Energy and Buildings 49 (2012) 560–567

Contents lists available at SciVerse ScienceDirect

Energy and Buildings

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

Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools

Athanasios Tsanas[a,*], Angeliki Xifara[b]

[a] Oxford Centre for Industrial and Applied Mathematics (OCIAM), Mathematical Institute, University of Oxford, 24-29 St. Giles', OX1 3LB Oxford, UK
[b] Architectural Science Group, Welsh School of Architecture, Cardiff University, UK

ARTICLE INFO

Article history:
Received 9 August 2011
Received in revised form 16 February 2012
Accepted 3 March 2012

Keywords:
Building energy evaluation
Heating load
Cooling load
Non-parametric statistics
Statistical machine learning

ABSTRACT

We develop a statistical machine learning framework to study the effect of eight input variables (relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution) on two output variables, namely heating load (HL) and cooling load (CL), of residential buildings. We systematically investigate the association strength of each input variable with each of the output variables using a variety of classical and non-parametric statistical analysis tools, in order to identify the most strongly related input variables. Then, we compare a classical linear regression approach against a powerful state of the art nonlinear non-parametric method, random forests, to estimate HL and CL. Extensive simulations on 768 diverse residential buildings show that we can predict HL and CL with low mean absolute error deviations from the ground truth which is established using Ecotect (0.51 and 1.42, respectively). The results of this study support the feasibility of using machine learning tools to estimate building parameters as a convenient and accurate approach, as long as the requested query bears resemblance to the data actually used to train the mathematical model in the first place.

1. Introduction

There has been a considerable body of research on the topic of energy performance of buildings (EPB) recently due to growing concerns about energy waste and its perennial adverse impact on the environment [1,2]. Moreover, buildings in European countries are legally bound to conform to appropriate minimum requirements regarding energy efficiency following the European Directive 2002/91/EC [1]. Reports suggest that building energy consumption has steadily increased over the past decades worldwide [3,4], and heating, ventilation and air conditioning (HVAC), which have a catalytic role in regulating the indoor climate [5], account for most of the energy use in the buildings [6]. Therefore, one way to alleviate the ever increasing demand for additional energy supply is to have more energy-efficient building designs with improved energy conservation properties.

When it comes to efficient building design, the computation of the heating load (HL) and the cooling load (CL) is required to determine the specifications of the heating and cooling equipment needed to maintain comfortable indoor air conditions. In order to estimate the required cooling and heating capacities, architects and building designers need information about the characteristics of the building and of the conditioned space (for example occupancy and activity level), the climate, and the intended use (residential buildings have generally different requirements compared to industrial buildings).

Building energy simulation tools are currently widely used to analyse or forecast building energy consumption, in order to facilitate the design and operation of energy efficient buildings since practice has shown that the results of the simulations can often accurately reflect actual measurements [7]. Simulation tools are used extensively across diverse disciplines because they enable experimentation with parameters that would otherwise be infeasible, or at least very difficult to control in practice [8]. In the context of building energy design for example, simulations could facilitate the comparison of identical buildings where only a single parameter is modified across a range of possible values to investigate its effects on some observed quantity of interest. For an overview and comparison of building simulation tools we refer to Yezioro et al. [7] and to Crawley et al. [9].

Using advanced dedicated building energy simulation software may provide reliable solutions to estimate the impact of building design alternatives; however this process can be very time-consuming and requires user-expertise in a particular program. Moreover, the accuracy of the estimated results may vary across different building simulation software packages [7]. Hence, in practice many researchers rely on *machine learning tools* to study the effect of various building parameters (e.g. compactness) on some variables of interest (e.g. energy) because this is easier and

* Corresponding author. Tel.: +44 1865280603; fax: +44 1865270515.
E-mail addresses: tsanas@maths.ox.ac.uk, tsanasthanasis@gmail.com (A. Tsanas), angxifara@gmail.com (A. Xifara).

0378-7788/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.enbuild.2012.03.003

faster if a database of the required ranges of variables is available [2,10,11]. Using statistical and machine learning concepts has the distinct advantage that distilled expertise from other disciplines is brought in the EPB domain, and by using these techniques it is extremely fast to obtain answers by varying some building design parameters once a model has been adequately trained. Moreover, statistical analysis can enhance our understanding offering *quantitative expressions* of the factors that affect the quantity (or quantities) of interest that the building designer or architect may wish to focus on. Therefore, the integration of machine learning in EPB has sparked enormous interest lately.

