294	0.009	0.065	0.074
3-1	Median	2.875	0.637	3.054	12.29
	Difference	0.052	0.013	0.117	0.054
	IQR	0.184	0.007	0.041	0.045
3-2	Median	3.024	0.642	3.087	12.29
	Difference	0.097	0.018	0.084	0.054
	IQR	0.263	0.010	0.058	0.073

[a] For detailed descriptions of cases, please refer to Table 5.
[b] Median, median values of the estimated distributions for unknown factors.
[c] Difference: differences between the true values and the estimated median.
[d] IQR, interquartile range of unknown factors.

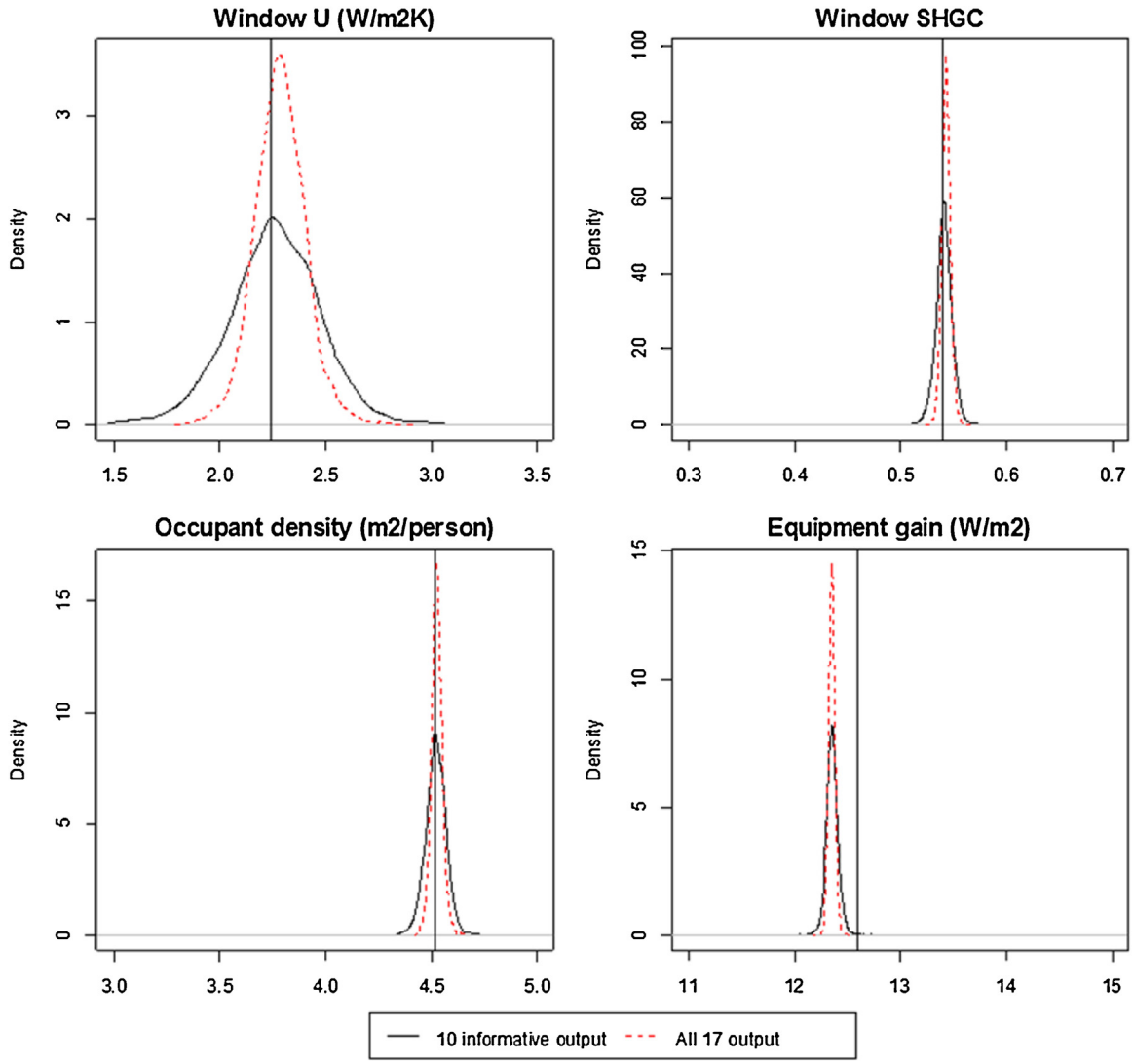

Fig. 11. Comparison of calibration results using both 10 informative and all the 17 outputs in the retail building (for detailed descriptions, please refer to the cases 3-1 and 3-2 in Table 5).

