Using probabilistic sampling-based sensitivity analyses for indoor air quality modelling

Payel Das [a],[*], Clive Shrubsole [a], Benjamin Jones [b], Ian Hamilton [c], Zaid Chalabi [d], Michael Davies [a], Anna Mavrogianni [a], Jonathon Taylor [a]

[a] Bartlett School of Graduate Studies, UCL, 14 Upper Woburn Place, London WC1H 0NN, UK
[b] Department of Architecture and Built Environment, University of Nottingham, NG7 2RD, UK
[c] UCL Energy Institute, UCL, 14 Upper Woburn Place, London WC1H 0NN, UK
[d] London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK

ABSTRACT

We develop a probabilistic framework for modelling indoor air quality in housing stocks, selecting appropriate sensitivity analyses to understand indoor air quality determinants, and constructing a reliable metamodel from the most relevant determinants to allow quick assessments of future intervention scenarios.

The replicated Latin Hypercube sampling method is shown to be efficient at propagating variations between model input and output variables. A comparison of a range of sample-based sensitivity methods shows that an initial visual assessment can help to select appropriate sensitivity analyses, as they test for different types of relations (i.e. linear, monotonic, and non-monotonic). An advantage of linear regression methods is that the total output can be apportioned to various input variables. The advantage of tests with correlation coefficients is that the associated *p*-values can be used to assess whether input variables are significant. An artificial neural network constructed from a reduced set of input variables selected at a 5% level of significance is able to accurately predict indoor air quality.

In the application of the framework to the modelling of winter indoor air quality in single-storey flats in England, the drivers for internally- and externally-generated $PM_{2.5}$ are found to be different, therefore allowing interventions that reduce both concentrations simultaneously. Principal determinants for externally-generated $PM_{2.5}$ are the internal deposition rate of $PM_{2.5}$, weather-corrected volumetric infiltration rate, and ambient concentration of $PM_{2.5}$, while for $PM_{2.5}$ produced by gas cooking, they are the kitchen window opening area, generation rate of $PM_{2.5}$, and indoor temperature.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Sensitivity analysis is the study of how variations in measurements or model outcomes (output variables) can be attributed to variations in potential determinants (input variables) [1]. They are carried out to identify key input variables that affect output variables as a result of natural variation, uncertainty, or if a design variable, as a result of possible future changes. These can then be used to help design future monitoring campaigns, effect changes by making appropriate alterations to the influential input variables, and create a regression model or metamodel that can be used to make predictions of the output variables as new data become available.

Local sensitivity analyses assume a linear relation between input and output variables and then vary them a small percentage around a baseline to make an estimate of the partial derivative at that point [e.g. Ref. [1]]. As each input variable is adjusted separately, interactions between them are not taken into account. In a global sensitivity analysis, the input variables are varied over the entire range and possibly also simultaneously to enable interactions between inputs to be explored. The Morris method is an example of a global sensitivity method, which makes an estimate of 'elementary effects' by estimating the mean and standard deviation of the distribution of each output obtained by sampling from each input one-at-a-time [e.g. Ref. [1]]. The means and standard deviations are then analysed jointly to determine the most important inputs. There are also sampling-based approaches that vary inputs simultaneously

and therefore take interactions between inputs into account. Regression, sample comparison, correlation, and variance decomposition methods can then be applied to the samples of input variables and associated output variables [e.g. Refs. [2,3]]. The samples in the global sensitivity analyses can be assembled either directly through data collection or generated using models, according to several experimental designs, e.g. non-probabilistic designs such as optimal designs [e.g. Ref. [4]], which are 'optimal' with respect to some statistical criterion such as variance. Experimental designs can also be generated probabilistically, using simple random sampling, Latin Hypercube sampling [5], maximin designs [6], uniform sampling [7,8], and Sobol sequences [9]. Latin Hypercube sampling is a type of stratified design, while the Sobol sequence is an example of quasi-random sampling [10]. Maximin designs maximize the minimum distances between pairs of sample points [6], and uniform designs search for sample points that are uniformly scattered [11].

In this paper, the results of the sensitivity analysis are used to construct a metamodel, which may be mathematically equivalent to a regression model, but approximates a model rather than data. Therefore as well as offering a better understanding of the relation between input and output variables, it also offers the benefit of reducing computational time, which is invaluable when multiple model simulations are required, for example, in the determination of optimal retrofit strategies [12]. When combined with sensitivity analysis, a metamodel also has the benefit of reducing the number of inputs that need to be measured to determine the outputs of interest. Several methods for metamodel construction have been reported in the literature. The simplest are polynomial models and splines [e.g. Refs. [12,13]], the latter of which consist of several connected polynomial 'segments'. Examples of other constructions include radial basis functions, in which the metamodel is a linear combination of radially symmetric functions [e.g. Refs. [13,14]], and artificial neural networks, which are interconnected structures of neurons, belonging to the class of machine learning techniques [e.g. Ref. [15]].