Various machine learning techniques such as polynomial regression [11], support vector machines (SVM) [10,12] artificial neural networks (ANN) [13,14], and decision trees [2] have been explored to predict various quantities of interest in the context of EPB. Machine learning tools have also been explicitly used in predicting HL and CL. Catalina et al. [11] used polynomial regression (including up to quadratic terms) to predict monthly heating demand for residential buildings. They used as inputs for the regression model the building shape factor, the envelope *U*-value, the window-to-floor area ratio, the building time constant, and climate. Wan et al. [15] studied the impact of climate change on HL and CL for office buildings in China. Schiavon et al. [16] focused on the influence of raised floor, structure type, window-to-wall ratio and the presence of carpet to determine CL for different zones, and reported that orientation and the presence of carpet are the most important predictors. Li et al. [12] forecast hourly building CL based mainly on preceding environmental parameters. Of particular interest to this study, HL and CL have been associated with variables such as relative compactness (RC) [17], climate [15], surface area, wall area, and roof area [16,17], orientation [16,17], and glazing [17]. The rationale for studying these variables is that designers and engineers have found that they are correlated with energy performance, and HL and CL in particular.

Many studies in the general research area of EPB have made rigid simplifying mathematical assumptions relying on linear correlations and classical least squares regression techniques, tools which are known to be ill-suited for many complicated applications where normality assumptions do not hold. Other studies have used complicated machine learning tools, but have failed to rigorously examine the available data (*data mining*), for example to report which variables are the most important for the particular problem addressed, thus failing to leverage on important information that can be inferred when using statistical tools.

In this study, we investigate the effect of eight *input variables*: (RC), surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution, to determine the *output variables* HL and CL of residential buildings. Those eight variables have been frequently used in the EPB literature to study energy-related topics in buildings, and this study builds on the work of Pessenlehner and Mahdavi [17] who used those particular eight variables to investigate their effect on HL. We statistically formally explore the data, provide meticulous statistical analysis to gain important insight of the underlying properties of input and output variables, and use robust classical regression and state of the art nonlinear and non-parametric statistical machine learning tools (random forests) to map the input variables to HL and CL.

2. Data

Taking the elementary cube ($3.5 \times 3.5 \times 3.5$) we generated 12 building forms where each building form is composed of 18 elements (elementary cubes). The simulated buildings were generated using Ecotect. All the buildings have the same volume, which is 771.75 m^3, but different surface areas and dimensions. The materials used for each of the 18 elements are the same for all building forms. The selection was made by the newest and most common materials in the building construction industry and by the lowest *U*-value. Specifically, we used the following building characteristics (the associated *U*-values appear in parenthesis): walls (1.780), floors (0.860), roofs (0.500), windows (2.260). The simulation assumes that the buildings are in Athens, Greece, residential with seven persons, and sedentary activity (70 W). The internal design conditions were set as follows: clothing: 0.6 clo, humidity: 60%, air speed: 0.30 m/s, lighting level: 300 Lux. The internal gains were set to sensible (5) and latent (2 W/m^2), while the infiltration rate was set to 0.5 for air change rate with wind sensitivity 0.25 air changer per hour. For the thermal properties we used mixed mode with 95% efficiency, thermostat range 19–24 °C, with 15–20 h of operation on weekdays and 10–20 h on weekends.

Table 1
Mathematical representation of the input and output variables to facilitate the presentation of the subsequent analysis and results.

Mathematical representation	Input or output variable	Number of possible values
$X1$	Relative compactness	12
$X2$	Surface area	12
$X3$	Wall area	7
$X4$	Roof area	4
$X5$	Overall height	2
$X6$	Orientation	4
$X7$	Glazing area	4
$X8$	Glazing area distribution	6
$y1$	Heating load	586
$y2$	Cooling load	636

Following the classical mathematical convention, we use X to denote input variables and y to denote output variables. Although 768 different buildings were simulated, in some cases the output variables of different buildings might coincide. The probability densities for each variable are shown in Fig. 1.

We used three types of glazing areas, which are expressed as percentages of the floor area: 10%, 25%, and 40%. Furthermore, five different distribution scenarios for each glazing area were simulated: (1) uniform: with 25% glazing on each side, (2) north: 55% on the north side and 15% on each of the other sides, (3) east: 55% on the east side and 15% on each of the other sides, (4) south: 55% on the south side and 15% on each of the other sides, and (5) west: 55% on the west side and 15% on each of the other sides. In addition, we obtained samples with no glazing areas. Finally, all shapes were rotated to face the four cardinal points.

Thus, considering twelve building forms and three glazing area variations with five glazing area distributions each, for four orientations, we obtained $12 \times 3 \times 5 \times 4 = 720$ building samples. In addition, we considered twelve building forms for the four orientations without glazing. Therefore, in total we studied $12 \times 3 \times 5 \times 4 + 12 \times 4 = 768$ buildings. Each of the 768 simulated buildings can be characterized by eight building parameters (to conform to standard mathematical notation and facilitate the analysis in this work, henceforth these building parameters will be called *input variables* and will be represented with X) which we are interested in exploring further. Also, for each of the 768 buildings we recorded HL and CL (henceforth these parameters will be called *output variables* and will be represented with y). Table 1 summarizes the input variables and the output variables in this study, introduces the mathematical representation for each variable, and indicates the number of possible values.