Table 7
Summary of Bayesian computation from six cases for data C.

Case	Item	WinU (W/m^{2}K)	SHGC	Occupant (m^2/person)	Equip (W/m^2)
	True values	1.607	0.475	4.478	13.854
1-1	Median	2.291	0.507	4.626	13.530
	Difference	0.684	0.032	0.148	0.324
	IQR	0.931	0.150	0.205	1.180
1-2	Median	2.273	0.493	4.629	13.580
	Difference	0.666	0.019	0.151	0.274
	IQR	0.966	0.037	0.214	0.261
2-1	Median	1.952	0.487	4.513	13.570
	Difference	0.345	0.012	0.035	0.284
	IQR	0.685	0.023	0.070	0.213
2-2	Median	1.875	0.484	4.542	13.590
	Difference	0.268	0.010	0.064	0.264
	IQR	0.245	0.007	0.054	0.049
3-1	Median	1.584	0.480	4.478	13.550
	Difference	0.023	0.006	0.000	0.304
	IQR	0.099	0.004	0.022	0.032
3-2	Median	1.668	0.483	4.497	13.560
	Difference	0.061	0.008	0.019	0.294
	IQR	0.200	0.008	0.044	0.058

Notes: 1: For detailed descriptions of cases, please refer to Table 5; 2: Median, median values of the estimated distributions for unknown factors; 3: Difference: differences between the true values and the estimated median; 4: IQR, interquartile range of unknown factors.

As for the calculation time of Bayesian analysis, the time for case 3-2 (i.e. using the 10 informative energy data) can be reduced by approximately 25% in comparison with case 3-1 (i.e. using all the 17 energy data). This advantage is very important, especially in the case of a large data set of energy data in which a lot of computational time can be significantly saved.

It is also interesting to note that the window U-value has the largest uncertainty for both cases (cases 1-1 and 1-2). This can be explained by the fact that the Window U-value is the least important factor based on the sensitivity analysis described in Section 3.1. Table 4 shows that the window U-value almost has no effects on the electricity use and this factor only has marginal effects on heating energy since the occupant density is the only dominant variable. Therefore, the large variation of the window U-value leads to a small uncertainty of energy data in this building. In contrast, the other three input values have similar variability as shown in Figs. 9–11.

Note that there are a number of factors that can affect the accuracy of estimated values in building energy calibration, including informative/uninformative data, model accuracy of outputs, and sensitivity indicators of outputs. The focus of this paper is whether monthly energy output data can be treated as informative or uninformative data in Bayesian calibration of energy models. Two other factors (model accuracy and sensitivity indicators) are not considered in this research. For instance, the occupant densities obtained from the case 1-2 and 1-1 are similar as shown in Fig. 9. This may be explained by the fact that the difference of ranking of variable importance for electricity use is significant by months as illustrated in Fig. 3. Therefore, the choice of various monthly electricity may be also dependent on sensitivity indictors. In fact, this is a very complicated issue since the order of inputs and outputs should be reversed in implementing sensitivity analysis. This is because the accuracy of unknown input factors depends on outputs based on the principle of Bayesian calibration in building energy models. This is an interesting area for further research.

4. Conclusions

This paper focuses on determining informative energy data in Bayesian calibration of building energy models. It comes as no surprise that the calibrated inputs values would be more accurate using more energy data. However, the results also indicate the different combinations of energy data have significant effects on the accuracy of inferred values for parameters in building energy models. The combination of energy data that leads to accurate values can be regarded as informative energy data, whereas the combination of uninformative energy data would result in unreliable results and its calculation cost may be also high. The methods of identifying the informative energy proposed in this study are correlation analysis and hierarchical cluster analysis. The correlation analysis is more concentrated on bivariate analysis, while the hierarchical cluster analysis would create a hierarchical data structure based on dissimilarity between group of observations (energy data in this case). The energy data allocated in the same group have similar energy trend and accordingly they can be treated as redundant data. In contrast, the energy data from different groups would represent different characteristics of energy use. In Bayesian calibration of energy models, the informative output data that are composed of energy data from different groups would be preferred to obtain reliable results with low computational cost.