Indoor environment modelling has been widely combined with sensitivity analyses to explore the determinants of energy performance, indoor air quality, good building design, and optimal retrofitting measures. Examples of methods used include local sensitivity analysis [e.g. Refs. [16,17]], one-at-a-time analyses in which each input is varied over the whole range [18–21], the full Morris method [22–24], multiple linear regression analysis, and analysis of variance - an example of a sample-comparison sensitivity method [25,26]. More complete reviews of the application of these methods in indoor environment modelling can be found in Refs. [27–29]. Examples of metamodels constructed of indoor environment models include polynomial approximations [30–32], splines [31], support vector machines (another example of a machine learning technique) [33], and artificial neural networks [34–36].

A comparison of determinants from a range of sensitivity analyses is however missing in the literature. The comparison between the various sensitivity analyses methods is especially important where there are non-linear correlations between the input and output variables. Probabilistic methods are also only used occasionally and have the benefit of allowing the total uncertainty to be easily determined. Finally, an application of the sensitivity results in the construction of a metamodel has only been carried out in a few instances [30,33,37]. In this paper, we propose a framework for combining probabilistic sampling, a range of sensitivity analyses, and the construction of a metamodel for the modelling of indoor air quality in a case-study sample of single-storey flats in the English housing stock. The paper is structured as follows. The selected indoor air quality modelling tool, probabilistic sensitivity methods,

and construction of the metamodel are described in Section 2. Section 3 describes the application of the framework to a case-study sample of single-storey flats in the English housing stock. The results are presented in Section 4, along with an illustration of the metamodel for making quick computational predictions of changes in the indoor environment following housing interventions. Section 5 adds interpretation to the results obtained, compares them to those in the literature, and discusses uncertainties in the analysis presented here. Finally, Section 6 summarizes the findings in this work.

2. Methodology

This section introduces the indoor air quality modelling tool selected for this work and describes the methods involved in the creation of a probabilistic housing stock model of indoor air quality, the application of a range of probabilistic sensitivity analyses, and the construction of a metamodel from a reduced set of input variables. Fig. 1 shows the chronology of the analytical steps.

2.1. Indoor air quality modelling tool

CONTAM [38] is a multizone indoor air quality and ventilation analysis tool that models airflows between multiple zones in the indoor environment, and between the indoor and external environments, as a result of mechanical ventilation systems, wind pressures acting on the building envelope and buoyancy effects. From the airflows, the ventilation rates in the building and their variation over time can be determined. The tool also calculates temporal profiles of concentrations of pollutants as a result of production via a range of mechanisms, transport by airflow, and transformation by chemical processes, adsorption and desorption to building materials, filtration, and deposition to building surfaces. Each simulation with CONTAM therefore requires the specification of hundreds of input variables including weather files, building dimensions, indoor temperatures, building envelope permeabilities, contaminant parameters, and those specifying occupant behaviour. The inputs will be denoted here by the vector $\boldsymbol{x} = (x_1 x_2, \ldots)$. The relevant post-processed outputs of each simulation will be denoted by the vector $\boldsymbol{y} = (y_1 y_2, \ldots)$. CONTAM has been extensively validated with regards to the modelled airflows, pressures, and pollutant concentrations against other models, measurements in a controlled environment, and measurements in field studies [e.g. Refs. [39–41]].

2.2. Generating a probabilistic housing stock model of indoor air quality

To create a probabilistic housing stock model of indoor air quality, firstly the inputs in CONTAM, which are subject to either uncertainty or variability over the housing stock need to be selected. Then the probability distribution functions of the selected input variables need to be specified for the housing stock, and multiple samples generated to replicate the input variations expected. The Latin Hypercube sampling method [5] is selected to generate the samples, a method considerably more efficient [e.g. Ref. [25]] at probing the probability space compared to simple random sampling. For a fixed number of samples, n, the Latin Hypercube method generates samples so that no value is generated more than once in each subspace (defined such that n equally probable subspaces cover the range of the hyperspace of each element in $\boldsymbol{x}$). The algorithm implemented in MATLAB [43] is applied here to generate a hypercube, where each dimension ranges uniformly between 0 and 1. The inverse cumulative distribution function of each element in $\boldsymbol{x}$ is used to convert the