Simulating building energy aspects is a widely used approach despite the fact that it is impossible to guarantee that the simulation findings will perfectly reflect actual data in the real world (here HL and CL). Nevertheless, the simulated results provide good indication of the likely percentage change and any underlying trend of the actual data, enabling energy comparisons of buildings [15]. That is, even if the data used in this study obtained via the

simulations could be biased in some way, they represent actual real data with high probability and as such will be considered as *ground truth*. Moreover, any inconsistency in the simulated data and actual real-world data does not affect whatsoever the methodology developed in this study.

3. Methods

This section briefly summarizes the data-driven statistical concepts and the machine learning techniques which are used to analyse the data.

3.1. Data exploration and statistical analysis

The first step in most data analysis applications is the exploration of the statistical properties of the variables. This is typically achieved by plotting the probability densities, which succinctly summarize each variable for visualization. One way to obtain an empirical non-parametric density estimate is by using histograms. Although histograms are considered crude for most advanced statistical applications, they have the great advantage of making no prior assumptions regarding the distribution of the examined variable and are very simple to compute. Often, this preliminary step can reveal whether the variable follows a Gaussian (normal) distribution, which is characterized by a unimodal peak in the middle of the variable's possible range of values, is completely symmetric, and is particularly useful because a large number of mathematical functions are applicable [18]. Moreover, we present scatter plots for each input variable with each of the two output variables. For simplicity, scatter plots often use *normalized* data (i.e. all the variables are normalized to lie between 0 and 1) to facilitate comparison between measures that possibly span orders of magnitude different ranges of values [22].

The data is non-Gaussian, so we used the Spearman rank correlation coefficient to obtain a statistical metric regarding the association strength of each input variable with each of the two outputs [19]. The Spearman rank correlation coefficient can characterize general *monotonic* relationships and lies in the range −1 to 1, where negative sign indicates inversely proportional and positive sign indicates proportional relationship, whilst the magnitude denotes how strong this relationship is. In addition, we evaluate whether this relationship is statistically significant using p-values, and check for significance at the 0.01 level. Moreover, we used the mutual information (MI) [20] which can be used to quantify *any arbitrary* relationships between the input and output variables. Because MI is not upper bounded we normalize it to lie in the range [0 to 1] (see Tsanas et al. [21] for details). The larger the MI value, the stronger the association strength between the two variables.

3.2. Statistical mapping of the input variables to the output variables

Given N samples (here N=*768*) and M input variables (here M=*8*), we construct a matrix $X \in \mathbb{R}^{N\times M}$ which includes the available information in compact matrix format:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NM} \end{bmatrix}$$

This is typically associated with a response variable vector $y \in \mathbb{R}^{N\times 1}$ and we need to find the *functional relationship* f to relate X and y (here y is either HL or CL) such that $y = f(X)$. For convenience in later analysis, we denote $y1$ to represent HL, and $y2$ to represent CL. The tool that performs the functional mapping is commonly referred to as *learner* in the machine learning literature. In this study, we approach the problem of inferring this functional relationship from two perspectives (i.e. we use two learners): a standard robust linear regression technique, and a powerful nonlinear non-parametric classifier. Given that the output variables span a continuous range of values, using a regression technique may seem obvious; the use of a classification tool may initially seem counterintuitive. However, in practice it may be convenient to discretize the output variable and treat the given application as a multi-class classification problem because there are powerful classification tools available. For a recent study that demonstrated the potential of this concept (discretizing an originally continuous valued output variable and using multi-class classification tools to determine the functional relationship) see Tsanas et al. [21].

3.2.1. Iteratively reweighted least squares

The simplest and most common regression method is the ordinary least squares method. However, in many applications the residuals often depart markedly from the Gaussian distribution (these points are known as *outliers*), and for this reason it is good practice to use a slightly more complicated method, the iteratively reweighted least squares (IRLS). In short, IRLS adjusts weights in the coefficients of the classical regression scheme to lessen the effect of the outliers in producing the fitting curve, and provide an improved least squares estimate. We refer to Bishop [18] and to Tsanas et al. [22] for more technical details.

3.2.2. Classification using random forests

In many practical applications it is possible that the inputs exhibit a complicated functional relationship to determine the output. The classification and regression tree (CART) method is a conceptually simple, yet powerful nonlinear method that often provides excellent results [22,23]. CART works by successively splitting the input feature space into smaller and smaller sub-regions. This procedure can be visualized as a tree that splits into successively smaller branches, each branch representing a sub-region of the input variable ranges. The tree grows until it is not possible to split it any more or a certain criterion has been met. A natural extension of CART is random forests (RF), which is simply a collection of many trees [24]. The training procedure is the same as in CART with the difference that a randomly chosen subset of candidate variables can be used to select the optimal variable for each split; practice has shown the RF algorithm works extremely well in many diverse applications [21,24].