The results from this research can be used to determine whether the reliable results can be obtained using Bayesian method in the case of missing energy data. If the available energy data is in the same group (identified using hierarchical cluster method) as the missing energy data, then Bayesian analysis can provide accurate calibration results. Moreover, the calculation time can be reduced significantly if using only informative energy data from different hierarchical groups. Furthermore, the methods used in this study (five steps shown in Fig. 1) can be used for calibrating energy models for buildings. The extra advantage from this calibration method is that the patterns of energy use in buildings can be explored thoroughly using both sensitivity analysis and correlation/clustering analysis. This would be also very useful to determine the effective measures of energy savings in building retrofitting projects.

Acknowledgements

This research is supported by the Tianjin Research Program of Application Foundation and Advanced Technology (No. 14JCYBJC42600) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry of China.

References

[1] IEA, Transition to Sustainable Buildings—Strategies and Opportunities to 2050, International Energy Agency (IEA) in 2013.
[2] C.K. Chau, T.M. Leung, W.Y. Ng, A review on life cycle assessment, life cycle energy assessment and life cycle carbon emissions assessment on buildings, Appl. Energy 143 (2015) 395–413.
[3] ASHRAE, Handbook of Fundamentals, Atlanta: American Society of Heating, Air-Conditioning and Refrigeration Engineers, in, Inc. (2013).
[4] N. Fumo, A review on the basics of building energy estimation, Renew. Sustain. Energy Rev. 31 (0) (2014) 53–60.
[5] G. Mustafaraj, D. Marini, A. Costa, M. Keane, Model calibration for building energy efficiency simulation, Appl. Energy 130 (0) (2014) 72–85.
[6] W. Shen, T. Hassan, Impact of occupant behaviour on the energy saving potential of retrofit measures for a public building in the UK, Intell. Build. Int. (2015), http://dx.doi.org/10.1080/17508975.2016.1139538 (in press).
[7] Z. Yang, B. Becerik-Gerber, A model calibration framework for simultaneous multi-level building energy simulation, Appl. Energy 149 (2015) 415–431.
[8] Y. Heo, R. Choudhary, G.A. Augenbroe, Calibration of building energy models for retrofit analysis under uncertainty, Energy Build. 47 (0) (2012) 550–560.
[9] Z. O'Neill, B. Eisenhower, S. Yuan, T. Bailey, S. Narayanan, V. Fonoberov, Modeling and calibration of energy models for a DoD building, ASHRAE Trans. 117 (2) (2011).
[10] Y. Pan, Z. Huang, G. Wu, Calibrated building energy simulation and its application in a high-rise commercial building in Shanghai, Energy Build. 39 (6) (2007) 651–657.
[11] D. Coakley, P. Raftery, M. Keane, A review of methods to match building energy simulation models to measured data, Renew. Sustain. Energy Rev. 37 (0) (2014) 123–141.
[12] R. Christensen, W.O. Johnson, A.J. Branscum, T.E. Hanson, Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians, CRC Press, 2011.
[13] Y. Heo, V.M. Zavala, Gaussian process modeling for measurement and verification of building energy savings, Energy Build. 53 (0) (2012) 7–18.
[14] W. Tian, R. Choudhary, A probabilistic energy model for non-domestic building sectors applied to analysis of school buildings in greater London, Energy Build. 54 (2012) 1–11.
[15] Y. Heo, G. Augenbroe, D. Graziano, R.T. Muehleisen, L. Guzowski, Scalable methodology for large scale building energy improvement: relevance of calibration in model-based retrofit analysis, Build. Environ. 87 (0) (2015) 342–350.
[16] W. Tian, Q. Wang, J. Song, S. Wei, Calibrating dynamic building energy models using regression model and Bayesian analysis in building retrofit projects, in: eSim, May 7 to 10, Ottawa, Canada, 2014.
[17] M. Manfren, N. Aste, R. Moshksar, Calibration and uncertainty analysis for computer models—a meta-model based approach for integrated building energy simulation, Appl. Energy 103 (0) (2013) 627–641.
[18] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc. Ser. B Stat. Methodol. (2001) 425–464.
[19] M. Riddle, R.T. Muehleisen, A guide to Bayesian calibration of building energy models, in: ASHRAE/IBPSA-USA Building Simulation Conference, Atlanta, GA, September 10–12, 2014, 2014.
[20] K. Campbell, Statistical calibration of computer simulations, Reliab. Eng. Syst. Saf. 91 (10–11) (2006) 1358–1363.
[21] Y. Heo, D.J. Graziano, L. Guzowski, R.T. Muehleisen, Evaluation of calibration efficacy under different levels of uncertainty, J. Build. Perform. Simul. 8 (3) (2015) 135–144.
[22] P. Congdon, Bayesian Statistical Modelling, Wiley, 2007.
[23] J. Liepe, P. Kirk, S. Filippi, T. Toni, C.P. Barnes, M.P.H. Stumpf, A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation, Nat. Protoc. 9 (2) (2014) 439–456.

