Energy and Buildings 127 (2016) 313–326

Contents lists available at ScienceDirect

Energy and Buildings

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

# A comparison of approaches to stepwise regression on variables sensitivities in building simulation and analysis

Mengchao Wang[a,*], Jonathan Wright[a], Alexander Brownlee[b], Richard Buswell[a]

[a] School of Civil & Building Engineering, Loughborough University, Loughborough LE11 3TU, UK
[b] Computing Science & Mathematics, University of Stirling, Stirling FK9 4LA, UK

ARTICLE INFO

*Article history:*
Received 28 August 2015
Received in revised form 4 May 2016
Accepted 19 May 2016
Available online 28 May 2016

*Keywords:*
Global sensitivity analysis
Stepwise regression
Sensitivity indexes
Standardized (rank) regression coefficients

ABSTRACT

Developing sensitivity analysis (SA) that reliably and consistently identify sensitive variables can improve building performance design. In global SA, a linear regression model is normally applied to sampled-based solutions by stepwise manners, and the relative importance of variables is examined by sensitivity indexes. However, the robustness of stepwise regression is related to the choice of procedure options, and therefore influence the indication of variables' sensitivities. This paper investigates the extent to which the procedure options of a stepwise regression for design objectives or constraints can affect variables global sensitivities, determined by three sensitivity indexes. Given that SA and optimization are often conducted in parallel, desiring for a combined method, the paper also investigates SA using both randomly generated samples and the biased solutions obtained from an optimization run. Main contribution is that, for each design objective or constraint, it is better to conclude the categories of variables importance, rather than ordering their sensitivities by a particular index. Importantly, the overall stepwise approach (with the use of bidirectional elimination, BIC, rank transformation and 100 sample size) is robust for global SA: the most important variables are always ranked on the top irrespective of the procedure options.



## 1. Introduction

Reducing the energy consumption in the building sector have a critical role in meeting energy and emission reduction targets in both developing and developed countries [15,21]. In order to improve building performance by implementing the optimal design solutions at a reasonable investment, both sensitivity analysis and model-based optimization can be used to inform design decisions. Compared to model-based optimization that is applied to find the combinations of variables values that optimize the objectives while satisfying the design constraints, Monte Carlo sensitivity analysis (SA) can support decision making by providing insight into the input variables that most influence design objectives, such as operational energy use or comfort metrics of the final building [16,24]. This makes SA a useful tool for designers. Consequently, developing SA methodologies that reliably and consistently identify sensitive variables is an important area of research.

SA can be generally grouped into local and global forms [23]. Global SA based on a linear regression is normally adopted to evaluate the relative importance of input variables [2,3,12]. When many input variables are involved, stepwise regression provides an alternative. Saltelli et al. [23] state that, when the input variables are ideally uncorrelated, the importance of variables sorted by any sensitivity indexes should be the same. This applies whether sorting by: order of addition to the regression model; size of the $R^2$ changes attributable to individual variables; absolute values of variables' standardized (rank) regression coefficients (SRCs/SRRCs); absolute values of correlations coefficients (CCs); or absolute values of partial correlation coefficients (PCCs). Therefore, in previous researches, the global sensitivities of variables are normally evaluated by a particular index, e.g. De Wilde et al. [7] using SRRCs to determine the major contributors for heating energy use; Hopfe and Hensen [16] using both SRRCs and the change of $R^2$ to explore variables influence on building performance simulation.

However, according to our previous researches [27,28], the ordering of variables' importance with respect to design objectives and constraints could be switched by applying the same global SA method to different sets of random samples. Furthermore, the robustness and effectiveness of a stepwise regression depends on the choice of procedure options, as each option has its advantages

* Corresponding author.
*E-mail address:* wangmengchao@hotmail.com (M. Wang).

http://dx.doi.org/10.1016/j.enbuild.2016.05.065
0378-7788/

and weakness [26]. For instance, the F-test is often used as default criterion to stop the stepwise regression process, but it has been shown to perform poorly relative to other criteria, e.g. the corrected AIC [17]. It has also been shown that it is possible to conduct global SA in parallel with optimization by an evolutionary algorithm [25]. This further supports decision making by providing suggested optimal solutions as well as variables' sensitivities. Solutions generated during the optimization run can be reused for the global SA to save computational efforts, despite the bias in the samples arising from operation of the evolutionary algorithm.