Moreover, RF have the desirable ability of promoting the most important input variables towards predicting the output variable as part of their inherent learning strategy [23]. This *wrapper* aspect is particularly useful in practical applications pointing the input variables that are particularly well suited with the RF learner for the designed problem. In this study we do not explore the wrapper aspect of RF to actually *select* features in order to use the selected subset as input into the learner (for a recent application of this see [25]), but merely use this property to report the most strongly associated input variables with the output variables. We stress that variable importance is not assessed for each variable *independently*; instead, it is *jointly* assessed for the feature subset used in the RF, making use of *relevance* (association strength of variable and response), *redundancy* (association strength between variables), and *complementarity* (joint association strength of variables with the response) concepts. Effectively, this means that highly correlated variables (which exhibit high correlations between/amongst variables) are penalised and hence the redundant variables are not assigned large importance even though they may be highly correlated with the response [24]. Further particulars on CART and RF can be found in Hastie et al. [23].

3.3. Cross validation and model generalization

Having trained the learner, it is necessary to test its generalization performance, i.e. the performance we can expect in a new dataset with similar characteristics. We use *cross validation* (CV), a standard statistical re-sampling technique. Specifically, the dataset is split into a training subset with which the learner is trained, and a testing subset which is used to assess the learner's generalization performance. Typically some percentage of the data is left out for testing the learner, and this is known as K-fold CV, where K is usually 5 or 10 [23]. In this study, we used 10-fold CV. The model parameters are derived using the training subset, and errors are computed using the testing subset (*out-of-sample* error or *testing* error). For statistical confidence, the training and testing process is repeated 100 times with the dataset randomly permuted in each run prior to splitting in training and testing subsets. On each test repetition, we record the mean absolute error (MAE), the mean square error (MSE), and the mean relative error (MRE) for both training and testing subsets. In all cases we report the out-of sample error.

$$\mathrm{MAE} = \frac{1}{S}\sum_{i\in Q}|y_i - \hat{y}_i| \quad (1)$$

$$\mathrm{MSE} = \frac{1}{S}\sum_{i\in Q}|y_i - \hat{y}_i|^2 \quad (2)$$

$$\mathrm{MRE} = 100\cdot\frac{1}{S}\sum_{i\in Q}\frac{|y_i - \hat{y}_i|}{y_i} \quad (3)$$

where $\hat{y}_i$ is the predicted output variable and y_i is the actual output variable for the ith entry in the training or testing subset, S is the number of samples in the training or testing subset, and Q contains the indices of that set. Errors over the 100 CV realisations were averaged. The MAE has often been used in recent studies that relied on decision trees and RF because of its ease of interpretation [21,22], whilst the MSE is commonly used in domains relying on minimizing the least squares (e.g. in IRLS). Moreover, the MAE has often been used in some EPB-related studies, e.g. [7].

4. Results

This section applies the methodology outlined in Section 3 for the input and output variables.

4.1. Statistical analysis

Fig. 1 presents the empirical probability distributions of all the input and output variables. These distributions demonstrate that none of the variables follows the normal distribution. Fig. 2 displays the scatter plots for each of the (normalized) input variables with each of the two output variables. These scatter plots show that any functional relationship of the input variables and the output variables is not trivial. This suggests that we can reasonably expect that classical learners such as linear regression may fail to find an accurate mapping of the input variables to the output variables. Therefore, these plots intuitively justify the need to experiment with complicated learners such as RF.

Tables 2 and 3 report the association strength (quantified using MI and the rank correlation coefficient) for each input variable with HL and with CL, respectively. From these results we infer that the first five input variables appear reasonably strongly associated with the output variables. Table 4 presents the covariance matrix which denotes the rank correlations between input variables. The results in this Table indicate that $X1$ (RC) and $X2$ (surface area) are inversely proportional, which is because in our simulations we have assumed that the volume of the buildings is constant (there is an analytic formula linking the surface area to RC and volume). Interestingly, the results in Table 4 reveal that some of the input variables are also highly correlated, for example $X4$ (roof area) and $X5$ (height). As we intuitively expected, these variables are almost inversely proportional, which is revealed from the sign and the magnitude of the rank correlation coefficient (−0.937).

Table 2
Association strength estimated using the mutual information and the Spearman rank correlation coefficient of the eight input variables ($X1\ldots X8$) with HL ($y1$).

Input variable	Mutual information (normalized)	Spearman rank correlation coefficient	p-value
$X1$	0.605	0.622	<0.001
$X2$	0.602	−0.622	<0.001
$X4$	0.567	−0.804	<0.001
$X5$	0.548	0.861	<0.001
$X3$	0.402	0.471	<0.001
$X7$	0.149	0.323	<0.001
$X8$	0.051	0.068	fail ($p>0.05$)
$X6$	0	−0.004	fail ($p>0.05$)

Table 3
Association strength estimated using the mutual information and the Spearman rank correlation coefficient of the input variables ($X1\ldots X8$) with CL ($y2$).