[24] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2015 http://www.R-project.org/.
[25] MOC, GB50189-2005, Energy Conservation Design Regulation for Public Buildings. Ministry of Construction (MOC) of P.R. China, China Planning Press, 2005 (in Chinese).
[26] DOE, EnergyPlus V8.1, October 2013, Department of Energy, USA, in, 2013.
[27] M. Konstantoglou, A. Tsangrassoulis, Dynamic operation of daylighting and shading systems: a literature review, Renew. Sustain. Energy Rev. 60 (2016) 268–283.
[28] Y.K. Yi, Adaptation of Kriging in daylight modeling for energy simulation, Energy Build. 111 (2016) 479–496.
[29] T. Hong, H. Sun, Y. Chen, S.C. Taylor-Lange, D. Yan, An occupant behavior modeling tool for co-simulation, Energy Build. 117 (2016) 272–281.
[30] T. Méndez Echenagucia, A. Capozzoli, Y. Cascone, M. Sassone, The early design stage of a building envelope: multi-objective search through heating, cooling and lighting energy performance analysis, Appl. Energy 154 (2015) 577–591.
[31] Y. Kwak, J.-H. Huh, C. Jang, Development of a model predictive control framework through real-time building energy management system data, Appl. Energy 155 (2015) 1–13.
[32] W. Tian, P. de Wilde, Uncertainty and sensitivity analysis of building performance using probabilistic climate projections: a UK case study, Autom. Constr. 20 (8) (2011) 1096–1109.
[33] S. Kucherenko, D. Albrecht, A. Saltelli, Comparison of latin hypercube and quasi Monte Carlo sampling techniques, in: IBPSA (International Building Performance Simulation Association) conference 2011, November 14–16, Sydney, 2011.
[34] S. Levy, D.M. Steinberg, Computer experiments: a review, AStA Adv. Stat. Anal. 94 (4) (2010) 311–324.
[35] D. Christophe, S. Petr, R Package Randtoolbox: Generating and Testing Random Numbers. R package version 1.16, 2015 (accessed 01-06.15.) http://cran.r-project.org/web/packages/randtoolbox.
[36] W. Tian, A review of sensitivity analysis methods in building energy analysis, Renew. Sustain. Energy Rev. 20 (2013) 411–419.
[37] W. Tian, R. Choudhary, G. Augenbroe, S.H. Lee, Importance analysis and meta-model construction with correlated variables in evaluation of thermal performance of campus buildings, Build. Environ. 92 (0) (2015) 61–74.
[38] W. Tian, J. Song, Z. Li, P. de Wilde, Bootstrap techniques for sensitivity analysis and model selection in building thermal performance analysis, Appl. Energy 135 (2014) 320–328.
[39] J. Wang, Z. Zhai, Y. Jing, X. Zhang, C. Zhang, Sensitivity analysis of optimal model on building cooling heating and power system, Appl. Energy 88 (12) (2011) 5143–5152.
[40] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global Sensitivity Analysis: the Primer, Wiley-Interscience, 2008.
[41] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.
[42] B.I. Gilles Pujol, A. Janon, R package sensitivity V1.11: Sensitivity Analysis, 2015 http://cran.r-project.org/package=sensitivity/.
[43] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22.
[44] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[45] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer Publishing Company, Incorporated, 2009.
[46] J.J. Faraway, Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Chapman & Hall, 2006.
[47] Z. Qiao, L. Zhou, J.Z. Huang, Sparse linear discriminant analysis with applications to high dimensional low sample size data, Int. J. Appl. Math. 39 (1) (2009) 48–60.
[48] J. Song, L. Wei, Y. Sun, W. Tian, Implementation of meta-modelling for sensitivity analysis in building energy analysis, in: eSim, May 7 to 10, Ottawa, Canada, 2014.
[49] J.J. Faraway, Linear Models with R, CRC Press, London, 2005.
[50] M.K. Cowles, Applied Bayesian Statistics: With R and OpenBUGS Examples, Springer, 2013.
[51] A. Thomas, B. O'Hara, U. Ligges, S. Sturtz, Making BUGS open, R News 6 (1) (2006) 12–17.
[52] J.K. Marley, M.P. Wand, Non-standard semiparametric regression via BRugs, J. Stat. Softw. 37 (5) (2010) 1–30.