Therefore, this paper explores the impact of procedure options in a global SA for design objectives or constraints, providing insights that will enable more robust and more accurate assessments of variables' sensitivities to be made, through the relative magnitudes of variables sensitivity indexes. The procedure options explored are: different approaches to obtaining samples (i.e. randomly generated samples or the biased solutions obtained at the start of an optimization process, with a sample size of 100 or 1000); data forms of input variables (i.e. raw data with categorical variables, or rank-transformed data); selection approaches (i.e. the results in this paper are based on bidirectional elimination); and selection criteria (i.e. F-test, AIC or BIC).

## 2. Stepwise regression methodology

Stepwise regression analysis is usually used as an alternative of linear regression to do global SA. Since, when no impacted or correlated variables are included in the same regression model, it can avoid misleading regression of variables importance. It can also avoid overfitting of the data, as all of input variables are arbitrarily forced into the same regression model [23]. Overfitting occurs when the regression model in essence 'chases' the individual observations rather than following an overall pattern in the data, which can produce a spurious model, giving poor predictions of variables importance [23]. Thus, the overfitting is used as an important standard, to evaluate how well the linear model constructed by stepwise regression can fit to the data (generated from different samples).

Therefore, the global SA adopted here is based on a linear regression model in the stepwise manner, which is performed by R statistic software [20]. The idea is to add or remove variables in a linear model: at each iteration, selecting the variable which most increases the $R^2$ coefficient of the model. The robustness of stepwise regression analysis is dependent on the choice of procedure options, including the sampling method, sample size, data form of input variables, selection approach and selection criterion, with the accuracy being evaluated through the $R^2$ (coefficient of determination) and PRESS (predicted error sum of squares) values in the linear regression model.

In a stepwise regression analysis, the relative importance of the variables for a given output can be evaluated through sensitivity indexes, including variables' entry-order to the model, SRCs (standardized regression coefficients)/SRRCs (standardized rank regression coefficients, for rank-transformed data), and $R^2$ change attributable to the individual variables. The more important (sensitive) the variable is, the earlier it is selected into the linear model, the larger its SRC/SRRC is, the greater it is attributable to $R^2$ change [23].

### 2.1. Samples and sample size

The robustness of a sensitivity method is related to the choice of sample size and the manner in which the samples are generated [23]. For a sample size of 100 and above, the difference in the results from different sampling methods is decreased; thus, it is feasible to use a random sampling method and 100 samples in a Monte Carlo analysis for typical building simulation applications [19,22]. In this paper, the conclusion is further validated by comparing the SA resulting from a 100 random samples with those from a 1000 random samples; the 100 random samples are taken as being the first 100 samples of the 1000 randomly generated samples. The repeatability of the approach is investigated by repeating the analysis for two sets of random samples (Random Sample A and Random Sample B). Moreover, the first 100 solutions (being consistent with the smaller sample size in random samples) obtained from a multi-objective optimization process (based on NSGA-II) have also been used to do global SA, for design objectives and constraints (See Section 3.1). The aim is to explore the extent to which the biased samples can affect the robustness of variables global sensitivities, determined by the same method based on stepwise regression.

### 2.2. Input variables and rank-transformation

The input variables considered in most sensitivity analyses are real-valued quantities [14,16]. However, in this paper, the categorical variables for construction types are applied with others having physical representations (see Section 4.1). Such variables frequently appear in building design problems, so it is important to consider them. Furthermore, a non-linear relationship between the input variables and the output is possible, whether the input variables have real-valued quantities or not. A rank transformation of the variables based on a monotonic relationship can mitigate the problems associated with fitting linear models to nonlinear data [23]. The rank transformation is defined according to Spearman's rank correlation coefficients: raw data are replaced by their corresponding ranks, and then the ranks of input variables and outputs are used to do regression analysis. Particularly, the smallest rank *1* is assigned to the smallest value of each variable, and then the rank *2* is assigned to the next larger value, and so on until the largest rank *m* assigned to the largest value (i.e. *m* indicates the number of observations for each variable).

Thus, two alternative representations of the input variables are considered here:

- The input variables in their raw form.
- A rank-transformation of the variables (and outputs).

### 2.3. Selection approach

There are three model-selection approaches [6] as below. Due to identical results in this case study, the results from bidirectional elimination are only discussed here:

- Forward selection: which starts from an 'empty' model with no input variable but an intercept, and then adds the variable most improving the model one-at-a-time until no more added variables can significantly improve the model. This approach is based on a pre-defined selection criterion.
- Backward elimination: which starts from a 'full' model with all predictive input variables and an intercept, and then deletes the variable least improving the model one-at-a-time until no more deleted variables can significantly improve the model. This approach is based on a pre-selected selection criterion.
- Bidirectional elimination: which is essentially a forward selection procedure but with the possibility of deleting a selected variable at each stage, as in the backward elimination. This approach is commonly applied for stepwise regression, particularly when there are correlations between variables.