Input variable	Mutual information (normalized)	Spearman rank correlation coefficient	p-value
$X1$	0.616	0.651	<0.001
$X2$	0.615	−0.651	<0.001
$X4$	0.612	−0.803	<0.001
$X5$	0.59	0.865	<0.001
$X3$	0.423	0.416	<0.001
$X7$	0.092	0.289	<0.001
$X8$	0.028	0.046	fail ($p>0.05$)
$X6$	0.001	0.018	fail ($p>0.05$)

4.2. Cross validation results using IRLS and RF

Having completed the preliminary statistical analysis which provides important insight into the association strength of the input variables with the output variables, we study how accurate the actual statistical mapping is reporting out-of-sample errors. The IRLS coefficients for predicting HL are presented in Eq. (4), and the IRLS coefficients for predicting CL are presented in Eq. (5). We present the mean value of each IRLS coefficient over the 100 CV iterations: all coefficients were very stable across the 100 iterations.

$$\mathrm{IRLS_{HL}} = -4.75\cdot X_1 - 0.03\cdot X_2 + 0.07\cdot X_3 + 0\cdot X_4 - 3.44\cdot X_5 - 0.01\cdot X_6 + 18.13\cdot X_7 + 0.09\cdot X_8 \quad (4)$$

$$\mathrm{IRLS_{CL}} = -9.02\cdot X_1 - 0.01\cdot X_2 + 0.04\cdot X_3 + 0\cdot X_4 - 4.30\cdot X_5 - 0.12\cdot X_6 + 14.49\cdot X_7 + 0.03\cdot X_8 \quad (5)$$

Tables 5–7 present the out-of-sample MAE, MSE, and MRE of predicting the two output variables. Collectively, these results suggest that it is possible to estimate HL and CL very accurately simply using the eight variables that this study makes use of. Not unexpectedly, RF demonstrates *consistently superior* performance since the underlying relationships are quite complicated to be adequately captured by a simple linear learner. Finally, Table 8 presents the importance of the input variables estimated using RF. Interestingly, the importance results in Table 8 suggest that $X7$ (glazing area) is the most important predictor for both HL and CL. To verify these findings and further support our confidence on those

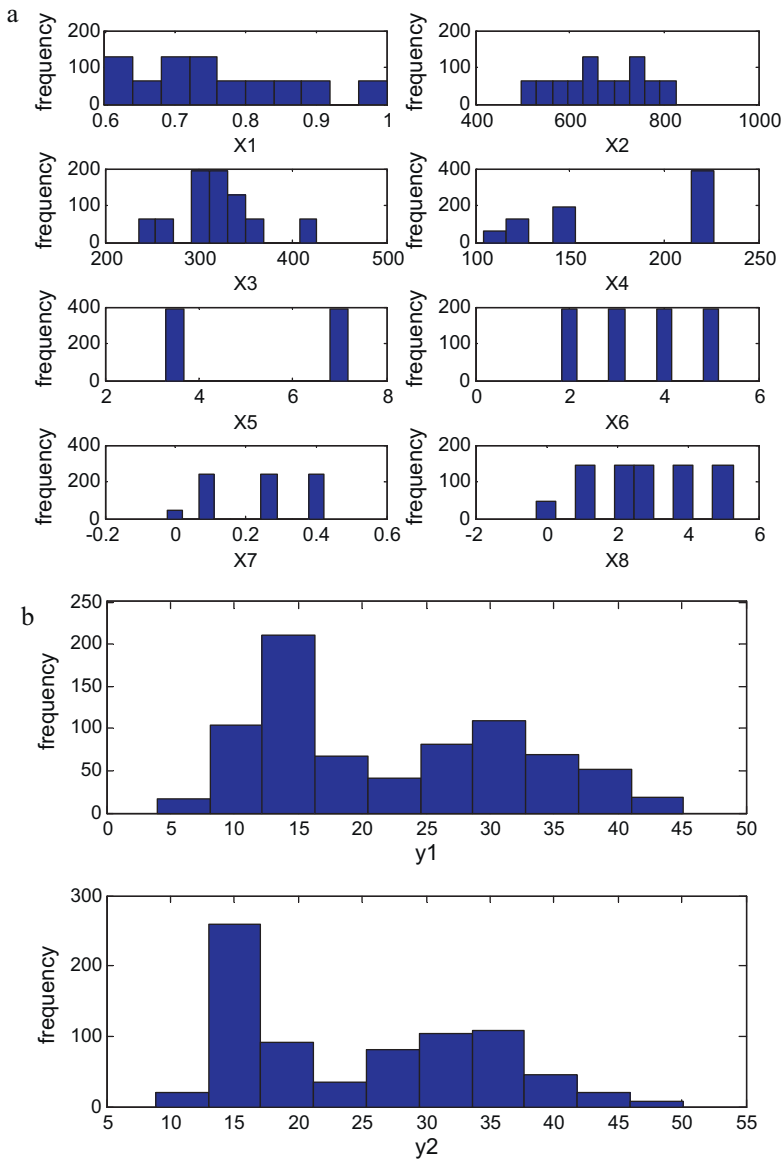

Fig. 1. Probability density estimates using histograms of (a) the eight input variables, and (b) the two output variables.

results, we introduce an artificial probe in the original design matrix (a random Gaussian variable) which is a standard methodology in the RF literature [26]. Hence, we now have the original eight variables and an additional (non-predictive) variable. We have repeated the learning process with the nine variables, and found that the artificial probe exhibits considerably lower importance score compared to the most predictive original variables (as expected).

The findings in Table 8 are in broad agreement with those inferred from observing the IRLS coefficients. Although we caution against the direct interpretation of variable importance in a linear regression setting due to collinearity problems, in some cases we can infer which variables may be most predictive of the response by looking at the associated IRLS coefficients. From Eqs. (4) and (5) we see that $X7$ appears to have the largest coefficient, verifying the results in Table 8 suggesting that this variable may be indeed most

Table 4
Correlation matrix using Spearman rank correlations between the eight input variables.

	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$
$X1$	1	−1	−0.256	−0.871	0.869	0	0	0
$X2$	−1	1	0.256	0.871	−0.869	0	0	0
$X3$	−0.256	0.256	1	−0.193	0.221	0	0	0
$X4$	−0.871	0.871	−0.193	1	−0.937	0	0	0
$X5$	0.869	−0.869	0.221	−0.937	1	0	0	0
$X6$	0	0	0	0	0	1	0	0
$X7$	0	0	0	0	0	0	1	0.188
$X8$	0	0	0	0	0	0	0.188	1

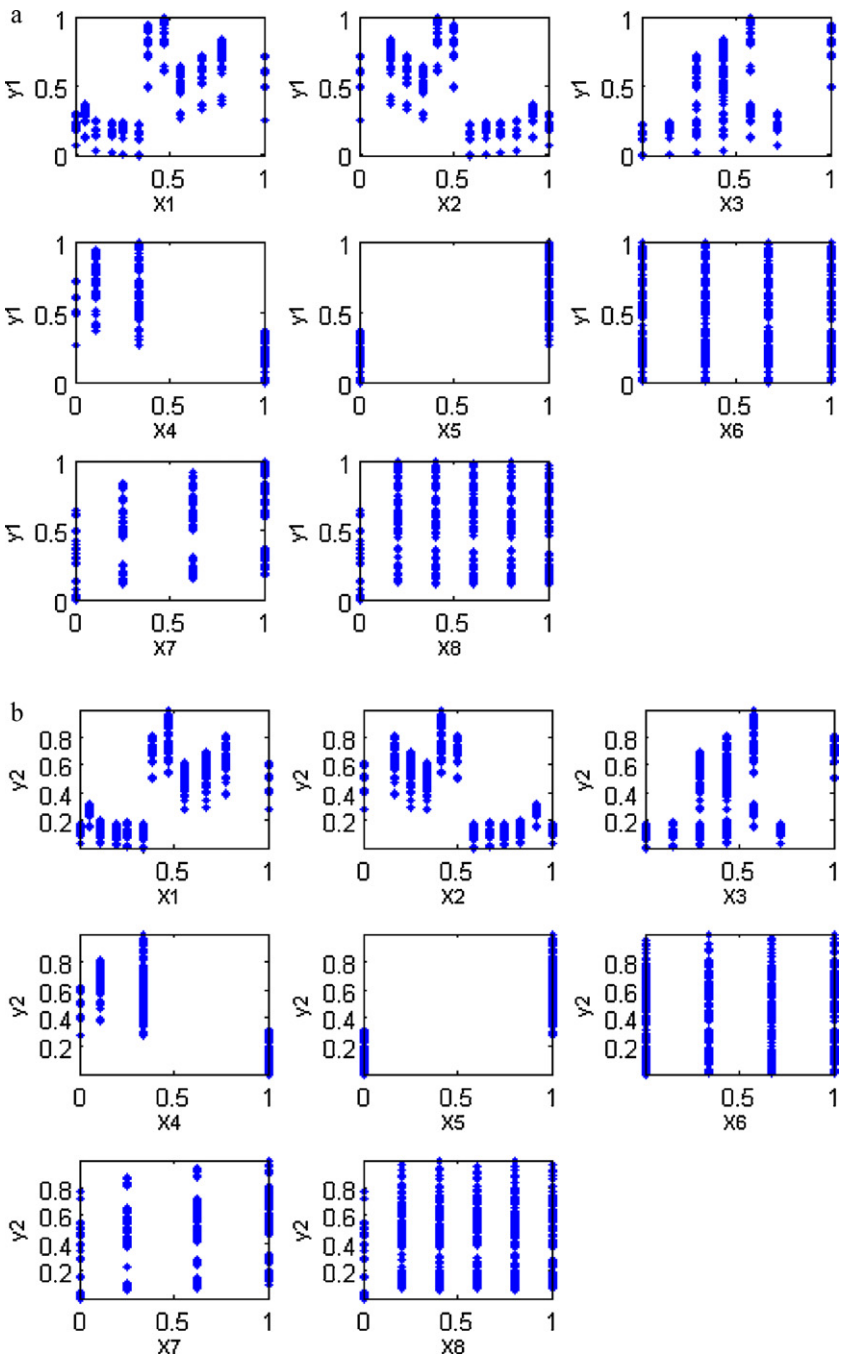

Fig. 2. Scatter plot demonstrating visually the relationship between each normalized input variable and the normalized outputs (a) the heating load, or (b) the cooling load.

predictive of both HL and CL. We elaborate further on this finding in Section 5.

5. Discussion

We have developed a comprehensive framework to study HL and CL using a range of diverse input variables which included compactness, orientation and glazing properties. We demonstrated that we can accurately estimate HL with only 0.5 points deviation and CL with 1.5 points deviation from the ground truth (the simulated results). These findings are particularly compelling given the accurate prediction, and also because we can easily infer the output variables in a matter of few seconds without requiring the painstaking design of a new building in a simulation tool such as Ecotect. We remark that the values provided by Ecotect for HL and CL are considered to reflect the true actual values; a detailed comparison of the provided output values from different simulation programs is beyond the scope of this study. Moreover, the presented methodology is applicable regardless of the simulation program that generates values which are believed to be accurate.

We explored the statistical relationship between eight input variables (RC, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution) and the two output variables (HL and CL). The statistical tools used here indicate that RC, wall area and roof area appear mostly associated with HL and CL. We remark that the surface area is inversely proportional to RC

Table 5
Out of sample MAE for predicting the output variables when all the input variables are inserted into the IRLS or RF learner, using 10-fold cross validation with 100 repetitions.

	IRLS	RF
y1	2.14 ± 0.24	0.51 ± 0.11
y2	2.21 ± 0.28	1.42 ± 0.25

MAE stands for *Mean Absolute Error*, a metric which is robust to outliers in the data, and which is easy to interpret. The results are given in the form mean ± standard deviation.

Table 6
Out of sample MSE for predicting the output variables when all the input variables are inserted into the IRLS or RF learner, using 10-fold cross validation with 100 repetitions.

	IRLS	RF
$y1$	9.87 ± 2.41	1.03 ± 0.54
$y2$	11.46 ± 3.63	6.59 ± 1.56

MSE stands for *Mean Squared Error*, the classical metric to quantify error in least squares regression settings. The results are given in the form mean ± standard deviation.

Table 7
Out of sample MRE for predicting the output variables when all the input variables are inserted into the IRLS or RF learner, using 10-fold cross validation with 100 repetitions.

	IRLS	RF
$y1$	10.09 ± 1.01	2.18 ± 0.64
$y2$	9.41 ± 0.80	4.62 ± 0.70

MRE stands for *Mean Relative Error*. The results are given in the form mean ± standard deviation.

Table 8
Importance of the input variables as determined by the RF for the output variables.

Measure	Importance for $y1$	Importance for $y2$
$X1$	50.51 ± 1.15	43.74 ± 1.11
$X2$	50.41 ± 1.41	43.55 ± 1.08
$X3$	40.16 ± 1.09	32.16 ± 0.83
$X4$	20.40 ± 0.95	20.12 ± 0.87
$X5$	8.97 ± 0.68	9.41 ± 0.72
$X6$	18.51 ± 0.44	22.03 ± 0.48
$X7$	93.12 ± 1.50	86.92 ± 1.58
$X8$	38.84 ± 0.94	39.07 ± 0.97

The importance was computed for each of the 100 cross-validation repetitions. The results are given in the form mean ± standard deviation.

in this study because of the assumption we made when performing the building simulations; hence, we do not elaborate further on discussing both variables.

We argue that the statistical analysis methodology presented in this study provides essential insight into the given problem, and is unfairly skipped in most papers in the literature in this discipline. For example, the density plots and the scatter plots give ample evidence that linear techniques are not appropriate for the available data in this application. However, it is well known that *statistical correlation* should not be conflicted with *causality*. For that reason, in addition to statistically formally exploring the data with rank correlation coefficients and MI, we have computed the importance of the variables using RF. Interestingly, the most important variable (glazing area) is not the most correlated with either output variable. From an engineering perspective, it can be intuitively understood that the glazing area is of paramount significance to determine EPB. This is because the amount of glazing determines the heat absorbed in a building due to the sun, and similarly glazing is a source of heat leakage from the building to the environment.

The findings of this study agree with those in the machine learning literature strongly endorsing the use of RF in complex applications [21]. The RF massively outperformed IRLS in finding an accurate functional relationship between the input and output variables. Classical regression settings (such as IRLS) may fail to account for multi-collinearity, where variables appear to have large magnitude but opposite side sign coefficients with regard to predicting the response [23]. On the contrary, the decision tree mechanism (and as an extension RF) optimizes the selection of the variable for each split, and thus can internally account for redundant and interacting variables [23,24,26]. Therefore, the problem of collinearity in RF is implicitly solved as part of the internal optimisation algorithm. The nature of the EPB topic where different authors introduce different input variables to study similar but different output variables in their simulations hinders direct comparisons amongst studies. Therefore, we tentatively approach this subject when referring to previously published works of other researchers. In general, the reported errors are similar to errors reported in the literature in the EPB domain, for example see [7,10,11]. Similarly to Wan et al. [15] (Table 7 in that study), HL can be estimated more accurately than CL (see Tables 5–7). This may be slightly surprising given that the univariate association strength of the eight variables with HL and CL is very similar (see Tables 2 and 3). We tentatively suggest that HL is estimated with considerably greater accuracy than CL because some of the variables in this study interact more efficiently to provide an estimate of HL. More formal tests need to be carried out to provide additional insight into this aspect of the dataset.

We believe the results of this study strongly caution against blindly using widely available mathematical tools which often rely on normality of the data. The methodology presented here is very general and could, in principle, be extended to encompass additional input variables (for example some of the parameters assumed constant such as the climate or occupancy could be introduced as input variables). Similarly, additional output variables could be studied using the approach developed in this study. We envisage the proposed method finding use as a simple, off-the-shelf approach to obtain accurate estimates of heating load and cooling load.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

A. Tsanas gratefully acknowledges the financial support of Intel Corporation and the Engineering and Physical Sciences Research Council (EPSRC). The funding sources had no involvement in this study.

References

[1] European Commission, Directive 2002/91/EC of the European Parliament and of the Council of 16th December 2002 on the energy performance of buildings, Official journal of the European Communities, L1/65–L1/71, 04/01/2003.
[2] Z. Yu, F. Haghigrat, B.C.M. Fung, H. Yoshimo, A decision tree method for building energy demand modeling, Energy and Buildings 42 (2010) 1637–1646.
[3] L. Perez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy and Buildings 40 (3) (2008) 394–398.
[4] W.G. Cai, Y. Wu, Y. Zhong, H. Ren, China building energy consumption: situation, challenges and corresponding measures, Energy Policy 37 (6) (2009) 2054–2059.
[5] G. Platt, J. Li, R. Li, G. Poulton, G. James, J. Wall, Adaptive HVAC zone modelling for sustainable buildings, Energy and Buildings 42 (2010) 412–421.
[6] R. Yao, B. Li, K. Steemers, Energy policy and standard for built environment in China, Renewable Energy 30 (2005) 1973–1988.
[7] A. Yezioro, B. Dong, F. Leite, An applied artificial intelligence approach towards assessing building performance simulation tools, Energy and Buildings 40 (2008) 612–620.

[8] A. Tsanas, J.Y. Goulermas, V. Vartela, D. Tsiapras, G. Theodorakis, A.C. Fisher, P. Sfirakis, The Windkessel model revisited: a qualitative analysis of the circulatory system, Medical Engineering and Physics 31 (2009) 581–588.
[9] D.B. Crawley, J.W. Hand, M. Kummert, B.T. Griffith, Contrasting the capabilities of building energy performance simulation programs, Building and Environment 43 (2008) 661–673.
[10] B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings 37 (2005) 545–553.
[11] T. Catalina, J. Virgone, E. Blanco, Development and validation of regression models to predict monthly heating demand for residential buildings, Energy and Buildings 40 (2008) 1825–1832.
[12] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Applying support vector machine to predict hourly cooling load in the building, Applied Energy 86 (2009) 2249–2256.
[13] J. Zhang, F. Haghighat, Development of artificial neural network based heat convection for thermal simulation of large rectangular cross-sectional area earth-to-earth heat exchanges, Energy and Buildings 42 (4) (2010) 435–440.
[14] S.S.K. Kwok, R.K.K. Yuen, E.W.M. Lee, An intelligent approach to assessing the effect of building occupancy on building cooling load prediction, Building and Environment (2011), doi:10.1016/j.buildenv.2011.02.008.
[15] K.K.W. Wan, D.H.W. Li, D. Liu, J.C. Lam, Future trends of building heating and cooling loads and energy consumption in different climates, Building and Environment 46 (2011) 223–234.
[16] S. Schiavon, K.H. Lee, F. Bauman, T. Webster, Influence of raised floor on zone design cooling load in commercial buildings, Energy and Buildings 42 (2010) 1182–1191.
[17] W. Pessenlehner, A. Mahdavi, A building morphology, transparency, and energy performance, in: Eighth International IBPSA Conference Proceedings, Eindhoven, Netherlands, 2003, pp. 1025–1032.
[18] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2007.
[19] C. Chatfield, The Analysis of Time Series: An Introduction, Chapman & Hall, 2004.
[20] T.M. Cover, J.A. Thomas, Elements of Information Theory, 2nd edition, Wiley-interscience, 2006.
[21] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity, Journal of the Royal Society Interface 8 (2011) 842–855.
[22] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests, IEEE Transactions on Biomedical Engineering 57 (2010) 884–893.
[23] T. Hastie, R. Tibshirani, J.J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer, New York, USA, 2009.
[24] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
[25] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity, in: International Symposium on Nonlinear Theory and its Applications, 2010, pp. 457–460.
[26] E. Tuv, A. Borisov, G. Runger, K. Torkkola, Feature selection with ensembles, artificial variables, and redundancy elimination, Journal of Machine Learning Research 10 (2009) 1341–1366.